COGNITIVE NEUROINTELLIGENCE

EDITED BY: Jia Liu, Si Wu, Ke Zhou and Yiying Song

PUBLISHED IN: Frontiers in Computational Neuroscience

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88971-340-0
DOI 10.3389/978-2-88971-340-0

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

COGNITIVE NEUROINTELLIGENCE

Topic Editors:
Jia Liu, Tsinghua University, China
Si Wu, Peking University, China
Ke Zhou, Beijing Normal University, China
Yiying Song, Beijing Normal University, China

Citation: Liu, J., Wu, S., Zhou, K., Song, Y., eds. (2021). Cognitive NeuroIntelligence. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-88971-340-0

Table of Contents

frontiers in Computational Neuroscience

EDITORIAL
published: 26 July 2021
doi: 10.3389/fncom.2021.718518

Editorial: Cognitive NeuroIntelligence

Yiying Song[1], *Si Wu*[2]*, *Ke Zhou*[1] *and Jia Liu*[3]*

[1] *Faculty of Psychology, Beijing Normal University, Beijing, China,* [2] *School of Electronics Engineering and Computer Science, Peking University, Beijing, China,* [3] *Tsinghua Laboratory of Brain and Intelligence, Tsinghua University, Beijing, China*

Keywords: artificial neural network, cognitive neuroscience, cognitive neurointelligence, cognitive-neuroscience-inspired AI, AI-inspired cognitive neuroscience

Editorial on the Research Topic

Cognitive NeuroIntelligence

In recent years, fascinating progresses have been made in utilizing artificial intelligence to solve a broad range of problems. AI systems today can match and even outperform human performance in certain challenging tasks, including visual cognition. Recent AI advances in deep learning have been largely inspired by neuroscience research on biological brain and guided by architectural constrains from biological neural networks. In this Research Topic, we advocate further interactions between the fields of AI and cognitive neuroscience to benefit both fields. The topic of Cognitive Neurointelligence, starting from May 25th, 2020 and ending on November 20th, 2020, was organized by Jia Liu, Si Wu, Ke Zhou, and Yiying Song.

On one hand, there is great potential for cognitive neuroscience to benefit AI (cognitive-neuroscience-inspired AI). The structures and functions of neural systems, which result from hundreds of millions of years evolution, are optimized for animals processing information in order to survive in natural environments. They naturally serve as the resources inspiring us to develop AI. In addition to this, cognitive neuroscience can also benefit AI research from a new aspect. The end-to-end learning strategy makes deep convolutional neural networks (DCNNs) remaining to be "black boxes," where the algorithms and computations of the networks are poorly understood. Techniques and approaches available in cognitive neuroscience, including experimental paradigms, data analysis techniques, and theoretical hypotheses, can serve as a repertoire of tools for unveiling the black boxes of DCNNs, illuminating the algorithms and computations inside the networks.

On the other hand, AI can make fundamental contributions to cognitive neuroscience as well (AI-inspired cognitive neuroscience). In addition to serving as advanced mathematical tools for analyzing big data in neuroscience, models of AI can also give us insight into understanding the inner mechanisms of biological brain and intelligence, for instance, DCNNs have offered the best models of biological visual systems to date. More importantly, biological brains are the result of evolution; and analogically we can manipulate loss functions, architectures, and datasets of DCNNs to "re-run" evolution and therefore to pry open secrets that lead to the emergence of the human brain and mind.

Therefore, the purpose of this Research Topic is to bring together research efforts from AI and cognitive neuroscience, seeking to integrate AI and cognitive neuroscience toward a new field of cognitive neurointelligence.

In the direction of cognitive-neuroscience-inspired AI, six papers in this special issue aimed to develop advanced information processing techniques inspired by biological systems. Motived by the unsupervised learning behavior of humans, Ji et al. proposed an unsupervised few-shot learning algorithm for object classification. Motivated by the rapid topology perception of humans, Wang et al. proposed a gap-junction network for fast topology detection in images. Inspired by the balance of excitation and inhibition (E-I) interactions in neural systems, Tian G. et al. proposed an E-I balanced network for fast signal detection. Based on the biologically plausible global feedback

OPEN ACCESS

Edited and reviewed by:
Yu-Guo Yu,
Fudan University, China

****Correspondence:***
Si Wu
siwu@pku.edu.cn
Jia Liu
liujiathu@tsinghua.edu.cn

Received: *01 June 2021*
Accepted: *30 June 2021*
Published: *26 July 2021*

Citation:
Song Y, Wu S, Zhou K and Liu J (2021) Editorial: Cognitive NeuroIntelligence. Front. Comput. Neurosci. 15:718518. doi: 10.3389/fncom.2021.718518

and the local STDP learning rule, Zhao et al. proposed a new method to train multi-layer spiking neural networks. Applying biological learning rules, Fang et al. developed spiking neural networks for sequence generation. Zhou et al. revealed that the function connectivity of the brain network accounts for critical dynamics, and the latter leads to efficient information processing. In addition, three papers in this special issue applied methods in cognitive neuroscience to understand inner representations in DCNNs. Liu et al. applied the concepts and measures of coding schemes from neuroscience studies to DCNNs and provided evidence that DCNNs adopted a hierarchically-evolved sparse coding scheme to represent objects as the biological brain does. Song et al. adopted a reverse-correlation approach in psychophysical studies and found that both DCNNs and humans utilized similar inner representations to perform the task of face gender classification. Tian J. et al. explored the phenomenon and mechanism of biased behaviors in DCNNs by borrowing the paradigms and theories from a classical race bias (i.e., the other race effect) in humans and found a human-like multidimensional face representation in DCNN. Together, these studies suggest the possibility that DCNNs and humans may use an implementation-independent representation to achieve the same computation goal.

In the direction of AI-inspired cognitive neuroscience, two studies in this issue used DCNNs as a model to inform our understanding of human cognition. Unlike studies on humans where perceptual experiences are always intermingled with conceptual guidance, Huang et al. used DCNNs to provide evidence that the semantic relatedness of object categories can automatically emerge from perceptual experiences without top-down conceptual guidance. In addition, investigation on the role of nature vs. nurture in the formation of domain-specific modules in biological brains cannot easily dissociate the effects of visual experience from genetic predisposition. To overcome this limitation, Xu et al. built a model of selective deprivation of the experience on faces with a DCNN and demonstrated that domain-specificity may evolve from non-specific experience without genetic predisposition, and is further fine-tuned by domain-specific experience. In other two studies, Zhang et al. applied AI algorithms to improve target detection in neural signals, and Zheng et al. applied DCNNs to investigate the transfer learning effects based on local and global invariant features.

Finally, to meet the objective of crosstalk between the AI and cognitive neuroscience, Chen et al. presented a Python-based toolbox, DNN Brain, which enables researchers from both fields to conveniently map the internal representations of DNNs and brain, and examine their correspondences.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers
in Computational Neuroscience

ORIGINAL RESEARCH
published: 03 September 2020
doi: 10.3389/fncom.2020.00079

Excitation-Inhibition Balanced Neural Networks for Fast Signal Detection

Gengshuo Tian[1], *Shangyang Li*[1,2], *Tiejun Huang*[1] *and Si Wu*[1,2]*

[1] *School of Electronics Engineering and Computer Science, Peking University, Beijing, China,* [2] *IDG/McGovern Institute for Brain Research, Peking-Tsinghua Center for Life Sciences, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing, China*

Excitation-inhibition (E-I) balanced neural networks are a classic model for modeling neural activities and functions in the cortex. The present study investigates the potential application of E-I balanced neural networks for fast signal detection in brain-inspired computation. We first theoretically analyze the response property of an E-I balanced network, and find that the asynchronous firing state of the network generates an optimal noise structure enabling the network to track input changes rapidly. We then extend the homogeneous connectivity of an E-I balanced neural network to include local neuronal connections, so that the network can still achieve fast response and meanwhile maintain spatial information in the face of spatially heterogeneous signal. Finally, we carry out simulations to demonstrate that our model works well.

Keywords: E-I balanced network, optimal noise structure, Fokker-Planck equation, fast tracking, asynchronous state

1. INTRODUCTION

To survive in natural environments, animals have developed, through millions of years evolution, the ability to process sensory inputs rapidly. For instance, studies have shown that human subjects can perform complex visual analyses within 150 ms (Thorpe et al., 1996), and the response latency of neurons in the visual cortex of monkeys is as short as tens of milliseconds (Raiguel et al., 1999; Sugase et al., 1999).

Meanwhile, many artificial engineering systems have high demands for real-time processing of rapidly varying signals. This is exemplified by the recently developed Spike Camera (Dong et al., 2017), which has a sampling rate of up to 40,000 frames per second (fps), far surpassing conventional cameras' 60 fps. This allows it to capture high-speed objects and their textual details, which can be used on real-time motion detection, tracking, and recognition if we have the appropriate algorithms and computing platforms. However, the processing speed of traditional algorithms often cannot meet such demands.

The balance of excitation and inhibition is a general property of neural systems. The excitation-inhibition (E-I) balanced neural network was first proposed to explain the irregular firing of cortical neurons widely observed in the cortex (Softky and Koch, 1993; Shadlen and Newsome, 1994), and was later confirmed by a large amount of experimental data (Haider et al., 2006; Okun and Lampl, 2008; Dorrn et al., 2010; Graupner and Reyes, 2013). Theoretical studies have found that the asynchronous irregular firing state spontaneously emerges in a network of excitatory and inhibitory neurons with random connections satisfying some very loose balancing conditions (van Vreeswijk and Sompolinsky, 1996; van Vreeswijk and Sompolinsky, 1998; Renart et al., 2010). The effects of this chaotic state on optimal coding (Denève and Machens, 2016), working memory (Lim and Goldman, 2014), and neuronal tuning (Hansel and van Vreeswijk, 2012), as well as its coexistence with attractor dynamics (Litwin-Kumar and Doiron, 2012) have been widely studied.

OPEN ACCESS

Edited by:
Pei-Ji Liang,
Shanghai Jiao Tong University, China

Reviewed by:
Robert Rosenbaum,
University of Notre Dame,
United States
Lianchun Yu,
Lanzhou University, China

****Correspondence:***
Si Wu
siwu@pku.edu.cn

Received: *27 June 2020*
Accepted: *27 July 2020*
Published: *03 September 2020*

Citation:
Tian G, Li S, Huang T and Wu S (2020)
Excitation-Inhibition Balanced Neural
Networks for Fast Signal Detection.
Front. Comput. Neurosci. 14:79.
doi: 10.3389/fncom.2020.00079

In the present study, we focus on the fast tracking ability of E-I balanced networks, where the population firing rate of the network is proportional to the input amplitude and tracks input changes rapidly (van Vreeswijk and Sompolinsky, 1996; Renart et al., 2010), and investigate how E-I balanced neural networks can be used for fast signal detection in brain-inspired computation. Neuromorphic computing, which mimics the structures and computational principles of the neural system, is receiving increasing attention in artificial intelligence (AI), as it has the potential to overcome the von Neumann bottleneck in modern computers that limits their processing speed (Indiveri and Liu, 2015). The fast response property of the E-I balanced network makes it a naturally compatible candidate to be implemented in neuromorphic systems to achieve rapid information processing.

In the following sections, we show that the asynchronous firing state of the network generates an optimal noise structure which enables the network to track input changes rapidly. We then extend the homogeneous connectivity of the classical E-I balanced neural network to include local neuronal connections, so that the network can achieve fast response and meanwhile maintain the spatial information when presented with spatially heterogeneous signals. Finally, we carry out simulations to demonstrate the performance of our model.

2. FAST RESPONSE OF A HOMOGENEOUS E-I BALANCED NETWORK

To illustrate the mechanism of the fast response property, we first investigate a homogeneously connected E-I balanced network.

2.1. Intuition on the Mechanism of Fast Response

The fast response property of an E-I balanced network is at the population level. To understand this, let us consider a non-leaky linear integrate-and-fire neuron, whose dynamics is given by

$$\tau \frac{\mathrm{d}v}{\mathrm{d}t} = I, \tag{1}$$

where τ is the integration time constant of the neuron, v the membrane potential, and I the input current. When v reaches the threshold θ, the neuron generates an action potential, and v is reset to the reset potential v_0. Thus, for a constant input I_0, the time it takes for a neuron to generate a spike starting from v_0 is

$$T = \tau \frac{\theta - v_0}{I_0}.$$

It can be seen that the response time of a single neuron is limited by τ (**Figure 1A**).

However, when a neural population receives a signal, if the noise in the system keeps membrane potentials of different neurons at different levels, there will always be a few neurons whose potentials are near the threshold that can quickly respond to input changes. In such a case, the network as a whole can respond to input changes very fast, whose reaction time is only restricted by insurmountable factors such as axonal conduction delays, rather than the membrane time constant τ of individual neurons (**Figure 1B**). The key of this mechanism is to prevent synchronous firing of neurons and maintain a stable distribution of membrane potentials in the neural population, and asynchronous firing happens to be one of the hallmarks of an E-I balanced network (Renart et al., 2010), which we shall discuss in more detail below.

2.2. The Balancing Condition

We first present the conditions for maintaining an E-I balanced neural network and the stationary population firing rates under those conditions in the large N limit, where N is the number of neurons (van Vreeswijk and Sompolinsky, 1998; Rosenbaum et al., 2017). Consider a network of size N, with $N_E = q_E N$ being excitatory and $N_I = q_I N$ inhibitory, where $q_E + q_I = 1$. The input current received by neuron i in population a ($a = E$ being excitatory and $a = I$ being inhibitory) can be written as

$$I_i^a(t) = F_i^a(t) + R_i^a(t), \quad a = E, I, \tag{2}$$

where F_i^a is the feedforward (i.e., external) input, and R_i^a is the recurrent input from other neurons in the network with the form

$$R_i^a(t) = \sum_{b=E,I} \sum_j J_{ij}^{ab} \sum_k \frac{1}{\tau_{b,s}} e^{-(t-t_{j,k})/\tau_{b,s}}, \quad a = E, I, \tag{3}$$

where j indexes presynaptic neurons, $\tau_{b,s}$ is the synaptic time constant of the presynaptic population b, and $t_{j,k}$ is the spike time of the k'th spike of neuron j.

Since in both the cortex and industrial applications, the number of neurons in a network is large, we may examine the balanced network in the $N \to \infty$ limit. Expressing the relevant quantities in orders of N can help elucidate the mechanism. Neurons in the network are connected randomly, with the connection probability determined solely by the neuron types. The probability that neuron j in population b connects to neuron i in population a is p_{ab} for all i, j. Note that here p_{ab} is constant, and does not tend to 0 as $N \to \infty$. This regime is usually referred to as dense connectivity, in contrast to sparse connectivity where the number of presynaptic neurons for each postsynaptic neuron is kept constant as $N \to \infty$ (van Vreeswijk and Sompolinsky, 1996; Brunel, 2000). If a connection exists, its strength is set to be $J_{ij}^{ab} = j_{ab}/\sqrt{N}$; otherwise $J_{ij}^{ab} = 0$. Here $j_{ab} \sim \mathcal{O}(1)$. ($\mathcal{O}$ denotes scaling with respect to $N \to \infty$ throughout this paper.) This scaling is a hallmark of balanced networks. Note that in some earlier works, especially those that employ a sparse connectivity regime (van Vreeswijk and Sompolinsky, 1996), this scaling is often written as $J \sim \mathcal{O}(\sqrt{K_{ab}})$, where K_{ab} is the average number of presynaptic inputs from population b for a neuron in population a. Here, since we have $K_{ab} = p_{ab} N$ and $p_{ab} \sim \mathcal{O}(1)$, these two scalings are essentially the same.

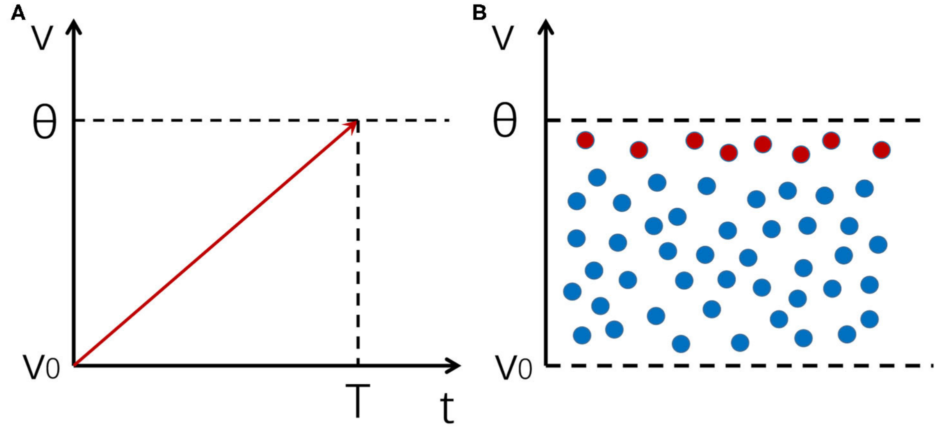


FIGURE 1 | An illustration of the mechanism of fast response for a neural population. **(A)** The integration and firing process of a neuron receiving a noiseless input. The integration time is constrained by the membrane time constant. **(B)** A distribution of membrane potentials across a neural population enables it to respond to input changes rapidly. Red dots represent neurons whose potentials are close to the firing threshold, which are the first ones to respond to input changes.

Using the mean-field approximation, the time- and population-averaged input current received by a neuron in population a can be written as,

$$\overline{I_a} = \overline{F_a} + \overline{R_a} = \sqrt{N}(f_a\mu_0 + w_{aE}r_E + w_{aI}r_I), \quad a = E, I, \tag{4}$$

where r_b is the mean firing rate of population b, $b = E, I$, and $w_{ab} = p_{ab}j_{ab}q_b \sim \mathcal{O}(1)$. Here, we have written $\overline{F_a}$ as $\overline{F_a} = \sqrt{N}f_a\mu_0$, where $f_a, \mu_0 \sim \mathcal{O}(1)$, because if we notice that long-distance projections are mainly excitatory, and assume that the feedforward inputs originated from another neural population of size $\mathcal{O}(N)$ and that the feedforward synaptic strength is also of order $\mathcal{O}(1/\sqrt{N})$, then $\overline{F_a} \sim \mathcal{O}(\sqrt{N})$ is a natural consequence. This is exactly the case in the Spike Camera data scenario that we shall examine later in section 3.

Therefore, to keep I (and thus r) bounded when $N \to \infty$, we must have

$$w_{aE}r_E + w_{aI}r_I + f_a\mu_0 \sim \mathcal{O}(\frac{1}{\sqrt{N}}), \quad a = E, I.$$

Letting $N \to \infty$, we get approximate firing rates in the large N limit

$$\begin{aligned} \lim_{N\to\infty} r_E &= \frac{f_E w_{II} - f_I w_{EI}}{w_{EI}w_{IE} - w_{EE}w_{II}}\mu_0, \\ \lim_{N\to\infty} r_I &= \frac{f_I w_{EE} - f_E w_{IE}}{w_{EI}w_{IE} - w_{EE}w_{II}}\mu_0. \end{aligned} \tag{5}$$

To keep the above limits positive and yield a stable solution, it is necessary and sufficient to let (van Vreeswijk and Sompolinsky, 1998)

$$\frac{f_E}{f_I} > \frac{w_{EI}}{w_{II}} > \frac{w_{EE}}{w_{IE}}.$$

This is the condition for the balanced firing state.

It is worth noting that whatever the neuronal transfer function is, the population firing rate in the large N limit is always linearly proportional to μ_0. That is, Equation (5) always holds. This is a direct result of Equation (4), where the total input current is the linear sum of the three $\mathcal{O}(\sqrt{N})$ order terms. The balanced firing state is a stable solution dynamically formed by the network (van Vreeswijk and Sompolinsky, 1998; Renart et al., 2010), and therefore requires no fine tuning of parameters such as j_{ab}, which is different from some other models that also try to recreate the asynchronous irregular firing state (e.g., Brunel, 2000).

It should be pointed out that Equation (5) only gives the $\mathcal{O}(1)$ order term of r_a. To satisfy the specific transfer function of neurons while maintaining the balance of the $\mathcal{O}(1)$ order term, the firing rates are adjusted by an $\mathcal{O}(1/\sqrt{N})$ term, which results in a $\mathcal{O}(1)$ order correction to I (Equation 4). We will come back to this in the specific case presented in the next section.

2.3. The Mechanism of Fast Response

As previously mentioned, the asynchronous firing of neurons is the key for fast response of the network. When the balancing conditions presented in the previous section are met, the network can achieve asynchronous irregular firing (Renart et al., 2010). We next use a network of non-leaky linear integrate and fire neurons to study the mechanism of fast response in more detail. Notably, this simple neuron model has already been implemented in a neuromorphic system (Fusi and Mattia, 1999). While not biologically realistic, this model captures the key characteristics of integrate-and-fire neurons crucial for neuromorphic computing.

The neuronal dynamics is given by Equation (1). For simplicity, let $v_0 = 0$. It can be easily seen that the transfer function of this neuron is threshold-linear, i.e.,

$$r = \begin{cases} \dfrac{\bar{I}}{\theta\tau}, & \bar{I} \geqslant 0, \\ 0, & \bar{I} < 0. \end{cases} \tag{6}$$

Substituting this into Equation (4) yields the population firing rates of excitatory and inhibitory neurons

$$r_E = \frac{(f_E w_{II} - f_I w_{EI}) - \frac{1}{\sqrt{N}} f_E \theta \tau_I}{(w_{EI} w_{IE} - w_{EE} w_{II}) + \frac{1}{\sqrt{N}} \theta (w_{EE} \tau_I + w_{II} \tau_E) - \frac{1}{N} \theta^2 \tau_I \tau_E} \mu_0,$$
$$r_I = \frac{(f_I w_{EE} - f_E w_{IE}) - \frac{1}{\sqrt{N}} f_I \theta \tau_E}{(w_{EI} w_{IE} - w_{EE} w_{II}) + \frac{1}{\sqrt{N}} \theta (w_{EE} \tau_I + w_{II} \tau_E) - \frac{1}{N} \theta^2 \tau_I \tau_E} \mu_0. \tag{7}$$

Comparing the above result with Equation (5), we can see that they are indeed $\mathcal{O}(1/\sqrt{N})$ order corrections to the $N \to \infty$ limit, as stated at the end of the last section. Note that the firing rates still linearly encode the external input, which is a result of the threshold-linear transfer function. We also check that even when the external input is small or the number of neurons is not large, the linear encoding property still holds, which expands the dynamic range of the network. However, for other non-linear neuron models, this linear encoding property may not hold.

Equation (7) is derived from the mean-field approximation, that is, it is the result of averaging over time and neurons when the system reaches a stable state. To study how the instantaneous firing rate of the population changes with time when external input changes, we need more detailed analysis. We shall use the Fokker-Planck equation (Risken, 1996) to study the membrane potential distribution $p_a(v,t)$ (Brunel and Hakim, 1999; Fusi and Mattia, 1999; Brunel, 2000; Huang et al., 2011).

First, we examine the input received by a single neuron as described in Equation (2). We consider an external input signal with additive white Gaussian noise

$$F_i^a(t) = \sqrt{N} f_a \mu_F(t) + \sigma_{aF}(t) \xi_i^{aF}(t), \quad a = E, I, \quad i = 1, \cdots, N_a, \tag{8}$$

where ξ_i^{aF} is a Gaussian white noise of magnitude 1 that is independent across neurons. Note that the signal mean is of order $\mathcal{O}(\sqrt{N})$, while the variance is of order $\mathcal{O}(1)$. This is because if we continue to use the settings considered before, and view the feedforward input as coming from Poisson spike trains generated by $\mathcal{O}(N)$ neurons firing at rates of order $\mathcal{O}(1)$, and transmitted through synapses with the strength of order $\mathcal{O}(1/\sqrt{N})$, then the resulting input's variance is the sum of $\mathcal{O}(N)$ number of terms with the same order as the square of synaptic strengths ($\mathcal{O}(1/N)$), and is therefore of order $\mathcal{O}(1)$. This characteristic is also present in the later analysis of recurrent inputs.

Next, we examine the recurrent inputs. When the network enters the balanced state, since the neurons fire asynchronously (Renart et al., 2010), and the effect of each spike is small, we could use Gaussian white noise to approximate the variations of recurrent inputs, and rewrite the second term in Equation (2) as (Brunel, 2000)

$$R_i^a(t) = \sqrt{N} \mu_{aR}(t) + \sigma_{aR}(t) \xi_i^{aR}(t), \quad a = E, I, \tag{9}$$

where

$$\mu_{aR} = w_{aE} r_E + w_{aI} r_I, \quad \sigma_{aR}^2 = j_{aE} w_{aE} r_E + j_{aI} w_{aI} r_I, \tag{10}$$

and ξ_i^{aR} is Gaussian noise of magnitude 1. The terms ξ_i^{aR} and ξ_i^{aF} are independent due to the asynchronous firing state, and can therefore be merged into one noise source. Thus, we transform Equation (2) into

$$I_i^a(t) = \mu_a(t) + \sigma_a(t) \xi_i^a(t), \quad a = E, I, \tag{11}$$

where

$$\begin{aligned} \mu_a &= \sqrt{N}(w_{aE} r_E + w_{aI} r_I + f_a \mu_F) \\ \sigma_a^2 &= j_{aE} w_{aE} r_E + j_{aI} w_{aI} r_I + \sigma_{aF}^2, \end{aligned} \tag{12}$$

and ξ_i^a is Gaussian white noise of magnitude 1. Also note that the mean of the signal is consistent with Equation (4), and the mean and variance are both of order $\mathcal{O}(1)$.

Since the balanced state implies asynchronous firing (Renart et al., 2010), the noise ξ_i^a of different neurons can be seen as independent. Then, the excitatory (inhibitory) population can be viewed as i.i.d. samples of the same random process. The membrane potential distribution of population a, $p_a(v,t)$, can thus be derived from Equation (11). We obtain the Fokker-Planck equation (Brunel, 2000; Huang et al., 2011)

$$\tau_a \frac{\partial p_a(v,t)}{\partial t} = -\mu_a \frac{\partial p_a(v,t)}{\partial v} + \frac{\sigma_a^2}{2\tau_a} \frac{\partial^2 p_a(v,t)}{\partial v^2}, \quad a = E, I. \tag{13}$$

A few boundary conditions can be naturally imposed (Brunel and Hakim, 1999; Brunel, 2000):

$$p_a(v,t) = 0, \quad \forall v \geqslant \theta. \tag{14}$$

$$p_a(0^-,t) = p_a(0^+,t), \tag{15}$$

$$\frac{\partial p_a(0^+,t)}{\partial v} - \frac{\partial p_a(0^-,t)}{\partial v} = \frac{\partial p_a(\theta,t)}{\partial v}. \tag{16}$$

$$\int_{-\infty}^{\theta} p_a(v,t) \mathrm{d}v = 1. \tag{17}$$

In Equation (13), letting $\partial p_a / \partial t = 0$, and using the above boundary conditions, we get the stationary solution

$$p_{a0}(v) = \begin{cases} \frac{1}{\theta}[1 - \exp(-2\tau_a \beta_a)] \exp\left(\frac{2\tau_a v}{\beta_a}\right), & v < 0 \\ \frac{1}{\theta}\left[1 - \exp\left(\frac{-2\tau_a(\theta - v)}{\beta_a}\right)\right], & 0 \leqslant v \leqslant \theta \\ 0, & v > \theta \end{cases} \tag{18}$$

where $\beta_a := \sigma_a^2 / \mu_a$ is the variance-to-mean ratio (VMR). This result is confirmed by simulations (**Figure 2A**).

The population firing rate, i.e., the flux at θ, is

$$r_a = -\frac{\sigma_a^2}{2\tau_a^2} \frac{\partial p_{a0}(v)}{\partial v}\bigg|_{\theta} = \frac{\mu_a}{\theta \tau_a}. \tag{19}$$

which is consistent with Equation (6).

It can be seen from Equation (18) that the membrane potential distribution is determined by the VMR β_a. The ideal noise structure is thus obtained when VMR stays constant

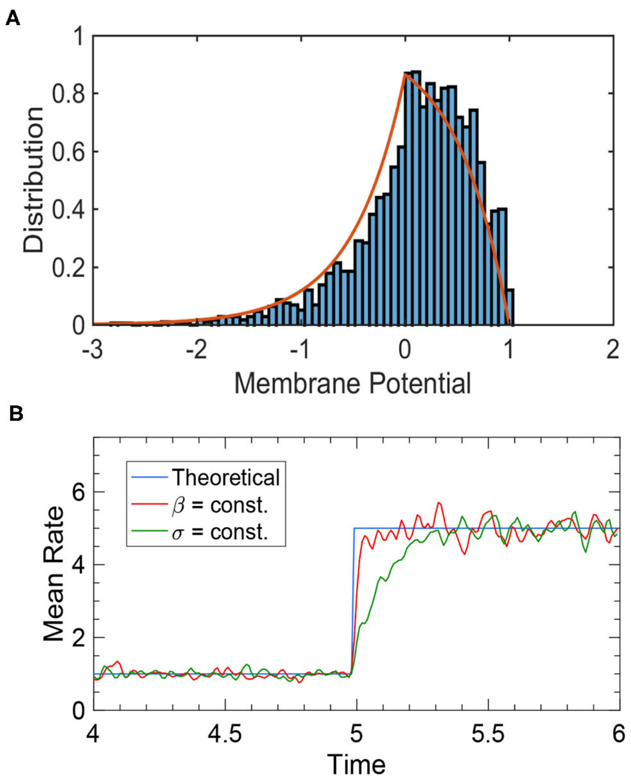


FIGURE 2 | Simulation results of an uncoupled neural population. **(A)** The membrane potential distribution of a neural population receiving independent white noise-corrupted signals with a constant VMR of 1. The red curve is the theoretical prediction given by Equation (18), and the blue histogram is the actual simulation result. **(B)** The tracking performance of a neural population depends on the input noise structure. The blue curve is the theoretical prediction of steady-state firing rate given by Equation (19). The red curve is the network performance when the VMR is constant ($\beta = 1$), which tracks the input change almost instantaneously. The green curve is the network performance when the noise variance, rather than the VMR, is constant ($\sigma \equiv 1$), where a significant delay is present. Other parameters are: $N = 2,500$, $\tau = 1$, $\theta = 1$, and μ changing from 1 to 5 at time $t = 5$.

(Huang et al., 2011), because it ensures that when the external input μ_F changes, the system remains in a stationary state where Equation (19), and thus Equation (7), always holds. In this way, the population rate can track input changes instantaneously and linearly encode μ_F at all times. **Figure 2B** illustrates how the response time of the population rate is determined by input noise structure.

From Equations (12) and (19), we know that when the network is at the stationary state,

$$\beta_a = \frac{\sigma_a^2}{\mu_a} = \frac{j_{aE}w_{aE}r_E + j_{aI}w_{aI}r_I + \sigma_{aF}^2}{\theta\tau_a r_a}, \qquad a = E, I.$$

From Equation (7), we know $r_E, r_I \propto \mu_F$. For the σ_{aF}^2 term, if we continue to assume that the external input comes from the Poisson spike trains of another population of neurons, and the changes in μ_F are due to the firing rate of that population, then we have $\sigma_{aF}^2 \propto \mu_F$. Thus, when μ_{aF} changes, β_a remains constant. This is the ideal noise structure, and the population rate of the network can track the external input instantaneously. In reality, the ideal noise structure can only be approximately satisfied, but the tracking speed of the network is still reasonably fast, as confirmed by **Figure 3**.

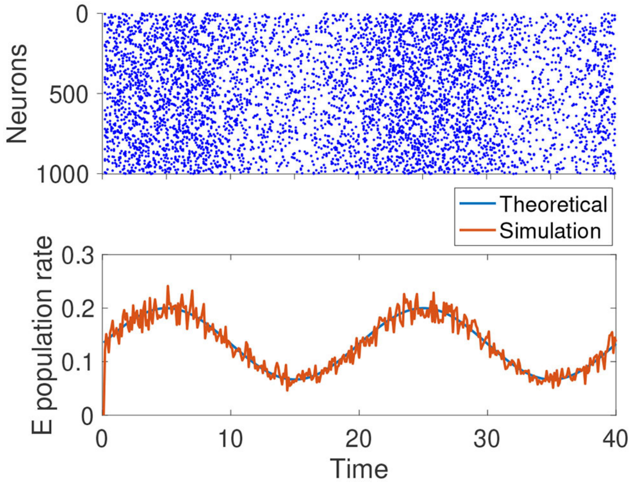


FIGURE 3 | Simulation results of a homogeneous E-I balanced network tracking a time-varying input. The network receives a sinusoidal input centered at $\mu_F = 0.1$ with an amplitude of 0.05. $\sigma_{aF}^2/\mu_F = 0.1$ remains constant. The blue curve is the theoretic prediction given by Equation (7). The red curve is the instantaneous average firing rate of excitatory neurons. The parameters are: $N = 1 \times 10^4, q_I = 0.2, p_{ab} = 0.25, \theta = 15, \tau_E = 15, \tau_I = 10, \tau_{E,s} = 6, \tau_{I,s} = 5, f_E = 3, f_I = 2, j_{EE} = 0.25, j_{EI} = -1, j_{IE} = 0.4, j_{II} = -1$.

It should be pointed out that the neuron model we used in this section does not have a lower bound to its membrane potential. In real applications, a reflecting barrier can be imposed at the reset potential v_0 (Fusi and Mattia, 1999). We verify that this does not affect our main results. The neuron model used in the following sections has a reflecting barrier.

3. PROCESSING SPATIALLY HETEROGENEOUS INPUT WITH LOCAL CONNECTIVITY

In the above, we have studied an E-I balanced neural network with homogeneous connectivity, which is able to track input changes rapidly. However, when the external input is spatially heterogeneous, that is, when different neurons receive inputs of different magnitudes, this homogeneous connectivity generates statistically equivalent recurrent inputs for each neuron that cannot balance the external inputs. The same R_a^i's cannot balance different F_a^i's, causing neurons to receive inputs of order $\mathcal{O}(\sqrt{N})$ and fire pathologically. In addition, the random long-range connections between neurons spread out local activities to the entire network, which blurs the spatial location of inputs. In applications, however, we often need to know not only when the signal occurs but also where it occurs. To solve this problem, we need to introduce local connectivity in the network. Previous studies have shown that if appropriate local connectivity is included, the network can maintain the balanced firing state as well as retain the spatial information of the input (Rosenbaum and Doiron, 2014; Rosenbaum et al., 2017), which enables the network to achieve both fast tracking and spatial location encoding. Below, we briefly introduce the balancing conditions and the response property of an E-I balanced neural network with local connectivity.

Here, each neuron is assigned a location (x, y) on the plane, and local connectivity is achieved by a connection probability that decays with the spatial distance between pairs of neurons instead of being homogeneous as in the previous sections, so that neurons closer to each other have higher probabilities to connect with each other. Specifically, the probability of a connection between neurons i and j follows

$$\mathbb{P}(j \text{ connects to } i) \propto G_b\left(d_{ij}\right), \tag{20}$$

where G_b is a 2-dimensional Gaussian shaped function whose spatial spread is determined by the presynaptic population b, and d_{ij} is the distance between the neurons.

Similar to Equation (4), we again utilize the mean-field approximation. Only this time, we do not average over the entire population, but rather approximate the neural activity of population a near location $\mathbf{x}$ with the neural field

$$\overline{I_a}(\mathbf{x}) = \overline{F_a}(\mathbf{x}) + \overline{R_a}(\mathbf{x}) = \sqrt{N}[f_a(\mathbf{x}) + w_{aE} * r_E(\mathbf{x}) - w_{aI} * r_I(\mathbf{x})], \quad a = E, I, \tag{21}$$

where the feedforward input $\overline{F}_a(\mathbf{x}) = \sqrt{N}f_a(\mathbf{x})$, $w_{ab}(\mathbf{x}) = q_b j_{ab} p_{ab} G_b(\mathbf{x})$ is the mean connectivity a neuron in population a receives from neurons in population b at location $\mathbf{x}$, and $r_a(\mathbf{x})$ is the firing rate. The symbol $*$ denotes the spatial convolution against $\mathbf{x}$.

Similar to section 2.2, we have

$$w_{aE} * r_E(\mathbf{x}) - w_{aI} * r_I(\mathbf{x}) + f_a(\mathbf{x}) \sim \mathcal{O}(1/\sqrt{N}), \quad a = E, I. \tag{22}$$

Let $N \to \infty$ and perform 2-dimensional Fourier transform against $\mathbf{x}$, and we get

$$\tilde{w}_{aE}\tilde{r}_E - \tilde{w}_{aI}\tilde{r}_I + \tilde{f}_a = 0, \quad a = E, I,$$

where the symbol $\tilde{}$ denotes the spatial Fourier transform. This gives

$$\tilde{r}_E = \frac{\tilde{f}_E\tilde{w}_{II} - \tilde{f}_I\tilde{w}_{EI}}{\tilde{w}_{EI}\tilde{w}_{IE} - \tilde{w}_{EE}\tilde{w}_{II}}, \quad \tilde{r}_I = \frac{\tilde{f}_E\tilde{w}_{IE} - \tilde{f}_I\tilde{w}_{EE}}{\tilde{w}_{EI}\tilde{w}_{IE} - \tilde{w}_{EE}\tilde{w}_{II}}. \tag{23}$$

To ensure that the above Fourier transform exists, it is necessary that $\tilde{r}_a$ tends to 0 as the frequency tends to infinity. This requires that the external input f be "wider" than recurrent input w. This can be understood intuitively from Equation (22), where we see convolution makes $w_{ab} * r_b(\mathbf{x})$ wider than $w_{ab}(\mathbf{x})$, so for the terms to balance each other, f has to be "wider" than w. Also, to get a positive stable solution, the following condition has to be met:

$$\frac{\overline{f}_E}{\overline{f}_I} > \frac{\overline{w}_{EI}}{\overline{w}_{II}} > \frac{\overline{w}_{EE}}{\overline{w}_{IE}}, \tag{24}$$

where the bar represents spatial average. Also, to make the solution stable, w_{aE} has to be "wider" than w_{aI}. For a more detailed account of these conditions, see Rosenbaum and Doiron, 2014; Pyle and Rosenbaum, 2017.

Rosenbaum et al. (2017) proved the asynchronous firing state of the network with local connections under the above conditions. Thus, with the premise of asynchronous firing satisfied, our results regarding the optimum noise structure in section 2.3 still holds. Let the total input variance of the neuron in population a at location $\mathbf{x}$ be $\sigma_a^2(\mathbf{x})$, and the VMR be $\beta_a(\mathbf{x})$, and we have

$$\sigma_a^2(\mathbf{x}) = j_{aE}w_{aE} * r_E(\mathbf{x}) + j_{aI}w_{aI} * r_I(\mathbf{x}).$$

The threshold-linear transfer function gives us $\overline{I_a}(\mathbf{x}) = \theta\tau_a r_a(\mathbf{x})$, so we have

$$\beta_a(\mathbf{x}) = \frac{j_{aE}w_{aE} * r_E(\mathbf{x}) + j_{aI}w_{aI} * r_I(\mathbf{x})}{\theta\tau_a r_a(\mathbf{x})},$$

Here the division is point-wise at each $\mathbf{x}$. If $\beta_a(\mathbf{x})$ is constant at each $\mathbf{x}$ for arbitrary external input $f_a(\mathbf{x})$, it must be spatially invariant, that is, $\beta_a(\mathbf{x}) \equiv \beta_a$. We can thus move the denominator on the r.h.s. to the left, and perform Fourier transform to get

$$\beta_a\theta\tau_a\tilde{r}_a = j_{aE}\tilde{w}_{aE} * \tilde{r}_E(\mathbf{x}) + j_{aI}\tilde{w}_{aI} * \tilde{w}_I(\mathbf{x}), \quad a = E, I.$$

Substituting it in Equation (23), we get

$$-\frac{j_{EE}\tilde{w}_{EE}}{j_{EI}\tilde{w}_{EI}} = \frac{\tilde{f}_E\tilde{w}_{IE} - \tilde{f}_I\tilde{w}_{EE}}{\tilde{f}_E\tilde{w}_{II} - \tilde{f}_I\tilde{w}_{EI}},$$

$$-\frac{j_{II}\tilde{w}_{II}}{j_{IE}\tilde{w}_{IE}} = \frac{\tilde{f}_E\tilde{w}_{II} - \tilde{f}_I\tilde{w}_{EI}}{\tilde{f}_E\tilde{w}_{IE} - \tilde{f}_I\tilde{w}_{EE}}.$$

The above equations cannot be satisfied for all f_a, so this network structure cannot maintain an optimum noise structure and track any input instantly. However, for input changes that only concerns magnitude and not the spatial shape, $\beta_a(\mathbf{x})$ can remain constant and allow instant tracking. For other kinds of input changes, although instantaneous tracking is not possible, the response speed of the network is still significantly smaller than what the neuronal time constant allows, as we shall explore in the next section.

4. SIMULATION RESULTS

One of the potential applications of the balanced network's fast response property is to process Spike Camera data in real time. Spike Camera is a newly developed neuromorphic hardware that encodes visual signals with spikes (Dong et al., 2017). It consists of artificial ganglion cells, each corresponding to a pixel, that linearly integrate the luminance intensity and fire a spike upon reaching the threshold, converting continuous visual information to discrete spikes. This event-based data transmission method significantly reduces the data volume and allows for a sampling rate of as high as 40,000 fps. Compared to another extremely

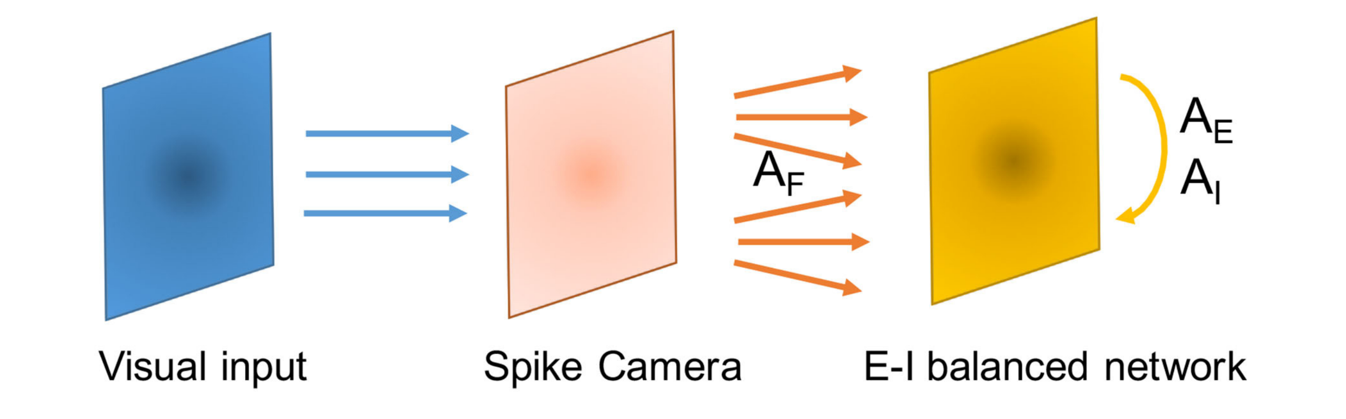


FIGURE 4 | Schematic of the network structure in the context of processing Spike Camera data.

high-speed camera, the Dynamic Vision Sensor (DVS) (Serrano-Gotarredona and Linares-Barranco, 2013), which only transmits changes in light intensity, Spike Camera can directly encode the absolute value of the luminance signal with its spiking rate while having an even higher sampling rate. In this section, we explore the tracking performance of our network under the setting of processing Spike Camera-like data.

4.1. Network Structure

We use a feedforward layer consisting of 50 × 50 non-leaky linear integrate-and-fire neurons to mimic the Spike Camera. Each neuron in this layer receives visual signal from its corresponding pixel location, and connects to the balanced network layer through feedforward connections $J_{ij}^{aF}, a = E, I$. The balanced network layer consists of 80 × 80 excitatory neurons and 40 × 40 inhibitory neurons. The neurons of each population is placed uniformly on a square area with a side length of 1. The neurons in the feedforward layer obeys Equation (1), and have a neuronal time constant of τ_F. To reflect the high sampling rate of Spike Camera, τ_F is set to be very small. The connection probability of the network obeys

$$\mathbb{P}(J_{ij}^{ab} = j_{ab}/\sqrt{N}) = p_{ab}G_b(d_{ij}^{ab}), \quad b = F, E, I, \quad a = E, I,$$

where F stands for the feedforward layer, G_b is a 2-dimensional Gaussian distribution centered at 0 with scale parameter A_b. To satisfy the balancing conditions, we let $A_F > A_E \geqslant A_I$ and make sure that Equation (24) holds. Since spatial location is discretized in the network, to keep the total connection probability from population b to population a at p_{ab}, we normalize G_b by letting $\sum_i G_b(d_{ij}^{ab}) = 1, \forall j$. **Figure 4** demonstrates this structure.

4.2. Tracking Time-Varying Stimuli

We test the tracking performance of our network with four example input stimuli. The first stimulus is the sudden appearance of an object, modeled as an abrupt change in input magnitude at the object's location. **Figure 5A** shows the network's response to this change summarized by the population rate of the excitatory neurons corresponding to the location of interest. We see that in this case, the network's activity tracks the stimulus change very quickly.

The second stimulus is similar to the previous one, except that the input magnitude continuously changes in a sinusoidal manner. **Figure 5B** shows the tracking performance of the network. It can be seen that the network can track the stimulus almost instantaneously, which is expected since β_E is constant here.

The third stimulus is an object moving quickly from left to right in the field of vision, which can be seen as a model of a typical motion tracking task. We use the coordinates of the center of the circular object to represent the location of the stimulus. The coordinates calculated from the Spike Camera data and the balanced network activity are then compared in **Figures 5C,D**. The network activity closely tracks the input, and the spatial information is preserved.

The last stimulus is similar to the previous one, except that the motion is circular instead of linear, which implies a constantly changing velocity. The same method is used to locate the stimulus, and the results are shown in **Figures 5E,F**. The performance is again very good.

4.3. Trackable Speeds

To explore the extent of the network's tracking ability, we next evaluate the temporal and spatial lags of the response. We first change the frequency of the sinusoidal signal in the second task in the previous section (**Figure 5B**) and calculate the phase lag of the balanced network's response. As can be seen in **Figure 6A**, while the phase lag $|\phi|$ increases when the signal frequency $1/T$ is higher, the delay is still very small overall.

Next, we vary the speed of the object's circular motion in the fourth task in the previous section (**Figures 5E,F**) and evaluate the spatial phase lag of the object location decoded from the balanced network activity compared to that of the Spike Camera layer. As shown in **Figure 6B**, the tracking error is small even when the object is moving very quickly.

Since the encoding happens at the population level, input changes have to be propagated through the population to be successfully tracked, and this process is mediated by synaptic

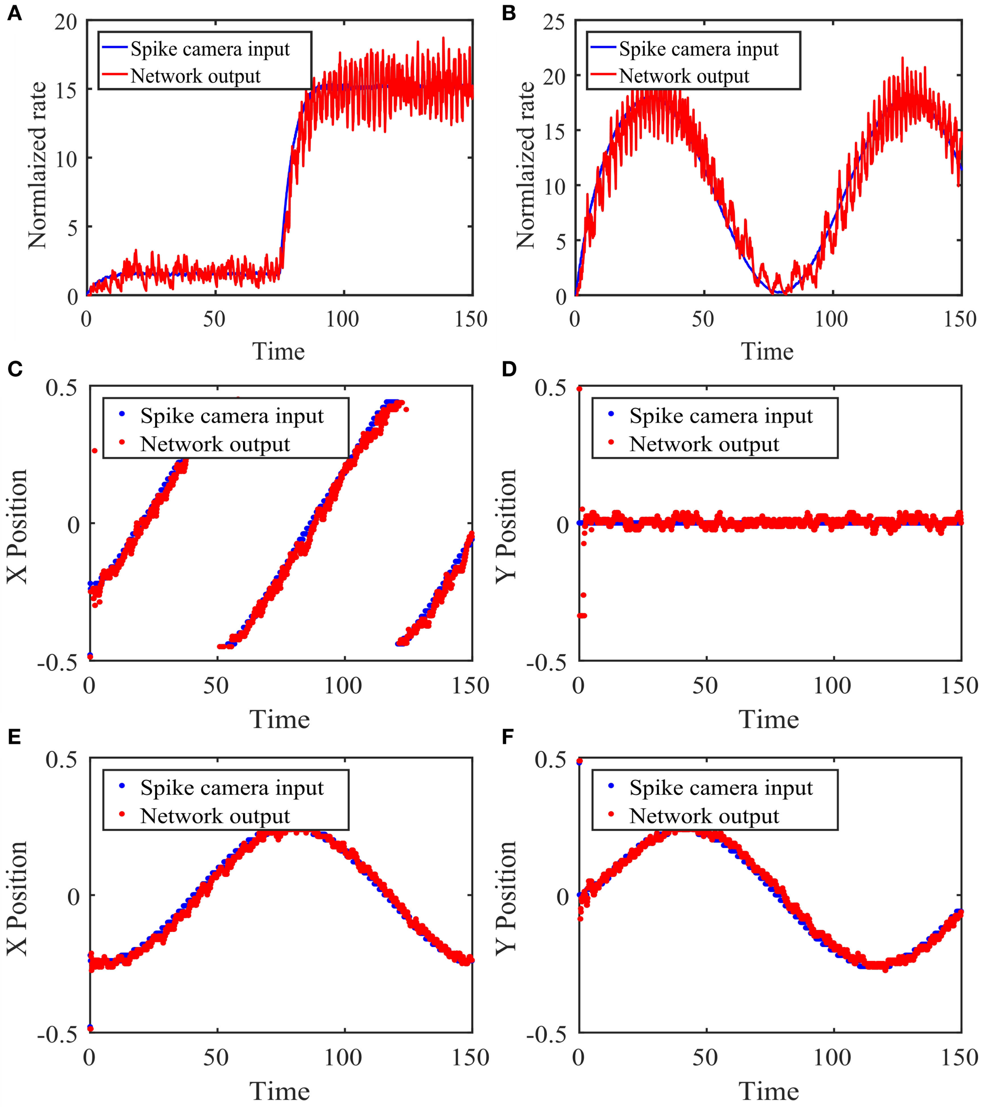


FIGURE 5 | Performance of the network with local connections in response to time-varying stimuli. **(A)** Network response to the sudden appearance of an object. The Spike Camera layer receives a disc-shaped visual input centered at (0.25, 0.5) with a radius of 0.05, whose magnitude changes abruptly from 1.5 to 15 at $t = 75$. A background noise is added. The blue curve is the firing rate of the area corresponding to the visual input in the Spike Camera layer. The red curve is the rate of the excitatory neurons at the same area in the balanced network layer, which is normalized for better comparison with the blue curve. **(B)** Same as panel **(A)**, except that the input amplitude follows the sinusoidal function $\mu(t) = A(\sin(B * 2\pi t/T)) + C, A = 30, B = 3/2, C = 30$. **(C,D)** The stimulus is an object moving across the visual field in constant velocity. The object has the same shape as panels **(A,B)**, with a magnitude of 10. Panels **(C,D)** show the tracking of the x and y coordinates, respectively. The blue curve is the object location decoded from the activity of the Spike Camera layer, and the red curve is that of the balanced network layer. **(E,F)** Same as panels **(C,D)**, except that the stimulus moves counterclockwise on a circle in constant speed. The network parameters are $\theta = 15, \tau_F = 1, \tau_E = 15, \tau_I = 10, \tau_{F,s} = \tau_{E,s} = 5, \tau_{I,s} = 2.5, p_{EF} = 0.05, p_{IF} = 0.025, p_{EE} = 0.02, p_{EI} = 0.08, p_{IE} = 0.06, p_{II} = 0.08, A_F = 0.05, A_E = 0.02, A_I = 0.02, j_{EF} = 140, j_{IF} = 93.3, j_{EE} = 80, j_{EI} = -320, j_{IE} = 40, j_{II} = -320$.

interactions. This lead us to suspect that the synaptic time constants $\tau_{b,s}$ could be a limiting factor for tracking performance. To study this, we varied $\tau_{b,s}$ in both the temporal and spatial tracking tasks. Indeed, as can be seen in **Figure 6**, a shorter synaptic time constant leads to better performance. In practice, the shape of the synaptic current can be designed to have a $\tau_{b,s}$ as small as possible. The real constricting factor is the synaptic transmission delay, which corresponds to the communication speed of the hardware, but this is expected to be insignificant given the highly compact nature of neuromorphic chips.

5. DISCUSSION AND CONCLUSION

This paper proposed an algorithm for fast response in neuromorphic systems based on E-I balanced networks,

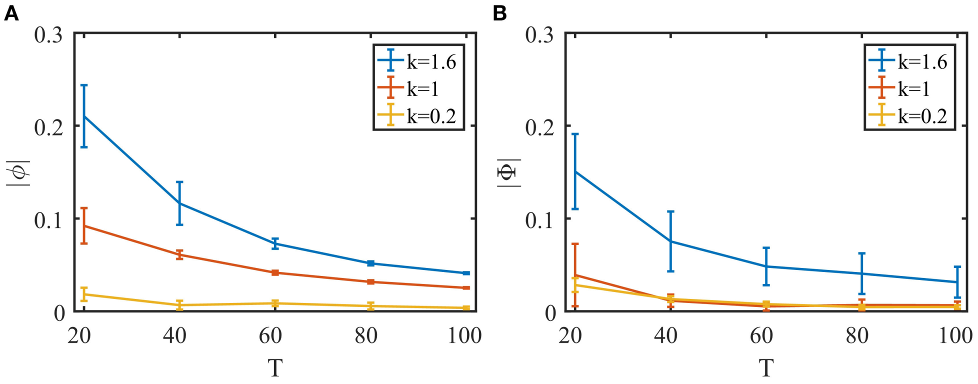


FIGURE 6 | Quantifying the network's performance with temporal and spatial phase lag. To examine the effect of the synaptic time constant on tracking performance, we define $\tau_{b,s} = k\tau_{b,s0}, b = F, E, I$, where $\tau_{b,s0}$ is the set of parameters used in **Figure 5**. **(A)** Temporal phase lag in the second task (**Figure 5B**) with different signal periods. **(B)** Spatial phase lag in the fourth task (**Figures 5E,F**) with different circular motion periods. Ten trials for each data point. Error bars show standard deviations.

systematically analyzed its fast response mechanism, and introduced local connections to maintain balance and retain spatial information in the face of spatially heterogeneous inputs. Simulations verified that the network indeed performs well with rapidly changing input stimuli.

There are still some questions left to explore. For instance, we have mentioned that the network cannot keep an optimal noise structure at all times, and thus the membrane potential distribution will change with the input. A study of the transient dynamics during such changes could help us further improve the network performance. As another example, notice that most of the theoretical analyses in the paper were conducted in the limit of $N \to \infty$. In real-world applications, we often have to track small objects, during which the number of neurons encoding it usually does not exceed a few hundred. Studying the finite-size effect could help us better understand the network dynamics.

Although we mainly discussed the case where the input comes from Spike Camera, the network structure we proposed is not limited to processing visual signal. The "location" of neurons can also correspond to tuning to different variables or representation of abstract features. To achieve real-time processing of high-frequency data, the fast response property is required for each computational process. There has been a lot of research discussing how to implement various computations on top of a balanced network (Barrett, 2012; Hansel and van Vreeswijk, 2012; Litwin-Kumar and Doiron, 2012; Lim and Goldman, 2014; Denève and Machens, 2016; Pyle and Rosenbaum, 2017). The asynchronous irregular state can be taken as a model of the spontaneous state in the cortex. With the spontaneous state as a global attractor, and the specific computations and memories as input-sensitive local attractors (Amit and Brunel, 1997; Litwin-Kumar and Doiron, 2012), the chaos in the network's balanced firing state can allow it to respond to specific inputs very rapidly and initiate the required computation. Besides the fast response property, the balanced state also has other computational advantages such as stochastic resonance (Barrett, 2012).

Neuromorphic computing systems colocalize computation and memory by mimicking neural structures like neurons and synapses. This allows it to circumvent the von Neumann bottleneck, granting it enormous potentials in processing speed (Indiveri and Liu, 2015). There has been a lot of work investigating possible mechanisms for fast neural response (e.g., Bharioke and Chklovskii, 2015; Yu et al., 2015) which could potentially complement the processing speed of neuromorphic systems, and the balance of excitation and inhibition we explored here is one of them. The model we proposed here, with its simple neuron model and connectivity structure, can be readily implemented in hardware and serve as a fast-responding module integrated in a general neuromorphic system for rapid information processing. This paper thus lays the groundwork for realizing various kinds of fast computation using balanced networks, especially in neuromorphic systems.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

SW designed the project. GT, SL, and SW wrote the paper. GT did the theoretical analyses. GT, SL, and SW carried out simulations and data analysis. TH contributed important ideas.

FUNDING

This work was supported by the National Natural Science Foundation of China (No: 61425025, TH) and Guangdong Province with Grant (No. 2018B030338001, SW). This work also received support from Huawei Technology Co., Ltd (YBN2019105137). The authors declare that this study received funding from Huawei Technology Co., Ltd. The funder was not involved in the study design, collection, analysis, interpretation of data, the writing of this article or the decision to submit it for publication.

REFERENCES

Amit, D. J., and Brunel, N. (1997). Model of global spontaneous activity and local structured activity during delay periods in the cerebral cortex. *Cereb. Cortex* 7, 237–252. doi: 10.1093/cercor/7.3.237

Barrett, D. G. T. (2012). *Computation in balanced networks* (Ph.D. thesis). University College London, London, United Kingdom.

Bharioke, A., and Chklovskii, D. B. (2015). Automatic adaptation to fast input changes in a time-invariant neural circuit. *PLoS Comput. Biol*. 11:e1004315. doi: 10.1371/journal.pcbi.1004315

Brunel, N. (2000). Dynamics of sparsely connected networks of excitatory and inhibitory spiking neurons. *J. Comput. Neurosci*. 8, 183–208. doi: 10.1023/A:1008925309027

Brunel, N., and Hakim, V. (1999). Fast global oscillations in networks of integrate-and-fire neurons with low firing rates. *Neural Comput*. 11, 1621–1671. doi: 10.1162/089976699300016179

Denéve, S., and Machens, C. K. (2016). Efficient codes and balanced networks. *Nat. Neurosci*. 19, 375–382. doi: 10.1038/nn.4243

Dong, S., Huang, T., and Tian, Y. (2017). "Spike camera and its coding methods," in *2017 Data Compression Conference (DCC)* (Snowbird, UT), 437. doi: 10.1109/DCC.2017.69

Dorrn, A. L., Yuan, K., Barker, A. J., Schreiner, C. E., and Froemke, R. C. (2010). Developmental sensory experience balances cortical excitation and inhibition. *Nature* 465, 932–936. doi: 10.1038/nature09119

Fusi, S., and Mattia, M. (1999). Collective behavior of networks with linear (VLSI) integrate-and-fire neurons. *Neural Comput*. 11, 633–652. doi: 10.1162/089976699300016601

Graupner, M., and Reyes, A. D. (2013). Synaptic input correlations leading to membrane potential decorrelation of spontaneous activity in cortex. *J. Neurosci*. 33, 15075–15085. doi: 10.1523/JNEUROSCI.0347-13.2013

Haider, B., Duque, A., Hasenstaub, A. R., and McCormick, D. A. (2006). Neocortical network activity *in vivo* is generated through a dynamic balance of excitation and inhibition. *J. Neurosci*. 26, 4535–4545. doi: 10.1523/JNEUROSCI.5297-05.2006

Hansel, D., and van Vreeswijk, C. (2012). The mechanism of orientation selectivity in primary visual cortex without a functional map. *J. Neurosci*. 32, 4049–4064. doi: 10.1523/JNEUROSCI.6284-11.2012

Huang, L., Cui, Y., Zhang, D., and Wu, S. (2011). Impact of noise structure and network topology on tracking speed of neural networks. *Neural Netw*. 24, 1110–1119. doi: 10.1016/j.neunet.2011.05.018

Indiveri, G., and Liu, S. (2015). Memory and information processing in neuromorphic systems. *Proc. IEEE* 103, 1379–1397. doi: 10.1109/JPROC.2015.2444094

Lim, S., and Goldman, M. S. (2014). Balanced cortical microcircuitry for spatial working memory based on corrective feedback control. *J. Neurosci*. 34, 6790–6806. doi: 10.1523/JNEUROSCI.4602-13.2014

Litwin-Kumar, A., and Doiron, B. (2012). Slow dynamics and high variability in balanced cortical networks with clustered connections. *Nat. Neurosci*. 15, 1498–1505. doi: 10.1038/nn.3220

Okun, M., and Lampl, I. (2008). Instantaneous correlation of excitation and inhibition during ongoing and sensory-evoked activities. *Nat. Neurosci*. 11, 535–537. doi: 10.1038/nn.2105

Pyle, R., and Rosenbaum, R. (2017). Spatiotemporal dynamics and reliable computations in recurrent spiking neural networks. *Phys. Rev. Lett*. 118:018103. doi: 10.1103/PhysRevLett.118.018103

Raiguel, S. E., Xiao, D.-K., Marcar, V. L., and Orban, G. A. (1999). Response latency of macaque area MT/V5 neurons and its relationship to stimulus parameters. *J. Neurophysiol*. 82, 1944–1956. doi: 10.1152/jn.1999.82.4.1944

Renart, A., de la Rocha, J., Bartho, P., Hollender, L., Parga, N., Reyes, A., et al. (2010). The asynchronous state in cortical circuits. *Science* 327, 587–590. doi: 10.1126/science.1179850

Risken, H. (1996). *Fokker-Planck Equation*. Berlin; Heidelberg: Springer. doi: 10.1007/978-3-642-61544-3_4

Rosenbaum, R., and Doiron, B. (2014). Balanced networks of spiking neurons with spatially dependent recurrent connections. *Phys. Rev. X* 4:021039. doi: 10.1103/PhysRevX.4.021039

Rosenbaum, R., Smith, M. A., Kohn, A., Rubin, J. E., and Doiron, B. (2017). The spatial structure of correlated neuronal variability. *Nat. Neurosci*. 20, 107–114. doi: 10.1038/nn.4433

Serrano-Gotarredona, T., and Linares-Barranco, B. (2013). A 128 × 128 1.5 mw asynchronous frame-free dynamic vision sensor using transimpedance preamplifiers. *IEEE J. Solid State Circ*. 48, 827–838. doi: 10.1109/JSSC.2012.2230553

Shadlen, M. N., and Newsome, W. T. (1994). Noise, neural codes and cortical organization. *Curr. Opin. Neurobiol*. 4, 569–579. doi: 10.1016/0959-4388(94)90059-0

Softky, W. R., and Koch, C. (1993). The highly irregular firing of cortical cells is inconsistent with temporal integration of random EPSPs. *J. Neurosci*. 13, 334–350. doi: 10.1523/JNEUROSCI.13-01-00334.1993

Sugase, Y., Yamane, S., Ueno, S., and Kawano, K. (1999). Global and fine information coded by single neurons in the temporal visual cortex. *Nature* 400, 869–873. doi: 10.1038/23703

Thorpe, S., Fize, D., and Marlot, C. (1996). Speed of processing in the human visual system. *Nature* 381, 520–522. doi: 10.1038/381520a0

van Vreeswijk, C., and Sompolinsky, H. (1996). Chaos in neuronal networks with balanced excitatory and inhibitory activity. *Science* 274, 1724–1726. doi: 10.1126/science.274.5293.1724

van Vreeswijk, C., and Sompolinsky, H. (1998). Chaotic balanced state in a model of cortical circuits. *Neural Comput*. 10, 1321–1371. doi: 10.1162/089976698300017214

Yu, L., Wang, L., Jia, F., and Jia, D. (2015). Stimulus-dependent frequency modulation of information transmission in neural systems. *arXiv [preprint]. arXiv:1507.08269*.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Computational Neuroscience

ORIGINAL RESEARCH
published: 14 October 2020
doi: 10.3389/fncom.2020.571982

A Neural Network Model With Gap Junction for Topological Detection

Chaoming Wang [1,2,3], Risheng Lian [1], Xingsi Dong [1], Yuanyuan Mi [4] and Si Wu [1,2]**

[1] Peking-Tsinghua Center for Life Sciences, School of Electronics Engineering and Computer Science, IDG/McGovern Institute for Brain Research, Peking University, Academy for Advanced Interdisceplinary Studies, Beijing, China, [2] Hefei Comprehensive National Science Center, Institute of Artificial Intelligence, Hefei, China, [3] Chinese Institute for Brain Research, Beijing, China, [4] Center for Neurointelligence, School of Medicine, Chongqing University, Chongqing, China

Visual information processing in the brain goes from global to local. A large volume of experimental studies has suggested that among global features, the brain perceives the topological information of an image first. Here, we propose a neural network model to elucidate the underlying computational mechanism. The model consists of two parts. The first part is a neural network in which neurons are coupled through gap junctions, mimicking the neural circuit formed by alpha ganglion cells in the retina. Gap junction plays a key role in the model, which, on one hand, facilitates the synchronized firing of a neuron group covering a connected region of an image, and on the other hand, staggers the firing moments of different neuron groups covering disconnected regions of the image. These two properties endow the network with the capacity of detecting the connectivity and closure of images. The second part of the model is a read-out neuron, which reads out the topological information that has been converted into the number of synchronized firings in the retina network. Our model provides a simple yet effective mechanism for the neural system to detect the topological information of images in ultra-speed.

Keywords: global first, topological perception, gap junction, electrical synapse, subcortical pathway, ipRGCs, alpha RGCs, superior colliculus

OPEN ACCESS

Edited by:
Pulin Gong,
The University of Sydney, Australia

Reviewed by:
Changsong Zhou,
Hong Kong Baptist University, Hong Kong
James A. Roberts,
The University of Queensland, Australia

****Correspondence:***
Yuanyuan Mi
miyuanyuan0102@cqu.edu.cn
Si Wu
siwu@pku.edu.cn

Received: *12 June 2020*
Accepted: *02 October 2020*
Published: *14 October 2020*

Citation:
Wang C, Lian R, Dong X, Mi Y and Wu S (2020) A Neural Network Model With Gap Junction for Topological Detection. Front. Comput. Neurosci. 14:571982. doi: 10.3389/fncom.2020.571982

1. INTRODUCTION

It has been a long-standing debate in the field concerning whether feature analysis in visual information processing goes from local to global, or from global to local (Palmer, 1999; Chen, 2005b). The former claims that the primitives of visual processing are local features of objects. This view has successfully explained a large number of experimental phenomena (Hubel and Wiesel, 1959; Treisman and Gelade, 1980; Marr, 1982; Hubel, 1988; DiCarlo et al., 2012), but failed to account for others where visual systems show superior sensitivity to global features, e.g., the topological perception (Chen, 1982, 2005b), the configural-superiority effect (Weisstein and Harris, 1974; Navon, 1977; Pomerantz et al., 1977), the holistic processing of face and objects (Farah et al., 1998; McKone et al., 2007; Goffaux et al., 2010; Taubert et al., 2011; Bona et al., 2016), and Gestalt psychology (Wagemans et al., 2012). On the other hand, the global-to-local view states that in the visual processing, global features of objects are processed first, which subsequently guide the processing of local features (Hegdé, 2008).

In the framework of global-to-local processing, Chen et al. went one step further to argue that the global nature of visual perception can be described by topological invariants and that the global precedence actually is topological primacy (see review Chen, 2005b). Topology is defined as the geometric properties which are preserved under continuous transformations, such as stretching and bending (Armstrong, 2013), and important topological properties include connectivity and the number of holes. Two shapes are called topologically different, as long as they differ in either the connectivity or the number of holes (**Figure 1**). Over decades, accumulating evidences on adults, infants and animals have demonstrated that visual systems are highly sensitive to topological features. The pioneering work of Chen (1982) first revealed that in the adult human visual system, the topological perception is prior to perceptions of other geometrical properties. Specifically, under 5-ms stimulus presentation, he found that subjects could discriminate a disc vs. a ring (which are topologically different) with a much higher accuracy than a disc vs. a square or a triangle (which are topologically same but different in other geometrical properties). Later, in other tasks, including multiple-object tracking (Zhou et al., 2010) and long-range apparent motion perception (Zhuo et al., 2003), Chen et al. further confirmed that the human visual perception is indeed sensitive to the connectivity or the hole of stimuli. The studies on infants also support the precedence of topological perception (Piaget and Inhelder, 1956; Darke, 1982; Chien et al., 2012; Kibbe and Leslie, 2016). It was found that newborns, even as young as few days old, display the preference of using the topological information to discriminate objects (Turati et al., 2003). Furthermore, animal studies provide more evidence to support the notion that topological perception is primitive in the visual processing. For example, Chen et al. (2003) found that honey bees with small brains have the ability to distinguish patterns that are topologically different after only a few trials learning, and they could even generalize the learned figure to novel patterns never seen before. Experiments from other researchers also demonstrated that chicks (Versace et al., 2016) and pigeons (Watanabe et al., 2019) use topological features as cues for discriminating objects.

Altogether, it suffices to say that topological properties are essential for visual perception, and very likely, they are the primitives of visual perception. Computationally, using topological features to represent and characterize objects has advantages, as it provides a relatively stable way to represent objects under transformations like stretching, rotation, or distortion. Although it is coarse, topology discrimination enables animals to detect the presence of objects rapidly without detailed local feature analysis, and this is crucial for animals to survive in natural environments.

Despite topological perception has been well-documented in the literature, the detail mechanism of how the neural system implements it remains largely unclear. It is a known fact that the conventional artificial feedforward neural network has difficulty to recognize the topology of images (McClelland et al., 1987; Minsky and Papert, 1987; Wang, 2000; Chen, 2005b). Recently, a number of experimental findings indicate that topology perception in the brain is carried out via the subcortical pathway from retina to superior colliculus (SC) and then to higher cortex. First, electrophysiological studies on retinal ganglia cells (RGCs) have revealed that there exists a type of RGCs, called alpha RGCs, which are specialized to encode the global features of stimuli (Neuenschwander and Singer, 1996; Roy et al., 2017). Specifically, they found that the presentation of a contiguous stimulus, rather than disjointed local features, produced long-range synchronization among widely separated alpha RGCs (Neuenschwander and Singer, 1996; Roy et al., 2017), and importantly, the occurrence of this kind of synchronization relies on gap junctions (also called electrical synapses) between neurons (Völgyi et al., 2013; Roy et al., 2017). Second, psychophysical and neuroimaging studies on humans have indicated that SC, rather than the primary visual cortex (V1), plays an important role in topological perception. For example, Turati et al. (2003) showed that despite of their immature visual cortex, newborns of 2–3 days old were able to detect and discriminate perceptual similarity based on the hole feature. Also, it was found that aging (Meng et al., 2019) and disruption of V1 (Du et al., 2011) significantly reduced human's ability of discriminating local geometric properties, but did not affect their topological discrimination. The neuroimaging study also showed that the neural responses in SC to hole stimuli were greater than that to no-hole stimuli under the low awareness condition (Meng et al., 2018). These findings are consistent with the electrophysiological studies on SC, which unveil that the functional role of neurons in the superficial layers of SC is to encode whether there is a new object in their receptive fields (Rizzolatti et al., 1980; Girman and Lund, 2007; Ito and Feldheim, 2018), and notably, their neuronal

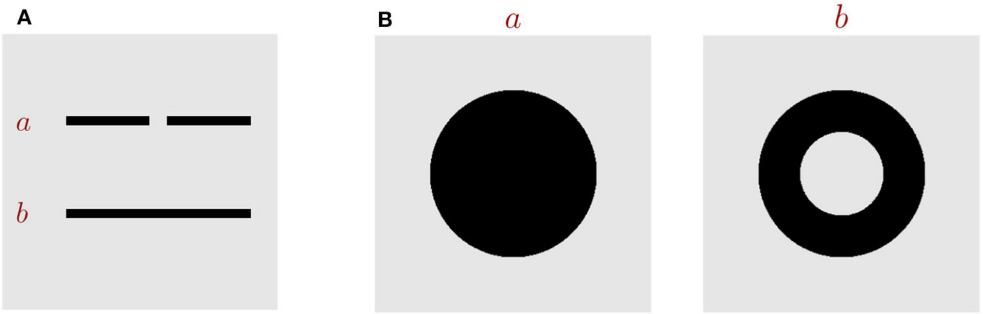


FIGURE 1 | Key topological properties. **(A)** Images *a* and *b* are topologically different in the property of connectivity. **(B)** Images *a* and *b* are topologically different in the number of holes.

responses to visual stimuli are irrelevant to specific features, such as direction, orientation or shape (Marrocco and Li, 1977; White et al., 2017a,b).

Inspired by the above experimental findings, we propose a simple computational model for topological perception in the brain. Specifically, the model consists of two parts. The first part is a neural network in which neurons are connected via gap junctions, and it models the neural circuit formed by alpha RGCs in the retina (Neuenschwander and Singer, 1996; Völgyi et al., 2013; Roy et al., 2017). The second part is a read-out neuron, which suggests a way for SC and higher cortical neurons (Marrocco and Li, 1977; Rizzolatti et al., 1980; Girman and Lund, 2007; White et al., 2017a,b; Ito and Feldheim, 2018) to read out the topological information extracted by the retina network. We elucidate the computational properties of the proposed network model, and demonstrate that the model is effective and robust for detecting holes in various visual stimuli as observed in human psychophysical experiments.

2. MATERIALS AND METHODS

We consider a two-layer spiking network model (see **Figures 2A,B** for the network architecture illustration). The first layer is the encoding layer, which is composed of 80 × 80 encoding neurons (ENs), and the second layer is the read-out layer, which consists of only one read-out neuron (RON). RON

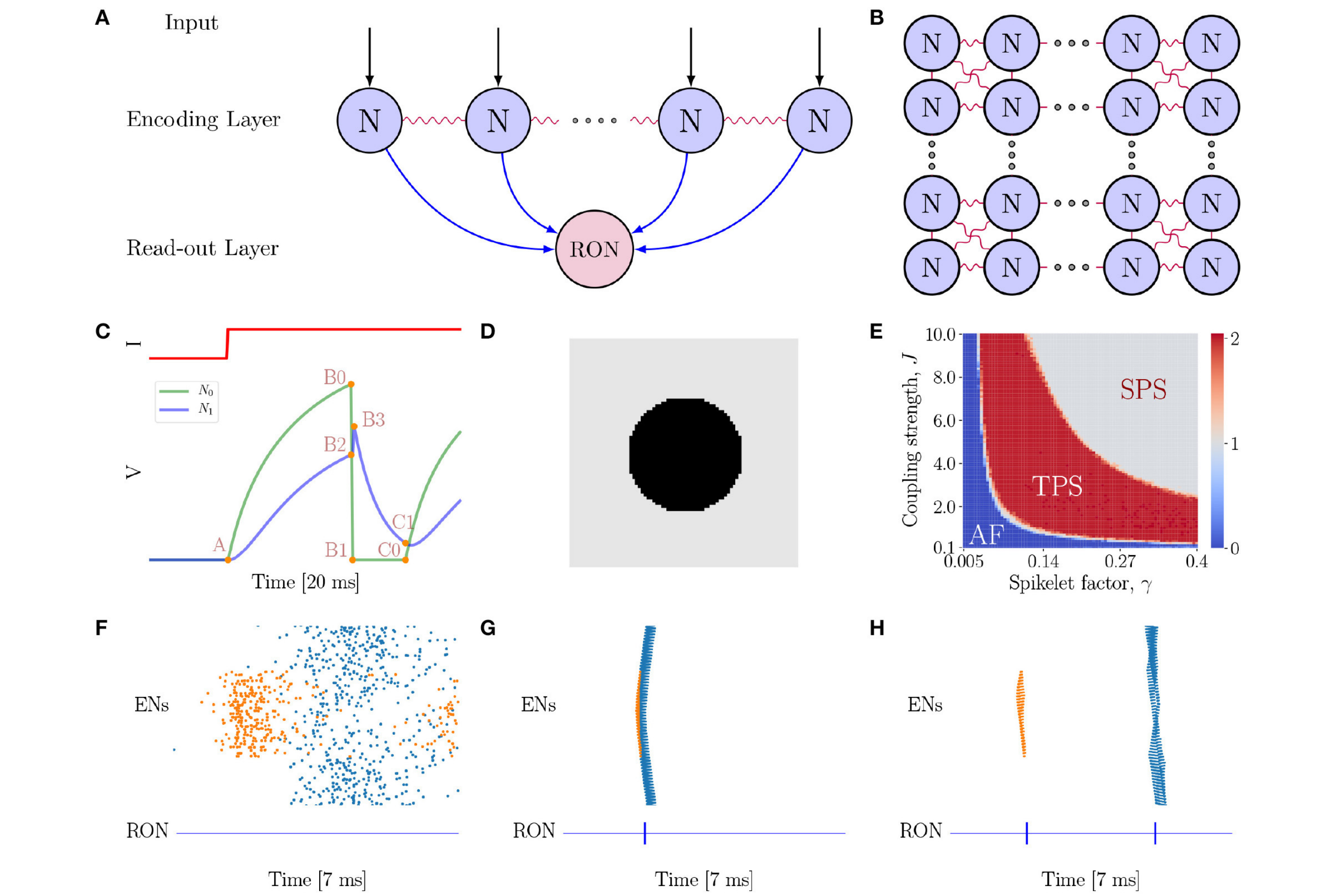


FIGURE 2 | The neural network model. **(A)** The model is composed of two layers. The first layer is the encoding layer which receives external inputs, and its function is to encode the connected regions in an image. The second layer is the read-out layer, whose function is to read-out neuronal activity patterns in the encoding layer. Notably, all neurons in the first layer project excitatory synapses to the neuron in the second layer. **(B)** Neurons in the encoding layer are uniformly distributed in the space and are coupled with eight nearest neighbors with gap junctions. **(C)** Simulation of a pair of electrically coupled neurons N_0 and N_1. The top panel shows the external input I to N_0, and the bottom panel presents the voltage dynamics of the neuron pair. N_0 exhibits excitation and inhibition effects on N_1 at different stages of the neuronal dynamics. At the A → B0 → B1 phase, N_0 shows an excitatory effect to N_1 (see N_1 rise phase A → B2 → B3); while at B1 → C0 phase (refractory period), N_0 exhibits a strong inhibitory effect to N_1 (see N_1 decay phase B3 → C1). **(D)** A full circle stimulus, containing two connected regions. **(E)** Parameter-space analysis of response behaviors of the network when the full circle stimulus (D) is presented. The phase plane shows three different spiking patterns which depend on the coupling strength J and spikelet factor γ. For each pair of (γ, J), the result is obtained by averaging over 10 trials. **(F)** The AF behavior of the network. $J = 0.5$ and $\gamma = 0.05$. **(G)** The SPS behavior of the network when the spikelet factor γ and the coupling strength J are too strong. $J = 3.0$ and $\gamma = 0.15$. **(H)** The TPS behavior of the network. $J = 6.0$ and $\gamma = 0.25$. **(F–H)** The top panel shows the raster plot of spikes in the encoding layer, while the bottom panel the spikes of RON. The abscissas and ordinates of both panels are time and neuron index, respectively. Colors indicating neurons in different groups. Specifically, coral denotes neurons on the circle, while blue denotes neurons on the background.

receives excitatory projections from all ENs, and hence can read out synchronized activities in the encoding layer.

2.1. Neuronal Dynamics

For simplicity, all neurons in the model are implemented as leaky integrate-and-fire (LIF) models. The encoding layer receives the external inputs, and each neuron is connected to its eight neighboring neurons by electrical synapses (**Figure 2B**). The dynamics of an encoding neuron is given by

$$\tau \frac{dV_i(t)}{dt} = -V_i(t) + \sum_{j \in N_G(i)} I_{ij}^{gap}(t) + I_i^{ext}(t), \tag{1}$$

where the subscript $i = (1, ..., N)$ refers to the neuron index, V_i the membrane potential of the neuron, τ the membrane time constant, I_{ij}^{gap} the current from neuron j transmitted through gap junction, $N_G(i)$ the set of neurons which are electrically coupled with the neuron i, and I^{ext} the external current from the image. Whenever $V_i(t)$ reaches a fixed threshold V_{th} (i.e., $V_i(t) \geq V_{th}$), the neuron generates a spike and its potential is reset to the rest value V_{reset}, followed by the refractory period τ^{arp}. At the onset of the simulation, membrane potentials of all neurons are randomly initialized.

The current mediated by electrical couplings is decomposed into two parts,

$$I_{ij}^{gap}(t) = I_{ij}^{gap,sub}(t) + I_{ij}^{gap,sup}(t), \tag{2}$$

where $I_{ij}^{gap,sub}$ denotes the sub-threshold current, and $I_{ij}^{gap,sup}$ the supra-threshold current, called as spikelet. The sub-threshold current mediated by electrical coupling is given by,

$$I_{ij}^{gap,sub}(t) = J[V_j(t) - V_i(t)], \tag{3}$$

where J is the coupling strength. The supra-threshold contribution is assumed to be proportional to the gap junction strength J and scaled by a spikelet factor γ, which is written as,

$$I_i^{gap,sup}(t) = \gamma J \delta(t - t_j^{spike}), \tag{4}$$

where t_j^{spike} represents the spiking moment of neuron j and γ is a parameter controlling the contribution of a spike to the increment of neuronal potential.

The external current I_i^{ext}, which conveys the luminance level of the image, is modeled as a continuous current with a Gaussian white noise, which is written as,

$$I_i^{ext}(t) = \mu_i^{ext} + \sigma^2 \eta_i(t), \tag{5}$$

where μ^{ext} is the mean of the external input, σ^2 the amplitude of input fluctuations, and $\eta_i(t)$ satisfies $\langle \eta_i(t) \rangle = 0$ and $\langle \eta_i(t)\eta_j(t') \rangle = \delta_{ij}\delta(t - t')$. Usually, the amplitude σ^2 in our simulations is set to be a value, so that the noise amplitude is around 10% compared to the mean external input.

The second layer in the model is a read-out neuron (RON) (see **Figure 2A**), which suggests a possible way for SC neurons to read out the topological information of an image that has been extracted by the encoding layer (see more discussions in Discussion section). Specifically, we consider RON receives projections from all neurons in the encoding layer, whose dynamics is given by

$$\tau_R \frac{dV_R(t)}{dt} = -V_R(t) + I_R^{chem}(t) + I_R^{noise}(t), \tag{6}$$

where V_R is the potential of RON, τ_R the time constant, I_R^{chem} the chemical synaptic current from the encoding neurons, and I_R^{noise} the background noise. Specifically, the current transmitted via chemical synapses is given by

$$I_R^{chem}(t) = \sum_{j \in N_C} J_R \delta(t - t_j - D), \tag{7}$$

where J_R denotes the chemical synaptic strength, t_j the spiking moment of the presynaptic neuron j, N_C the set of neurons in the encoding layer, and D the transmission delay of chemical synapses. For simplicity, we omit the rise and decay phases of post-synaptic currents. Since the function of the read-out layer in our model is coincidence detection, we set τ_R to be sufficiently small, such that RON will fire only when a sufficient number of neuronal spikes simultaneously arrive in a short-time window. Additionally, the background noise is set to be

$$I_R^{noise}(t) = \mu_R^{noise} + \Delta \eta_i(t), \tag{8}$$

with μ_R^{noise} and Δ are, respectively, the mean and the variance of the noise.

2.2. Simulation Experiments

In all simulations, the dynamical equations are integrated by using the Euler–Maruyama method with a fixed time-step $dt = 0.01$ ms. The network dynamics was simulated using Python, and the corresponding code the corresponding code can be available in the GitHub: https://github.com/chaoming0625/Gap_Junction_and_Topology. Parameters used in numerical simulations are reported in **Table 1**.

3. RESULTS

3.1. The Neural Network Model With Gap Junction

In our proposed model (**Figures 2A,B**), gap junction plays a key role for topological detection. The neuronal interaction mediated by gap junction exhibits two prominent properties, as illustrated in **Figure 2C**. Firstly, once a neuron fires, the spike generated by it will increase the potentials of the connected neurons rapidly, and this tends to synchronize coupled neurons in the network. Secondly, after firing, the neuron falls into the refractory period with a deep low potential, which induces strong negative currents to the connected neurons, and this tends to inhibit the firing of coupled neurons [note that $I_{ij}^{gap,sub}(t) = -V_i(t)$, when $V_j =$

TABLE 1 | Parameter of neurons, synapses, and simulation protocol.

Parameters of the encoding neurons	**Values**
V_{th}—Spike emission threshold	10 mV
V_{reset}—Reset potential	0 mV
τ—Membrane time constant	5 ms
τ^{arp}—Absolute refractory period	3.5 ms
σ^2—Variance of external current	1.0–2.0 mV
Parameters of the read-out neuron	**Values**
V_{th}—Spike emission threshold	10 mV
V_{reset}—Reset potential	0 mV
τ_R—Membrane time constant	0.05 ms
τ^{arp}—Absolute refractory period	0.5 ms
μ_R^{noise}—Mean background noise	4.0 mV
Δ—Variance of background noise	0.5 mV
Parameters of electrical couplings	**Values**
J—Gap junction strength	3.0
γ—Spikelet factor	0.15
Parameters of chemical synapses	**Values**
J_R—Chemical synaptic strength	0.15 mV
D—Chemical transmission delay	0.1 ms
Parameters of the stimuli	**Values**
I_b^{ext}—Value of black stimulus	20.0 mV
I_g^{ext}—Value of gray stimulus	12.0 mV

0]. As explained below, these two salient properties give rise to characteristic network responses which are differentiable with respect to connected and non-connected regions in an image.

As an example, consider a full black circle as in **Figure 2D** is presented to the network. The whole image consists of two connected regions, the circle and the background, which have different luminance levels. In our model, neurons covering a connected region (having the same luminance level) receive the same external input. We find that the network exhibits three response behaviors depending on the properties of gap junction (**Figure 2E**), which are: (1) Asynchronous Firing (AF, **Figure 2F**), i.e., all ENs fire independently and irregularly. This happens when both the spikelet factor γ and the coupling strength J are too small, and the neuronal interactions are very weak, leading to that neuronal firings are largely driven by external inputs with independent noises; (2) Single Population Spike (SPS, **Figure 2G**), i.e., all ENs are synchronized to generate a single population spike. This happens when the spikelet factor γ and the coupling strength J are both too large. In such a parameter regime, the synchronization effect of gap junction is too strong, leading to that all ENs are synchronized irrespective to the different external inputs they receive. (3) Two Population Spike (TPS, **Figure 2H**), i.e., ENs are synchronized but meanwhile clustered to generate two population spikes depending on the external inputs they receive. This happens when the spikelet factor γ and the coupling strength J have appropriate values, so that, on one hand, the synchronization effect of gap junction ensures that neurons covering the same connected region (receiving the same external input) are synchronized, and, on the other hand, the inhibitory effect of gap junction ensures that the synchronized firings of neuron groups covering different regions (having different luminance levels and hence receiving different external inputs) are well-separated in time. Computationally, this is due to that the neuron group receiving the larger external input will generate synchronized firing first; after that the neurons fall into the refractory period, and they will suppress and delay the synchronized firing of the other neuron group. To accomplish the topological detection task, we set the parameters of gap junction in the regime of TPS, such that the network can on one hand, generate synchronous firings to detect connected regions, and on the other hand, stagger synchronous firings of disconnected regions.

The synchronized responses of ENs can be easily detected by RON. Due to the small time constant, RON only responds to synchronized inputs from the encoding layer. As shown in **Figures 2F–H** (see the lower panels), each population spike of ENs generates a single spike of RON.

3.2. Topological Detection of the Network

The topology of an image has two fundamental features, connectivity and closedness (the existence and the number of holes). It is straightforward to understand that our model has the capability of detecting the connectivity of an image. In response to the inputs from a connected region, the responses of the neurons covering the connected region (they receive the similar external inputs) will become highly synchronized due to their electrical couplings (Bennett and Zukin, 2004), which provides a way to encode the connectivity of the image. This is also supported by the experimental evidence, which found that long-range synchronization occurred among widely separated alpha RGCs with electrical couplings in response to a continuous stimulus, rather than to disjointed local features (Neuenschwander and Singer, 1996; Roy et al., 2017).

Therefore, the focus of the present study is to demonstrate that our network model has the capability of detecting the existence and the number of holes in an image, another key property of topology (Pomerantz et al., 2003; He, 2008; Casati, 2009; Bertamini and Casati, 2015; Zhang et al., 2019). The stimuli we used, as shown in **Figures 3A,D,G**, are adapted from the materials in the human and animal experiments (Chen, 1982, 2005b; Chen et al., 2003; Chien et al., 2012; Zhang et al., 2019), where **Figure 3A** is a solid disk without hole, **Figure 3D** a stimulus containing a single hole, and **Figure 3G** a case of two holes. **Figures 3B,E,H** are the corresponding network responses to the stimuli, and **Figures 3C,F,I** are the illustrations of synchronized neuronal responses in ENs.

Overall, we show that the number of holes in an image is encoded by the number of synchronized responses (population spikes) in the encoding layer, which are further readout by the number of spikes of RON. For example, presentation of **Figure 3A** produces two population spikes of ENs and two spikes of RON (**Figures 3B,C**), while presentation of **Figure 3D** produces three population spikes of ENs and three RON spikes (**Figures 3E,F**). Notably, although the stimulation value (the luminance level) of the hole (inside the ring in **Figure 3D**) is the same as that of the background (outside the ring in

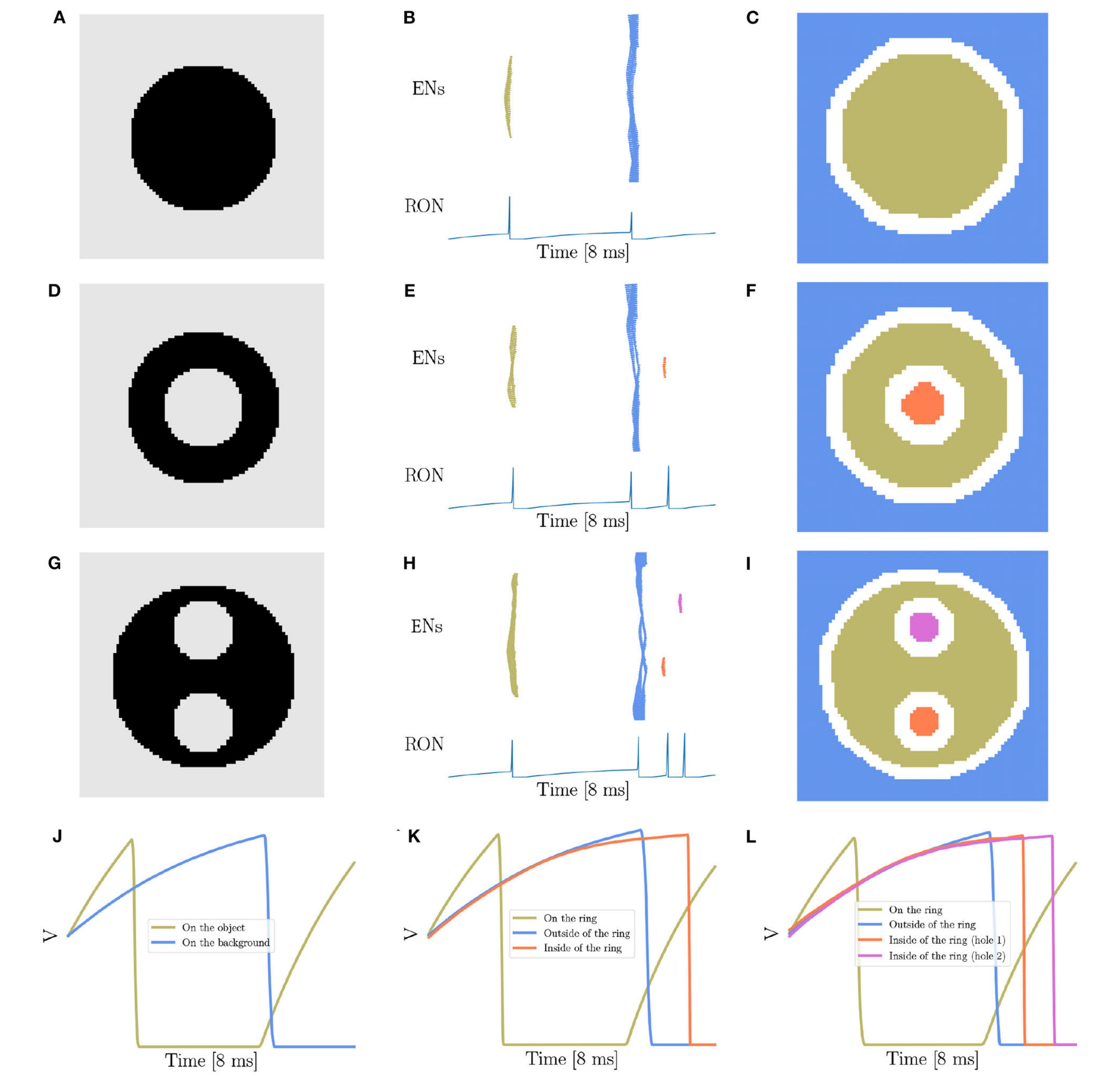


FIGURE 3 | Topological detection of the network. **(A,D,G)** The images with different number of holes. **(A)** contains no hole, **(D)** one hole, and **(G)** two holes. **(B,E,H)** The evolution of network activity. **(B,E,H)** Are results when stimuli **(A,D,G)** are presented, respectively. In each subfigure, the top panel shows the raster plot of the encoding layer, and the bottom the dynamics of the membrane potential of RON. The abscissas of both panels are time, and the ordinates of the top and bottom panel are neuron index and membrane potential, respectively. **(C,F,I)** The spatial mapping of EN spikes. **(C,F,I)** corresponds to **(B,E,H)**, respectively. Neurons in the same group are shown in the same color with **(B,E,H)**. Pixels in the white color denote neurons not firing in the whole process. **(J–L)** The averaged membrane potential traces of neurons inside, on or outside of the ring when stimuli **(A,D,G)** are presented, respectively. The orange line corresponds to the neurons on the ring, the blue line the neurons on the background, and the coral and orchid lines the neurons on the holes. Parameters: $J = 3.0$ and $\gamma = 0.15$.

Figure 3D), the synchronized response of the neurons covering the hole (the orange spikes in **Figures 3E,F**) always lags behind that of the neurons covering the background (the blue spikes in **Figures 3E,F**). This property comes from that compared to the neurons outside the ring, the neurons inside the ring receive stronger inhibition from the neurons on the ring (see more detailed analysis in the below). Moreover, we observe that presentation of **Figure 3G** (containing two holes) reliably produces four population spikes of ENs and four RON firings (**Figures 3H,I**).

To reveal the underling mechanism, we look at the dynamics of neurons inside, on, and outside the ring. Results are shown in **Figures 3J–L**. First, we see that because of receiving a stronger stimulation than those on the background or inside the ring, the neurons on the ring (black pixels) generate the first population spike; afterwards those neurons fall into a deep and relatively long-lasting refractory period (see the voltage trace in khaki color illustrated in **Figures 3J–L**). Second, during the refractory period of ring neurons, while the neurons inside and outside the ring all receive inhibitions from the ring neurons, inside neurons tend to receive stronger inhibitions than outside ones (see the voltage traces in blue and orange color shown in **Figure 3K**). Therefore, under the condition of receiving the same level of stimulation, the neurons inside the ring always generate a population spike before the neurons outside the ring. Third, for an image containing two holes having exactly the same size and surroundings, although the neurons inside two holes receive the same external input and lateral inhibition from surrounds, they still tend to fire at different moments due to receiving independent noises (see the average voltage dynamics in orange and orchid color in **Figure 3L**).

Notably, because of noises, the network response varies over trials. In the case of discriminating two holes from one hole, we observed a successful rate of 70%. This probabilistic behavior is in agreement with the observation of human psychophysical experiments, which showed that the topological detection of humans is also probabilistic when images are only briefly presented in <10 ms, e.g., the successful rate of discriminating hole from circle is about 64.5% (Chen, 2005b). For visualizing the detailed spatio-temporal voltage dynamics when the stimuli (**Figures 3A,D,G**) are presented, please refer to **Supplementary Videos 1–3**. Note that, for simplicity, we have only presented the results for images with shape luminance level changes. We check that our model works equally well when the luminance intensity of the image changes smoothly (see **Supplementary Figure 1**).

In summary, we demonstrate that the synchronization and lateral inhibition effects mediated by gap junctions enable the network to encode the number of holes in an image into different numbers of population spikes of ENs, which provides a reliable cue for the neural system to read out the topology information of an image.

3.3. Topological Detection Is Invariant to Variations of Shape and Spatial Frequency

To confirm that our network model can really detect the topological property of closedness, we vary the stimulus to various forms, while keeping their topological property unchanged.

From our intuitive experience, circle, square, triangle, and cross are quite different figures, but from the viewpoint of topology, they are equivalent. Therefore, the characteristic of network responses for topological detection should be the same. We first conduct experiments on a solid (**Figure 4A**) and a hollow squares (**Figure 4D**), and find that the network responses are exactly the same as when the disk (**Figure 3A**) and the ring (**Figure 3D**) are presented, that is, two population spikes of ENs and two RON spikes are generated for the stimuli without hole (comparing **Figures 3B,C** with **Figures 4B,C**),

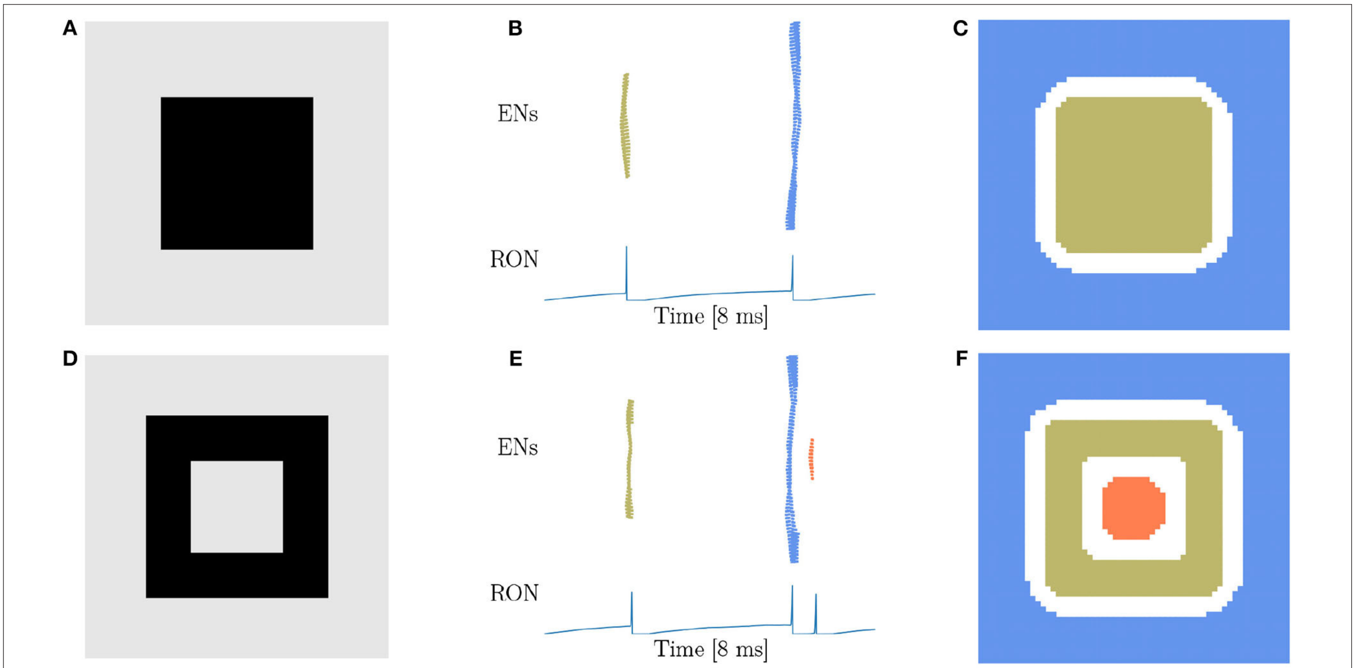


FIGURE 4 | Topological detection with respect to shape variation of images. **(A,D)** Image of square shape. **(A)** A solid square. **(D)** A hollow square. **(B,E)** Population spikes of ENs (top panels) and the voltage dynamics of RON (bottom panels). **(C,F)** Spatial activities of EN neurons. Figure legends are the same as in **Figure 3**. Parameters: $J = 3.0$ and $\gamma = 0.15$.

and three population spikes of ENs and three RON spikes are generated for the stimuli with one hole (comparing **Figures 3E,F** with **Figures 4E,F**). Furthermore, we perform experiments on a solid triangle (**Supplementary Figure 2A**), a hollow triangle (**Supplementary Figure 2B**), and a cross (**Supplementary Figure 2C**), and get the same result. Overall, these results confirm that the network response varies with the topology, rather than the shape of the stimulus.

Based on the finding of Carlson et al. (1984) that geometrical illusions are not primarily a consequence of low spatial frequencies and the suggestion of Chen (2005a) that low spatial frequencies are not likely to be critical to perceptual organization in general, we try to figure out whether the spatial frequency will affect the network behavior. Considering that the stimuli used above are all in low spatial frequencies (LSF), we construct new stimuli (**Figures 5A,D,G**) in high spatial frequencies (HSF), which are adapted from the materials used in human experiments (Carlson et al., 1984; Chen, 2005b). **Figures 5A,D** are made of exactly the same four line segments, while they are topologically different. We find that the network response doesn't vary with the spatial frequency. Specifically, the stimulus without hole persistently produces two population spikes of ENs and two RON spikes (**Figures 5B,C**), whereas the stimulus with one hole reliably generates three population spikes of ENs and three RON spikes (**Figures 5E,F**). We also try stimuli of triangle-shape and obtain the same result, see **Supplementary Figure 3**. Furthermore, we generate a stimulus composed of discrete dots (**Figure 5G**), which is similar to the figures in Carlson et al.

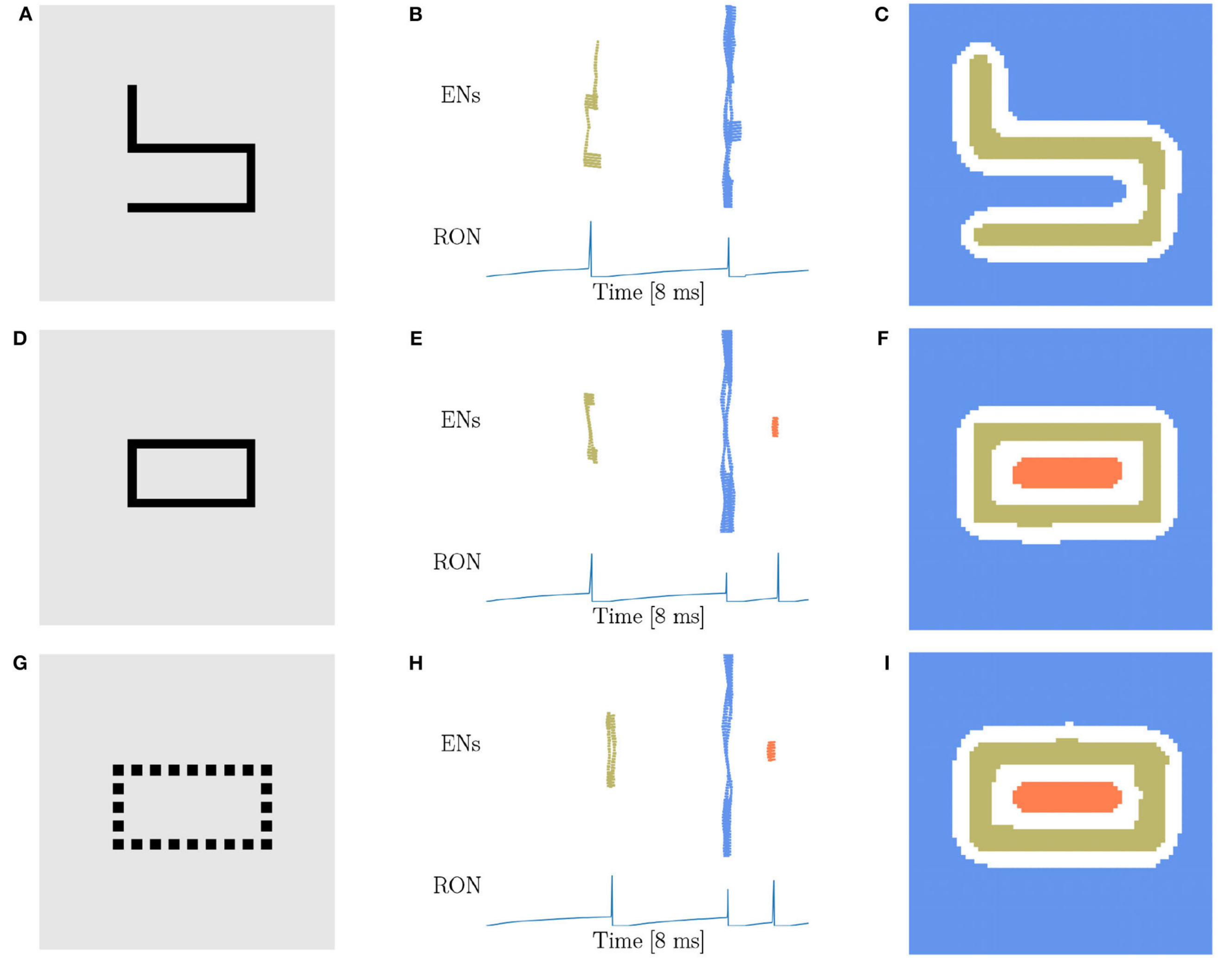


FIGURE 5 | Topological detection with respect to variations of spatial frequency of images. **(A,D,G)** Images with different spatial frequencies. **(A)** An image made of four line segments without hole. **(D)** An image made of the same four line segments as in **(A)** but containing one hole. **(G)** An image shaped like **(D)** but comprised of discrete dots. **(B,E,H)** Population spikes of ENs (top panels) and the voltage dynamics of RON (bottom panels). **(C,F,I)** Spatial activity of EN neurons. Figure legends are the same as in **Figure 3**. Parameters: $J = 3.0$ and $\gamma = 0.15$.

(1984) and is free of low spatial frequencies. We observe that the network model displays the same response property as when the continuous line is presented (comparing **Figures 5H,I** with **Figures 5E,F**). Altogether, these results indicate that the hole detection property of our model is rather robust to the variation of spatial frequencies of images.

In above, we demonstrate that the topological detection of our network model is rather robust to the variations of shape and spatial frequency of images. It is also straightforwardly understandable that our network model is invariant with respect to the position shift, rotation, and distortion of an image, as they all generate the same number of population spikes of ENs depending only on the number of holes in the image. Thus, our network model does have the capability of detecting the topological property of an image.

3.4. Sensitivity of Topological Detection

In above, we have demonstrated that our network model is able to detect the existence of holes in an image, i.e., the closure of a region. In practice, there always exists a threshold of gap below which we perceive disconnected segments as connected. Therefore, we are going to investigate how our network model is sensitive to the gap size in topological detection. We present incomplete rings with different degrees of breach (**Figure 6A**) to the network, and observe that with the small size of breach, the network outputs three RON spikes (**Figures 6B,C**). However, when the breach size θ gradually increases, the network suddenly "recognizes" that the image has no hole (see **Figure 6D**), i.e., ENs only generate two population spikes (**Figures 6E,F**). This is straightforwardly understandable, as the breach increases, the activities of the neurons inside and outside the ring

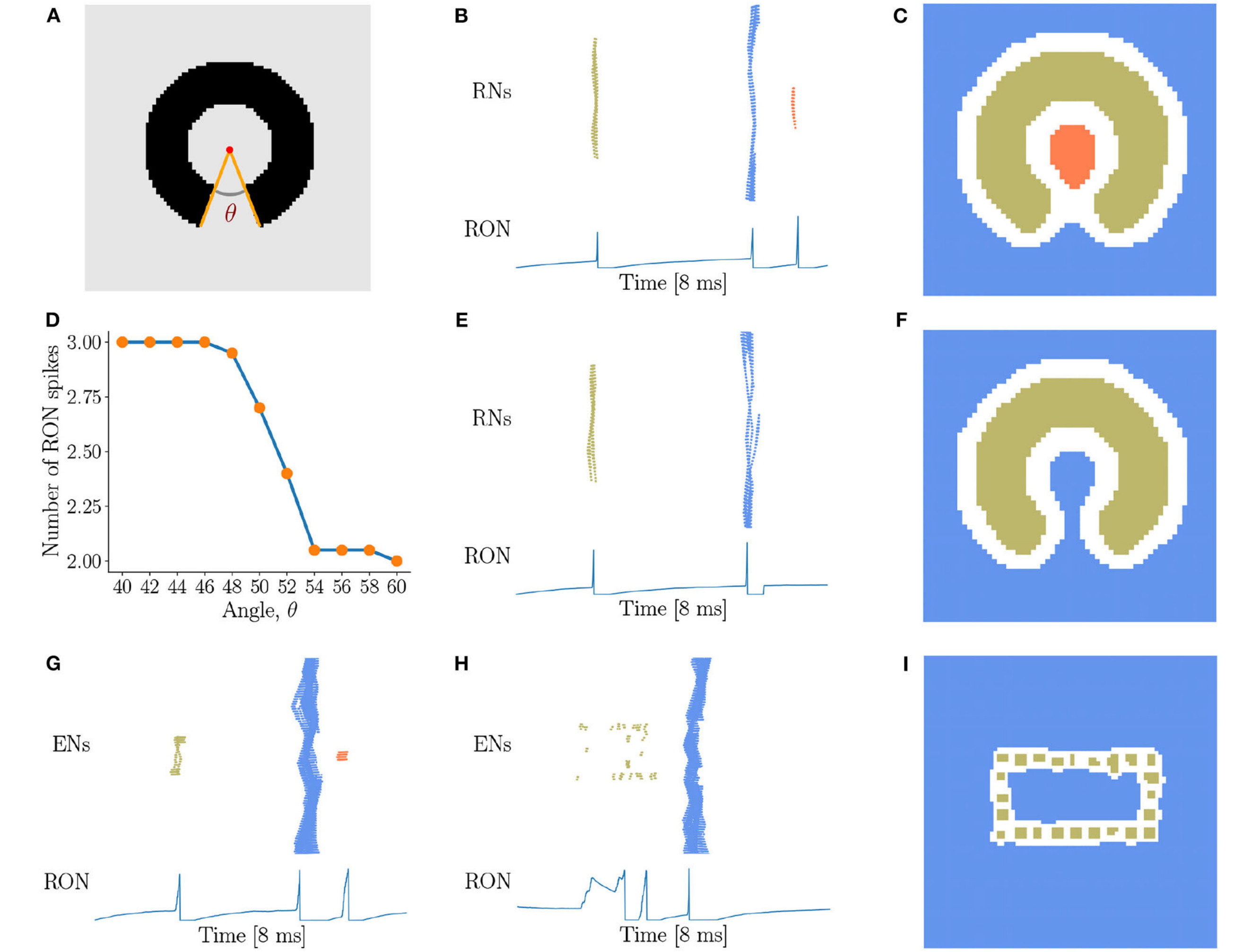


FIGURE 6 | Sensitivity of topological detection. **(A)** An example of a ring with a breach, whose degree is θ. $\theta = 40°$ is shown. **(B,C)** The network activity in response to a ring with a small breach, where ENs generate three population spikes and RON produces three spikes. $J = 3.0$, $\gamma = 0.15$, $\theta = 40°$. **(D)** The average number of RON spikes vs. the breach size. The transition occurs sharply around 50°. The results are obtained by averaging over 20 trials. **(E,F)** The network activity in response to a ring with a big breach, where ENs generate two population spikes and RON produces two pulses. $J = 3.0$, $\gamma = 0.15$, $\theta = 54°$. **(G–I)** The response properties of the network with a varied coupling range, where each neuron is connected to its four nearest neighbors. **(G)** The image of **Figure 5D** is presented. **(H,I)** The image of **Figure 5G** is presented. Parameters: $J = 3.0$, $\gamma = 0.20$. **(B,C,E–I)** Figure legends are the same as **Figure 3**.

become more and more synchronized due to more and more direct interactions between them, and eventually the population spikes they generate merges to a single one (see **Figures 6E,F**). Interestingly, we find that this transition occurs sharply, which is around the breach size of 50° at the current parameter setting (see **Figure 6D**). We confirm that although the value of the transition point may vary with the parameters, this sharp transition behavior always holds (see **Supplementary Figure 4**). This property can serve as a prediction of our model testable in human psychophysical experiments.

Furthermore, we test how the coupling range of gap junction affects the sensitivity of topological detection. We construct a network model in which each neuron is connected with its four nearest neighbors. We first confirm that the model has the capability of detecting a hole in an image, see the network response in **Figure 6G** when the stimulation of **Figure 5D** is presented. However, we also observe that when the image composed of dotted lines as shown in **Figure 5G** is presented, the network is unable to generate synchronous firing, but is rather in the state of irregular firing (see **Figure 6H**), and the network response can no longer stagger the hole and the background. This result tells us that the coupling range of gap junctions between neurons strongly affects the sensitivity of topological detection in reality.

4. DISCUSSION

In the present study, we have proposed a spiking neural network with gap junction for topological detection. Our results show that gap junction-coupled neural networks are intrinsically sensitive to the topological properties, such as connectivity, closure (**Figures 3–5**) or semi-closure (**Figure 6**) of an image. A prominent computational property of gap junction is that it promotes neuron synchronization, which endows the network with the ability of detecting connected regions in an image. Another prominent computational property of gap junction is that it mediates strong lateral inhibition between connected neurons after one of them fires. Together with the fact that neurons within a closure receive much stronger inhibition than neurons outside, the network is able to stagger the moments of neuron firings within and outside a closure, and hence produces different numbers of synchronized firings corresponding to an image having or not having holes. Overall, our model provides a simple yet effective mechanism for topological detection in neural systems. Importantly, our model captures a key behavioral characteristic of object vision, i.e., the ultra-speed object detection (Thorpe et al., 1996; Kirchner and Thorpe, 2006). It has been suggested that the human visual system has the ability of getting "gist" of a scene when the stimulus is presented as briefly as 10 ms (Hegdé, 2008). In the case of topological perception, Chen (1982) demonstrated even the stimulation duration is <10 ms, adult humans are able to discriminate the global topological difference. Our proposed model provides a simple mechanistic explanation for this kind of ultra-speed topological perception: a gap junction-coupled neural network can rapidly group those distant neurons covering the same connected region and meanwhile segregate different neuron groups covering different regions, forming a stable topological visual representation in <10 ms.

4.1. Biological Plausibility

Our model uses electrical synapses to synchronize distant neurons corresponding to a connected region. This is consistent with the recent experimental works which found that gap junction is important for long-range synchronization among neurons over long distances (Neuenschwander and Singer, 1996; Völgyi et al., 2013; Roy et al., 2017). Particularly, Roy et al. (2017) found that electrical couplings between ON alpha RGCs and polyaxonal amacrine cells are responsible to produce the long-range correlated activity critical for global object perception. Specifically, they found that presentation of large stimuli of various shapes always produced long-range synchronization between distant ON alpha RGC pairs under electrical coupling, whereas presentation of discontinuous stimuli of several segments could not. Moreover, blockade of gap junctions diminished such kind of coherent firing. These results indicate that electrical couplings are essential for the neural representation of the image connectivity.

We propose that a retina network with electrical coupling is capable of encoding global topological features. This is in line with the functional roles of ON alpha RGC network (Schmidt et al., 2014; Allen et al., 2019). ON alpha RGCs found by Roy et al. (2017) are actually one type of ipRGCs, i.e., M4 ipRGCs (Schmidt et al., 2011, 2014). Recently, M4 ipRGCs are found essential for full contrast sensitivity in mouse visual functions (Schmidt et al., 2014). Deletion of ON alpha RGCs in mice caused severe deficits in contrast sensitivity. Meanwhile, by constructing special patterns that are distinguishable for cones but contain significant contrast for melanopsin, Allen et al. (2019) found that M4 ipRGCs in human have the capacity to encode coarse patterns and influence the appearance of everyday images. Hence, it is evident that M4 ipRGCs, which are crucial for the coarse pattern encoding and contrast sensitivity, should also be able to encode global topological patterns. However, it was reported that M4 cells have rich dendrites and exhibit non-linear spatial summation (Estevez et al., 2012). The simplified biophysics of our neurons does not capture this effect, and the functional role of dendritic computation in the M4 cells should be investigated in the future work.

If retina RGCs are able to encode global topological patterns, where and how these topological information extracted in the retina are further processed? The candidate brain area is SC. It has been long suggested that there is a type of SC neurons which is capable of global visual processing (Rizzolatti et al., 1980; Bender and Davidson, 1986). For example, Rizzolatti et al. (1980) found that some neurons in SC respond very poorly to simple visual stimuli, while produce strong and sustained discharges for all complex stimuli. In the primate, compared with the role of "feature detector" of neurons in visual cortex (like V1), this type of SCs neurons is now thought to be a class of "event detector" (Ito and Feldheim, 2018), because their responses to the visual stimuli within their receptive fields are irrelevant to the specific stimulus features, such as direction, orientation or shape (Girman and Lund, 2007; White et al., 2017a,b, 2019). One example is the recent study done by White et al. (2017a,b, 2019), in which they found that SC neurons in monkeys are capable of encoding visual saliency in a featureless manner (Marrocco and Li, 1977). Inspired by these neurobiological findings, we used a single neuron to read out each event that ENs produce coherent activity for a connected region in an image. However, our implementation of the read-out mechanism is over-simplified, because despite the existence of wide-field SC cells receiving hundreds of RGC projections (Gabbiani et al., 2001; Wang et al., 2010; Gale and Murphy, 2014), a SC neuron receiving global RGC projections is rare. Future work will consider the detailed connections between retina and SC.

4.2. Gap Junctions Mediate Retinal Lateral Inhibition

Lateral inhibition in the retina is thought to be crucial for visual perception (Kramer and Davenport, 2015). It has been suggested these inhibition activities are the results of retinal microcircuits which involve two inhibitory interneurons: horizontal cells (HCs) in the outer retina and amacrine cells (ACs) in the inner retina. First synaptic mechanism of lateral inhibition results from the feedback regulation mediated by HCs, which alters the neurotransmitter release in rods and cones (Wu, 1991). Later, lateral inhibition due to AC GABAergic inhibitory feedback to bipolar cells has also been observed (Feigenspan et al., 1993; Dong and Werblin, 1998; Roska et al., 2000). Furthermore, recent works suggested lateral inhibition occurs among RGCs which are indirectly mediated by spiking GABAergic wide-field ACs (Chen et al., 2016; Johnson et al., 2018). Overall, all three levels of lateral inhibition are produced by interneurons and have been shown to be closely involved in various visual processes, such as edge (contrast) enhancement (Campbell and Robson, 1968; Kramer and Davenport, 2015), spatial induction (Cook and McReynolds, 1998; Yeonan-Kim and Bertalmío, 2016), direction selectivity (Chen et al., 2016), and color processing (Schnaitmann et al., 2018). In this paper, our modeling study suggests that through gap junctions, RGCs can provide direct lateral inhibition to the coupled cells without the involvement of interneurons. This is due to that when a RGC briefly spikes, it will enter into a long refractory period, during which its connected cells via gap junctions will be strongly inhibited. This kind of lateral inhibition has been observed in Golgi cells in the cerebellar input layer (Vervaeke et al., 2010), in which a relatively deep and protracted afterhyperpolarization (one of the processes that contribute to the refractory period) in Golgi cells mediated a robust form of surround depression.

To further highlight the crucial role of gap junction-mediated lateral inhibition in topological detection, we carry out experiments by adding local GABAergic AC feedback inhibitions in the model (see **Supplementary Figure 5A**). Since the chemical transmission is too slow in reality, we set the synapse delay to be 0.1 ms. With such unrealistic fast feedback AC inhibition, we observe that the network behaves similarly to that without AC inhibitions (compare **Supplementary Figures 5B,C** with **Figures 3E,F**). Furthermore, to ablate the lateral inhibition of gap junctions while preserve their synchronization effect, we artificially block gap junctions when neurons are in their refractory period (setting $J = 0$). In such a way, the contribution of local chemical inhibitions is isolated. We find that: (1) when the receptive field of AC is not big enough to cover most of the hole, synchronous firings of neurons on the hole cannot be segregated from that of neurons on the background (**Supplementary Figures 5D,E**); (2) when the receptive field of AC is big enough to cover most of the hole, synchronous firings of neurons on the hole and the background can be well-segregated in the first 10 ms but are mixed together later on (**Supplementary Figures 5F,G**). Overall, our ablation study reveals that gap junction-mediated lateral inhibition is the necessary and sufficient requirement for rapid topological detection. Certainly, AC-mediated and other chemical inhibitions are also important for neural information processing, but they tend to work at different time scales and are more likely responsible for non-topological feature analysis, such as edge detection. It will be interesting to explore how different inhibitory mechanisms cooperate together to solve the coarse-to-fine feature analysis.

4.3. Global-to-Local Visual Processing Starts From Early Topological Detection

It is now widely agreed that visual perception takes place in a predominantly global-to-local or coarse-to-fine procedure (Bullier, 2001; Bar, 2004, 2007; Hegdé, 2008). Supporting evidence comes from the experiments using various materials, ranging from the simple stimuli [like lines, dots, gratings, and letters (Weisstein and Harris, 1974; Navon, 1977; Pomerantz et al., 1977; Watt, 1987; Hughes et al., 1996)] to complex images [such as faces (Farah et al., 1998; McKone et al., 2007; Goffaux et al., 2010; Taubert et al., 2011) and natural scenes (Parker et al., 1992, 1997; Schyns and Oliva, 1994; Lu et al., 2018)]. In this framework, the global and coarse information is processed first and subsequently activates the high-level visual cortex rather than primary visual cortex; whereafter, a feedback signal is generated and further guides the processing of the conventional local feature analysis (Bar, 2003; Bar et al., 2006). The bottom-up local feature analysis has so far been well-established, in which the visual processing begins from extracting the local features in the low visual areas followed by integrating such local features to extract more global features in the higher visual areas (Hubel and Wiesel, 1959; Treisman and Gelade, 1980; Marr, 1982; Hubel, 1988; DiCarlo et al., 2012). Later, more and more researches begin to emphasize the role of top-down facilitation in visual perception (Bar et al., 2006; Gilbert and Li, 2013). However, several questions remain elusive in this framework: how and where is such top-down facilitation ignited (Bar, 2003; Goffaux et al., 2010)? In particular, at the early visual stage, how global features are rapidly extracted?

In the case of topological perception, it has been found that the neural substrate of topological perception in humans lies in the final stage of the ventral cortical visual system, i.e., the temporal lobe (Zhuo et al., 2003; Wang et al., 2007). Moreover, on monkeys, a single-unit recording study unveiled there exists a subset of inferior temporal neurons responding selectively to hole patterns with a short latency (<100 ms) (Komatsu and Ideura, 1993). Similarly, how are such topological features extracted? What pathway does it route through to ignite the temporal lobe? Here, we hypothesize that the topological features (like "holes") begin to be extracted in the retina. Specifically, we propose that in the retina, the alpha RGC network coupled through electrical couplings is capable of producing the topologically discriminable neural representations in a short time interval of <10 ms. We also demonstrate that such rapid and stable topological representations can be easily read-out by the SC or higher visual cortex. Our hypothesis can be partially supported by earlier two experiments (Ölveczky et al., 2003; Baccus et al., 2008). Specifically, they found that there exists a subset of RGCs specialized to distinguish local motion within the scene from the global retinal image drift due to fixational eye movements. In other words, the global motion detection begins in the retina, which supports the notion of the retinal representation of global information. In future, further detailed investigations should be carried out.

4.4. Related Works

The most relevant work is a pioneering model called LEGION (Wang and Terman, 1995), which was designed using the mechanisms of local excitation and global inhibition. Wang (2000) demonstrated that LEGION exhibits sensitivity to the topological connectivity, but did not investigate the detection of holes. Our model differs from LEGION in two fundamental aspects. First, the computational mechanisms are different. LEGION achieves synchronization via chemical excitatory synapses between nearby oscillators and employs a global chemical inhibitory synapse to deactivate different groups of oscillators, which are not feasible in retina; whereas, our model relies on gap functions which widely exist in the retina to synchronize and differentiate neuron groups. Second, the time courses are different. The time for LEGION to detect the topological connectivity is too slow, as the emergence of stable phase differences between objects needs multiple cycles. In contrary, our model has the ability to detect the topological property rapidly as briefly as <10 ms. Overall, our model better captures the computational nature of the retina.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/**Supplementary Material**. The code of the models is available on GitHub: https://github.com/chaoming0625/Gap_Junction_and_Topology.

AUTHOR CONTRIBUTIONS

CW, SW, and YM built the model. SW and YM supervised the project. CM, RL, and XD did the simulation and algorithm implementation. CM, SW, and RL wrote the paper. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by Guangdong Province with Grant (No. 2018B030338001, SW & YM), the National Natural Science Foundation of China (No: 31771146, 11734004, YM), Beijing Nova Program (No: Z181100006218118, YM), and the Fundamental Research Funds for the Central Universities (2020CDJQY-A073, YM). This work also received support from Huawei Technology Co., Ltd (YBN2019105137).

ACKNOWLEDGMENTS

The authors would like to thank Longping Liu, Zilong Ji, Xiaolong Zou, Ang Li, and Liutao Yu for valuable discussions and comments on the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fncom.2020.571982/full#supplementary-material

REFERENCES

Allen, A. E., Martial, F. P., and Lucas, R. J. (2019). Form vision from melanopsin in humans. *Nat. Commun.* 10:2274. doi: 10.1038/s41467-019-10113-3

Armstrong, M. A. (2013). *Basic Topology.* New York, NY: Springer Science & Business Media.

Baccus, S. A., Ölveczky, B. P., Manu, M., and Meister, M. (2008). A retinal circuit that computes object motion. *J. Neurosci.* 28, 6807–6817. doi: 10.1523/JNEUROSCI.4206-07.2008

Bar, M. (2003). A cortical mechanism for triggering top-down facilitation in visual object recognition. *J. Cogn. Neurosci.* 15, 600–609. doi: 10.1162/089892903321662976

Bar, M. (2004). Visual objects in context. *Nat. Rev. Neurosci.* 5, 617–629. doi: 10.1038/nrn1476

Bar, M. (2007). The proactive brain: using analogies and associations to generate predictions. *Trends Cogn. Sci.* 11, 280–289. doi: 10.1016/j.tics.2007.05.005

Bar, M., Kassam, K. S., Ghuman, A. S., Boshyan, J., Schmid, A. M., Dale, A. M., et al. (2006). Top-down facilitation of visual recognition. *Proc. Natl. Acad. Sci. U.S.A.* 103, 449–454. doi: 10.1073/pnas.0507062103

Bender, D., and Davidson, R. (1986). Global visual processing in the monkey superior colliculus. *Brain Res.* 381, 372–375. doi: 10.1016/0006-8993(86)90092-2

Bennett, M. V., and Zukin, R. S. (2004). Electrical coupling and neuronal synchronization in the mammalian brain. *Neuron* 41, 495–511. doi: 10.1016/S0896-6273(04)00043-1

Bertamini, M., and Casati, R. (2015). Figures and holes. in *The Oxford Handbook of Perceptual Organization*, ed J. Wagemans (Oxford, UK: Oxford University Press), 281–293.

Bona, S., Cattaneo, Z., and Silvanto, J. (2016). Investigating the causal role of rofa in holistic detection of mooney faces and objects: an fMRI-guided tms study. *Brain Stimul.* 9, 594–600. doi: 10.1016/j.brs.2016.04.003

Bullier, J. (2001). Integrated model of visual processing. *Brain Res. Rev.* 36, 96–107. doi: 10.1016/S0165-0173(01)00085-6

Campbell, F. W., and Robson, J. G. (1968). Application of fourier analysis to the visibility of gratings. *J. Physiol.* 197:551. doi: 10.1113/jphysiol.1968.sp008574

Carlson, C., Moeller, J., and Anderson, C. (1984). Visual illusions without low spatial frequencies. *Vis. Res.* 24, 1407–1413. doi: 10.1016/0042-6989(84)90196-2

Casati, R. (2009). Does topological perception rest on a misconception about topology? *Philos. Psychol.* 22, 77–81. doi: 10.1080/09515080802703711

Chen, L. (1982). Topological structure in visual perception. *Science* 218, 699–700. doi: 10.1126/science.7134969

Chen, L. (2005a). Author's response: where to begin? *Visual Cogn.* 12, 691–701. doi: 10.1080/13506280444000364

Chen, L. (2005b). The topological approach to perceptual organization. *Visual Cogn.* 12, 553–637. doi: 10.1080/13506280444000256

Chen, L., Zhang, S., and Srinivasan, M. V. (2003). Global perception in small brains: topological pattern recognition in honey bees. *Proc. Natl. Acad. Sci. U.S.A.* 100, 6884–6889. doi: 10.1073/pnas.0732090100

Chen, Q., Pei, Z., Koren, D., and Wei, W. (2016). Stimulus-dependent recruitment of lateral inhibition underlies retinal direction selectivity. *Elife* 5:e21053. doi: 10.7554/eLife.21053

Chien, S. H.-L., Lin, Y.-L., Qian, W., Zhou, K., Lin, M.-K., and Hsu, H.-Y. (2012). With or without a hole: young infants' sensitivity for topological versus geometric property. *Perception* 41, 305–318. doi: 10.1068/p7031

Cook, P. B., and McReynolds, J. S. (1998). Lateral inhibition in the inner retina is important for spatial tuning of ganglion cells. *Nat. Neurosci.* 1:714. doi: 10.1038/3714

Darke, I. (1982). A review of research related to the topological primacy thesis. *Educ. Stud. Math.* 13, 119–142. doi: 10.2307/3482369

DiCarlo, J. J., Zoccolan, D., and Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron* 73, 415–434. doi: 10.1016/j.neuron.2012.01.010

Dong, C.-J., and Werblin, F. S. (1998). Temporal contrast enhancement via gabac feedback at bipolar terminals in the tiger salamander retina. *J. Neurophysiol.* 79, 2171–2180. doi: 10.1152/jn.1998.79.4.2171

Du, X., Zhou, K., and Chen, L. (2011). Different temporal dynamics of topological and projective geometrical perceptions in primary visual cortex: a tms study. *J. Vis.* 11, 863–863. doi: 10.1167/11.11.863

Estevez, M. E., Fogerson, P. M., Ilardi, M. C., Borghuis, B. G., Chan, E., Weng, S., et al. (2012). Form and function of the m4 cell, an intrinsically photosensitive retinal ganglion cell type contributing to geniculocortical vision. *J. Neurosci.* 32, 13608–13620. doi: 10.1523/JNEUROSCI.1422-12.2012

Farah, M. J., Wilson, K. D., Drain, M., and Tanaka, J. N. (1998). What is "special" about face perception? *Psychol. Rev.* 105:482. doi: 10.1037/0033-295X.105.3.482

Feigenspan, A., Wässle, H., and Bormann, J. (1993). Pharmacology of gaba receptor ci- channels in rat retinal bipolar cells. *Nature* 361:159. doi: 10.1038/361159a0

Gabbiani, F., Mo, C., and Laurent, G. (2001). Invariance of angular threshold computation in a wide-field looming-sensitive neuron. *J. Neurosci.* 21, 314–329. doi: 10.1523/JNEUROSCI.21-01-00314.2001

Gale, S. D., and Murphy, G. J. (2014). Distinct representation and distribution of visual information by specific cell types in mouse superficial superior colliculus. *J. Neurosci.* 34, 13458–13471. doi: 10.1523/JNEUROSCI.2768-14.2014

Gilbert, C. D., and Li, W. (2013). Top-down influences on visual processing. *Nat. Rev. Neurosci.* 14, 350–363. doi: 10.1038/nrn3476

Girman, S. V., and Lund, R. D. (2007). Most superficial sublamina of rat superior colliculus: neuronal response properties and correlates with perceptual figure–ground segregation. *J. Neurophysiol.* 98, 161–177. doi: 10.1152/jn.00059.2007

Goffaux, V., Peters, J., Haubrechts, J., Schiltz, C., Jansma, B., and Goebel, R. (2010). From coarse to fine? Spatial and temporal dynamics of cortical face processing. *Cereb. Cortex* 21, 467–476. doi: 10.1093/cercor/bhq112

He, S. (2008). Holes, objects, and the left hemisphere. *Proc. Natl. Acad. Sci. U.S.A.* 105, 1103–1104. doi: 10.1073/pnas.0710631105

Hegdé, J. (2008). Time course of visual perception: coarse-to-fine processing and beyond. *Prog. Neurobiol.* 84, 405–439. doi: 10.1016/j.pneurobio.2007.09.001

Hubel, D. H. (1988). *Eye, Brain, and Vision.* New York, NY: Scientific American Library.

Hubel, D. H., and Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *J. Physiol.* 148, 574–591. doi: 10.1113/jphysiol.1959.sp006308

Hughes, H. C., Nozawa, G., and Kitterle, F. (1996). Global precedence, spatial frequency channels, and the statistics of natural images. *J. Cogn. Neurosci.* 8, 197–230. doi: 10.1162/jocn.1996.8.3.197

Ito, S., and Feldheim, D. A. (2018). The mouse superior colliculus: an emerging model for studying circuit formation and function. *Front. Neural Circuits* 12:10. doi: 10.3389/fncir.2018.00010

Johnson, K. P., Zhao, L., and Kerschensteiner, D. (2018). A pixel-encoder retinal ganglion cell with spatially offset excitatory and inhibitory receptive fields. *Cell Rep.* 22, 1462–1472. doi: 10.1016/j.celrep.2018.01.037

Kibbe, M. M., and Leslie, A. M. (2016). The ring that does not bind: topological class in infants' working memory for objects. *Cogn. Dev.* 38, 1–9. doi: 10.1016/j.cogdev.2015.12.001

Kirchner, H., and Thorpe, S. J. (2006). Ultra-rapid object detection with saccadic eye movements: Visual processing speed revisited. *Vis. Res.* 46, 1762–1776. doi: 10.1016/j.visres.2005.10.002

Komatsu, H., and Ideura, Y. (1993). Relationships between color, shape, and pattern selectivities of neurons in the inferior temporal cortex of the monkey. *J. Neurophysiol.* 70, 677–694. doi: 10.1152/jn.1993.70.2.677

Kramer, R. H., and Davenport, C. M. (2015). Lateral inhibition in the vertebrate retina: the case of the missing neurotransmitter. *PLoS Biol.* 13:e1002322. doi: 10.1371/journal.pbio.1002322

Lu, Y., Yin, J., Chen, Z., Gong, H., Liu, Y., Qian, L., et al. (2018). Revealing detail along the visual hierarchy: neural clustering preserves acuity from v1 to v4. *Neuron* 98, 417–428. doi: 10.1016/j.neuron.2018.03.009

Marr, D. (1982). *Vision: A Computational Investigation Into the Human Representation and Processing of Visual Information.* Cambridge, MA: MIT Press.

Marrocco, R., and Li, R. (1977). Monkey superior colliculus: properties of single cells and their afferent inputs. *J. Neurophysiol.* 40, 844–860. doi: 10.1152/jn.1977.40.4.844

McClelland, J. L., Rumelhart, D. E., and Group, P. R. (1987). *Parallel Distributed Processing*, Vol. 2. Cambridge, MA: MIT Press.

McKone, E., Kanwisher, N., and Duchaine, B. C. (2007). Can generic expertise explain special processing for faces? *Trends Cogn. Sci.* 11, 8–15. doi: 10.1016/j.tics.2006.11.002

Meng, Q., Huang, Y., Cui, D., He, L., Chen, L., Ma, Y., et al. (2018). The dissociations of visual processing of "hole" and "no-hole" stimuli: an functional magnetic resonance imaging study. *Brain Behav.* 8:e00979. doi: 10.1002/brb3.979

Meng, Q., Wang, B., Cui, D., Liu, N., Huang, Y., Chen, L., et al. (2019). Age-related changes in local and global visual perception. *J. Vis.* 19:10. doi: 10.1167/19.1.10

Minsky, M., and Papert, S. A. (1987). *Perceptrons: An Introduction to Computational Geometry, Expanded Edition.* Cambridge, MA: MIT Press.

Navon, D. (1977). Forest before trees: the precedence of global features in visual perception. *Cogn. Psychol.* 9, 353–383. doi: 10.1016/0010-0285(77)90012-3

Neuenschwander, S., and Singer, W. (1996). Long-range synchronization of oscillatory light responses in the cat retina and lateral geniculate nucleus. *Nature* 379:728. doi: 10.1038/379728a0

Ölveczky, B. P., Baccus, S. A., and Meister, M. (2003). Segregation of object and background motion in the retina. *Nature* 423, 401–408. doi: 10.1038/nature01652

Palmer, S. E. (1999). *Vision Science: Photons to Phenomenology.* Cambridge, MA: MIT Press.

Parker, D. M., Lishman, J. R., and Hughes, J. (1992). Temporal integration of spatially filtered visual images. *Perception* 21, 147–160. doi: 10.1068/p210147

Parker, D. M., Lishman, J. R., and Hughes, J. (1997). Evidence for the view that temporospatial integration in vision is temporally anisotropic. *Perception* 26, 1169–1180. doi: 10.1068/p261169

Piaget, J., and Inhelder, B. (1956). *The Child's Conception of Space.* London: Routledge and Kegan Paul.

Pomerantz, J. R., Agrawal, A., Jewell, S. W., Jeong, M., Khan, H., and Lozano, S. C. (2003). Contour grouping inside and outside of facial contexts. *Acta Psychol.* 114, 245–271. doi: 10.1016/j.actpsy.2003.08.004

Pomerantz, J. R., Sager, L. C., and Stoever, R. J. (1977). Perception of wholes and of their component parts: some configural superiority effects. *J. Exp. Psychol. Huma. Percept. Perform.* 3:422. doi: 10.1037/0096-1523.3.3.422

Rizzolatti, G., Buchtel, H., Camarda, R., and Scandolara, C. (1980). Neurons with complex visual properties in the superior colliculus of the macaque monkey. *Exp. Brain Res.* 38, 37–42. doi: 10.1007/bf00237928

Roska, B., Nemeth, E., Orzo, L., and Werblin, F. S. (2000). Three levels of lateral inhibition: a space–time study of the retina of the tiger salamander. *J. Neurosci.* 20, 1941–1951. doi: 10.1523/JNEUROSCI.20-05-01941.2000

Roy, K., Kumar, S., and Bloomfield, S. A. (2017). Gap junctional coupling between retinal amacrine and ganglion cells underlies coherent activity integral to global object perception. *Proc. Natl. Acad. Sci. U.S.A.* 114, E10484–E10493. doi: 10.1073/pnas.1708261114

Schmidt, T. M., Alam, N. M., Chen, S., Kofuji, P., Li, W., Prusky, G. T., et al. (2014). A role for melanopsin in alpha retinal ganglion cells and contrast detection. *Neuron* 82, 781–788. doi: 10.1016/j.neuron.2014.03.022

Schmidt, T. M., Chen, S.-K., and Hattar, S. (2011). Intrinsically photosensitive retinal ganglion cells: many subtypes, diverse functions. *Trends Neurosci.* 34, 572–580. doi: 10.1016/j.tins.2011.07.001

Schnaitmann, C., Haikala, V., Abraham, E., Oberhauser, V., Thestrup, T., Griesbeck, O., et al. (2018). Color processing in the early visual system of drosophila. *Cell* 172, 318–330. doi: 10.1016/j.cell.2017.12.018

Schyns, P. G., and Oliva, A. (1994). From blobs to boundary edges: evidence for time-and spatial-scale-dependent scene recognition. *Psychol. Sci.* 5, 195–200. doi: 10.1111/j.1467-9280.1994.tb00500.x

Taubert, J., Apthorp, D., Aagten-Murphy, D., and Alais, D. (2011). The role of holistic processing in face perception: evidence from the face inversion effect. *Vis. Res.* 51, 1273–1278. doi: 10.1016/j.visres.2011.04.002

Thorpe, S., Fize, D., and Marlot, C. (1996). Speed of processing in the human visual system. *Nature* 381:520. doi: 10.1038/381520a0

Treisman, A. M., and Gelade, G. (1980). A feature-integration theory of attention. *Cogn. Psychol.* 12, 97–136. doi: 10.1016/0010-0285(80)90005-5

Turati, C., Simion, F., and Zanon, L. (2003). Newborns' perceptual categorization for closed and open geometric forms. *Infancy* 4, 309–325. doi: 10.1207/S15327078IN0403_01

Versace, E., Schill, J., Nencini, A. M., and Vallortigara, G. (2016). Naïve chicks prefer hollow objects. *PLoS ONE* 11:e0166425. doi: 10.1371/journal.pone.0166425

Vervaeke, K., Lőrincz, A., Gleeson, P., Farinella, M., Nusser, Z., and Silver, R. A. (2010). Rapid desynchronization of an electrically coupled interneuron network with sparse excitatory synaptic input. *Neuron* 67, 435–451. doi: 10.1016/j.neuron.2010.06.028

Völgyi, B., Pan, F., Paul, D. L., Wang, J. T., Huberman, A. D., and Bloomfield, S. A. (2013). Gap junctions are essential for generating the correlated spike activity of neighboring retinal ganglion cells. *PLoS ONE* 8:e69426. doi: 10.1371/journal.pone.0069426

Wagemans, J., Feldman, J., Gepshtein, S., Kimchi, R., Pomerantz, J. R., Van der Helm, P. A., et al. (2012). A century of gestalt psychology in visual perception: II. Conceptual and theoretical foundations. *Psychol. Bull.* 138:1218. doi: 10.1037/a0029334

Wang, B., Zhou, T. G., Zhuo, Y., and Chen, L. (2007). Global topological dominance in the left hemisphere. *Proc. Natl. Acad. Sci. U.S.A.* 104, 21014–21019. doi: 10.1073/pnas.0709664104

Wang, D., and Terman, D. (1995). Locally excitatory globally inhibitory oscillator networks. *IEEE Trans. Neural Netw.* 6, 283–286. doi: 10.1109/72.363423

Wang, D. L. (2000). On connectedness: a solution based on oscillatory correlation. *Neural Comput.* 12, 131–139. doi: 10.1162/089976600300015916

Wang, L., Sarnaik, R., Rangarajan, K., Liu, X., and Cang, J. (2010). Visual receptive field properties of neurons in the superficial superior colliculus of the mouse. *J. Neurosci.* 30, 16573–16584. doi: 10.1523/jneurosci.3305-10.2010

Watanabe, A., Fujimoto, M., Hirai, K., and Ushitani, T. (2019). Pigeons discriminate shapes based on topological features. *Vis. Res.* 158, 120–125. doi: 10.1016/j.visres.2019.02.012

Watt, R. (1987). Scanning from coarse to fine spatial scales in the human visual system after the onset of a stimulus. *JOSA A* 4, 2006–2021. doi: 10.1364/JOSAA.4.002006

Weisstein, N., and Harris, C. S. (1974). Visual detection of line segments: an object-superiority effect. *Science* 186, 752–755. doi: 10.1126/science.186.4165.752

White, B. J., Berg, D. J., Kan, J. Y., Marino, R. A., Itti, L., and Munoz, D. P. (2017a). Superior colliculus neurons encode a visual saliency map during free viewing of natural dynamic video. *Nat. Commun.* 8:14263. doi: 10.1038/ncomms14263

White, B. J., Itti, L., and Munoz, D. P. (2019). Superior colliculus encodes visual saliency during smooth pursuit eye movements. *Eur. J. Neurosci.* doi: 10.1111/ejn.14432. [Epub ahead of print].

White, B. J., Kan, J. Y., Levy, R., Itti, L., and Munoz, D. P. (2017b). Superior colliculus encodes visual saliency before the primary visual cortex. *Proc. Natl. Acad. Sci. U.S.A.* 114, 9451–9456. doi: 10.1073/pnas.1701003114

Wu, S. M. (1991). Input-output relations of the feedback synapse between horizontal cells and cones in the tiger salamander retina. *J. Neurophysiol.* 65, 1197–1206. doi: 10.1152/jn.1991.65.5.1197

Yeonan-Kim, J., and Bertalmío, M. (2016). Retinal lateral inhibition provides the biological basis of long-range spatial induction. *PLoS ONE* 11:e0168963. doi: 10.1371/journal.pone.0168963

Zhang, J., Wu, J., Liu, X., Jin, Z., Li, L., and Chen, L. (2019). Hole superiority effect with 3D figures formed by binocular disparity. *J. Vis.* 19:2. doi: 10.1167/19.2.2

Zhou, K., Luo, H., Zhou, T., Zhuo, Y., and Chen, L. (2010). Topological change disturbs object continuity in attentive tracking. *Proc. Natl. Acad. Sci. U.S.A.* 107, 21920–21924. doi: 10.2307/25756984

Zhuo, Y., Zhou, T. G., Rao, H. Y., Wang, J. J., Meng, M., Chen, M., et al. (2003). Contributions of the visual ventral pathway to long-range apparent motion. *Science* 299, 417–420. doi: 10.1126/science.1077091

Conflict of Interest: The authors declare that this study received funding from Huawei Technology Co., Ltd. The funder was not involved in the study design, collection, analysis, interpretation of data, the writing of this article or the decision to submit it for publication.

frontiers in Computational Neuroscience

ORIGINAL RESEARCH
published: 14 October 2020
doi: 10.3389/fncom.2020.00083

Unsupervised Few-Shot Feature Learning via Self-Supervised Training

Zilong Ji[1], *Xiaolong Zou*[2], *Tiejun Huang*[2] *and Si Wu*[2,3]*

[1] State Key Laboratory of Cognitive Neuroscience & Learning, Beijing Normal University, Beijing, China, [2] School of Electronics Engineering & Computer Science, Peking University, Beijing, China, [3] IDG/McGovern Institute for Brain Research, PKU-Tsinghua Center for Life Sciences, Peking University, Beijing, China

Learning from limited exemplars (few-shot learning) is a fundamental, unsolved problem that has been laboriously explored in the machine learning community. However, current few-shot learners are mostly supervised and rely heavily on a large amount of labeled examples. Unsupervised learning is a more natural procedure for cognitive mammals and has produced promising results in many machine learning tasks. In this paper, we propose an unsupervised feature learning method for few-shot learning. The proposed model consists of two alternate processes, progressive clustering and episodic training. The former generates pseudo-labeled training examples for constructing episodic tasks; and the later trains the few-shot learner using the generated episodic tasks which further optimizes the feature representations of data. The two processes facilitate each other, and eventually produce a high quality few-shot learner. In our experiments, our model achieves good generalization performance in a variety of downstream few-shot learning tasks on Omniglot and MiniImageNet. We also construct a new few-shot person re-identification dataset FS-Market1501 to demonstrate the feasibility of our model to a real-world application.

Keywords: unsupervised, few-shot learning, clustering, pseudo labels, episodic learning

Edited by:
Yu-Guo Yu,
Fudan University, China

Reviewed by:
Songting Li,
Shanghai Jiao Tong University, China
Yina Wei,
Allen Institute for Brain Science, United States

****Correspondence:***
Si Wu
siwu@pku.edu.cn

Received: *08 July 2020*
Accepted: *05 August 2020*
Published: *14 October 2020*

Citation:
Ji Z, Zou X, Huang T and Wu S (2020) Unsupervised Few-Shot Feature Learning via Self-Supervised Training. Front. Comput. Neurosci. 14:83. doi: 10.3389/fncom.2020.00083

1. INTRODUCTION

Few-shot learning, which aims to accomplish a learning task by using very few training examples, is receiving increasing attention in both of the machine learning and cognitive science community. The challenge of few-shot learning lies on the fact that traditional techniques such as fine-tuning would normally incur overfitting (Wang et al., 2018). To overcome this, an episodic training paradigm was proposed (Vinyals et al., 2016). In such a paradigm, episodic training replaces the conventional mini-batch training, such that a batch of episodic tasks, each of which have the same setting as the testing environment, are presented to the learning model; and in each episodic task, the model learns to predict the classes of unlabeled points (the query set) using very few labeled examples (the support set). By this, the learning model acquires the transferable knowledge across tasks, and due to the consistency between the training and testing environments, the model is able to generalize to novel but related downstream tasks. Although this set-to-set few-shot learning paradigm has made great progress, in its current supervised form, it requires a large number of labeled examples for constructing episodic tasks, which is often infeasible or too expensive in practice. So, can we build up a few-shot learner in the paradigm of episodic training using only unlabeled data?

It is well-known that humans have the remarkable ability to learn a concept when given only several exposures to its instances, for example, young children can effortlessly learn and generalize the concept of "giraffe" after seeing a few pictures of giraffes. While the specifics of the human learning process are complex (trial-based, perpetual, multi-sourced, and simultaneous for multiple tasks) and yet to be solved, previous works agree that its nature is progressive and unsupervised in many cases (Dupoux, 2018). Given a set of unlabeled items, humans are able to organize them into different clusters by comparing one with another. The comparing or associating process follows a *coarse-to-fine* manner. At the beginning of learning, humans tend to group items based on fuzzy-rough knowledge such as color, shape, or size. Subsequently, humans build up associations between items using more fine-grained knowledge, i.e., stripes of images, functions of items, or other domain knowledge. Furthermore, humans can extract representative representations across categories and apply this capability to learn new concepts (Kemp et al., 2010; Wang et al., 2014; Gopnik and Bonawitz, 2015).

In the present study, inspired by the unsupervised and progressive characteristics of human learning, we propose an unsupervised model for few-shot learning via a self-supervised training procedure (UFLST). Different from previous unsupervised learning methods, our model integrates unsupervised learning and episodic training into a unified framework, which facilitates feature extraction and model training iteratively. Basically, we adopt the episodic training paradigm, taking advantage of its capability of extracting transferable knowledge across tasks, but we use an unsupervised strategy to construct episodic tasks. Specifically, we apply progressive clustering to generate pseudo labels for unlabeled data, and this is done alternatively with feature optimization via few-shot learning in an iterative manner (**Figure 1**). Initially, unlabeled data points are assigned into several clusters, and we sample a few training examples from each cluster together with their pseudo labels (the identities of clusters) to construct a set of episodic tasks having the same setting as the testing environment. We then train the few-shot learner using the constructed episodic tasks and obtain improved feature representations for the data. In the next round, we use the improved features to re-cluster data points, generating new pseudo labels and constructing new episodic tasks, and train the few-shot learner again. The above two steps are repeated till a stopping criterion is reached. After training, we expect that the few-shot learner has acquired the transferable knowledge (the optimized feature representations) suitable for a novel task of the same setting as in the episodic training. Using benchmark datasets, we demonstrate that our model outperforms other unsupervised few-shot learning methods and approaches to the performances of fully supervised models.

1.1. Related Works

In the paradigm of episodic training, few-shot learning algorithms can be divided into two main categories: "learning to optimize" and "learning to compare." The former aims to develop a learning algorithm which can adapt to a new task efficiently using only few labeled examples or with only few steps of parameter updating (Andrychowicz et al., 2016; Ravi and Larochelle, 2016; Finn et al., 2017; Mishra et al., 2017; Nichol and Schulman, 2018; Rusu et al., 2018), and the latter aims to learn a proper embedding function, so that prediction is based on the distance (metric) of a novel example to the labeled instances (Vinyals et al., 2016; Snell et al., 2017; Liu et al., 2018; Ren et al., 2018; Sung et al., 2018). In the present study, we focus on the "learning to compare" framework, although methods belonging to the other framework can also be integrated into our model.

A number of unsupervised few-shot learning models have been developed recently. Hsu et al. (2018) proposed a method called CACTUs, which constructs tasks from unlabeled data by partitioning features extracted by some prior unsupervised feature learning methods, e.g., ACAI, BiGAN, and DeepCluster in an automatic way and performs meta-learning over the constructed tasks. Khodadadeh et al. (2018) proposed a method called UMTRA, which utilizes the statistical diversity properties and domain-specific augmentations to generate training and validation data. Antoniou and Storkey (2019) proposed a similar model called AAL, which uses data augmentations of the unlabeled support set to generate the query data. All these methods construct episodic tasks with the aid of unsupervised

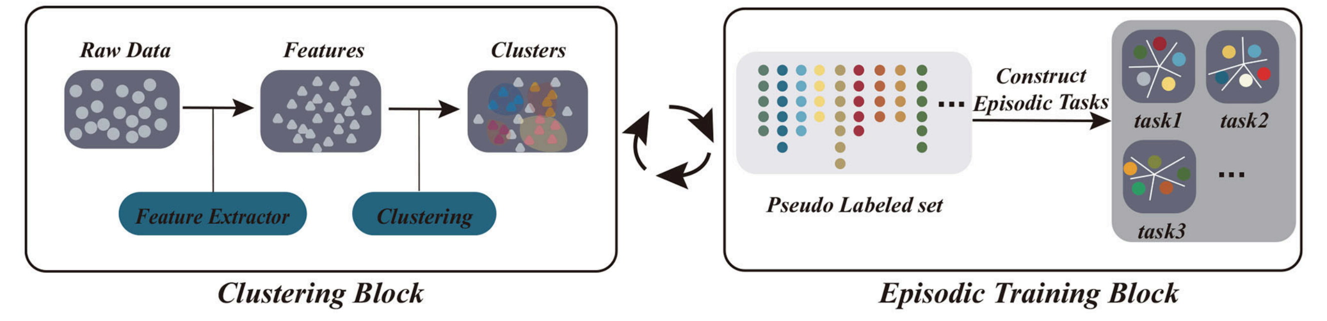


FIGURE 1 | The scheme of our model UFLST, which integrates two iterative processes: clustering and episodic training. At each iteration, unlabeled datapoints are clustered based on the extracted features, and pseudo labels are assigned according to the cluster identities. After clustering, a set of episodic tasks are constructed by sampling from the pseudo labeled data, and the few-shot learner is trained, which further optimizes feature representations. The two processes are repeated.

feature embedding or data augmentation; whereas in our method, the construction of episodic tasks and model training are performed iteratively within the same few-shot embedding network, and they facilitate each other.

The idea of iterative training used in our model is a type of self-supervised training, which aims to artificially generate pseudo labels for unlabeled data and then perform feature learning as in the supervised manner iteratively. It is quite useful when supervisory signals are not available or too expensive (de Sa, 1994). This idea was first applied in NLP tasks, which aims to self-train a two-phase parser-reranker system using unlabeled data (McClosky et al., 2006). Xie et al. (2016) proposed a Deep Embedded Clustering network to jointly learn cluster centers and network parameters. Caron et al. (2018) further proposed strategies to solve the degenerated solution problem during deep clustering. Fan et al. (2018) and Song et al. (2018) applied the iterative training idea to the person re-identification task, both of which aim to transfer the extracted feature representations to an unseen domain. However, none of these studies have considered integrating iterative clustering and episodic training in unsupervised few-shot learning as we do in this work.

2. MATERIALS AND METHODS

2.1. Preliminaries

In this section, we introduce the proposed model UFLST in detail. Consider a M-way K-shot classification task. Our goal is to train a few-shot learner based on the unlabeled data set $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^{N}$, where N is the total number of unlabeled datapoints. The previous studies have demonstrated that by matching the training and testing paradigms, episodic learning can extract transferable knowledge across tasks suitable for few-shot classification (Vinyals et al., 2016). In the supervised setting, one can easily construct a set of episodic tasks, with each task having K training examples $\{(\mathbf{x}_k, y_k)\}$ per class to learn the few-shot classifier and Q query examples per class to evaluate the learned classifier. Totally, there are $K + Q$ examples for each of M classes in each episodic task. In the unsupervised setting, however, we do not have labeled data to construct episodic tasks directly. Therefore, we consider using pseudo labels generated by a clustering algorithm to support episodic learning. Different from the previous work (Hsu et al., 2018) which uses a prior trained feature embedding network to extract fixed representations of data, data representations in our model are dynamically fine-tuned along with the episodic training.

Let us denote the embedding function in UFLST as f_θ, which takes $\mathcal{X}$ as the input and outputs the corresponding feature vector $\mathcal{Z} = \{\mathbf{z}_i\}$, for $i = 1, \ldots, N$, where θ represents the network parameters. Firstly, we cluster the unlabeled data based on the embedding features $\mathcal{Z}$ and obtain the pseudo labels of data $\{y_i\}$, for $i = 1, \ldots, N$. Secondly, using the pseudo labeled data, we construct a set of episodic tasks $\mathcal{T} = \{T_1, T_2, ..., T_S\}$, with S the number of constructed tasks in the current iteration, and carry out episodic learning, which improves the embedding features $\mathcal{Z}$ further. Notably, each episodic task T_s has the same setting as the application, i.e., it is a M-way K-shot classification. The above two steps are performed iteratively until a stopping criterion is reached. Below describes the two training processes in more detail.

2.2. Data Clustering

2.2.1. Distance Metric for Clustering

To cluster data, the first is to choose a suitable metric measuring the distance between data points. For constructing a large number of episodic tasks, an over-complete partition of data points is preferred, leading to a large number of classes with a small number of examples in each class. In such a situation, the conventional Euclidean distance or the Cosine distance is no longer optimal. Inspired by the re-ranking idea used in object retrieval as a post-processing tool to improve the retrieval accuracy, we propose to use the k-reciprocal Jaccard distance (KRJD) metric (Qin et al., 2011; Zhong et al., 2017) as the distance measurement between two feature points $\mathbf{z}_i$ and $\mathbf{z}_j$, which is written as

$$J_{ij} = 1 - \frac{|R(\mathbf{z}_i, k) \cap R(\mathbf{z}_j, k)|}{|R(\mathbf{z}_i, k) \cup R(\mathbf{z}_j, k)|}. \tag{1}$$

Here, $R(\mathbf{z}, k)$ counts the k-reciprocal nearest neighbors of a feature point $\mathbf{z}$ and is given by

$$R(\mathbf{z}, k) = \left\{ \mathbf{z}_j \mid \left(\mathbf{z}_j \in N(\mathbf{z}, k)\right) \cap \left(\mathbf{z} \in N(\mathbf{z}_j, k)\right) \right\}, \tag{2}$$

where $N(\mathbf{z}, k)$ denotes the k nearest neighbors of $\mathbf{z}$. $R(\mathbf{z}, k)$ imposes the condition that $\mathbf{z}$ and each element of $R(\mathbf{z}, k)$ are mutually the k nearest neighbors of each other.

Compared to the Euclidean distance, KRJD takes into account the reciprocal relationship between data points, and hence is a stricter metric measuring whether two feature points match or not. Given a query probe, we find that the results of nearest neighbors based on the KRJD is more accurate than that of the Euclidean distance (i.e., the k-nearest neighbors) as demonstrated in **Figure 2** (see **Appendix 1** for more detail).

2.2.2. Density-Based Spatial Clustering

To partition feature points and generate pseudo labels, we adopt a clustering method called density-based spatial clustering algorithm (DBSCAN) (Ester et al., 1996). This method regards clusters as the areas of high density separated by low density regions, that is, a cluster is composed of a set of core points (i.e., those points in a high density region close to each other) and a set of non-core points (i.e., those points in the surrounding low density regions close to the core points but not to themselves). Compared to the conventional Kmeans algorithm, DBSCAN has a number of appealing properties: (1) it applies to any shape of clusters, as opposed to the Kmeans algorithm assuming that clusters are convex; (2) it requires no assumption of the number of clusters; (3) it can detect outliers, which is extremely useful for iterative training, as data points are typically intertwined in the first few iterations.

After applying DBSCAN, we get the pseudo label set (the cluster identity), which is expressed as

$$\{y_i\} = DBSCAN\left(ms, \epsilon, \{\mathbf{z_i}\}\right), \tag{3}$$

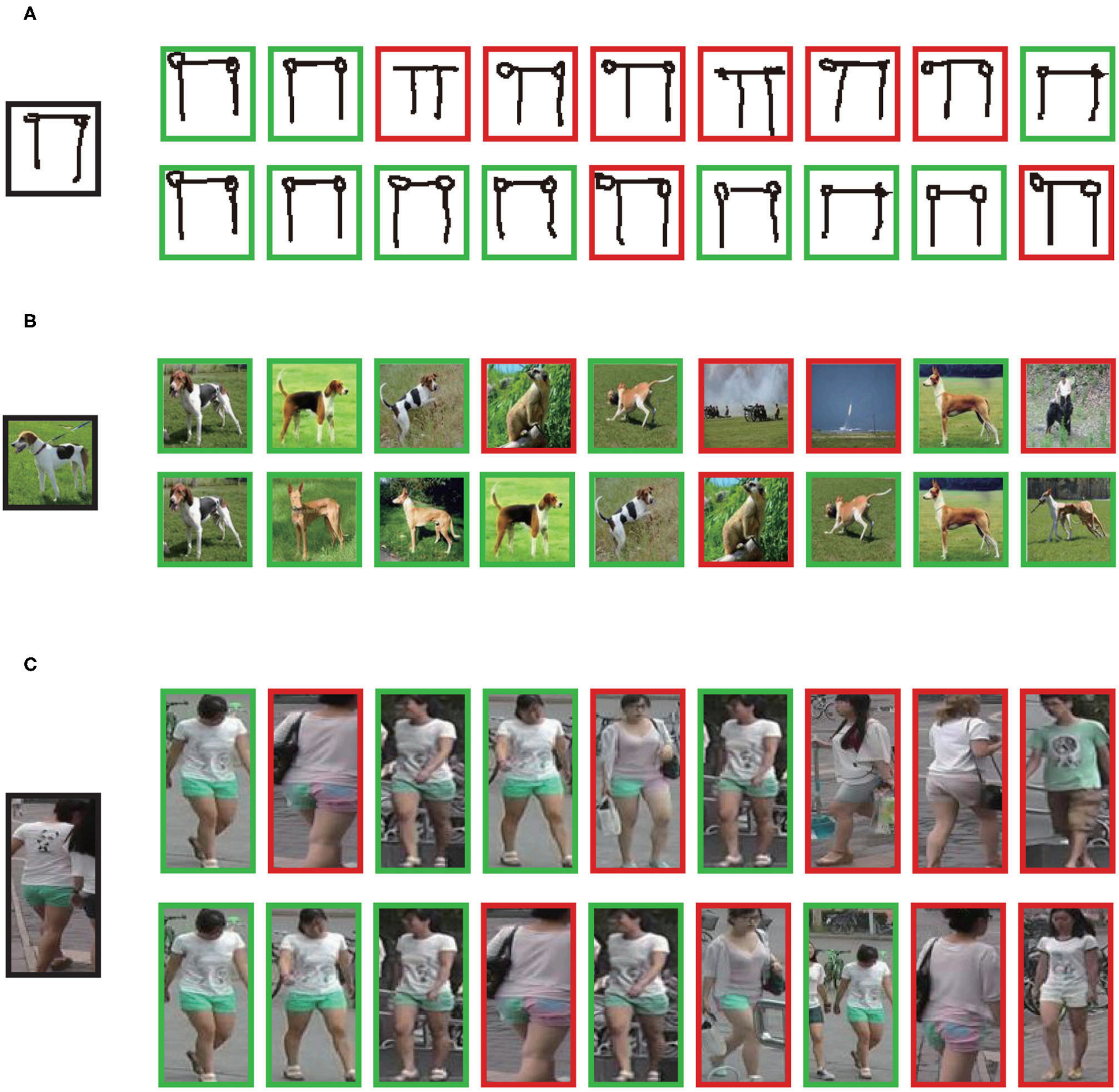


FIGURE 2 | Comparison between k-nearest neighbors and k-reciprocal nearest neighbors. Given an probe (in the black box), nearest neighbors of the example are shown. Examples in green boxes are those in the same class and examples in red boxes are those in different classes. **(A–C)** Examples from Omniglot, MiniImageNet, and FS-Market1501, respectively. The upper row in each panel is the result of k-nearest neighbors and the lower row in each panel is the result of k-reciprocal nearest neighbors. By adopting KRJD, more positive examples (those in the same class) appear in the nearest neighborhood of the probe.

where the parameter ms defines the minimum sample value, i.e., the minimum number of points huddled together for a region to be considered as dense, and the parameter ϵ defines the distance threshold, i.e., the maximum distance for two points to be considered as in the same neighborhood. Higher ms or lower ϵ indicate higher density is necessary to form a cluster. Both ms and ϵ affect the cluster numbers and the size of clusters. In general, we want the constructed episodic tasks $\mathcal{T}$ to be diverse, so that transferable knowledge can be acquired by the few-shot learner. This corresponds to setting small ms and ϵ. We will discuss the choice of ms and ρ in section 2.5.

2.3. Episodic Training

After removing outliers (i.e., those data points in low density regions in the feature space) in DBSCAN, we construct episodic tasks using the remaining pseudo labeled data $\{(\widetilde{\mathbf{x}}_i, \widetilde{y}_i)\}_{i=1}^{\widetilde{N}}$, with

$\widetilde{N}$ the number of remaining points. For each episodic task T_i, we randomly sample M classes and $K + Q$ examples per class as described in section 2.1, with $K + Q \leq ms$.

A number of metric loss functions can be used in our model, including the prototypical loss (Snell et al., 2017), the triplet loss (Weinberger and Saul, 2009; Hermans et al., 2017), the contrastive loss (Hadsell et al., 2006), and the center loss (Wen et al., 2016). To save space, here we mainly describe the prototypical loss. More results of using other metric loss functions can be found in **Appendix 2**. The prototypical loss aims to learn a prototype for each class and then discriminate a novel example based on its distance to all M prototypes, which is written as

$$L_{proto}(\mathbf{z}, \mathbf{c}_p; \theta) = \frac{\exp(-\|\mathbf{z} - \mathbf{c}_p\|_2^2)}{\sum_m^M \exp(-\|\mathbf{z} - \mathbf{c}_m\|_2^2)}, \tag{4}$$

where $\mathbf{z}$ is a data point from the query set of class p, and $\mathbf{c}_m$ is the prototype of class m given by $\mathbf{c}_m = \sum_{\mathbf{z}_i \in S_m}(\mathbf{z}_i)/K$, with S_m the support set of class m. In practice, we choose to minimize the negative log value of Equation 4, i.e., $L_{proto}^{\log}(\mathbf{z}, \mathbf{c}_p; \theta) = -\log L_{proto}(\mathbf{z}, \mathbf{c}_p; \theta)$, as the log value better reflects the geometry of the loss function, making it easier to select a suitable learning rate to minimize the loss function.

In summary, the above two steps for data clustering and episodic training are performed iteratively. They facilitate each other, similar to the EM-style algorithm: data clustering frequently generates pseudo labeled data for episodic learning, and the latter improves the feature representations of data, which in return further improve the clustering quality and few-shot learning (see section 4 for more discussions on why the iterative learning works). The pseudo code of UFLST is summarized in Algorithm 1.

Algorithm 1: Unsupervised Few-shot Feature Learning via Self-supervised Training (UFLST)

```
Input: Unlabeled data set $\mathcal{X} = \{\mathbf{x}_i\}$, the few-shot feature embedding $f_{\theta^0}$, the training iteration $T$.
Output: Trained few-shot embedding $f_{\theta^T}$
$t = 0$
repeat
    Clustering:
    Extracting features $\{\mathbf{z}_i\}$ of $\{\mathbf{x}_i\}$ using the feature extractor $f_{\theta^t}$.
    Calculating KRJD $J_{ij}$ based on the K-reciprocal nearest neighbors of any data pairs $\mathbf{z}_i$ and $\mathbf{z}_j$.
    Clustering data using DBSCAN and generating pseudo labels $\{y_i\}$.
    Removing outliers and obtaining the pseudo labeled data set $\{(\widetilde{\mathbf{x}}_i, \widetilde{y}_i)\}$.
    Episodic Training:
    Constructing a set of episodic tasks $\{\mathcal{T}_s\}$; for each task, randomly sampling $M$ classes with $K+Q$ examples per class from $\{(\widetilde{\mathbf{x}}_i, \widetilde{y}_i)\}$.
    Updating model parameters $\theta^t$ by training the few-shot learner on the series of episodic tasks $\{\mathcal{T}_s\}$.
    $t = t + 1$
until $t = T$
```

2.4. Datasets

Omniglot contains 1,623 different handwritten characters from 50 different alphabets. There are 20 examples per class and each of them was drawn by a different human subject via Amazon's Mechanical Turk. Following Vinyals et al. (2016), we split the data into two parts: 1,200 characters for training and 423 for testing, and we resize the images to 32 × 32, instead of 28 × 28.

MiniImageNet is derived from the ILSVRC-12 dataset. We follow the data split as suggested in Ravi and Larochelle (2016), which contains 100 classes including 64 for training, 16 for validating, and 20 for testing. Each class contains 600 colored images of size 84 × 84.

FS-Market1501 is a person re-identification (Re-ID) dataset modified from the Market1501 dataset (Zheng et al., 2015). The training set contains 12,936 images with 751 pedestrian identities and the testing set contains 16,483 images with the remaining 750 pedestrian identities. All images were resized to 256 × 128. For more details of how to construct FS-Market1501, see **Appendix 3**.

2.5. Implementation Details

When training on Omniglot and MiniImageNet, we set the model architecture to be the same as in the previous works for fair comparison. The model consists of four stacked layers, and each layer comprises 64-filter 3 × 3 convolution, followed by a batch normalization, a ReLU nonlinearity, and 2 × 2 max-pooling. When training on FS-Market1501, due to high variance in pedestrian pose and image illumination, we use Resnet50 pretrained on ImageNet as the backbone, followed by a global max-pooling layer and a batch normalization layer. Omniglot is relatively easy compared to the other two datasets, and therefore we only pre-process data with normalization. For MiniImageNet and FS-Market1501, we randomly flip images horizontally and crop them with random sizes, and then normalize them with the channel-wise mean and standard deviation of the whole dataset. Color information is important to partition images in FS-Market1501 (pedestrians with the same ID vary in pose, view angle, and illumination but not in the color), while it is not that important to partition images in MiniImageNet (Caron et al., 2018). Hence, we discard color information and increase local contrast by adding a linear transformation based on Sobel filters as proposed in Bojanowski and Joulin (2017) and Paulin et al. (2015). For the clustering method DBSCAN, we set $ms = 2$ and ϵ to be the mean of top P values of distance pairs, with $P = \rho N(N-1)/2$ and $\rho = 0.0015$. The values of ms and ϵ are set to be relatively small to ensure that feature points are well-separated, so that diverse episodic tasks can be constructed (for more details of the choice of ms and ϵ, see **Appendix 4**). For the prototype loss, we used a higher "way" value ($M = 60$) during training, which leads to better performances as empirically observed in Snell et al. (2017). Since it is possible that the numbers of points in some clusters are too small, we only train the model in the M-way 1-shot learning scenario, i.e., $K = Q = 1$. The total number of iterations during training is set to be 100, and in each iteration, 500 episodic tasks are constructed. We used Adam with momentum to update model parameters, and the learning rate is set to be 0.001.

3. RESULTS

3.1. Comparison With Non-episodic Learning Methods

Episodic learning plays a key role in leveraging unsupervised few-shot feature learning. To demonstrate this, we first compare our model with other unsupervised feature learning methods without employing episodic learning. Three such methods are chosen, which are (Denoising) AutoEncoder (Vincent et al., 2008), InfoGAN (Chen et al., 2016), and DeepClustering (Caron et al., 2018) (for the detailed training process of these methods, see **Appendix 5**). These methods are the typical approaches used to learn useful feature representations, covering a wide range of unsupervised feature learning strategies including reconstruction (prediction), two-player games, discriminative clustering, and so on. For comparison, we use the features extracted by these methods to calculate the prototype of each class directly and perform the M-way K-shot classification. The results are presented in **Table 1**, which shows that: (1) compared to other unsupervised feature learning methods whose learning objective is different from ours, iterative data clustering and episodic learning improves the few-shot learning performance significantly, even when the Kmeans clustering with the Euclidean distance is used in our model; (2) by applying DBSCAN with the KRJD metric, the performance of our model is improved further to a large extent. Notably, DeepClustering also jointly learns the parameters of a neural network and the cluster assignments of the resulting features. However, it optimizes the feature representations with a relatively simple learning objective (softmax classification) which is not suitable for few-shot classification.

3.2. The Effect of Iterative Training

In our model, iterative training will gradually improve the clustering quality and the performance of the few-shot learner. To demonstrate this, we randomly select 10 hand-written characters from the Futurama alphabets in Omniglot and visualize clustering behaviors over iteration with T-SNE (Maaten and Hinton, 2008). As shown in **Figure 3**, initially all data points are intertwined with each other and no clear cluster structure exists. Over training, clusters gradually emerge, in the sense that data points from the same class are grouped together and the margins between different classes are enlarged. This indicates that our model gradually "discovers" the underlying semantic structure of the data. We quantify the clustering quality by computing the normalized Mutual Information (NMI) between the pseudo labels generated by the clustering algorithm $\{\widetilde{y_i}\}$ and the ground truth of real labels $\{\hat{y}_i\}$, which is given by,

$$NMI\left(\{\hat{y}_i\}, \{\widetilde{y_i}\}\right) = \frac{I(\{\hat{y}_i\}, \{\widetilde{y_i}\})}{\sqrt{H(\{\hat{y}_i\})H(\{\widetilde{y_i}\})}}, \tag{5}$$

where $I(\cdot,\cdot)$ is the mutual information between $\{\hat{y}_i\}$ and $\{\widetilde{y_i}\}$, and $H(\cdot)$ the entropy. The value of NMI lies in $[0,1]$, with 1 standing for the perfect alignment between two sets. Note that NMI is independent of the permutation of labeling orders. As shown in **Figure 4** (left), the value of NMI increases with the training iterations and gradually reaches a high value close to 1. Remarkably, the value of NMI well predicts the classification accuracy of the few-shot learning (**Figure 4**, right). These results demonstrate that iterative data clustering and episodic training are able to discover the underlying structure of data manifold, and extract the representative features of data necessary for the few-shot classification task.

3.3. Comparison With State-of-the-Art Unsupervised Few-Shot Learning Methods

We compare our model with other state-of-the-art unsupervised few-shot learning methods, including CACTUs (Hsu et al.,

TABLE 1 | Performances of our model compared to other non-episodic unsupervised feature learning methods on Omniglot and MiniImageNet.

			Omniglot				MiniImageNet			
Methods (M, K)	**Clustering**	**Metric**	**(5,1)**	**(5,5)**	**(20,1)**	**(20,5)**	**(5,1)**	**(5,5)**	**(5,20)**	**(5,50)**
Baseline	N/A	N/A	57.97	79.25	34.17	59.33	25.91	32.38	37.01	38.95
AutoEncoder	N/A	N/A	53.63	77.34	32.98	55.01	26.17	33.01	37.98	39.39
Denoising autoEncoder	N/A	N/A	59.63	79.89	34.78	60.88	27.81	34.19	39.01	40.11
InfoGAN	N/A	N/A	51.49	76.38	31.01	53.99	29.81	36.47	40.17	42.46
BiGAN+KNN	N/A	N/A	49.55	68.06	27.37	46.70	25.56	31.10	37.31	43.60
BiGAN+LC	N/A	N/A	-	-	-	-	27.08	33.91	44.00	50.41
DeepClustering	Kmeans	Euclidean	59.07	79.81	34.05	60.12	28.91	36.01	39.29	41.98
UFLST	Kmeans	Euclidean	69.54	86.18	47.11	69.19	31.77	43.03	51.35	55.72
UFLST	BSCAN	KRJD	**96.51**	**99.23**	**90.27**	**97.22**	**37.75**	**50.95**	**59.18**	**62.27**

Baseline performance means training from scratch. Results based on BiGAN are adapted from Hsu et al. (2018). For complete results with confidence intervals, see ***Appendix 6****. The best performances are in bold.*

2018), UMTRA (Khodadadeh et al., 2018), and AAL (Antoniou and Storkey, 2019), as shown in **Table 2**. On Omniglot, our model outperforms them to a large extent. Remarkably, the best performances of our model approaches that of two supervised methods, which are the upper bounds for unsupervised learning. Our model also achieves significant improvement on MiniImageNet (note that we only test the model under the 5-way few-shot learning scenario). For example, in the 5-way 1-shot scenario, our model achieves 37.75%, which is significant compared to the baseline performance 25.91%.

We also note that some methods outperform our model on MiniImageNet, e.g., DeepCluster-CACTUs-ProtoNets and

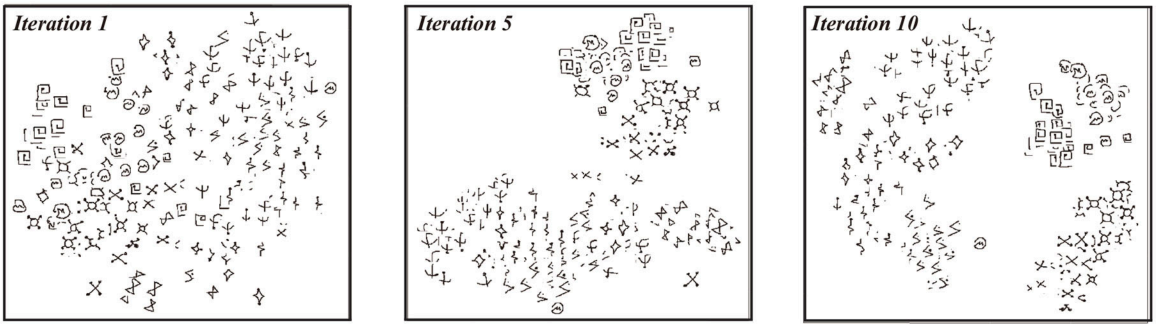


FIGURE 3 | Visualizing clustering results during iterative training with T-SNE. 10 characters from the Futurama alphabets in Omniglot are were selected and results from iteration 1, iteration 5, and iteration 10 are showed here.

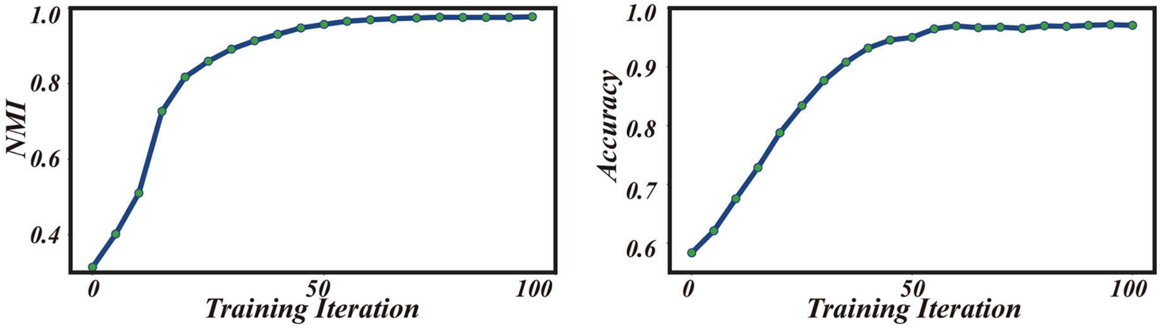


FIGURE 4 | Performances of iterative training under the 5-way 1-shot learning scenario on the Omniglot dataset. **(Left)** NMI vs. training iteration. **(Right)** Classification accuracy vs. training iteration.

TABLE 2 | Comparison to state-of-the-art unsupervised few-shot learning models on Omniglot and MiniImageNet under different settings.

	Omniglot				MiniImageNet			
Methods (M, K)	**(5,1)**	**(5,5)**	**(20,1)**	**(20,5)**	**(5,1)**	**(5,5)**	**(5,20)**	**(5,50)**
ACAI/DC-CACTUs-MAML (Hsu et al., 2018)	68.84	87.78	48.09	73.36	39.90	**53.97**	**63.84**	**69.64**
ACAI/DC-CACTUs-ProtoNets (Hsu et al., 2018)	68.12	83.58	47.75	66.27	39.18	53.36	61.54	63.55
BiGAN-CACTUs-MAML (Hsu et al., 2018)	58.18	78.66	35.56	58.62	36.24	51.28	61.33	66.91
BiGAN-CACTUs-ProtNets (Hsu et al., 2018)	54.74	71.69	33.40	50.62	36.62	50.16	59.56	63.27
UMTRA+AutoAugment (Khodadadeh et al., 2018)	83.80	95.43	74.25	92.12	**39.93**	50.73	61.11	67.15
AAL-MAML++ (Antoniou and Storkey, 2019)	88.40	97.96	70.21	88.32	33.30	49.18	–	–
AAL-ProtoNets (Antoniou and Storkey, 2019)	84.66	89.14	68.79	74.28	37.67	40.29	–	–
UFLST+Kmeans+Euclidean (ours)	69.54	86.18	47.11	69.19	31.77	43.03	51.35	55.72
UFLST+DBSCAN+KRJD (ours)	**96.51**	**99.23**	**90.27**	**97.22**	37.75	50.95	59.18	62.27
MAML (Finn et al., 2017) (supervised)	98.7	99.9	95.8	98.9	46.81	62.13	71.03	75.54
ProtoNets (Snell et al., 2017) (supervised)	98.8	99.7	96.0	98.9	46.56	62.29	70.05	72.04

*Results based on BiGAN are adapted from Hsu et al. (2018). For complete results with confidence intervals, see **Appendix 7**. The best performances are in bold.*

UMTRA-AutoAugment achieve 39.18 and 39.93% in the 5-way 1-shot scenario, respectively. The reasons we believe are due to three aspects. Firstly, for the convenience of comparing to other (un)supervised few-shot learning methods, we have used the 4-layer convnet as the few-shot embedding network. Such a simple network is unable to adequately extract the semantic meanings of images under the unsupervised setting, especially as the in-class variations of MiniImageNet are large but the total size of the dataset is small (only 64 classes with 600 images per class in the training set). Secondly, for constructing diverse episodic tasks, our model prefers to over-segment the data into hundreds of clusters, whereas the ground truth cluster number of MiniImageNet is only 64. This induces mismatch between the constructed episodic tasks and the ground truth. Thirdly, the methods outperforming our model adopt either powerful prior unsupervised feature learning to partition data points (the CACTU-based model) or complicated data augmentation strategies to construct the episodic tasks (the UMTRA-based model and the AAL-based model), while our model partitions data points with the features directly extracted from the few-shot embedding network and only adopts a simple data augmentation strategy to avoid overfitting. One solution is to use deeper feature embedders, e.g., Resnet12, AlexNet in our model to improve the performance (see **Appendix 9**). Even so, our model still achieves competitive results compared to other unsupervised few-shot learning methods.

3.4. Results on FS-Market1501

In order to show the applicability of our model to a real-world few-shot learning problem, we apply our model on the FS-Market1501 dataset which has been described in section 2.4. In reality, labeled data is extremely lacking for person Re-ID, and unsupervised learning becomes crucial. Results in **Table 3** show that our UFLST model performs very well on the 1-shot learning problem on this dataset. Note that the 1-shot learning problem we demonstrate here is to mimic the typical single query setting in person Re-ID. For example, 50-way 1-shot means the model needs to identify a pedestrian from one of 50 unknown persons by training a classifier with only one image per person. To compare our model with the supervised results as described in section 3.3, we train a supervised model with the same model architecture, i.e., the Resnet50 backbone pretrained on ImageNet as described in section 2.5. Overall, we observe that our model achieves encouraging performances compared to the supervised methods, in particular, in the scenario of low-way classification. This suggests that our model is potentially feasible in practice for person Re-ID when annotated labels are unavailable.

TABLE 3 | Performances of our model on FS-Market1501 with different settings.

	5-way	10-way	15-way	20-way	50-way	100-way
Baseline	48.8	35.7	29.7	27.8	20.9	16.4
UFLST-Tripetloss	72.8	63.0	56.2	53.4	42.5	35.4
UFLST-Prototypeloss	88.3	81.2	75.8	73.0	62.5	54.0
UFLST-HardTripletloss	**91.4**	**86.9**	**81.6**	**80.4**	**70.1**	**62.1**
Supervised upper bound	96.8	94.7	92.5	91.1	83.7	77.3

Only 1-shot learning is considered to mimic the typical single query evaluation condition in person Re-ID. We adopt three metric losses to optimize the model, see ***Appendix 8*** *for detail. The best performances are in bold.*

4. CONCLUSION AND DISCUSSION

In this study, we have proposed a model UFLST for unsupervised few-shot learning. Different from other unsupervised feature learning methods, such as the prediction-based and the GAN-based ones, our model exploits the paradigm of episodic training, which is a more effective way to implement few-shot learning. Recently, a few unsupervised few-shot learning models based on episodic learning were proposed, and they have taken different strategies to construct episodic tasks from unlabeled data. For instance, CACTUs constructs episodic tasks by partitioning the features extracted by a prior-trained unsupervised feature embedding network with different objective functions and then train the few-shot learner (Hsu et al., 2018). UMTRA utilizes a domain-specific data augmentation strategy to generate synthetic tasks for the meta-learning phase, while in such a way, the constructed episodic tasks are restricted by the data augmentation strategy (Khodadadeh et al., 2018). Different from the above methods, we propose a simple yet effective way to construct episodic tasks, that is, we partition the features directly from the few-shot embedding network and do this in an iterative manner along with the training of the few-shot learner; and by this, the construction of episodic tasks and the training of few-shot learner are improved concurrently. Furthermore, to improve the clustering quality, we have proposed to use the k-reciprocal Jaccard distance metric to reduce false positive examples during the clustering.

We have demonstrated encouraging performances of our model on two benchmark datasets, Omniglot, and MiniImageNet. We also constructed a new dataset called FS-Market1501 adapted from Market1501 to test our model, and demonstrated the feasibility of our model to real-world applications. The high efficiency of our model also prompts us to think about why it works. The key of our model is the iterative implementation of data clustering and episodic training, and they tend to facilitate each other as the EM-style algorithm. At the beginning of training, the few-shot embedding network is randomly initialized, and the embedded features are intertwined with each other, making the constructed episodic tasks very noisy. However, even in such a situation, the embedded features are not completely random as observed in Noroozi and Favaro (2016), which showed that the performance of a randomly initialized convnet is above the chance level. For example, a simple multilayer perceptron built on top of the last convolutional layer of a random AlexNet achieves 12% accuracy on ImageNet, while the chance level is only 0.1%. This implies that this weak signal can be exploited to bootstrap the discriminative power of our model through iterative training. As shown in **Figures 3**, **4**, data clustering and feature extraction in our model facilitate each other, which eventually produces a well-performed few-shot learner. To our knowledge, our work is the first one that integrates progressive clustering and episodic training for unsupervised few-shot learning. Notably, the idea of unsupervised iterative learning of our model agrees with the self-learning nature of humans. It will be interesting to further explore the relationship between human learning and machine learning on unsupervised few-shot learning.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

ZJ designed the study, performed the experiments, and wrote the first draft of the manuscript. XZ helped with integrating algorithms and conducting experiments. TH and SW contributed to the conception and design of the study and revision. ZJ and SW wrote the final manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by Guangdong Province with Grant (No. 2018B030338001, SW). This work also received support from Huawei Technology Co., Ltd (YBN2019105137).

ACKNOWLEDGMENTS

This manuscript has been released as a pre-print at: https://arxiv.org/abs/1912.12178 (Ji et al., 2019).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fncom.2020.00083/full#supplementary-material

REFERENCES

Andrychowicz, M., Denil, M., Gomez, S., Hoffman, M. W., Pfau, D., Schaul, T., et al. (2016). "Learning to learn by gradient descent by gradient descent," in *Advances in Neural Information Processing Systems*, 3981–3989.

Antoniou, A., and Storkey, A. (2019). Assume, augment and learn: unsupervised few-shot meta-learning via random labels and data augmentation. *arXiv preprint arXiv:1902.09884.*

Bojanowski, P., and Joulin, A. (2017). "Unsupervised learning by predicting noise," in *Proceedings of the 34th International Conference on Machine Learning-Vol. 70*, 517–526.

Caron, M., Bojanowski, P., Joulin, A., and Douze, M. (2018). "Deep clustering for unsupervised learning of visual features," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 132–149. doi: 10.1007/978-3-030-01264-9_9

Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., and Abbeel, P. (2016). "Infogan: interpretable representation learning by information maximizing generative adversarial nets," in *Advances in Neural Information Processing Systems*, 2172–2180.

de Sa, V. R. (1994). "Learning classification with unlabeled data," in *Advances in Neural Information Processing Systems*, 112–119.

Dupoux, E. (2018). Cognitive science in the era of artificial intelligence: a roadmap for reverse-engineering the infant language-learner. *Cognition* 173, 43–59. doi: 10.1016/j.cognition.2017.11.008

Ester, M., Kriegel, H.-P., Sander, J., Xu, X. (1996). "A density-based algorithm for discovering clusters in large spatial databases with noise," in *KDD, Vol. 96*, 226–231.

Fan, H., Zheng, L., Yan, C., and Yang, Y. (2018). Unsupervised person re-identification: clustering and fine-tuning. *ACM Trans. Multim. Comput. Commun. Appl*. 14:83. doi: 10.1145/3243316

Finn, C., Abbeel, P., and Levine, S. (2017). "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proceedings of the 34th International Conference on Machine Learning-Vol*. 70, 1126–1135.

Gopnik, A., and Bonawitz, E. (2015). Bayesian models of child development. *Wiley Interdisc. Rev*. 6, 75–86. doi: 10.1002/wcs.1330

Hadsell, R., Chopra, S., and LeCun, Y. (2006). "Dimensionality reduction by learning an invariant mapping," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), Vol. 2*, 1735–1742. doi: 10.1109/CVPR.2006.100

Hermans, A., Beyer, L., and Leibe, B. (2017). In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*.

Hsu, K., Levine, S., and Finn, C. (2018). Unsupervised learning via meta-learning. *arXiv preprint arXiv:1810.02334*.

Ji, Z., Zou, X., Huang, T., and Wu, S. (2019). Unsupervised few-shot learning via self-supervised training. *arXiv preprint arXiv:1912.12178*.

Kemp, C., Goodman, N. D., and Tenenbaum, J. B. (2010). Learning to learn causal models. *Cogn. Sci*. 34, 1185–1243. doi: 10.1111/j.1551-6709.2010.01128.x

Khodadadeh, S., Bölöni, L., and Shah, M. (2018). Unsupervised meta-learning for few-shot image and video classification. *arXiv preprint arXiv:1811.11819*.

Liu, Y., Lee, J., Park, M., Kim, S., Yang, E., Hwang, S. J., et al. (2018). Learning to propagate labels: transductive propagation network for few-shot learning. *arXiv preprint arXiv:1805.10002*.

Maaten, L. V. D., and Hinton, G. (2008). Visualizing data using T-SNE. *J. Mach. Learn. Res*. 9, 2579–2605.

McClosky, D., Charniak, E., and Johnson, M. (2006). "Effective self-training for parsing," in *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics* (Association for Computational Linguistics), 152–159. doi: 10.3115/1220835.1220855

Mishra, N., Rohaninejad, M., Chen, X., and Abbeel, P. (2017). A simple neural attentive meta-learner. *arXiv preprint arXiv:1707.03141*.

Nichol, A., and Schulman, J. (2018). Reptile: a scalable metalearning algorithm. *arXiv preprint arXiv:1803.02999*.

Noroozi, M., and Favaro, P. (2016). "Unsupervised learning of visual representations by solving jigsaw puzzles," in *European Conference on Computer Vision* (Springer), 69–84. doi: 10.1007/978-3-319-4 6466-4_5

Paulin, M., Douze, M., Harchaoui, Z., Mairal, J., Perronin, F., and Schmid, C. (2015). "Local convolutional features with unsupervised training for image retrieval," in *Proceedings of the IEEE International Conference on Computer Vision*, 91–99. doi: 10.1109/ICCV.2015.19

Qin, D., Gammeter, S., Bossard, L., Quack, T., and Van Gool, L. (2011). "Hello neighbor: accurate object retrieval with k-reciprocal nearest neighbors," in *CVPR 2011*, 777–784. doi: 10.1109/CVPR.2011.5995373

Ravi, S., and Larochelle, H. (2016). Optimization as a model for few-shot learning.

Ren, M., Triantafillou, E., Ravi, S., Snell, J., Swersky, K., Tenenbaum, J. B., et al. (2018). Meta-learning for semi-supervised few-shot classification. *arXiv preprint arXiv:1803.00676*.

Rusu, A. A., Rao, D., Sygnowski, J., Vinyals, O., Pascanu, R., Osindero, S., and Hadsell, R. (2018). Meta-learning with latent embedding optimization. *arXiv preprint arXiv:1807.05960*.

Snell, J., Swersky, K., and Zemel, R. (2017). "Prototypical networks for few-shot learning," in *Advances in Neural Information Processing Systems*, 4077–4087.

Song, L., Wang, C., Zhang, L., Du, B., Zhang, Q., Huang, C., et al. (2018). Unsupervised domain adaptive re-identification: theory and practice. *arXiv preprint arXiv:1807.11334*.

Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P. H., and Hospedales, T. M. (2018). "Learning to compare: Relation network for few-shot learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1199–1208. doi: 10.1109/CVPR.2018.00131

Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P.-A. (2008). "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the 25th International Conference on Machine Learning*, 1096–1103. doi: 10.1145/1390156.1390294

Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al. (2016). "Matching networks for one shot learning," in *Advances in Neural Information Processing Systems*, 3630–3638.

Wang, R., Zhang, J.-Y., Klein, S. A., Levi, D. M., and Yu, C. (2014). Vernier perceptual learning transfers to completely untrained retinal locations after double training: a "piggybacking" effect. *J. Vis*. 14:12. doi: 10.1167/14.13.12

Wang, Y.-X., Girshick, R., Hebert, M., and Hariharan, B. (2018). "Low-shot learning from imaginary data," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7278–7286. doi: 10.1109/CVPR.2018.00760

Weinberger, K. Q., and Saul, L. K. (2009). Distance metric learning for large margin nearest neighbor classification. *J. Mach. Learn. Res.* 10, 207–244.

Wen, Y., Zhang, K., Li, Z., and Qiao, Y. (2016). "A discriminative feature learning approach for deep face recognition," in *European Conference on Computer Vision* (Springer), 499–515. doi: 10.1007/978-3-319-46478-7_31

Xie, J., Girshick, R., and Farhadi, A. (2016). "Unsupervised deep embedding for clustering analysis," in *International Conference on Machine Learning*, 478–487.

Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., and Tian, Q. (2015). "Scalable person re-identification: a benchmark," in *Proceedings of the IEEE International Conference on Computer Vision*, 1116–1124. doi: 10.1109/ICCV.2015.133

Zhong, Z., Zheng, L., Cao, D., and Li, S. (2017). "Re-ranking person re-identification with k-reciprocal encoding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1318–1327. doi: 10.1109/CVPR.2017.389

Conflict of Interest: The authors declare that this study received funding from Huawei Technology Co., Ltd (YBN2019105137). The funder was not involved in the study design, collection, analysis, interpretation of data, the writing of this article or the decision to submit it for publication.

frontiers in Computational Neuroscience

ORIGINAL RESEARCH
published: 12 November 2020
doi: 10.3389/fncom.2020.576841

GLSNN: A Multi-Layer Spiking Neural Network Based on Global Feedback Alignment and Local STDP Plasticity

Dongcheng Zhao [1,2†], Yi Zeng [1,2,3,4*†], Tielin Zhang [1], Mengting Shi [1,2] and Feifei Zhao [1]

[1] Research Center for Brain-Inspired Intelligence, Institute of Automation, Chinese Academy of Sciences, Beijing, China, [2] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China, [3] Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Beijing, China, [4] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China

Spiking Neural Networks (SNNs) are considered as the third generation of artificial neural networks, which are more closely with information processing in biological brains. However, it is still a challenge for how to train the non-differential SNN efficiently and robustly with the form of spikes. Here we give an alternative method to train SNNs by biologically-plausible structural and functional inspirations from the brain. Firstly, inspired by the significant top-down structural connections, a global random feedback alignment is designed to help the SNN propagate the error target from the output layer directly to the previous few layers. Then inspired by the local plasticity of the biological system in which the synapses are more tuned by the neighborhood neurons, a differential STDP is used to optimize local plasticity. Extensive experimental results on the benchmark MNIST (98.62%) and Fashion MNIST (89.05%) have shown that the proposed algorithm performs favorably against several state-of-the-art SNNs trained with backpropagation.

Keywords: SNN, plasticity, brain, local STDP, global feedback alignment

Edited by:
Ke Zhou,
Beijing Normal University, China

Reviewed by:
Jiang Wang,
Tianjin University, China
Radwa Khalil,
Jacobs University Bremen, Germany

****Correspondence:***
Yi Zeng
yi.zeng@ia.ac.cn

† These authors have contributed equally to this work

Received: *27 June 2020*
Accepted: *12 October 2020*
Published: *12 November 2020*

Citation:
Zhao D, Zeng Y, Zhang T, Shi M and Zhao F (2020) GLSNN: A Multi-Layer Spiking Neural Network Based on Global Feedback Alignment and Local STDP Plasticity.
Front. Comput. Neurosci. 14:576841.
doi: 10.3389/fncom.2020.576841

1. INTRODUCTION

Deep neural networks (DNNs) have been advancing the state-of-the-art performance in many domain-specific tasks, such as image classification (He et al., 2016), visual object tracking (Danelljan et al., 2015), visual object segmentation (Chen et al., 2017), etc. However, they are still far from the performance of efficiency and accuracy of information processing in the biological system. The structural connections (e.g., long-term feedback loops in the cortex) and functional plasticity (e.g., neighborhood plasticity based on discrete spikes) are carefully designed by the million years of evolution in the biological brain. This phenomenon has lead to the research of biologically plausible Spiking Neural Networks (SNNs). SNNs have received extensive research in recent years, and have a wide range of applications in various domains, such as brain function modeling (Durstewitz et al., 2000; Levina et al., 2007; Izhikevich and Edelman, 2008; Potjans and Diesmann, 2014; Zenke et al., 2015; Breakspear, 2017; Khalil et al., 2017a,b, 2018), image classification (Zhang et al., 2018a; Gu et al., 2019), decision making (Héricé et al., 2016; Zhao et al., 2018), object detection (Kim et al., 2019), and visual tracking (Luo et al., 2020). The discrete spike activation and high dimension information representation in SNNs make it more biologically plausible and energy-efficient. However, due to the non-differentiable characteristics, how to properly optimize the strength of synapses to improve the performance of the whole-brain network is still an open question.

Hebbian theory (Amit et al., 1994) could be considered as the first principle to demonstrate the relations between neurons, with the description of fire together, wire together. Later, Spiking Time Dependent Plasticity (STDP) (Bi and Poo, 1998) was proposed to model the synaptic plasticity. All the methods mentioned above are based on local adjustments without introducing global plasticity information.

Learning and inference in the brain are based on the interactions of feedforward connections and mutual feedback connections across the hierarchy of cortical areas, as shown in **Figure 1A**. Both anatomical and physiological evidences point to the feedback connections in the brain (Felleman and Van, 1991; Sporns and Zwi, 2004). A large number of feedback connections in the cortex connect the feedforward series in the reverse order, thereby bringing global information from the higher cortex to the early cortical areas during perceptual inference. Feedback connections from higher layers will make predictions represented by the lower layers, and the feedforward path will get the state of neurons in the entire hierarchy. Therefore, combining global long-term feedback connections with local plasticity rules to train the SNNs is an urgent problem to be explored.

In this paper, we proposed an SNN training method that combines global feedback connections and local differential STDP learning rule and performs favorably against several existing state-of-the-art methods. The contributions of this paper are summarized as follows:

- We introduce the feedback connections in SNNs, which will help to introduce global plasticity information. The feedback connections are random, and no additional calculations are introduced.
- The global feedback connections combined with the local STDP plasticity rule are combined to directly optimize the synaptic strengths of all layers, instead of transferring error layer by layer as Back-Propagation. Compared with other methods, it provides an alternative method for training deeper SNNs.
- Extensive experimental results on different datasets indicated that the proposed algorithm could significantly improve the learning ability of SNNs.

2. BACKGROUND

The success of DNNs is attributed mainly to the Back-Propagation algorithm (BP) (Rumelhart et al., 1986), which can take great advantage of the multilayer structure of neural networks to learn features related to a given task. However, firstly, the feedback path will have the symmetric weight of the forward path, which does not exist in biological systems, calling the weight transport problem (Lillicrap et al., 2016). Secondly, the precise derivatives of the operating point used in the corresponding feedforward path are needed. While for SNNs, information is transmitted in discrete spikes, and it is difficult to get the precise derivative of the operating point. Thirdly, the errors propagate layer by layer, which can easily lead to the problem of gradient vanishes or explosion. To tackle the problems mentioned above, many other learning rules are proposed to train the ANNs and further extended to train SNNs. In this section, we will review several of these approaches and several SNN frameworks in recent years.

2.1. Biologically Plausible Methods in ANNs

Recently, non-BP methods used to train neural networks can be roughly divided into three categories.

One family of promising approaches is Contrastive Hebbian Learning (Movellan, 1991). Equilibrium Propagation approaches (Scellier and Bengio, 2017) can be seen as a particular case of Contrastive Hebbian Learning. These kinds of energy-based models consist of two phases, the free phase is used

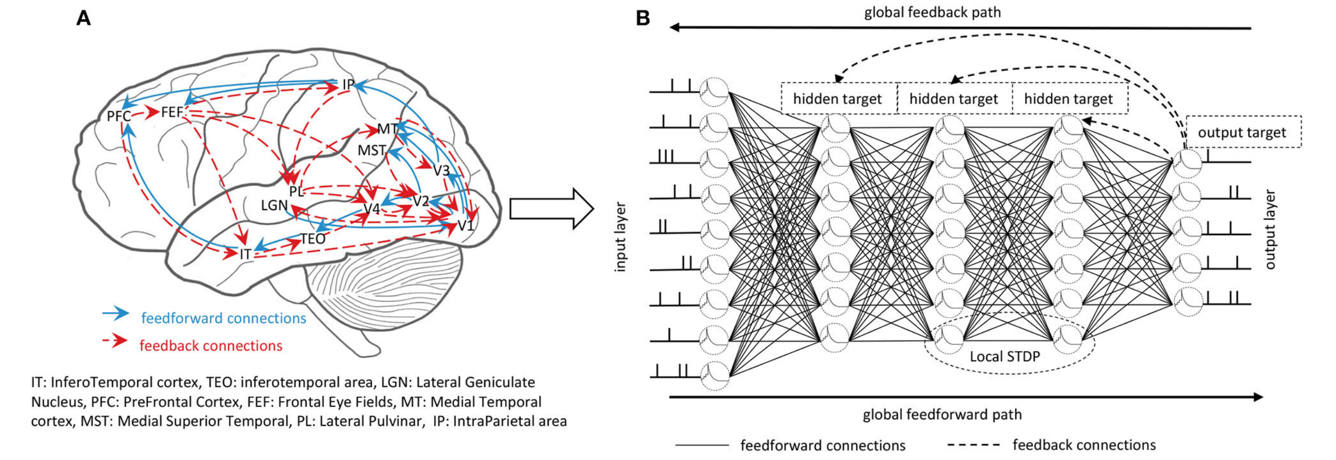


FIGURE 1 | (A) The feedforward and feedback interactions in the brain. The massive feedback connections interact with feedforward connections contributing to the learning and inference of the brain. **(B)** The whole training process of the GLSNN. The global feedforward path uses the LIF spiking neuron model to get the forward state. The global feedback path uses the direct connection between the output layer and the hidden layers to propagate the target. The local STDP learning rule helps to update the weight of the neighborhood layers.

to achieve the stationary distribution, and the clamp phase is used to update the network toward the target. Through the iteration of these two phases, the energy of the network can reach convergence gradually. However, due to the indirect feedforward process, the network state is obtained by minimizing the energy function. When the network becomes deeper, the entire algorithm will be unstable and therefore, difficult to train. We will give the experimental results below. Similarly, the free phase (feedforward propagation) and the clamp phase (feedback propagation) use the same weights, and the weight transpose problem still exists, as mentioned in backpropagation.

In order to solve the weight transport problem, the Random Feedback Alignment (RFA) algorithm (Lillicrap et al., 2016) uses a fixed random matrix B instead of the transposition of synaptic weights W, which can enable the network to converge to the optimal solution efficiently. Subsequent work DFA (Nøkland, 2016) propagates error signals through the direct connection matrix between the output layer and hidden layers. However, the error feedback does not influence the neural activity, which has not been confirmed by known biofeedback mechanisms based on neural communication.

In the Target Propagation (TP) family, for Difference Target Propagation (DTP) (Lee et al., 2015), targets for each hidden layer are passed through feedback connections, which avoids the weight transport problem, as the feedback connections are different from feedforward connections. The error-driven local representation alignment (LRA-E) (Ororbia and Mali, 2019), attempt to calculate the local target with the local error loss. Random feedback connections are utilized to transmit errors. However, the error is calculated and propagated layer by layer, and as the network deepens, performance will deteriorate.

2.2. Spiking Neural Networks

Much effort has been put into training SNNs, which can be roughly divided into three categories. First, directly convert the well-trained ANNs to SNNs. Second, SNNs are processed in some unique methods so that they can be trained with BP. Third, training SNNs with STDP and other biologically plausible methods.

For the conversion methods, SDBN (O'Connor et al., 2013) mapped an offline-trained deep belief network (DBN) onto an efficient event-driven SNN based on the Siegert approximation. The LIF response function is softened to lead to the bounded derivative value, which helps SDN (Hunsberger and Eliasmith, 2015) to convert the trained static network to a dynamic spiking network. WTSNN (Diehl et al., 2015) converted the DBNs into SNNs through weight and threshold balancing. Although these networks achieve good performance, the good results came from the well-trained ANNs, which does not reflect the characteristics of SNNs well.

For the BP training methods, DSN (O'Connor and Welling, 2016) proposed that SNN is equivalent to a deep network of ReLU units, and could be directly trained with BP. Event-SNN (Neftci et al., 2017) demonstrated an event-driven random BP rule for learning deep representations. SCSNN (Wu et al., 2019) used spike count as a surrogate for gradient backpropagation. BPSNN (Lee et al., 2016) treated the membrane potentials of spiking neurons as differentiable signals, which enabled the backpropagation. HM2-BP (Jin et al., 2018) proposed a hybrid macro/micro level backpropagation algorithm for training multi-layer SNNs. Temporal SNN (Mostafa, 2017) trained the SNN with temporal coding. STBP (Wu et al., 2018) trained the SNNs with BP both in spatial and temporal domains. The excellent performance of these methods came from BP, which turns out to not existed in the brain.

For STDP and other biologically plausible methods, Unsupervised-SNN (Diehl and Cook, 2015) trained an SNN with STDP, lateral inhibition, and an adaptive spiking threshold with a poor little performance 95% on the MNIST dataset. LIF-BA (Samadi et al., 2017) approximated dynamic input-output relations with piecewise-smooth functions based on fixed feedback weights. STCA (Gu et al., 2019) trained SNNs with credit assignments both in spatial and temporal domains. Both of them update the weights layer by layer. VPSNN (Zhang et al., 2018a) and Balance-SNN (Zhang et al., 2018b) trained the SNNs with Equilibrium Propagation, Balance-SNN is an improved version of VPSNN, which introduced much more learning rules to get the training balance of SNNs. However, as they trained with Equilibrium Propagation, the problems in Equilibrium Propagation also exist in both of them.

To sum up, a model to propagate the global plasticity information with a random feedback connection directly to each layer combined with the local plasticity learning rule to train SNNs has so far been rarely studied.

3. METHODS

The pipeline of our model is shown in **Figure 1B**. First, we will introduce the spiking neuron model used in our framework. Second, the global and local plasticity learning process will be introduced. Third, the whole framework will be introduced to understand our model better.

3.1. The Basic LIF Neuron Model

The spiking neuron model we use for temporal information processing is the Leaky integrate-and-fire (LIF) model, which is widely used in most SNN frameworks. As can be seen in **Figure 2**, for the LIF model, the neuron will accumulate the potential from the input, once its potential reaches the threshold, the neuron will be fired with a spike.

Generally, the membrane potential V can be calculated with Equation (1)

$$I(t) - \frac{V(t)}{R_m} = C_m \frac{dV(t)}{dt} \tag{1}$$

R_m is the membrane resistance and C_m denotes the membrane capacitance. $I(t)$ denotes the total input current from pre-synaptic neurons. For simplicity, we denote $V(t)$ with V, $I(t)$ with I, g_L and V_L denote leaky conductance and leaky potential. In a network with a more realistic synapse model, the input current I is generated as a change in conductance, which is caused by spikes of presynaptic neurons. The excitatory conductance g_E will be non-linearly increased by the number of the input spikes δ_j (Gerstner et al., 2014). V_E is the reversal potential from neuron

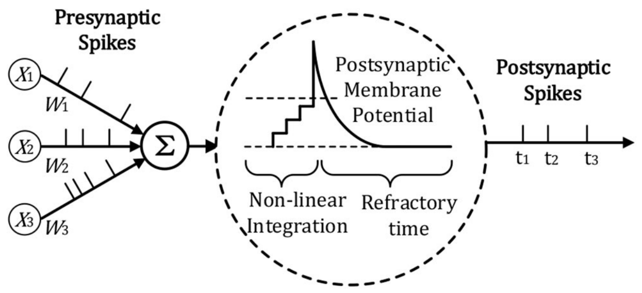


FIGURE 2 | Illustration of LIF Neuron Model adopted from Lee et al. (2019) and Zhang et al. (2018a).

i to neuron j. When the membrane reaches the threshold, the neuron will produce a spike, and the membrane will be reset to V_{reset}. $\tau_m = \frac{C_m}{g_L}$, τ_E is the conductance decay of excitatory neurons, $w_{j,i}$ is the synapse weight from neuron j to neuron i.

$$\begin{cases} \tau_m \frac{dV_i}{dt} = -(V_i - V_L) - \frac{gE}{gL}(V_i - V_E) \\ \tau_E \frac{dg_E}{dt} = -gE + \sum_j^N w_{j,i}\delta_j \end{cases} \tag{2}$$

3.2. The Global Plasticity Learning Process of Our Model

The global plasticity learning process is applied to a multi-layer feedforward neural network to illustrate better our learning algorithm, in which neurons in the previous layer are fully connected to the subsequent layer. In the adjacent layers, information from pre-synaptic neurons will be transferred to the post-synaptic neurons. For a deep spiking neural network, if only the spike is used, it will take a long time for the information transfer to the subsequent deeper layers, which will make the network hard to converge. To solve the problems, Diehl and Cook (2015) has used the spike trace to adjust the network weights, Zhang et al. (2018a) and Lee et al. (2016)'s work use voltage-based weight adjustments. Inspired by the residual neural network (He et al., 2016), which transfers the information as $x + f(x)$, here we think that in addition to the spikes output by the LIF neuron can be used to regulate the weight, the input to the LIF neuron also contains a wealth of information. The final output of the neuron is denoted as $S_j(t+1)$. To convert Equation (2) into discrete form, the whole process is shown in Equation (3):

$$\begin{cases} V_i(t+1) = V_i(t) - \frac{dt}{\tau_m}[V_i(t) - V_L + \frac{gE}{gL}(V_i(t) - V_E)] \\ g_E(t+1) = g_E(t) + \frac{dt}{\tau_E}(-g_E(t) + \sum_j^N w_{j,i}S_j(t+1)) \\ \delta_i(t+1) = 1 \qquad V_i = V_{reset} \qquad if \quad V_i > V_{th} \\ S_i(t+1) = \sum_j^N w_{j,i}S_j(t+1) + \tau\delta_i(t+1) \end{cases} \tag{3}$$

τ is the constant to control the magnitude of the output. To accelerate the calculation, we only calculate the loss at the end of the simulation to update the target and weight. We denote the target with S^T, S_{out} denotes the output of the last layer, M is the number of the samples. For the output layer, the loss function we choose here is the L2 norm so that the prediction error can be written as Equation (4):

$$\begin{cases} loss = \sum_{i=1}^M ||S_{out} - S^T||^2 \\ e = 2 * \sum_{i=1}^M |S_{out} - S^T| \end{cases} \tag{4}$$

Supposing a network with L layers. The output of the l_{th} layer is denoted with S_l. For supervised learning, the target of the penultimate layer $\hat{S}_{L-1}$ can be directly calculated, as shown in Equation (5), W_l denotes the forward weight between the l_{th} layer and the $(l+1)_{th}$. η_t represents the learning rate of the target.

$$\hat{S}_{L-1} = S_{L-1} - \eta_t \Delta S = S_{L-1} - \eta_t W_{L-1}^T e \tag{5}$$

For the target of the other hidden layers, the target can not be directly calculated as Equation (5). By introducing the feedback connections, the prediction error can be easily transmitted to the hidden layers, and we denote the feedback layer as G_l. Moreover, the target of the hidden layer can be written as Equation (6):

$$\begin{cases} \hat{S}_l = S_l - G_l(e) \\ G_l(e) = B_l * e + b_l \end{cases} \tag{6}$$

B_l denotes the random feedback weight of the l_{th} layer, and b_l represents the random feedback bias. With the operation of all layers, we can directly get the target of each layer.

3.3. The Local Learning Process of Our Model

STDP can be seen as the leading learning rule in the brain, and it can simulate the expected change of synaptic weights depending on states between pre-synaptic and post-synaptic (Bi and Poo, 1998), which can be regarded as a local learning rule. As introduced in (Xie and Seung, 2000; Hinton, 2007), STDP is associated with the change of postsynaptic activity. Here we use the difference between the feedforward state and feedback state to denote the change, as shown in Equation (7).

$$\Delta W \propto S_j S_i' = S_j(S_i - \hat{S}_i) \tag{7}$$

where S_j and S_i indicate the pre-synaptic and post-synaptic output in the forward learning process. $\hat{S}_i$ denotes the target of the i_{th} layer calculated in Equation (6).

3.4. The Whole Learning Framework

For a multi-layer feedforward SNN, global plasticity information should be introduced so that STDP can train the whole network to obtain the desired result. Firstly, the feedforward process is used to obtain the feedforward state of the network, and then the feedback is used to obtain the targets of different hidden layers. Then, the change of weights in different neighborhood layers are calculated by local STDP plasticity rule in Equation (7). Finally, the weight of the forward propagation is updated with Equation (8):

$$W = W - \eta_w \Delta W \tag{8}$$

η_w denotes the learning rate of weight.

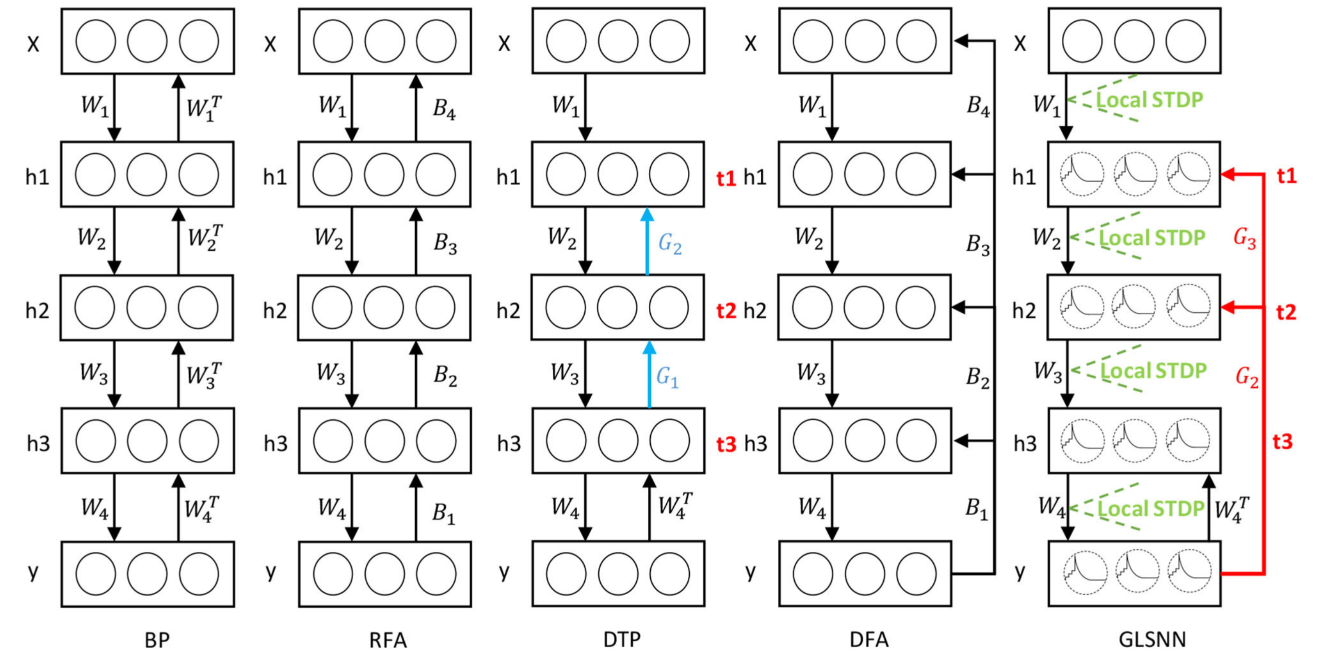


FIGURE 3 | The learning process of our GLSNN compared with BP, RFA, DTP, and DFA. B in RFA and DFA means the random matrix to transfer the error directly. Blue connection G_l in DTP means the feedback layer needs to update. Red connection G_L in GLSNN means the feedback layer without updates.

Inspired by FAs (Lillicrap et al., 2016; Nøkland, 2016), random weights can be used to transmit the error in the network. In this paper, we use the random feedback layer to get the target of the hidden layers. As shown in **Figure 3**, in our model, feedback connections are directly connected from the output layer to the hidden layers, which means that the neural network can update the parameters of all hidden layers simultaneously, and the random feedback connections do not introduce extra computations. The details are shown in Algorithm 1.

Algorithm 1 The whole learning process of our GLSNN.

```
Require: Initialize a multi-layer neural network with $L$ layers
         Feedforward process in Equation (3) $F_i$
         Feedback process in Equation (6) $G_i$
         t = 0, simulation interval $dt$, and simulation time T, max
iteration $EPO$
for epoch = 1 to $EPO$ do
    while $t \leq T$ do
        for i = 1 to L-1 do
            $S_i = F_i(S_{i-1})$.
            t = t + dt
        end for
    end while
    Get the prediction error $e$ with Equation (4).
    Get the target of the penultimate layer $\hat{S}_{L-1}$ with
Equation (5).
    for i = 1 to L-2 do
        $\hat{S}_i = S_i - G_i(e)$
    end for
    for i = 1 to L-1 do
        Update synapse weights Equations (7) and (8)
    end for
end for
```

4. EXPERIMENTS

In this section, we experimentally evaluate the performance of our model on two benchmark datasets, basic MNIST (LeCun, 1998) and Fashion MNIST (Xiao et al., 2017). The experiments are performed with PyTorch on TITAN RTX. To fully reflect the performance of our algorithm, the fully connected network is considered to carry out the experiment without batch normalization or weight regularization. The update method of the weight is the Stochastic Gradient Descent (SGD) method. In addition, we compare our GLSNN with other state-of-the-art biological plausible methods. The initiation method of the weight is the same as DTP (Lee et al., 2015). Also, the ablation studies are performed to study the effect of the feedback layers. For the parameters of the network, the learning rate for the target $\eta_t = 0.5$, the learning rate for the weight $\eta_w = 0.015$. The batchsize is 10. For the hyper-parameter of the LIF neuron as described in section 3, we set $V_E = 0.2$, $V_I = 0$, $V_L = 0$, $V_{th} = 0.0009$, $V_{reset} = 0$, $\tau_m = 0.5$, $\tau_E = 0.2$, $\tau = 0.01$, $g_{leak} = 20$, the simulated time interval $dt = 0.01$, and the total simulation time $T = 0.1$.

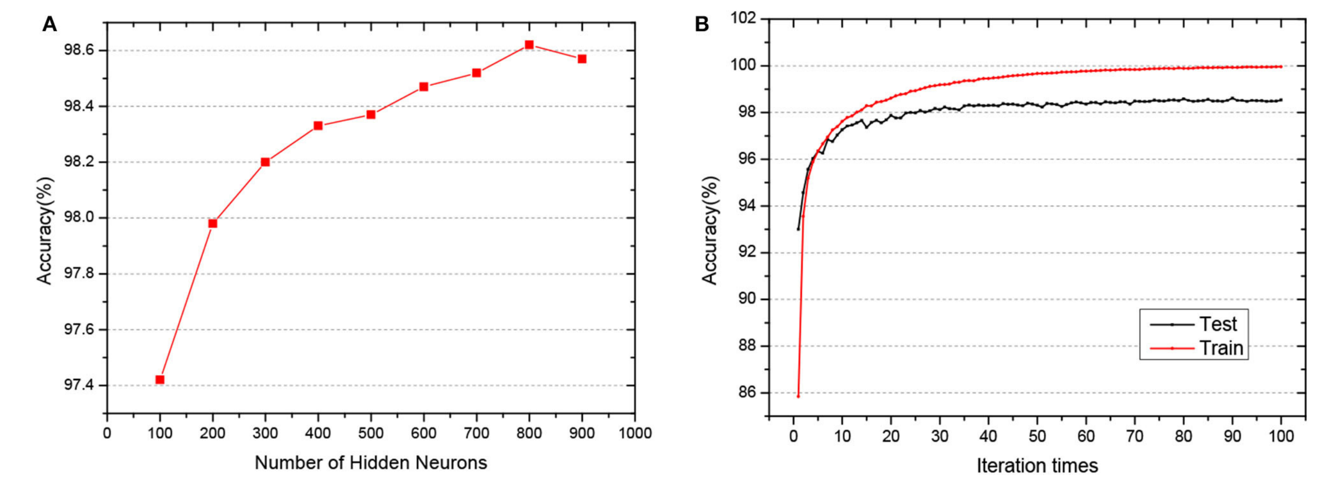


FIGURE 4 | (A) The test accuracy of GLSNN of different hidden neurons of 3 hidden layers. **(B)** The train and test accuracy when the hidden layer is set with 800*3.

TABLE 1 | Comparison of classification accuracies of GLSNN with other SNN frameworks on the MNIST dataset.

Model	Structure	Neural coding	Learning rule	Acc
SDBN (O'Connor et al., 2013)	FC	Spike	ANN to SNN	94.09
Unsupervised-SNN (Diehl and Cook, 2015)	FC	Spike	STDP	95
LIF-BA (Samadi et al., 2017)	FC	Spike	Broadcast Alignment	97.05
SN (O'Connor and Welling, 2016)	FC	Rate	BP	97.93
Event-SNN (Neftci et al., 2017)	FC	Rate	BP	97.98
Temporal SNN (Mostafa, 2017)	FC	Spike	BP with Temporal Coding	98
SDN (Hunsberger and Eliasmith, 2015)	FC	Spike	ANN to SNN	98.37
VPSNN (Zhang et al., 2018a)	FC	Spike	Equi-prop + STDP	98.52
STCA (Gu et al., 2019)	FC	Spike	Spatial + Tempral Credit Assignment	98.6
GLSNN (This study)	**FC**	**Spike**	**Global Feedback + STDP**	**98.62**
Balance-SNN (Zhang et al., 2018b)	FC	Spike	Equi-Prop + Multiple Balance Rules	98.64
SCSNN (Wu et al., 2019)	FC	Rate	BP	98.66
BPSNN (Lee et al., 2016)	FC	Rate	BP	98.71
HM2-BP (Jin et al., 2018)	FC	Rate	Macro/Micro level BP	98.88
STBP (Wu et al., 2018)	FC	Rate	Spatial + Tempral BP	98.89

TABLE 2 | The average training time (seconds) per epoch.

500 * 1	500 * 2	500 * 3	500 * 4	500 * 5	500 * 6
80.86	124.99	133.33	178.25	200.08	256.98

4.1. MNIST

MNIST is the most widely used dataset to measure the performance of the algorithm in machine learning. It consists of 60,000 training samples and 10,000 test samples, used to describe the hand-written digits from 0 to 9. The sample size is 28*28. The number of epochs is set with 100. We wonder how our model fares in this benchmark as the model goes deeper in that target is directed computed from the output layer. To that end, we have trained a network of 3 hidden layers of different hidden neurons to evaluate the performance of the network.

As shown in **Figure 4**, when the network structure is set with [784-800-800-800-10], the test accuracy is the highest at 98.62%. To demonstrate the superiority of our GLSNN, we compare our methods with several different SNN frameworks, as can be seen in **Table 1**, our GLSNN has surpassed all other SNN frameworks trained without BP, such as Unsupervised-SNN (Diehl and Cook, 2015), VPSNN (Zhang et al., 2018a), and so on. Moreover, for the BP trained SNNs, we have exceeded most of them. For the Balance-SNN (Zhang et al., 2018b), in addition to the STDP learning rule, several other rules were introduced, such as LTP, LTD, STF, STD, however only 0.2% accuracy improved compared to our GLSNN. For SCSNN (Wu et al., 2019), BPSNN (Lee et al., 2016), HM2-BP (Jin et al.,

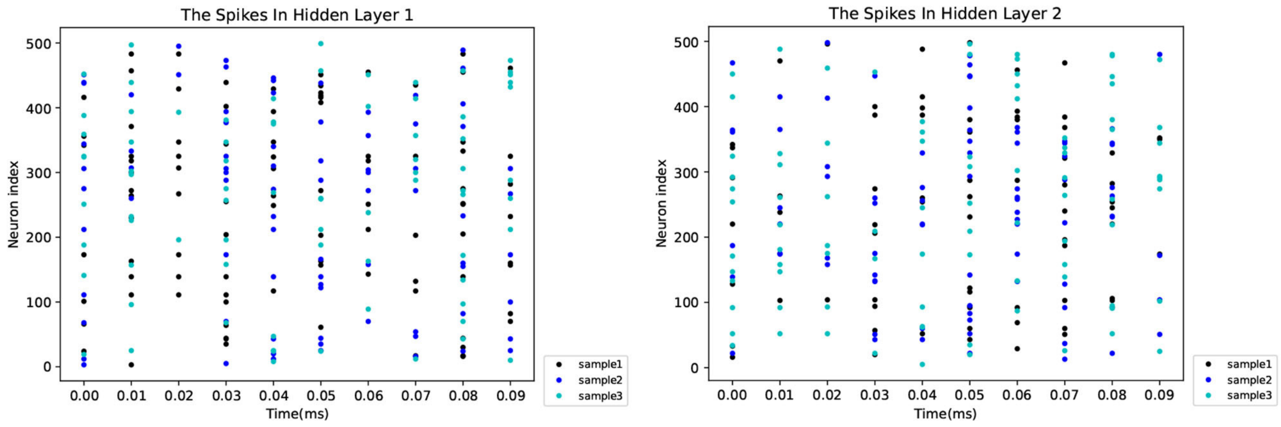


FIGURE 5 | The spikes in the hidden layer of the three randomly chosen samples.

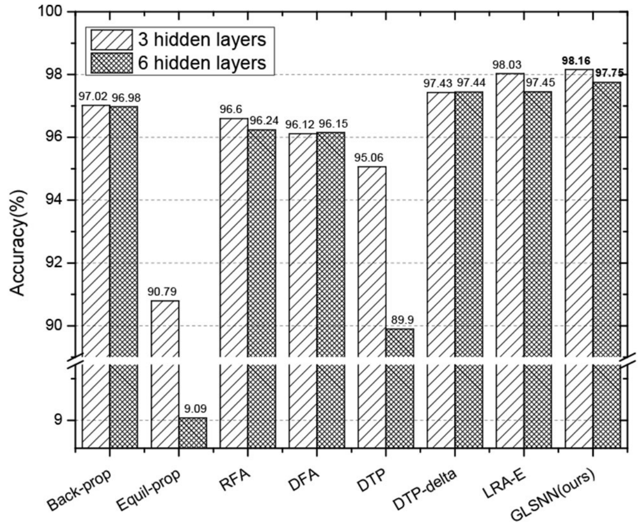


FIGURE 6 | The test accuracy on MNIST dataset of GLSNN compared with ANNs trained with BP, Equil-Prop, RFA, DFA, DTP, DTP-delta, and LRA-E with different hidden layers.

2018), and STBP (Wu et al., 2018), the different levels of backpropagation was connected to contribute to their superior performance, however, which is non-existent in the human brains. To the best of our knowledge, our result could be a new record for the SNNs trained with STDP. The spike transfer process is shown in **Figure 5**, as the network structure is set with [784-500-500-10].

Also, to prove that our algorithm still performs well when the network is going deeper, we test the results with different hidden layers, whose hidden neurons are set with 256 for consistency with the paper (Ororbia and Mali, 2019). As can be seen in **Figure 6**, for Equil-prop methods, the accuracy quickly drops down when the network is deeper. Also, the accuracy of the DTP method begins to struggle from 95.06 to 89.9%, which shows the instability of them. Compared with other stable methods, our GLSNN outperforms better than them both for the five hidden layers and the eight hidden layers, which indicates the stability and superiority of our algorithm.

Also, to measure the computation speed of our model, we test the average runtime per epoch with different hidden layers as shown in **Table 2**.

To demonstrate the underlying mechanism of our GLSNN model, the t-SNE method (Maaten and Hinton, 2008) was used to visualize the model's clustering ability of different layers. The network structure is set with [784-500-500-10], as shown in **Figure 7**, for the original input, samples of different categories are very close to each other, and some clusters contain samples from other categories. After the training of SNN, the separability of the output information of the hidden layer shows more vital clustering ability than the input layer as the interval between the class clusters is coming larger. For the output layer, different categories are distinguished, which has shown that the learning process of our GLSNN has helped the network to perform better clustering and classification performances.

4.2. Fashion MNIST

Fashion-MNIST is a more complex version compared to MNIST, consisting of gray-scale images of clothing items. Since the dataset is more complicated compared with MNSIT, the training epoch is set with 200, and we tried networks of different hidden layers, as shown in **Figure 8**. When the network structure is set with five hidden layers of 200 hidden neurons each layer, the network achieves the best performance with 89.05% accuracy on the test dataset. Also, we compare our GLSNN with other biologically plausible methods shown in **Table 3**. We have chosen the best results of each method as recorded in (Ororbia and Mali, 2019). Our GLSNN exceeds all of them.

4.3. Ablation Studies

To study the effect of the feedback layers of the network, we create four networks with 7, 8, 10, and 12 layers separately. All of the hidden neurons are set with 200. First, we remove all the feedback connections of the network, which means only the weight of the last two-layers could be updated. Then we incrementally add

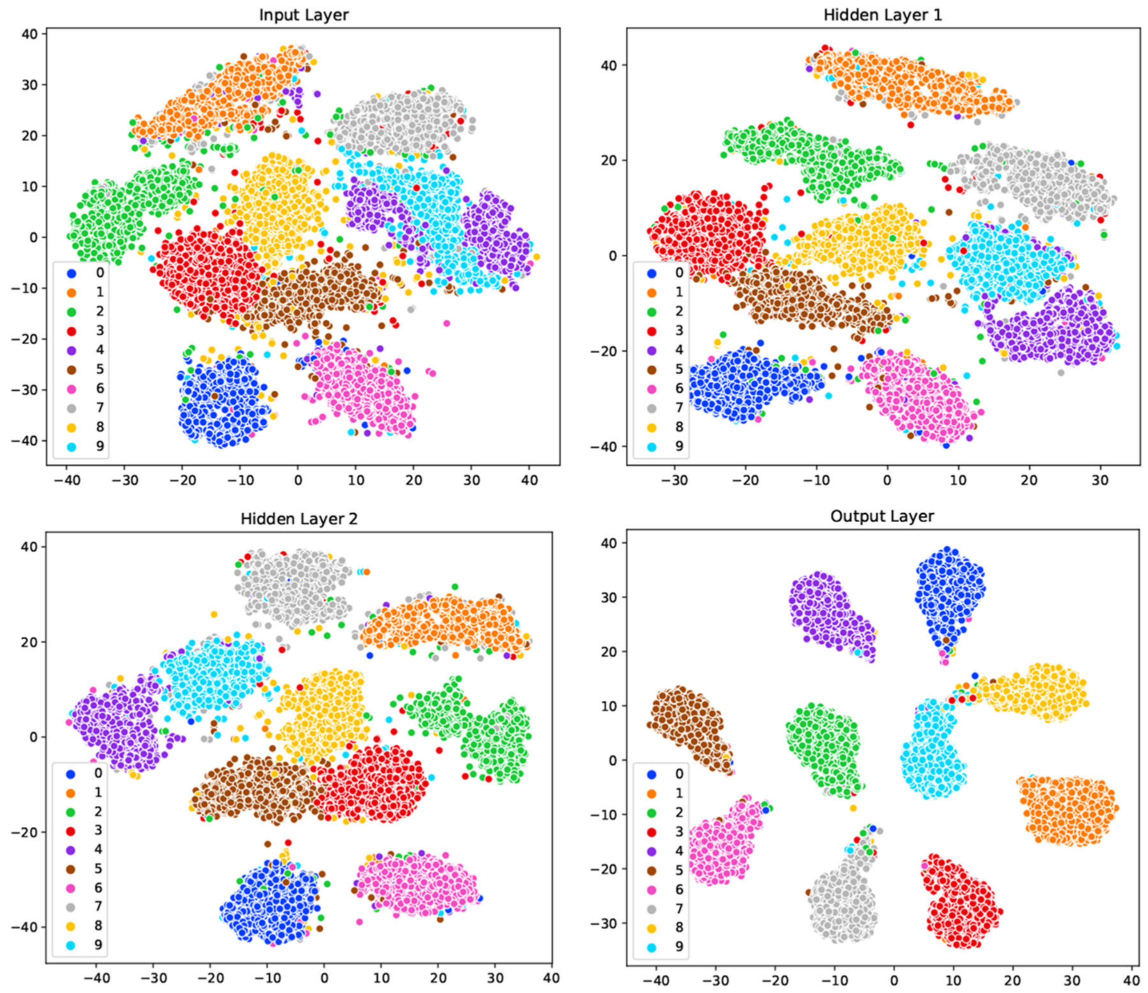


FIGURE 7 | The visualization on the input layer, hidden layer 1, hidden layer 2, and output layer in GLSNN with t-SNE.

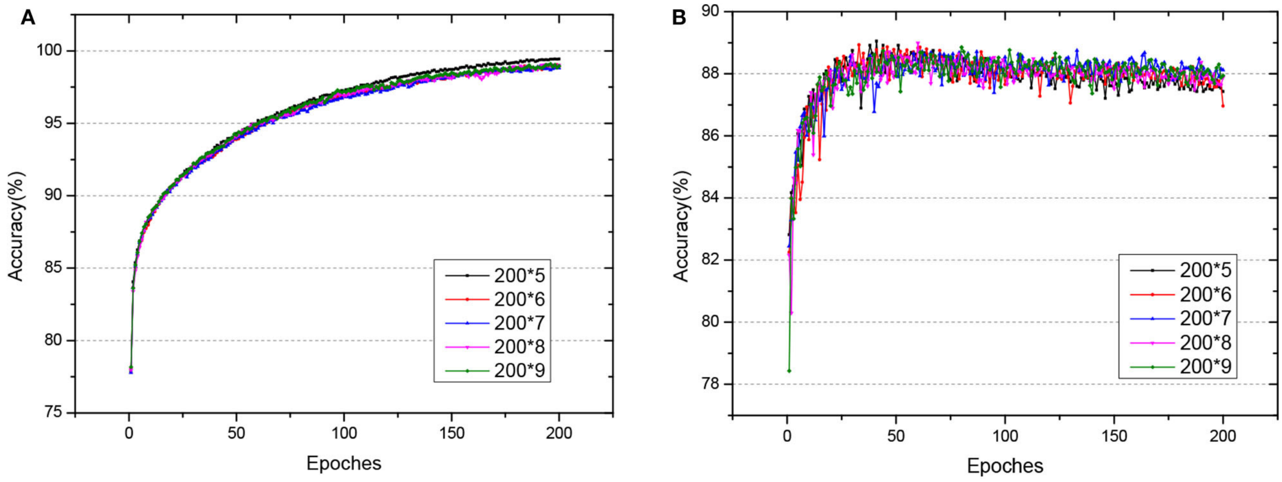


FIGURE 8 | (A,B) The train and test accuracy of GLSNN of different hidden layers of Fashion MNIST, the 200*n, means n hidden layers with 200 neurons each hidden layer.

the feedback layers in the network to see the performance of the network.

As shown in **Figure 9**, with the increase of the number of feedback layers, the performance of the network gradually improves. When all the feedback layers are added, the SNN reaches the highest accuracy. The performance of the network did not improve linearly with the increase of the feedback layers. The variation in accuracy can be roughly divided into three steps:

- In step 1, the linear increment of accuracy with weights tuning in only top layers.
- In step 2, the non-increment or stabilization of accuracy with weight tuning in both top and mid-layers.
- In step 3, the prominent increment toward the best accuracy with only adding into the weight tuning in the bottom layer.

The deeper layers play a role in decision-making, while the former layers play a role in feature extraction. That is to say, the feedback connections play a significant role in the perceptual inference, which is consistent with neurophysiology (Harris and Shepherd, 2015).

TABLE 3 | The test accuracy on the Fashion MNSIT dataset of GLSNN compared with VPSNN and other ANNs trained with BackProp, Equi-Prop, RFA, DFA, DTP, DTP-delta, and LRA-E.

Model	Structure	Type	Performance
VPSNN (Zhang et al., 2018a)	FC	SNN	82.69
Equiprop (Scellier and Bengio, 2017)	FC	ANN	85.99
DTP (Lee et al., 2015)	FC	ANN	86.4
DTP_delta (Ororbia and Mali, 2019)	FC	ANN	87.01
LRA-E (Ororbia and Mali, 2019)	FC	ANN	87.69
RFA (Lillicrap et al., 2016)	FC	ANN	88.01
DFA (Nøkland, 2016)	FC	ANN	88.41
Backprop (Rumelhart et al., 1986)	FC	ANN	88.45
GLSNN (This study)	**FC**	**SNN**	**89.05**

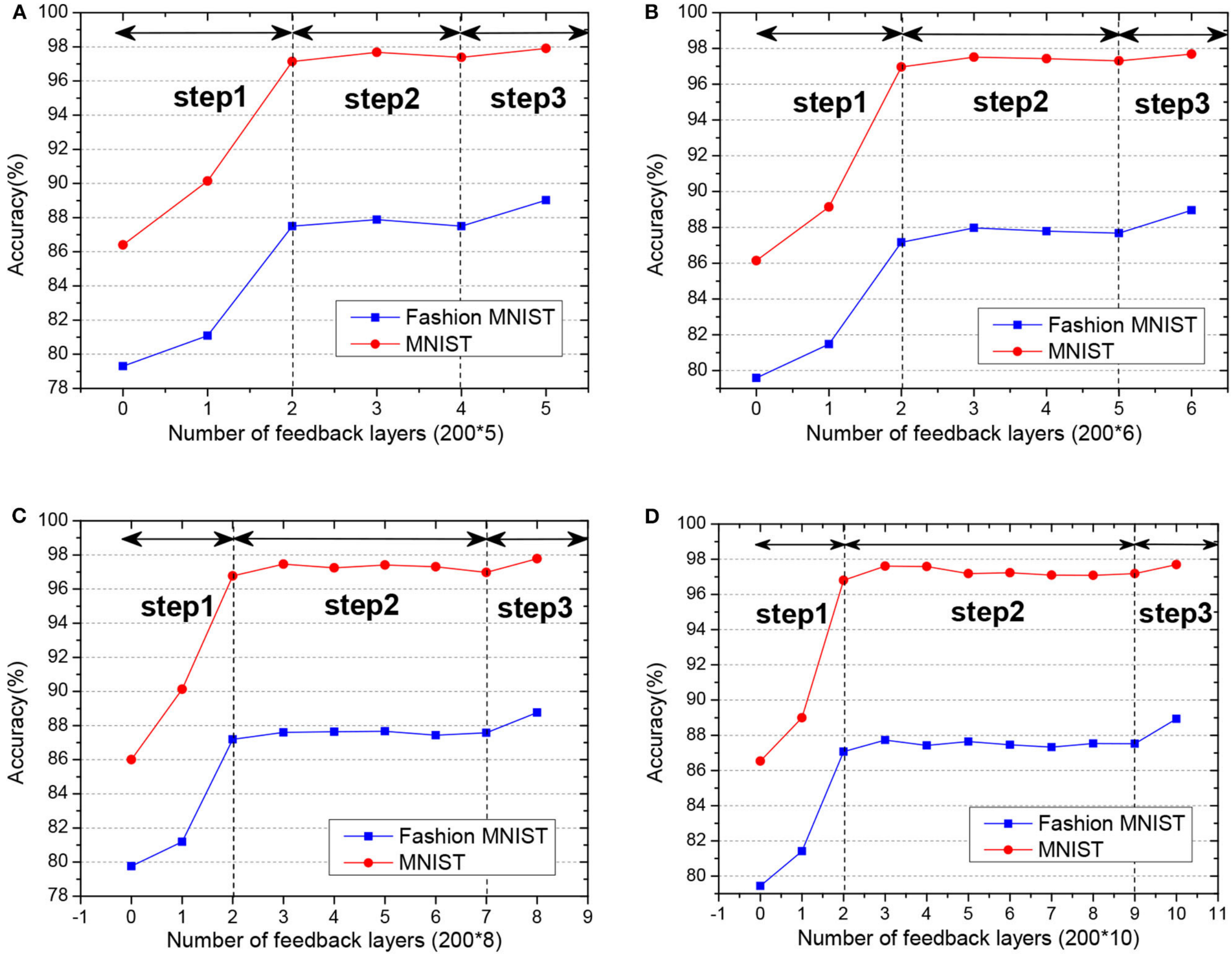


FIGURE 9 | (A–D) The test accuracy of GLSNN with different feedback layers on MNIST and Fashion MNIST, and the variation in accuracy can be roughly divided into three steps.

4.4. Comparison With Other Traditional SNNs Trained With STDP

For the SNNs trained with STDP, the problem is how to introduce global information. The success of the BP algorithm in deep neural networks training is mainly due to the chain rules, which introduce the global error. Traditional SNNs trained with STDP often sidestep this problem, that is they avoid multi-layer training. For Diehl's unsupervised SNN (Diehl and Cook, 2015), only the weight between the input and excitatory neurons is trained with STDP. The extension (Hao et al., 2020) modified the last clustering layer to a supervised classification layer. Masquelier (Masquelier and Thorpe, 2007) introduced a multi-layer SNN combined with convolutional/pooling layer, feature discovery layer and a classification layer. However, the first convolutional layer is set with the Gabor filters, and only the feature discovery layer is trained with STDP. To solve this, Tavanaei (Tavanaei and Maida, 2017) introduced a sparse coding model to replace the handcrafted features in Masquelier and Thorpe (2007). However, the training is layer-wise, the feature discovery layer can only be trained after the first convolutional layer is completed training. Recently, Zhang's work (Zhang et al., 2018a) introduced the equilibrium propagation, the forward and feedback process in SNNs are implicitly defined in the negative and positive phase in equilibrium propagation, which solved the multi-layer training in SNNs to a certain extent. However, due to the implicit definition, when the network went deeper, it becomes hard to converge to a stable situation. Our GLSNN explicitly introduced the global feedback connections, which provides a feasible solution to the training of the multi-layer SNN.

5. CONCLUSION AND FUTURE WORK

In this paper, we propose an SNN training method, which takes full advantage of the global and local plasticity information. We mimic the global feedback connections and the local STDP learning rules in the brain, providing a powerful way to train a multi-layer SNN. The global random feedback connections help to propagate the target from the output layer to the hidden layers. The local STDP learning rule is utilized to optimize the local synaptic strength of the network with the obtained target. Our GLSNN offers an alternative way to solve the weight transpose problem in BP, as well as the feedback layers are directly connected to the hidden layers, leading the weight of each layer can be directly updated without the error transmitted layer by layer. Experiments indicate that our GLSNN model has performed favorably against several state-of-the-art SNNs on the standard benchmark MNIST and Fashion MNIST dataset.

In terms of future work, the authors intend to study more biologically inspired learning rules in this work, as we only use the STDP local learning rule. The dynamic combination of different learning rules and different types of spiking neurons may further enhance the learning performance of the network. Also, we only verify the performance on the fully connected network structures, in the following work, we would consider more complex network structures such as convolutional neural network and recurrent neural network to accommodate more complex visual perception tasks, such as video object detection and visual tracking.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: http://yann.lecun.com/exdb/mnist/; https://github.com/zalandoresearch/fashion-mnist.

AUTHOR CONTRIBUTIONS

DZ and YZ designed the study, performed the experiments and the analyses. MS and FZ participated in the biological background discussion and refined the paper. DZ, YZ, and TZ were involved in algorithm discussion, result analysis, and wrote the paper. All authors contributed to the article and approved the submitted version.

FUNDING

This work is supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant No. XDB32070100), the new generation of artificial intelligence major project of the Ministry of Science and Technology of the People's Republic of China (Grant No. 2020AAA0104305), the Beijing Municipal Commission of Science and Technology (Grant No. Z181100001518006), the CETC Joint Fund (Grant No. 6141B08010103), and the Beijing Academy of Artificial Intelligence (BAAI).

REFERENCES

Amit, D. J., Brunel, N., and Tsodyks, M. (1994). Correlations of cortical Hebbian reverberations: theory versus experiment. *J. Neurosci.* 14, 6435–6445. doi: 10.1523/JNEUROSCI.14-11-06435.1994

Bi, G.-Q., and Poo, M.-M. (1998). Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. *J. Neurosci.* 18, 10464–10472. doi: 10.1523/JNEUROSCI.18-24-10464.1998

Breakspear, M. (2017). Dynamic models of large-scale brain activity. *Nat. Neurosci.* 20, 340–352. doi: 10.1038/nn.4497

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* 40, 834–848. doi: 10.1109/TPAMI.2017.2699184

Danelljan, M., Hager, G., Shahbaz Khan, F., and Felsberg, M. (2015). "Convolutional features for correlation filter based visual tracking," in *Proceedings of the IEEE International Conference on Computer Vision Workshops* (Boston), 58–66. doi: 10.1109/ICCVW.2015.84

Diehl, P. U., and Cook, M. (2015). Unsupervised learning of digit recognition using spike-timing-dependent plasticity. *Front. Comput. Neurosci.* 9:99. doi: 10.3389/fncom.2015.00099

Diehl, P. U., Neil, D., Binas, J., Cook, M., Liu, S.-C., and Pfeiffer, M. (2015). "Fast-classifying, high-accuracy spiking deep networks through weight and threshold balancing," in *2015 International Joint Conference on Neural Networks (IJCNN)* (Killarney: IEEE), 1–8. doi: 10.1109/IJCNN.2015.7280696

Durstewitz, D., Seamans, J. K., and Sejnowski, T. J. (2000). Neurocomputational models of working memory. *Nat. Neurosci*. 3, 1184–1191. doi: 10.1038/81460

Felleman, D. J., and Van, D. E. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cereb. Cortex* 1, 1–47. doi: 10.1093/cercor/1.1.1

Gerstner, W., Kistler, W. M., Naud, R., and Paninski, L. (2014). *Neuronal Dynamics: From Single Neurons to Networks and Models of Cognition*. Cambridge, MA: Cambridge University Press. doi: 10.1017/CBO9781107447615

Gu, P., Xiao, R., Pan, G., and Tang, H. (2019). "STCA: spatio-temporal credit assignment with delayed feedback in deep spiking neural networks," in *Proceedings of the 28th International Joint Conference on Artificial Intelligence* (Macao: AAAI Press), 1366–1372. doi: 10.24963/ijcai.2019/189

Hao, Y., Huang, X., Dong, M., and Xu, B. (2020). A biologically plausible supervised learning method for spiking neural networks using the symmetric STDP rule. *Neural Netw*. 121, 387–395. doi: 10.1016/j.neunet.2019.09.007

Harris, K. D., and Shepherd, G. M. (2015). The neocortical circuit: themes and variations. *Nat. Neurosci*. 18:170. doi: 10.1038/nn.3917

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas), 770–778. doi: 10.1109/CVPR.2016.90

Héricé, C., Khalil, R., Moftah, M., Boraud, T., Guthrie, M., and Garenne, A. (2016). Decision making under uncertainty in a spiking neural network model of the basal ganglia. *J. Integr. Neurosci*. 15, 515–538. doi: 10.1142/S021963521650028X

Hinton, G. (2007). "How to do backpropagation in a brain," in *Invited Talk at the NIPS'2007 Deep Learning Workshop* (Vancouver).

Hunsberger, E., and Eliasmith, C. (2015). Spiking deep networks with lif neurons. *arXiv preprint arXiv:1510.08829*.

Izhikevich, E. M., and Edelman, G. M. (2008). Large-scale model of mammalian thalamocortical systems. *Proc. Natl. Acad. Sci. U.S.A*. 105, 3593–3598. doi: 10.1073/pnas.0712231105

Jin, Y., Zhang, W., and Li, P. (2018). "Hybrid macro/micro level backpropagation for training deep spiking neural networks," in *Advances in Neural Information Processing Systems*, eds S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, and N. Cesa-Bianchi (Montreal: Curran Associates Inc.), 7005–7015.

Khalil, R., Karim, A. A., Khedr, E., Moftah, M., and Moustafa, A. A. (2018). Dynamic communications between GABAA switch, local connectivity, and synapses during cortical development: a computational study. *Front. Cell. Neurosci*. 12:468. doi: 10.3389/fncel.2018.00468

Khalil, R., Moftah, M. Z., Landry, M., and Moustafa, A. A. (2017a). Models of dynamical synapses and cortical development. *Comput. Models Brain Behav*. 321. doi: 10.1002/9781119159193.ch23

Khalil, R., Moftah, M. Z., and Moustafa, A. A. (2017b). The effects of dynamical synapses on firing rate activity: a spiking neural network model. *Eur. J. Neurosci*. 46, 2445–2470. doi: 10.1111/ejn.13712

Kim, S., Park, S., Na, B., and Yoon, S. (2019). Spiking-yolo: Spiking neural network for real-time object detection. *arXiv preprint arXiv:1903.06530*.

LeCun, Y. (1998). *The MNIST Database of Handwritten Digits*. Available online at: http://yann.lecun.com/exdb/mnist/

Lee, C., Sarwar, S. S., and Roy, K. (2019). Enabling spike-based backpropagation in state-of-the-art deep neural network architectures. *arXiv preprint arXiv:1903.06379*. doi: 10.3389/fnins.2020.00119

Lee, D.-H., Zhang, S., Fischer, A., and Bengio, Y. (2015). "Difference target propagation," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (Coimbra: Springer), 498–515. doi: 10.1007/978-3-319-23528-8_31

Lee, J. H., Delbruck, T., and Pfeiffer, M. (2016). Training deep spiking neural networks using backpropagation. *Front. Neurosci*. 10:508. doi: 10.3389/fnins.2016.00508

Levina, A., Herrmann, J. M., and Geisel, T. (2007). Dynamical synapses causing self-organized criticality in neural networks. *Nat. Phys*. 3, 857–860. doi: 10.1038/nphys758

Lillicrap, T. P., Cownden, D., Tweed, D. B., and Akerman, C. J. (2016). Random synaptic feedback weights support error backpropagation for deep learning. *Nat. Commun*. 7:13276. doi: 10.1038/ncomms13276

Luo, Y., Xu, M., Yuan, C., Cao, X., Xu, Y., Wang, T., et al. (2020). SiamSNN: spike-based siamese network for energy-efficient and real-time object tracking. *arXiv preprint arXiv:2003.07584*.

Maaten, L. V. d., and Hinton, G. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res*. 9, 2579–2605.

Masquelier, T., and Thorpe, S. J. (2007). Unsupervised learning of visual features through spike timing dependent plasticity. *PLoS Comput. Biol*. 3:e31. doi: 10.1371/journal.pcbi.0030031

Mostafa, H. (2017). Supervised learning based on temporal coding in spiking neural networks. *IEEE Trans. Neural Netw. Learn. Syst*. 29, 3227–3235. doi: 10.1109/TNNLS.2017.2726060

Movellan, J. R. (1991). "Contrastive Hebbian learning in the continuous hopfield model," in *Connectionist Models*, eds D. S. Touretzky, J. L. Elman, T. J. Sejnowski, and G. E. Hinton (San Mateo, CA: Elsevier), 10–17. doi: 10.1016/B978-1-4832-1448-1.50007-X

Neftci, E. O., Augustine, C., Paul, S., and Detorakis, G. (2017). Event-driven random back-propagation: enabling neuromorphic deep learning machines. *Front. Neurosci*. 11:324. doi: 10.3389/fnins.2017.00324

Nøkland, A. (2016). "Direct feedback alignment provides learning in deep neural networks," in *Advances in Neural Information Processing Systems*, eds D. D. Lee, U. von Luxburg, R. Garnett, M. Sugiyama, and I. Guyon (Barcelona: Curran Associates Inc.), 1037–1045.

O'Connor, P., Neil, D., Liu, S.-C., Delbruck, T., and Pfeiffer, M. (2013). Real-time classification and sensor fusion with a spiking deep belief network. *Front. Neurosci*. 7:178. doi: 10.3389/fnins.2013.00178

O'Connor, P., and Welling, M. (2016). Deep spiking networks. *arXiv preprint arXiv:1602.08323*.

Ororbia, A. G., and Mali, A. (2019). "Biologically motivated algorithms for propagating local target representations," in *Proceedings of the AAAI Conference on Artificial Intelligence* (Honolulu), 4651–4658. doi: 10.1609/aaai.v33i01.33014651

Potjans, T. C., and Diesmann, M. (2014). The cell-type specific cortical microcircuit: relating structure and activity in a full-scale spiking network model. *Cereb. Cortex* 24, 785–806. doi: 10.1093/cercor/bhs358

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature* 323, 533–536. doi: 10.1038/323533a0

Samadi, A., Lillicrap, T. P., and Tweed, D. B. (2017). Deep learning with dynamic spiking neurons and fixed feedback weights. *Neural Comput*. 29, 578–602. doi: 10.1162/NECO_a_00929

Scellier, B., and Bengio, Y. (2017). Equilibrium propagation: bridging the gap between energy-based models and backpropagation. *Front. Comput. Neurosci*. 11:24. doi: 10.3389/fncom.2017.00024

Sporns, O., and Zwi, J. D. (2004). The small world of the cerebral cortex. *Neuroinformatics* 2, 145–162. doi: 10.1385/NI:2:2:145

Tavanaei, A., and Maida, A. S. (2017). "Multi-layer unsupervised learning in a spiking convolutional neural network," in *2017 International Joint Conference on Neural Networks (IJCNN)* (Anchorage: IEEE), 2023–2030. doi: 10.1109/IJCNN.2017.7966099

Wu, J., Chua, Y., Zhang, M., Yang, Q., Li, G., and Li, H. (2019). Deep spiking neural network with spike count based learning rule. *arXiv preprint arXiv:1902.05705*. doi: 10.1109/IJCNN.2019.8852380

Wu, Y., Deng, L., Li, G., Zhu, J., and Shi, L. (2018). Spatio-temporal backpropagation for training high-performance spiking neural networks. *Front. Neurosci*. 12:331. doi: 10.3389/fnins.2018.00331

Xiao, H., Rasul, K., and Vollgraf, R. (2017). Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*.

Xie, X., and Seung, H. S. (2000). "Spike-based learning rules and stabilization of persistent neural activity," in *Advances in Neural Information Processing Systems*, eds T. K Leen, T. Dietterich, and V. Tresp (Denver: MIT Press), 199–208.

Zenke, F., Agnes, E. J., and Gerstner, W. (2015). Diverse synaptic plasticity mechanisms orchestrated to form and retrieve memories in spiking neural networks. *Nat. Commun*. 6, 1–13. doi: 10.1038/ncomms7922

Zhang, T., Zeng, Y., Zhao, D., and Shi, M. (2018a). "A plasticity-centric approach to train the non-differential spiking neural networks," in *Thirty-Second AAAI Conference on Artificial Intelligence* (New Orleans).

Zhang, T., Zeng, Y., Zhao, D., and Xu, B. (2018b). "Brain-inspired balanced tuning for spiking neural networks," in *IJCAI* (Stockholm), 1653–1659. doi: 10.24963/ijcai.2018/229

Zhao, F., Zeng, Y., and Xu, B. (2018). A brain-inspired decision-making spiking neural network and its application in unmanned aerial vehicle. *Front. Neurorobot*. 12:56. doi: 10.3389/fnbot.2018.00056

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Computational Neuroscience

METHODS
published: 30 November 2020
doi: 10.3389/fncom.2020.580632

DNNBrain: A Unifying Toolbox for Mapping Deep Neural Networks and Brains

Xiayu Chen[1], Ming Zhou[1], Zhengxin Gong[1], Wei Xu[1], Xingyu Liu[1], Taicheng Huang[2], Zonglei Zhen[1]* and Jia Liu[1]*

[1] Beijing Key Laboratory of Applied Experimental Psychology, Faculty of Psychology, Beijing Normal University, Beijing, China,
[2] State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University, Beijing, China

Deep neural networks (DNNs) have attained human-level performance on dozens of challenging tasks via an end-to-end deep learning strategy. Deep learning allows data representations that have multiple levels of abstraction; however, it does not explicitly provide any insights into the internal operations of DNNs. Deep learning's success is appealing to neuroscientists not only as a method for applying DNNs to model biological neural systems but also as a means of adopting concepts and methods from cognitive neuroscience to understand the internal representations of DNNs. Although general deep learning frameworks, such as PyTorch and TensorFlow, could be used to allow such cross-disciplinary investigations, the use of these frameworks typically requires high-level programming expertise and comprehensive mathematical knowledge. A toolbox specifically designed as a mechanism for cognitive neuroscientists to map both DNNs and brains is urgently needed. Here, we present DNNBrain, a Python-based toolbox designed for exploring the internal representations of DNNs as well as brains. Through the integration of DNN software packages and well-established brain imaging tools, DNNBrain provides application programming and command line interfaces for a variety of research scenarios. These include extracting DNN activation, probing and visualizing DNN representations, and mapping DNN representations onto the brain. We expect that our toolbox will accelerate scientific research by both applying DNNs to model biological neural systems and utilizing paradigms of cognitive neuroscience to unveil the black box of DNNs.

Keywords: deep neural network, brain imaging, neural representation, neural encoding and decoding, representational similarity analysis (RSA), feature visualization

OPEN ACCESS

Edited by:
Arpan Banerjee,
National Brain Research Centre (NBRC), India

Reviewed by:
Madhura Ingalhalikar,
Symbiosis International University, India
Raju Surampudi Bapi,
International Institute of Information Technology, Hyderabad, India

****Correspondence:***
Zonglei Zhen
zhenzonglei@bnu.edu.cn
Jia Liu
liujia@bnu.edu.cn

Received: *06 July 2020*
Accepted: *27 October 2020*
Published: *30 November 2020*

Citation:
Chen X, Zhou M, Gong Z, Xu W, Liu X, Huang T, Zhen Z and Liu J (2020) DNNBrain: A Unifying Toolbox for Mapping Deep Neural Networks and Brains.
Front. Comput. Neurosci. 14:580632.
doi: 10.3389/fncom.2020.580632

INTRODUCTION

Over the past decade, artificial intelligence (AI) has been able to make dramatic advances because of the rise of deep learning (DL) techniques. DL makes use of deep neural networks (DNNs) to model complex non-linear relationships and thus is able to solve real-life problems. A DNN often consists of an input layer, multiple hidden layers, and an output layer. Each layer generally implements some non-linear operations that transform the representation at one level into another representation at a more abstract level. In one particular example, deep convolutional neural network (DCNN) architecture stacks multiple convolutional layers hierarchically, inspired by the hierarchical organization of the primate ventral visual stream. A supervised learning algorithm is generally used to tune the parameters of the network to minimize errors between the network output and the target label in an end-to-end manner (LeCun et al., 1998; Rawat and Wang, 2017). As a result, DL is able to automatically discover multiple levels of representations that are needed for a given task (LeCun et al., 2015; Goodfellow et al., 2016). With this built-in architecture and learning from large external datasets, DCNNs have achieved human-level performance on a variety of challenging object (Krizhevsky et al., 2012; Simonyan and Zisserman, 2015; Szegedy et al., 2015; He et al., 2016) and speech recognition tasks (Hinton et al., 2012; Sainath et al., 2013; Hannun et al., 2014).

In addition to these achievements in engineering, DNNs provide a potentially rich interaction between studies on both biological and artificial information processing systems. On the one hand, DNNs offer the best models of biological intelligence to date (Cichy and Kaiser, 2019; Richards et al., 2019). In particular, good correspondence between DNNs and the visual systems has been identified (Yamins and DiCarlo, 2016; Kell and McDermott, 2019; Serre, 2019; Lindsay, 2020). First, DNNs exhibit behavioral patterns similar to those of human and non-human primate observers on some object recognition tasks (Jozwik et al., 2017; Rajalingham et al., 2018; King et al., 2019). Second, DCNNs appear to recapitulate the representation of visual information along the ventral stream. That is, early stages of the ventral visual stream (e.g., V1) are well-predicted by early layers of DNNs optimized for visual object recognition, whereas intermediate stages (e.g., V4) are best predicted by intermediate layers and late stages (e.g., IT) are best predicted by late layers (Khaligh-Razavi and Kriegeskorte, 2014; Yamins et al., 2014; Güçlü and van Gerven, 2015; Eickenberg et al., 2017). Finally, DNNs designated for object recognition spontaneously generate many well-known behavioral and neurophysiological signatures of cognitive phenomena such as shape tuning (Pospisil et al., 2018), numerosity (Nasr et al., 2019), and visual illusions (Watanabe et al., 2018). Thus, DNNs provide a new perspective to study the origin of intelligence. Indeed, neuroscientists have already used DNNs to model the primate visual system (Schrimpf et al., 2018; Lindsey et al., 2019; Lotter et al., 2020).

Alternatively, the end-to-end DL strategy makes DNN a black box, without any explanation of its internal representations. Experimental paradigms and theoretical approaches from cognitive neuroscience have significantly advanced our understanding of how DNNs work (Hasson and Nusbaum, 2019). First, concepts and hypotheses from cognitive neuroscience, such as sparse coding and modularity, provide a hands-on terminology to describe the internal operations of DNNs (Agrawal et al., 2014; Ritter et al., 2017). Second, a variety of methods of manipulating stimuli, such as stimulus degradation and simplification, have been used to characterize unit response dynamics (Baker et al., 2018; Geirhos et al., 2019). Finally, the rich data analysis techniques from cognitive neuroscience, such as ablation analysis (Morcos et al., 2018; Zhou et al., 2018), activation maximization (Nguyen et al., 2016), and representation similarity analysis (Khaligh-Razavi and Kriegeskorte, 2014; Jozwik et al., 2017), provide a powerful arsenal for exploring the computational mechanisms of DNNs.

Such a crosstalk between cognitive neuroscience and AI needs an integrated toolbox that meets the objectives of both fields. However, the most commonly used DL frameworks such as PyTorch[1] and TensorFlow[2] are developed for AI researchers. The use of these frameworks typically requires advanced programming expertise and comprehensive mathematical knowledge of DL. To our knowledge, there is no software package, specifically designed for both AI scientists and cognitive neuroscientists, that is able to interrogate DNNs and brains at the same time. Therefore, it would be of great value to have a unifying toolbox that maximally integrates DNN software packages and well-established brain mapping tools.

In this paper, we present DNNBrain, a Python-based toolbox specifically designed for exploring representations of both DNNs and brains. The toolbox has five major features.

- Versatility: DNNBrain supports a diverse range of applications for exploring DNN and brain representations. These include accessing DNN representations, building an encoding/decoding model for external stimuli, analyzing representational similarity between DNN and brain, transfer learning from pretrained models on study-specific stimuli, and visualizing DNN representations. Moreover, DNNBrain supports multiple modalities of input stimulus including image, audio, and video.
- Usability: DNNBrain provides a command line interface (CLI) and an application programming interface (API) for the user's convenience. At the application level, users can directly run commands to conduct typical representation analysis for both DNN and brain without any programming needed. At the programming level, all algorithms and computational pipelines are encapsulated into objects with high-level interface in the experimental design and data analysis language of neuroscientists. Users can easily program their own pipelines on these encapsulated algorithms objects.
- Transparent input/output (IO): DNNBrain transparently reads and writes multimodal neuroimaging data and multiple customized meta-data. As a result, DNNBrain spares users from the need to have specific knowledge about different data formats.
- Open source: DNNBrain is freely available in source. Users can access every detail of DNNBrain implementation. This improves the reproducibility of experimental results, leads to efficient debugging, and allows for accelerated scientific progress.
- Portability: DNNBrain, implemented in Python, runs on all major systems (e.g., Windows, Mac, and Linux). It is easy to set up, as it has no complicated dependencies on external libraries and packages.

As follows, we first introduce the functionalities of DNNBrain and then describe its framework (i.e., building blocks). Finally, with a typical application example, we demonstrate the versatility and usability of DNNBrain in characterizing both DNNs and brains as well as in examining the correspondences between

[1] https://pytorch.org

[2] https://www.tensorflow.org

DNNs and brains. The toolbox is freely available for download[3] and complemented with an expandable online documentation.[4]

Functionalities of DNNBrain

The primary aim of DNNBrain is to provide a framework that makes it easy to explore the internal representations of DNNs and brains, and the representational similarity between them. To do this, DNNBrain integrates a diverse range of tools such as encoding/decoding models to reveal stimuli or behavioral relevance of the representations, encoding/decoding models to map DNNs representations to those of brains, representational similarity analysis (RSA) between DNNs and brains, visualizing DNN representations, and transfer learning from pretrained models on study-specific stimuli.

Encoding and Decoding Model

Information processing in the brain and DNNs can generally be divided into two stages: (1) the neural code is generated from the stimuli (i.e., map stimuli to neural responses), and (2) the neural code is used to produce behavior (i.e., map neural responses to behavioral responses; Kriegeskorte and Douglas, 2019). In DNNBrain, neural (artificial) encoding models are implemented to do the former, whereas neural (artificial) decoding models are used for the latter (**Figure 1**).

Encoding models are implemented as linear models because the manner in which features of stimuli are represented in an explicit format by a neuron/voxel is a primary concern of neuroscientists (Yamins et al., 2014; Wen et al., 2018). Two kinds of linear models were introduced into DNNBrain to support encoding models (**Figure 2A**). First, univariate linear models (e.g., GLM, ridge, and lasso regression) were adopted to find linear combinations of stimuli features to predict the response of a neuron/voxel (Naselaris et al., 2011). The univariate encoding model describes how information is encoded in the activity of the individual neuron/voxel; however, it ignores interactions between different neurons/voxels. Second, multivariate partial least squares (PLS) linear models were introduced to find linear relations in two sets of multivariate variables (i.e., stimulus features and neural responses) by maximizing covariance of the transformed variables (Bilenko and Gallant, 2016; O'Connell and Chun, 2018). PLS models the covariance structures of stimuli features and neural responses, and thus provides information on how individual features and their interactions contribute to predicting responses from multiple neurons/voxels. Decoding models, which predict behavioral responses based on neural responses, work in the opposite direction of encoding models. Therefore, univariate linear models used for encoding models can serve as decoding models by simply exchanging response variables for predictor variables of the encoding models (**Figure 2B**).

DNNBrain uses cross-validation (CV) techniques (e.g., k-fold and leave-one-out CV) to evaluate the generalization performance of encoding/decoding models. The CV techniques divide a dataset into several non-overlapping subsets. Each subset is held back in turn as the test set, whereas all other subsets are collectively used as a training dataset. The accuracy (i.e., the fraction of correct predictions) and explained variance are generally used to measure performance for classification- and regression-based encoding/decoding models, respectively. Permutation testing is utilized to test the significance of the model performance. The null distribution is generated by deriving the performance measure multiple times using original data samples, but with permuted targets.

Analyzing Representational Similarity

Another focus of DNNBrain is to provide tools to examine representational similarities between DNNs and brains (i.e., describe the relationships between neural responses from DNNs and those from brains) (**Figure 1**). First, encoding models can be used to examine the representational similarity between DNNs and brains if internal representations of DNNs are considered as extracted features of external stimuli (**Figure 2A**). Second, representational similarity analysis (RSA) was implemented in DNNBrain to evaluate the similarity between two representations (Kriegeskorte et al., 2008) (**Figure 2C**). RSA differs from encoding/decoding models, which measure the representational similarity between DNNs and brains by examining how brain responses could be directly predicted from DNN responses, or vice versa. In contrast, RSA utilizes pairwise comparison of stimuli in representation space to characterize their representation. Representational dissimilarity, which is often calculated as Euclidean distance or correlation distance between two multivariate response patterns, is first created for every pair of stimuli or conditions, and then summarized in a representational dissimilarity matrix (RDM) which characterizes the geometry of the set of points in the multivariate response space. Finally, the correlation between RDMs from DNNs and brains is calculated to measure their representational similarity. Multiple correlation metrics are supported by DNNBrain including the Pearson correlation, Kendall's tau correlation, and Spearman's correlation. Permutation tests were integrated in DNNBrain to estimate significance of the representational similarity between DNNs and brains. The permutation test randomizes the stimulus labels multiple times to generate the null distribution.

Transfer Learning From Pretrained Models on Study-Specific Stimuli

Training a DNN from scratch often requires a large amount of computational demand that results in significant time and energy costs. Moreover, there usually is not enough existing data available to train a DNN *de novo*. Fortunately, it turns out that representations from pretrained DNNs on large datasets (e.g., ImageNet) often work well for related new tasks. Therefore, instead of training a DNN from scratch, it can be trained to solve a new task by fine-tuning the weights of a pretrained model using just a very few training examples. This is known as transfer learning. Clearly, transfer learning is of great value in the study of representational similarities between DNNs and brains because it is often not possible to collect large-scale neural datasets. DNNBrain provides a set of utilities that assists users in

[3] http://github.com/BNUCNL/dnnbrain

[4] http://dnnbrain.readthedocs.io

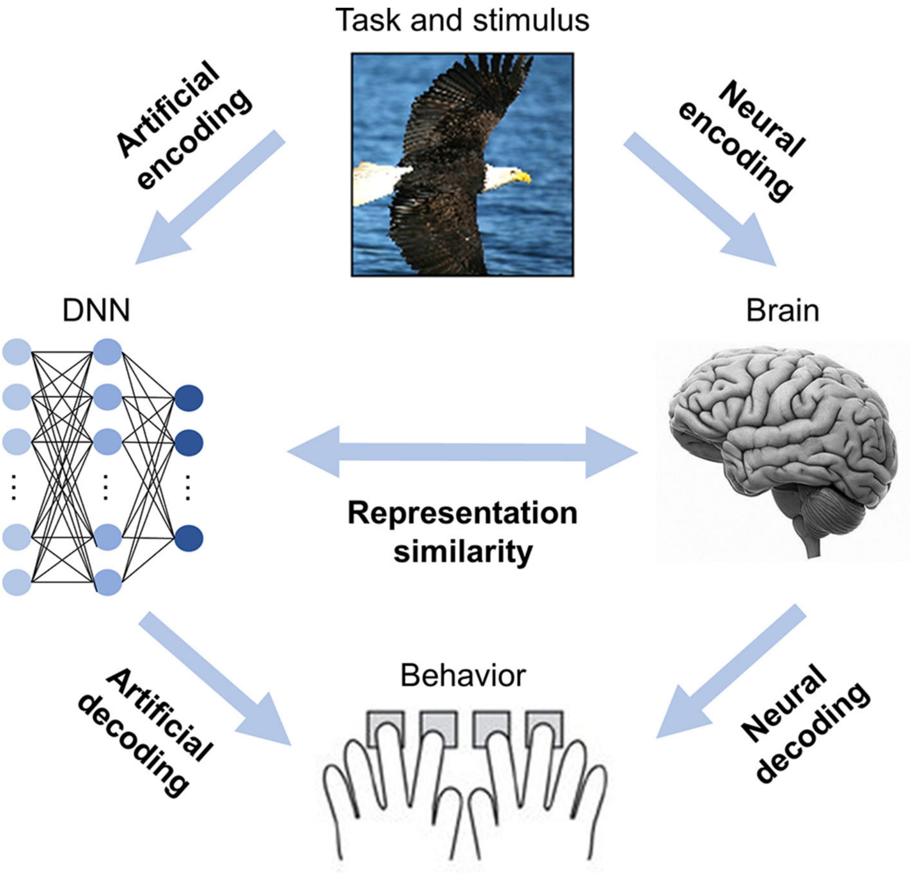


FIGURE 1 | DNNBrain is designed as an integrated toolbox that characterizes artificial representations of DNNs and neural representations of brains. After stimuli are submitted to both DNNs and brains, the artificial neural activities, and the biological neural activities are acquired. By assembling the stimuli, the artificial activity data, and the biological neural activity data together with custom-designed auxiliary IO files, DNNBrain allows users to easily characterize, compare, and visualize representations of DNNs and brains.

transfer learning from pretrained DNNs on their study-specific dataset. Users can easily specify which target layers/channels to be fine-tuned and customize the new task layers.

Visualizing Features From DNNs

DNNs are a kind of complex non-linear transformation that does not provide explicit explanation of their internal workings. Identifying the relevant features that contribute most to the responses of an artificial neuron is central to the understanding of precisely what each neuron has learned (Montavon et al., 2018; Nguyen et al., 2019). Three approaches have been implemented in DNNBrain to assist users in examination of the stimulus features that an artificial neuron prefers. The first approach is known as *top stimulus discovering*. The top images with the highest activations for a specific neuron (or unit) are identified from a large image collection (Zeiler and Fergus, 2014; Yosinski et al., 2015). The second approach, known as *saliency mapping*, computes gradients on the input images relative to the target unit, utilizing a backpropagation algorithm. It highlights pixels of the image that change the unit's activation most when its value changes (Simonyan et al., 2014; Springenberg et al., 2015). The third approach is termed *optimal stimulus synthesizing*. This approach synthesizes the visual stimulus from initial random noise, guided by increasing activation of the target neuron (Erhan et al., 2009; Nguyen et al., 2016).

Other Utilities Provided by DNNBrain

In addition to the functionalities described previously, DNNBrain provides additional flexible pipelines for neuroscience-orientated analysis of DNNs. These include ablation analysis of individual units (Morcos et al., 2018; Zhou et al., 2018) and estimation of the empirical receptive field of a unit (Zhou et al., 2014). It also comes with a variety of utilities, such as image processing tools used for converting different data structures (e.g., PyTorch tensor, NumPy array, and PIL image objects), translating and cropping images, and more. Details can be found on the DNNBrain documentation page.[4]

Implementation of DNNBrain

DNNBrain is a modular Python toolbox that consists of four modules: IO, Base, Model, and Algorithm (**Figure 3**). The Python language was selected for DNNBrain because it provides an ideal

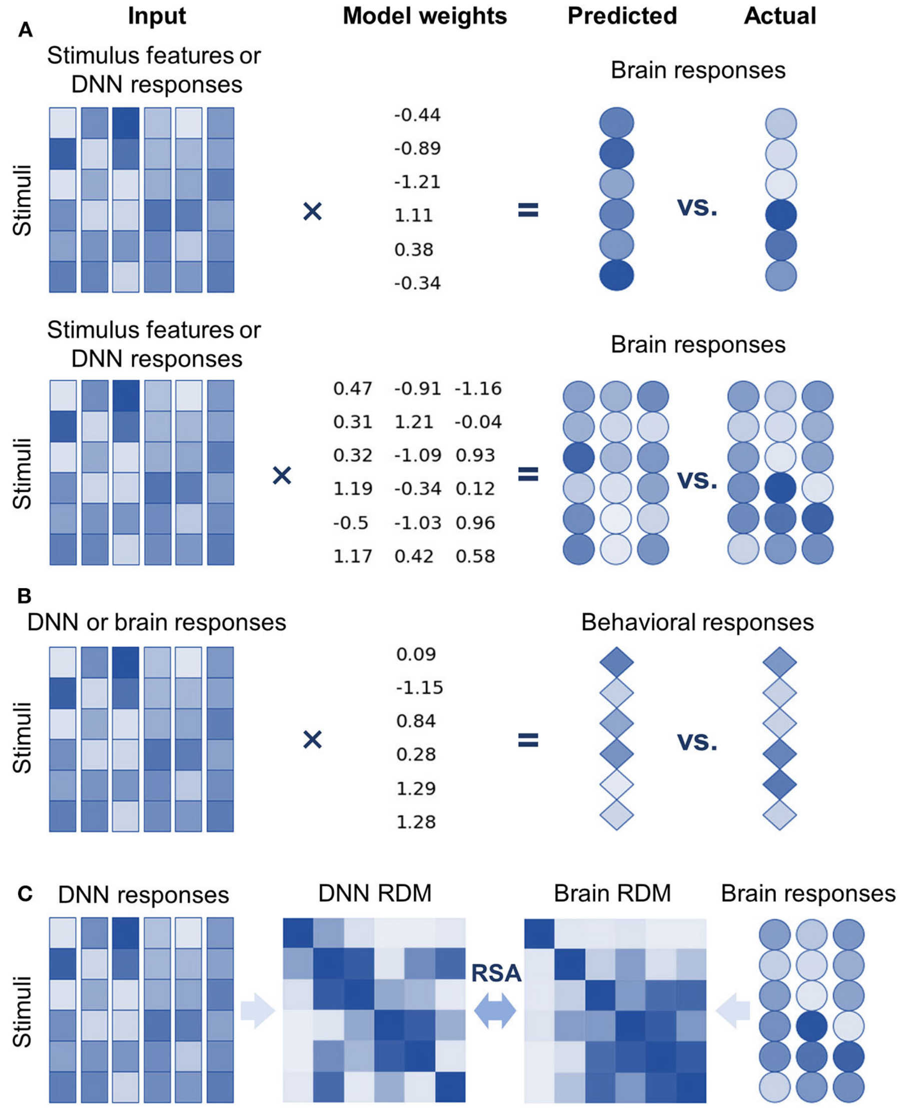


FIGURE 2 | DNNBrain provides multiple approaches to explore internal representations of DNNs and the brain, and the representational similarities between them. **(A)** Top: univariate linear encoding models find optimal linear combinations of multiple stimulus features (or DNN responses) to predict the response of a neuron/voxel. Bottom: multivariate linear models search optimal linear combinations of multiple stimulus features (or DNN responses) to predict the responses from multiple neurons/voxels by maximizing their covariance. **(B)** In the opposite direction of encoding models, linear decoding models find optimal linear combinations of neural responses (or DNN responses) to predict behavior responses. **(C)** Representational similarity analysis evaluates the similarity of two representations by comparing representational dissimilarity matrices obtained from them.

environment for the research on DNNs and brains. First, Python is currently the most commonly used programming language for scientific computing. Many excellent Python libraries have been developed for scientific computing. The libraries used in the DNNBrain are as follows: NumPy for numerical computation,[5] SciPy for general-purpose scientific computing,[6] scikit-learn for machine learning,[7] and Python imaging library (PIL) for image processing.[8] Second, Python is increasingly used in the field of brain imaging. Many Python libraries for brain imaging data analysis have been developed such as NiPy[9] (Millman and Brett, 2007) and fMRIPrep[10] (Esteban et al., 2019). Finally, Python is the most popular language in the field of DL. Python is well-supported by the two most popular DNN libraries (i.e., PyTorch[1] and TensorFlow[2]).

[5] https://numpy.org
[6] https://www.scipy.org
[7] https://scikit-learn.org
[8] http://pythonware.com/products/pil
[9] https://nipy.org
[10] https://fmriprep.org

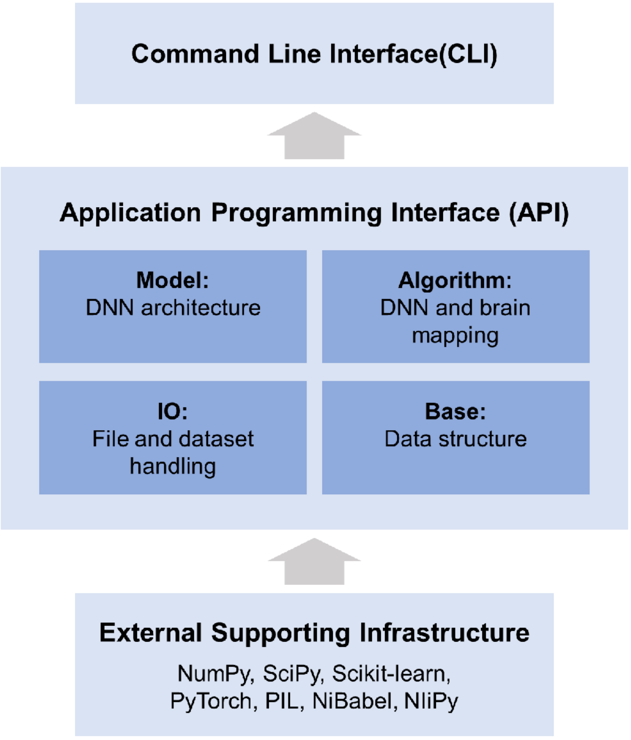

Command Line Interface(CLI)
Application Programming Interface (API)
Model:
DNN architecture
Algorithm:
DNN and brain mapping
IO:
File and dataset handling
Base:
Data structure
External Supporting Infrastructure
NumPy, SciPy, Scikit-learn,
PyTorch, PIL, NiBabel, NIiPy

FIGURE 3 | DNNBrain is a modular framework which consists of four modules: IO, Base, Model, and Algorithm. The IO module provides facilities for managing file-related input and output operations. The Base module defines base level classes for array computing and data transforming. The Model module holds a variety of DNN models. The Algorithm module defines various algorithms for exploring DNNs and the brain. All modules provide user-friendly APIs. A set of CLIs was developed for a variety of research scenarios.

Supported by a large variety of existing software packages, DNNBrain was designed with a high-level API in the domain language of cognitive neuroscience. All algorithms and computational pipelines are encapsulated into classes in an object-oriented programming manner. All modules provide user-friendly APIs. On these APIs, a set of CLIs was developed for a variety of research scenarios.

Of note, neuroimaging data preprocessing pipelines are not included in DNNBrain. The data need to be preprocessed before they are input into DNNBrain. This separation between the DNNBrain representation analysis pipeline and the data preprocessing pipeline provides users with maximum flexibility to utilize different neuroimaging toolboxes to preprocess their data.

IO Module: Organizing Datasets in DNNBrain

DNNBrain introduces auxiliary file formats to handle various types of scientific data and supporting metadata. These include stimulus files, DNN mask files, and DNN activation files. With these file formats, users can easily organize their inputs and outputs. The stimulus file is a comma separated values (CSV) text file designed to configure stimulus information including the stimulus type (image, audio, and video), stimulus directory, stimulus ID, stimulus duration, stimulus conditions, and other possible stimulus attributes. The DNN mask file is also a CSV text file designed for users to specify channels and units of interest when analyzing DNNs. Both the stimulus file and the DNN mask file can be easily configured with a text editor. The DNN activation file is a HDF5 (Hierarchical Data Format) file in which activation values from specified channels are stored. In addition, DNNBrain uses NiBabel[11] to access brain imaging files. Almost all common MRI file formats are supported, including GIFTI, NIfTI, CIFTI, and MGH.

Base Module: Defining the Basic Data Structure

The base module defines base level objects for data structure and data transformations. Specifically, a set of objects is defined to organize either data from the input stimulus or the output activation data from the DNN. The data objects were designed to be as simple as possible, while retaining necessary information for further representation analysis. The stimulus object contains stimulus paths and associated attributes (e.g., category label), which are read from stimulus files. The activation object holds DNN activation patterns and associated location information (e.g., layer, channel, and unit). Aside from these data objects, several encoding/decoding models were developed, including popular classification and regression models such as generalized linear models, logistic regression, and lasso. Each of these models was wrapped from the widely used machine learning library, scikit-learn.[7]

Model Module: Encapsulating DNNs

In DNNBrain, a DNN model is implemented as a neural network model from PyTorch. Each DNN model is a sequential container which holds the DNN architecture (i.e., connection pattern of units) and associated connection weights. The DNN model is equipped with a suite of methods that access attributes of the model and update states of the model. PyTorch has become the most popular DL framework because of its simplicity and ease of use in creating and deploying DL applications. At present, several well-known PyTorch DCNN models[12] pretrained for different stimulus modalities have been adopted into DNNBrain, including AlexNet (Krizhevsky et al., 2012), VGG (Simonyan and Zisserman, 2015), GoogLeNet (Szegedy et al., 2015), and ResNet (He et al., 2016) for image classification; VGGish for audio classification (Hershey et al., 2017); and R3D for video classification (Tran et al., 2018).

Algorithm Module: Characterizing DNNs and Brains

The algorithm module defines various algorithms objects for exploring DNNs. An algorithm object contains a DNN model and corresponding methods that allow the study of specific properties of the model. Three types of algorithms are implemented in DNNBrain. The first type is the gradient descent algorithm for DNN model training, which is wrapped from PyTorch.[13] The second type of algorithm comprises tools for extracting and summarizing the activation of a

[11] https://nipy.org/nibabel

[12] https://github.com/pytorch/vision

[13] https://pytorch.org/docs/stable/optim.html

DNN model, such as principal component analysis (PCA) and clustering. The third type is made up of algorithms that visualize representations of a DNN, including discovering the top stimulus, mapping saliency features of a stimulus, and synthesizing the maximum activation stimulus for a specific DNN channel. Each algorithm takes a DNN model, as well as a stimulus object, as input.

Command Line Interface

At the application level, DNNBrain provides several workflows as command line interface, including those that access DNN representations, visualize DNN representations, evaluate the behavioral relevance of the representations, and map DNN representations to brains. Users can conveniently run commands to perform typical representation analysis on their data.

Extension of DNNBrain

Along with the modules and algorithms that have already been implemented in DNNBrain, the user can extend DNNBrain in the following ways. First, any PyTorch model can be easily wrapped into DNNBrain by inheriting DNN Class and overriding its few methods. Second, any linear or non-linear model can conveniently be introduced into DNNBrain as either an encoding/decoding model, as long as they have the same interface as the scikit-learn Classifier/Regression object. Finally, users can write their own scripts to develop customized pipelines by reusing the algorithms and dataset objects.

METHODS

DNN Model: AlexNet

AlexNet is used as an example to illustrate the functionality of DNNBrain. AlexNet is one of the most influential DCNNs. In the 2012 ImageNet challenge (Krizhevsky et al., 2012), it demonstrated for the first time that DCNNs can increase ImageNet classification accuracy by a significant stride. AlexNet is composed of five convolutional (Conv) layers and three fully connected (FC) layers that receive inputs from all units in the previous layer (**Figure 4A**). Each Conv layer is generally composed of a convolution, a rectified linear unit function (ReLU), and max pooling operations. These operations are repeatedly applied across the image. In this paper, when we refer to Conv layers, we mean the output after the convolution and ReLU operations.

Because AlexNet contains thousands of units in each layer, the dimension (i.e., the number of units) of the activation patterns

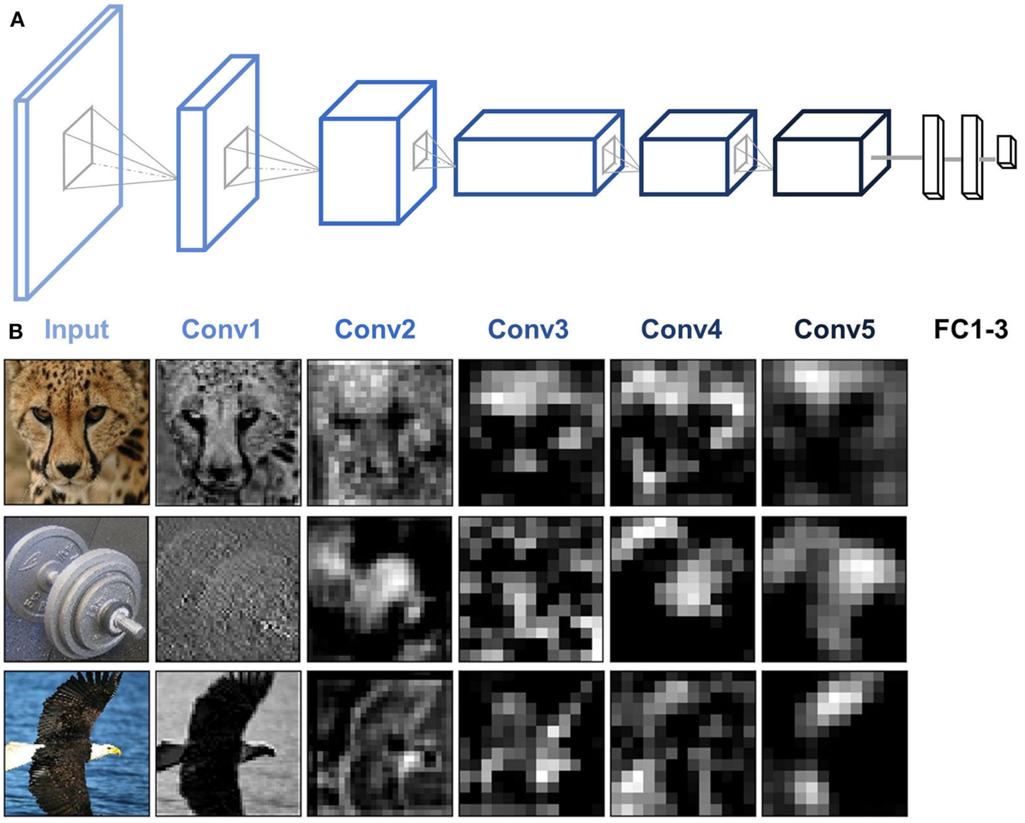


FIGURE 4 | AlexNet architecture and activity patterns from example units. **(A)** AlexNet consists of five Conv layers followed by three FC layers. **(B)** The activation maps from each of the five Conv layers of AlexNet were extracted for three example images (cheetah, dumbbell, and bald eagle). Presented channels are those showing maximal mean activation for that example image within each of the five Conv layers.

from each layer was reduced to 100 via PCA to avoid the risk of overfitting the models in further analyses of DNN and brain representation.

BOLD5000: Stimulus and Neuroimaging Data

BOLD5000 is a large-scale publicly available human functional MRI (fMRI) dataset in which four participants underwent slow event-related BOLD fMRI while viewing ∼5,000 distinct images depicting real-world scenes (Chang et al., 2019). The stimulus images were drawn from the three most commonly used computer vision datasets: 1,000 hand-curated indoor and outdoor scene images from the Scene UNderstanding dataset (Xiao et al., 2010), 2,000 images of multiple objects from the Common Objects in Context dataset (Lin et al., 2014), and 1,916 images of mostly singular objects from the ImageNet dataset (Deng et al., 2009). Each image was presented for 1 s followed by a 9-s fixation cross. Functional MRI data were collected using a T2*-weighted gradient recalled echo planar imaging multi-band pulse sequence (In-plane resolution = 2 × 2 mm; 106 × 106 matrix size; 2 mm slice thickness, no gap; TR = 2,000 ms; TE = 30 ms; flip angle = 79°). The scale, diversity, and naturalness of the stimuli, combined with a slow event-related fMRI design, make BOLD5000 an ideal dataset to explore the DNNs and brain representations of a wide range of visual features and object categories. The raw fMRI data were preprocessed utilizing the fMRIPrep pipeline including motion correction, linear detrending, and spatial registration to native cortical surface via boundary-based registration (Esteban et al., 2019). No additional spatial or temporal filtering was applied. For a complete description of the experimental design, fMRI acquisition, and preprocessing pipeline, see Chang et al. (2019).

The preprocessed individual fMRI data were firstly transformed into 32k_fs_LR space using ciftify (Dickie et al., 2019). BOLD response maps for each image were then estimated from the fMRI data using the general linear model (GLM) from HCP Pipelines (Glasser et al., 2013). The response maps of each image were finally averaged across four subjects in the fsLR space and used for further analyses. Moreover, we constrained our analysis to the ventral temporal cortex (VTC), a critical region for object visual recognition. The VTC region was defined by merging the areas V8, FFC (fusiform face complex), PIT (posterior inferotemporal complex), VVC (ventral visual complex), and VMV (ventromedial visual areas) from HCP MMP 1.0 (Glasser et al., 2016). DNNBrain pipelines support both surface and volume data. Here, we preferred to use surface-based preprocessed data instead of volume-based preprocessed data because previous studies have shown that surface-based analysis can increase the specificity of cortical activation patterns (Van Essen et al., 1998; Brodoehl et al., 2020).

RESULTS

We demonstrated the functionality of DNNBrain on AlexNet and BOLD5000 dataset. Specifically, we accessed DNN activation of the images from BOLD5000, probed the category information represented in each DNN layer, mapped the DNN representations onto the brain, and visualized the DNN representations. We do not aim to illustrate the full functionalities that are available from DNNBrain, but rather to sketch out how DNNBrain can be easily used to examine DNN and brain representations in a realistic study. All the analyses were implemented in both API and CLI levels. The code can be found in the DNNBrain online documentation.[4]

Scanning DNNs

To examine the artificial representations of DNNs, we needed to scan the DNN to obtain its neural activities, just as we scan the human brain using brain imaging equipment. DNNBrain

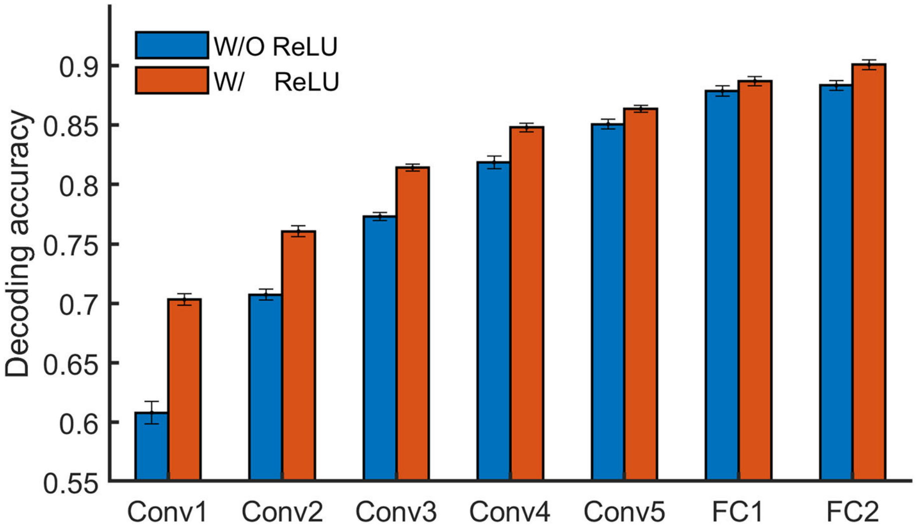


FIGURE 5 | DNNBrain provides linear decoding models to probe the explicit representation contents of layers of interest in a DNN. On BOLD5000 stimuli, a logistic regression model revealed that the higher a layer is, the more animate information is encoded within it.

provides both API and CLI to extract activation states for user-specified channels of a DNN. **Figure 4** shows the activation patterns of three example images (cheetah, dumbbell, and bald eagle) from the channels of AlexNet which showed the maximal mean activation within each of the five Conv layers. The activation patterns revealed that DNN representations of the images became more abstract along the depth of the layers.

Revealing Information Presented in DNN Layers

To learn whether specific stimuli attributes or behavioral performances are explicitly encoded in a certain layer of a DNN, one direct approach is to measure to what degree the representation from the layer is useful for decoding them. Linear decoding models (classifier or regression) were implemented in DNNBrain to enable this. Here, we manually sorted BOLD5000 stimulus images into binary categories (animate vs. inanimate) according to salient objects located in each image, and then examined how animate information is explicitly encoded in AlexNet. In total, 2,547 images were labeled as animate and 2,369 as inanimate. We trained a logistic regression model on the artificial representations to decode the stimulus category for each Conv layer of AlexNet. The accuracy of the model was evaluated with a 10-fold cross-validation. As shown in **Figure 5**, the classification accuracy progressed with the depth of Conv layers, indicating higher layers encoded more animate information than lower layers. Moreover, the ReLU operation within each Conv layer played a significant role in improving the representation capacity for animate information.

Mapping Representations Between a DNN and the Brain

A growing body of literature is investigating the potential of DNNs to work as models of brain information processing. Several recent studies found that internal representations of object recognition DNNs provided the best current models of representations of visual images in the inferior temporal cortex of both humans and monkeys (for a recent review, see Lindsay, 2020). Here, we adopted the univariate encoding model, multivariate encoding model, and RSA on BOLD5000 dataset to map artificial representations from Conv layers of AlexNet to neural representations from the VTC of the brain. On the artificial representation from each Conv layer of AlexNet, a univariate GLM encoding model was constructed for each voxel within the VTC, and a multivariate PLS encoding model was built for the whole VTC. Encoding accuracy was evaluated with the Pearson correlation between the measured responses and the predicted responses from the encoding model using a 10-fold cross-validation procedure. For RSA, RDM was derived using the correlation distance between each pair of stimuli, and a Pearson correlation was used to measure the similarity between two representations. Four main findings were revealed (**Figure 6**). First, the encoding accuracy of the VTC gradually increased for the hierarchical layers of AlexNet, indicating that as the complexity of the visual representations increases along the DNN hierarchy, the representations become increasingly VTC-like. Second, the encoding accuracy varied greatly across voxels within the VTC for the artificial representations of each AlexNet layer, suggesting the VTC may organize in distinct functional modules, each preferring different kinds of features. Third, the univariate encoding model and the multivariate encoding model produced similar results, indicating that interactions between different voxels encode little representation information from each DNN Conv layer. Finally, RSA also showed results similar to those of encoding models, suggesting that the encoding model and RSA are likely to be equally useful for comparing representations from DNNs and brains.

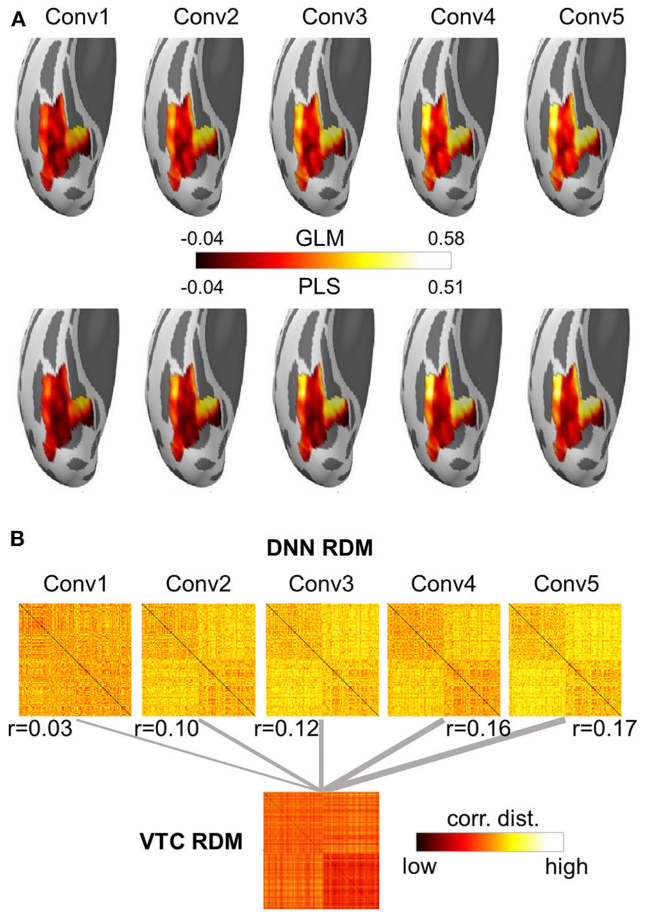


FIGURE 6 | Both the encoding model and the representational similarity analysis are implemented in DNNBrain to help researchers to examine the correspondence between the DNN and brain representations. **(A)** Encoding accuracy maps from univariate GLM encoding models of predicting VTC BOLD responses using artificial representation from the Conv layers of AlexNet (top), and encoding accuracy maps from multivariate PLS encoding models of predicting VTC BOLD responses using artificial representation from the Conv layers of AlexNet (bottom). **(B)** RDMs for BOLD5000 stimuli computed on artificial representations from Conv layers of AlexNet and brain activation patterns from the human VTC. The representation distance between each pair of images was quantified as the correlation distance between their representations. The representational similarity between the DNN and the brain is further calculated as the Pearson correlation between their RDMs.

Visualizing Features From DNNs

Visualization of critical features of a stimulus that cause the responses of an artificial neuron is central to the understanding of precisely what each neuron has learned. As an example, we

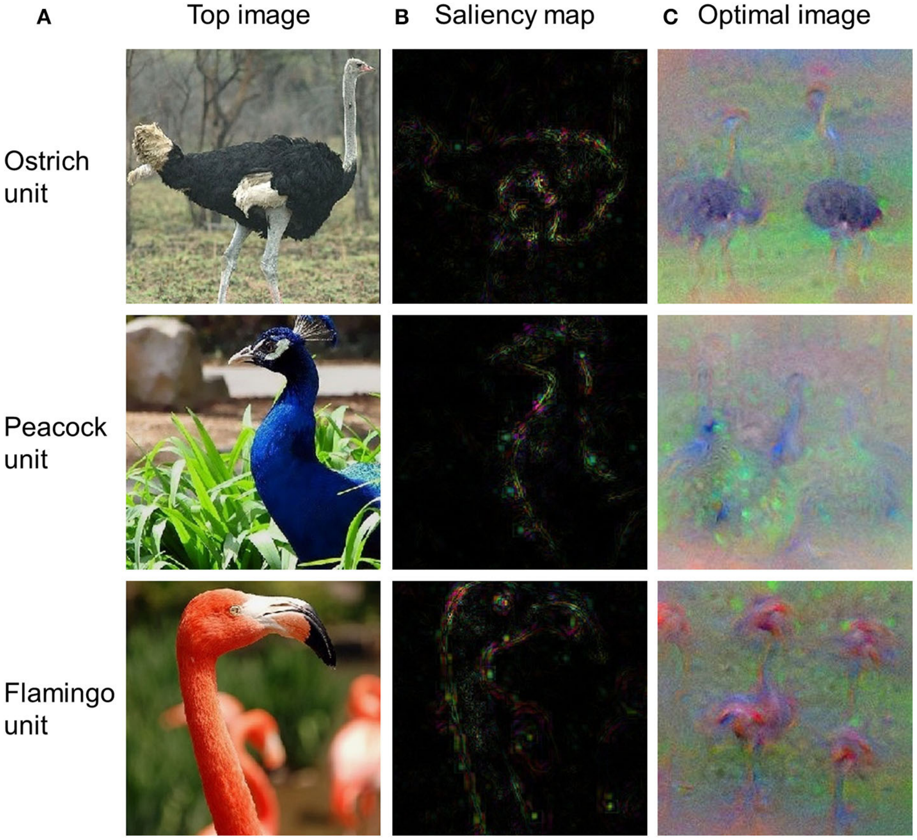


FIGURE 7 | The top stimuli, saliency maps, and optimal images for three output units of AlexNet. **(A)** Top stimuli discovered from the BOLD5000 dataset. **(B)** Saliency maps computed for the top stimuli presented in **(A)**. **(C)** Optimal images synthesized from initial random noise guided by increasing the activation of corresponding neurons.

used three DNN visualization approaches from DNNBrain (i.e., top stimulus, saliency map, and optimal stimulus) to visualize the preferred features for three output units of AlexNet (i.e., ostrich, peacock, and flamingo). The output units were selected as examples because produced features for them are easy to check (i.e., each unit corresponds to a unique category). These procedures essentially work for any unit in a DNN. As shown in **Figure 7A**, the top stimulus was correctly found from 4,916 BOLD5000 images for each of three units: every top stimulus contains the object in the correct category. Saliency maps highlight the pixels in the top stimuli that contribute most to the activation of the neurons (**Figure 7B**). Finally, the optimal images synthesized from initial random noise correctly produced objects of the corresponding category (**Figure 7C**). In summary, these three approaches are able to reveal the visual patterns that a neuron has learned on various levels and thus provide a qualitative guide to neural interpretations.

DISCUSSION

DNNBrain integrates well-established DNN software and brain imaging packages to enable researchers to conveniently map the representations of DNNs and brains, and examine their correspondences. DNN models provide a biologically plausible account of biological neural systems, and show great potential for generating novel insights into the neural mechanisms of brains. On the other hand, experimental paradigms from cognitive neuroscience provide powerful approaches to pry open the black boxes of DNNs. DNNBrain, as a toolbox that is specifically tailored toward mapping the representations of DNNs and brains, has good potential to accelerate the merge of these two trends.

There are some issues that we would like to target in future development. First, DNNBrain integrates many of the currently most popular pretrained DCNN models. With the advance of the interplay between neuroscience and DNN communities, new DNN models are constantly emerging, and will be included into future iterations of DNNBrain. For example, generative adversarial networks could be introduced into DNNBrain to help users reconstruct external stimuli (Shen et al., 2019; VanRullen and Reddy, 2019) or synthesize preferred images for either neurons or brain areas (Ponce et al., 2019). Second, DNNBrain, up until now, only supports DNN models from PyTorch, which limits the study of DNNs constructed under other frameworks. We would like to put significant effort toward integrating other DNN frameworks into DNNBrain, especially TensorFlow. Third, only fMRI data are currently well-supported in DNNBrain. The magnetoencephalography (MEG), electroencephalography (EEG), multiunit recordings, and local field potentials can capture the temporal dynamics of neural representations which fMRI cannot provide. Support for these modalities is forthcoming according to recently published data standardization of electrophysiology (Niso et al., 2018; Pernet et al., 2019). Finally, DNNBrain mainly supports the exploration of pretrained DNN models, trained on large-scale external stimuli. It would be a good idea in the future to equip DNNBrain with tools that fuse brain activities and external tasks/stimuli to create DNN models that more closely resemble the human brain. Recent advances demonstrate that brain representations provide additional and efficient constraints on DNN constructions (McClure and Kriegeskorte, 2016; Fong et al., 2018). The brain has acquired a robust representation that generalizes across many tasks. As a result, while training DNNs to solve behavioral tasks, co-training DNNs to match the brain's latent representations observed from massive neural recordings will move the representation of DNNs toward these neural representations, and make them more closely resemble the human brain.

DATA AVAILABILITY STATEMENT

DNNBrain is freely available via github.[3] The code and data used in this article is available from readthedocs[4] and OSF[14], respectively. Further inquiries can be directed to the corresponding authors.

[14] https://osf.io/hy5m7

ETHICS STATEMENT

All procedures followed the principles in the Declaration of Helsinki. Participants all provided written informed consent. The experimental MRI protocols for BOLD5000 were approved by the Institutional Review Board (IRB) of Carnegie Mellon University. The reported analyses in this study were approved by the IRB of the Beijing Normal University.

AUTHOR CONTRIBUTIONS

XC, MZ, ZG, WX, XL, TH, and ZZ developed the toolbox and designed the validations. XC, ZZ, and JL wrote the paper. MZ, ZG, WX, XL, and TH revised and approved the paper. All authors contributed to the article and approved the submitted version.

FUNDING

This study was funded by the National Key R&D Program of China (Grant No. 2019YFA0709503), the National Natural Science Foundation of China (Grant Nos. 31861143039 and 31771251), and the National Basic Research Program of China (Grant No. 2018YFC0810602).

REFERENCES

Agrawal, P., Girshick, R., and Malik, J. (2014). "Analyzing the performance of multilayer neural networks for object recognition," in *European Conference on Computer Vision*, eds D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars (Zurich: Springer), 329–344. doi: 10.1007/978-3-319-10584-0_22

Baker, N., Lu, H., Erlikhman, G., and Kellman, P. J. (2018). Deep convolutional networks do not classify based on global object shape. *PLoS Comput. Biol.* 14:e1006613. doi: 10.1371/journal.pcbi.1006613

Bilenko, N. Y., and Gallant, J. L. (2016). Pyrcca: regularized kernel canonical correlation analysis in python and its applications to neuroimaging. *Front. Neuroinform.* 10:49. doi: 10.3389/fninf.2016.00049

Brodoehl, S., Gaser, C., Dahnke, R., Witte, O. W., and Klingner, C. M. (2020). Surface-based analysis increases the specificity of cortical activation patterns and connectivity results. *Sci. Rep.* 10:5737. doi: 10.1038/s41598-020-62832-z

Chang, N., Pyles, J. A., Marcus, A., Gupta, A., Tarr, M. J., and Aminoff, E. M. (2019). BOLD5000, a public fMRI dataset while viewing 5000 visual images. *Sci. Data* 6:49. doi: 10.1038/s41597-019-0052-3

Cichy, R. M., and Kaiser, D. (2019). Deep neural networks as scientific models. *Trends Cogn. Sci.* 23, 305–317. doi: 10.1016/j.tics.2019.01.009

Deng, J., Dong, W., Socher, R., Li, L., Kai, L., and Fei-Fei, L. (2009). "ImageNet: a large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition* (Miami, FL: IEEE), 248–255. doi: 10.1109/CVPR.2009.5206848

Dickie, E. W., Anticevic, A., Smith, D. E., Coalson, T. S., Manogaran, M., Calarco, N., et al. (2019). Ciftify: a framework for surface-based analysis of legacy MR acquisitions. *Neuroimage* 197, 818–826. doi: 10.1016/j.neuroimage.2019.04.078

Eickenberg, M., Gramfort, A., Varoquaux, G., and Thirion, B. (2017). Seeing it all: Convolutional network layers map the function of the human visual system. *Neuroimage* 152, 184–194. doi: 10.1016/j.neuroimage.2016.10.001

Erhan, D., Bengio, Y., Courville, A., and Vincent, P. (2009). *Visualizing higher-layer features of a deep network*. Montreal, QC: Tech. Report. Univ. Montr.

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., et al. (2019). fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nat. Methods* 16, 111–116. doi: 10.1038/s41592-018-0235-4

Fong, R. C., Scheirer, W. J., and Cox, D. D. (2018). Using human brain activity to guide machine learning. *Sci. Rep.* 8:5397. doi: 10.1038/s41598-018-23618-6

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. (2019). *"ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and* robustness," in *International Conference on Learning Representations* (New Orleans).

Glasser, M. F., Coalson, T. S., Robinson, E. C., Hacker, C. D., Harwell, J., Yacoub, E., et al. (2016). A multi-modal parcellation of human cerebral cortex. *Nature* 536, 171–178. doi: 10.1038/nature18933

Glasser, M. F., Sotiropoulos, S. N., Wilson, J. A., Coalson, T. S., Fischl, B., Andersson, J. L., et al. (2013). The minimal preprocessing pipelines for the human connectome project. *Neuroimage* 80, 105–124. doi: 10.1016/j.neuroimage.2013.04.127

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. Cambridge, MA: MIT Press.

Güçlü, U., and van Gerven, M. A. J. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *J. Neurosci.* 35, 10005–10014. doi: 10.1523/JNEUROSCI.5023-14.2015

Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., et al. (2014). Deep Speech: Scaling up end-to-end speech recognition. *arXiv [preprint]* arXiv:1412.5567.

Hasson, U., and Nusbaum, H. C. (2019). Scientific life emerging opportunities for advancing cognitive neuroscience. *Trends Cogn. Sci.* 23, 363–365. doi: 10.1016/j.tics.2019.02.007

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, (Las Vegas, NV), 770–778. doi: 10.1109/CVPR.2016.90

Hershey, S., Chaudhuri, S., Ellis, D. P. W., Gemmeke, J. F., Jansen, A., Moore, R. C., et al. (2017). "CNN architectures for large-scale audio classification," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, (New Orleans, LA), 131–135. doi: 10.1109/ICASSP.2017.7952132

Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A., Jaitly, N., et al. (2012). Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. *IEEE Signal Process. Mag.* 29, 82–97. doi: 10.1109/MSP.2012.2205597

Jozwik, K. M., Kriegeskorte, N., Storrs, K. R., and Mur, M. (2017). Deep convolutional neural networks outperform feature-based but not categorical models in explaining object similarity judgments. *Front. Psychol.* 8:1726. doi: 10.3389/fpsyg.2017.01726

Kell, A. J., and McDermott, J. H. (2019). Deep neural network models of sensory systems: windows onto the role of task constraints. *Curr. Opin. Neurobiol.* 55, 121–132. doi: 10.1016/j.conb.2019.02.003

Khaligh-Razavi, S. M., and Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Comput. Biol.* 10:e1003915. doi: 10.1371/journal.pcbi.1003915

King, M. L., Groen, I. I. A., Steel, A., Kravitz, D. J., and Baker, C. I. (2019). Similarity judgments and cortical visual responses reflect different properties of object and scene categories in naturalistic images. *Neuroimage* 197, 368–382. doi: 10.1016/j.neuroimage.2019.04.079

Kriegeskorte, N., and Douglas, P. K. (2019). Interpreting encoding and decoding models. *Curr. Opin. Neurobiol.* 55, 167–179. doi: 10.1016/j.conb.2019.04.002

Kriegeskorte, N., Mur, M., and Bandettini, P. (2008). Representational similarity analysis – connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* 2:4. doi: 10.3389/neuro.06.004.2008

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems,* eds F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Lake Tahoe, CA: Curran Associates Inc.), 1097–1105.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 2278–2324. doi: 10.1109/5.726791

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., et al. (2014). "Microsoft COCO: common objects in context," in *European Conference on Computer Vision* (Zurich), 740–755. doi: 10.1007/978-3-319-10602-1_48

Lindsay, G. W. (2020). Convolutional neural networks as a model of the visual system: Past, present, and future. *J. Cogn. Neurosci.* doi: 10.1162/jocn_a_01544. [Epub ahead of print].

Lindsey, J., Ocko, S. A., Ganguli, S., and Deny, S. (2019). "A unified theory of early visual representations from retina to cortex through anatomically constrained deep CNNs," in *International Conference on Learning Representations* (New Orleans). doi: 10.1101/511535

Lotter, W., Kreiman, G., and Cox, D. (2020). A neural network trained for prediction mimics diverse features of biological neurons and perception. *Nat. Mach. Intell.* 2, 210–219. doi: 10.1038/s42256-020-0170-9

McClure, P., and Kriegeskorte, N. (2016). Representational distance learning for deep neural networks. *Front. Comput. Neurosci.* 10:131. doi: 10.3389/fncom.2016.00131

Millman, K. J., and Brett, M. (2007). Analysis of functional magnetic resonance imaging in python. *Comput. Sci. Eng.* 9, 52–55. doi: 10.1109/MCSE.2007.46

Montavon, G., Samek, W., and Müller, K. R. (2018). Methods for interpreting and understanding deep neural networks. *Digit. Signal Process. A Rev. J.* 73, 1–15. doi: 10.1016/j.dsp.2017.10.011

Morcos, A. S., Barrett, D. G. T., Rabinowitz, N. C., and Botvinick, M. (2018). *"On the importance of single directions for generalization*," in *International Conference on Learning Representations* (Vancouver, Canada).

Naselaris, T., Kay, K. N., Nishimoto, S., and Gallant, J. L. (2011). Encoding and decoding in fMRI. *Neuroimage* 56, 400–410. doi: 10.1016/j.neuroimage.2010.07.073

Nasr, K., Viswanathan, P., and Nieder, A. (2019). Number detectors spontaneously emerge in a deep neural network designed for visual object recognition. *Sci. Adv.* 5:eaav7903. doi: 10.1126/sciadv.aav7903

Nguyen, A., Dosovitskiy, A., Yosinski, J., Brox, T., and Clune, J. (2016). "Synthesizing the preferred inputs for neurons in neural networks via deep generator networks," in *Advances in Neural Information Processing Systems*, eds D. D. Lee, U. V. Luxburg, R. Garnett, M. Sugiyama, and I. Guyon (Barcelona: Curran Associates Inc.), 3395–3403.

Nguyen, A., Yosinski, J., and Clune, J. (2019). "Understanding neural networks via feature visualization: a survey," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, eds W. Samek, G. Montavon, A. Vedaldi, L. Hansen, and K. R. Müller (Cham: Springer), 55–76. doi: 10.1007/978-3-030-28954-6_4

Niso, G., Gorgolewski, K. J., Bock, E., Brooks, T. L., Flandin, G., Gramfort, A., et al. (2018). MEG-BIDS, the brain imaging data structure extended to magnetoencephalography. *Sci. Data* 5:180110. doi: 10.1038/sdata.2018.110

O'Connell, T. P., and Chun, M. M. (2018). Predicting eye movement patterns from fMRI responses to natural scenes. *Nat. Commun.* 9:5159. doi: 10.1038/s41467-018-07471-9

Pernet, C. R., Appelhoff, S., Gorgolewski, K. J., Flandin, G., Phillips, C., Delorme, A., et al. (2019). EEG-BIDS, an extension to the brain imaging data structure for electroencephalography. *Sci. Data* 6:103. doi: 10.1038/s41597-019-0104-8

Ponce, C. R., Xiao, W., Schade, P. F., Hartmann, T. S., Kreiman, G., and Livingstone, M. S. (2019). Evolving images for visual neurons using a deep generative network reveals coding principles and neuronal preferences. *Cell 177*, 999–1009.e10. doi: 10.1016/j.cell.2019.04.005

Pospisil, D. A., Pasupathy, A., and Bair, W. (2018). 'Artiphysiology' reveals V4-like shape tuning in a deep network trained for image classification. *Elife* 7:e38242. doi: 10.7554/eLife.38242

Rajalingham, R., Issa, E. B., Bashivan, P., Kar, K., Schmidt, K., and DiCarlo, J. J. (2018). Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *J. Neurosci.* 38, 7255–7269. doi: 10.1523/JNEUROSCI.0388-18.2018

Rawat, W., and Wang, Z. (2017). Deep convolutional neural networks for image classification: a comprehensive review. *Neural Comput.* 29, 2352–2449. doi: 10.1162/neco_a_00990

Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., et al. (2019). A deep learning framework for neuroscience. *Nat. Neurosci.* 22, 1761–1770. doi: 10.1038/s41593-019-0520-2

Ritter, S., Barrett, D. G. T., Santoro, A., and Botvinick, M. M. (2017). "Cognitive psychology for deep neural networks: a shape bias case study," in *International Conference on Machine Learning,* (Sydney, NSW), 2940–2949.

Sainath, T. N., Mohamed, A. R., Kingsbury, B., and Ramabhadran, B. (2013). "Deep convolutional neural networks for LVCSR," in *IEEE International Conference on Acoustics, Speech and Signal Processing* (Vancouver, BC: IEEE), 8614–8618. doi: 10.1109/ICASSP.2013.6639347

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., et al. (2018). Brain-score: Which artificial neural network for object recognition is most brain-like? *bioRxiv*. doi: 10.1101/407007

Serre, T. (2019). Deep learning: the good, the bad, and the ugly. *Annu. Rev. Vis. Sci.* 5, 399–426. doi: 10.1146/annurev-vision-091718-014951

Shen, G., Horikawa, T., Majima, K., and Kamitani, Y. (2019). Deep image reconstruction from human brain activity. *PLoS Comput. Biol.* 15:e1006633. doi: 10.1371/journal.pcbi.1006633

Simonyan, K., Vedaldi, A., and Zisserman, A. (2014). *"Deep inside convolutional networks: Visualising image classification models and saliency maps*," in *International Conference on Learning Representations* (Banff, Canada).

Simonyan, K., and Zisserman, A. (2015). *"Very deep convolutional networks for large-scale image recognition*," in *International Conference on Learning* Representations (San Diego, CA).

Springenberg, J. T., Dosovitskiy, A., Brox, T., and Riedmiller, M. (2015). *"Striving for simplicity: The all convolutional net*," in *International Conference on Learning Representations* (San Diego, CA).

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., et al. (2015). "Going deeper with convolutions," in *IEEE Conference on Computer Vision and Pattern Recognition* (Boston, MA), 1–9. doi: 10.1109/CVPR.2015.7298594

Tran, D., Wang, H., Torresani, L., Ray, J., Lecun, Y., and Paluri, M. (2018). "A closer look at spatiotemporal convolutions for action recognition," in *IEEE*

Conference on Computer Vision and Pattern Recognition, (Salt Lake City, *UT*: IEEE*)*, 6450–6459. doi: 10.1109/CVPR.2018.00675

Van Essen, D. C., Drury, H. A., Joshi, S., and Miller, M. I. (1998). Functional and structural mapping of human cerebral cortex: solutions are in the surfaces. *Proc. Natl. Acad. Sci. U. S. A.* 95, 788–795. doi: 10.1073/pnas.95.3.788

VanRullen, R., and Reddy, L. (2019). Reconstructing faces from fMRI patterns using deep generative neural networks. *Commun. Biol.* 2:193. doi: 10.1038/s42003-019-0438-y

Watanabe, E., Kitaoka, A., Sakamoto, K., Yasugi, M., and Tanaka, K. (2018). Illusory motion reproduced by deep neural networks trained for prediction. *Front. Psychol.* 9:345. doi: 10.3389/fpsyg.2018.00345

Wen, H., Shi, J., Zhang, Y., Lu, K., Cao, J., and Liu, Z. (2018). Neural encoding and decoding with deep learning for dynamic natural vision. *Cereb. Cortex* 28, 4136–4160. doi: 10.1093/cercor/bhx268

Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., and Torralba, A. (2010). "SUN database: large-scale scene recognition from abbey to zoo," in *IEEE Conference on Computer Vision and Pattern Recognition*, (IEEE: San Francisco, *CA)*, 3485–3492. doi: 10.1109/CVPR.2010.5539970

Yamins, D. L. K., and DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.* 19, 356–365. doi: 10.1038/nn.4244

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., and DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci. U.S.A.* 111, 8619–8624. doi: 10.1073/pnas.1403112111

Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., and Lipson, H. (2015). Understanding neural networks through deep visualization. *arXiv [preprint]* arXiv:1412.6856.

Zeiler, M. D., and Fergus, R. (2014). "Visualizing and understanding convolutional networks," in *European Conference on Computer Vision* (Zurich), 818–833. doi: 10.1007/978-3-319-10590-1_53

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2014). Object detectors emerge in deep scene CNNs. *arXiv [preprint]* arXiv:1412.6856.

Zhou, B., Sun, Y., Bau, D., and Torralba, A. (2018). Revisiting the importance of individual units in CNNs via ablation. *arXiv [preprint]* arXiv:1806.02891.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Computational Neuroscience

ORIGINAL RESEARCH
published: 09 December 2020
doi: 10.3389/fncom.2020.578158

Hierarchical Sparse Coding of Objects in Deep Convolutional Neural Networks

Xingyu Liu[1], *Zonglei Zhen*[1]** and *Jia Liu*[2]*

[1] Beijing Key Laboratory of Applied Experimental Psychology, Faculty of Psychology, Beijing Normal University, Beijing, China,
[2] Department of Psychology & Tsinghua Laboratory of Brain and Intelligence, Tsinghua University, Beijing, China

Recently, deep convolutional neural networks (DCNNs) have attained human-level performances on challenging object recognition tasks owing to their complex internal representation. However, it remains unclear how objects are represented in DCNNs with an overwhelming number of features and non-linear operations. In parallel, the same question has been extensively studied in primates' brain, and three types of coding schemes have been found: one object is coded by the entire neuronal population (distributed coding), or by one single neuron (local coding), or by a subset of neuronal population (sparse coding). Here we asked whether DCNNs adopted any of these coding schemes to represent objects. Specifically, we used the population sparseness index, which is widely-used in neurophysiological studies on primates' brain, to characterize the degree of sparseness at each layer in representative DCNNs pretrained for object categorization. We found that the sparse coding scheme was adopted at all layers of the DCNNs, and the degree of sparseness increased along the hierarchy. That is, the coding scheme shifted from distributed-like coding at lower layers to local-like coding at higher layers. Further, the degree of sparseness was positively correlated with DCNNs' performance in object categorization, suggesting that the coding scheme was related to behavioral performance. Finally, with the lesion approach, we demonstrated that both external learning experiences and built-in gating operations were necessary to construct such a hierarchical coding scheme. In sum, our study provides direct evidence that DCNNs adopted a hierarchically-evolved sparse coding scheme as the biological brain does, suggesting the possibility of an implementation-independent principle underling object recognition.

Keywords: deep convolutional neural network, sparse coding, coding scheme, object recognition, object representation, hierarchy

OPEN ACCESS

Edited by:
Matjaž Perc,
University of Maribor, Slovenia

Reviewed by:
Tai Sing Lee,
Carnegie Mellon University,
United States
Laurent U. Perrinet,
UMR7289 Institut de Neurosciences
de la Timone (INT), France

****Correspondence:***
Zonglei Zhen
zhenzonglei@bnu.edu.cn
Jia Liu
liujiaTHU@tsinghua.edu.cn

Received: *30 June 2020*
Accepted: *17 November 2020*
Published: *09 December 2020*

Citation:
Liu X, Zhen Z and Liu J (2020)
Hierarchical Sparse Coding of Objects
in Deep Convolutional Neural
Networks.
Front. Comput. Neurosci. 14:578158.
doi: 10.3389/fncom.2020.578158

INTRODUCTION

One spectacular achievement of human vision is that we can accurately recognize objects at a fraction of a second in the complex visual world (Thorpe et al., 1996). In recent years, deep convolutional neural networks (DCNNs) have achieved human-level performances in object recognition tasks (He et al., 2015; Simonyan and Zisserman, 2015; Szegedy et al., 2015). The success is primarily credited to the architecture that generic DCNNs compose of a stack of convolutional layers and fully-connected layers, each of which has multiple units with different filters (i.e., "neurons" in DCNNs), similar to the hierarchical organization of primates' ventral visual stream. With such hierarchical architecture and supervised learning on a large number of object exemplars, DCNNs are thought to construct complex internal representations for external objects. However, little is known about how exactly objects are represented in DCNNs.

This question has already puzzled neuroscientists for a long time. To understand how primates' visual system encodes the external world, three types of coding schemes are proposed to describe how neurons are integrated together to represent an object. At one extreme is distributed coding, by which the whole neuronal population is involved, whereas at the other extreme is local coding, by which one neuron is designated to represent one object. The distributed coding scheme is superior in large coding capacity, easy generalization, and high robustness, while the local coding scheme is good at information compression, energy conservation and better interpretability. In between lies the sparse coding that different subsets of neurons in the population participate in coding different objects. As a trade-off, sparse coding possesses advantages of both local coding and distrusted coding (Barlow, 1972; Thorpe, 1989; Berkes et al., 2009; Rolls, 2017; Thomas and French, 2017; Beyeler et al., 2019). Neurophysiological studies have revealed that the sparse coding scheme is adopted in some areas in primate visual cortex for object recognition (Olshausen and Field, 1996; Lehky et al., 2011; Barth and Poulet, 2012; Rolls, 2017).

Following the studies on biological intelligent systems, several pioneer studies started to characterize DCNNs' representation with coding scheme (Szegedy et al., 2013; Agrawal et al., 2014; Li et al., 2016; Wang et al., 2016; Morcos et al., 2018; Casper et al., 2019; Parde et al., 2020). Studies using the ablation approach show that the processing of objects usually requires the participation of multiple units, but only 10–15% of units in a layer are actually needed to achieve 90% of the full performance (Agrawal et al., 2014). Even when half of the units in all layers are ablated, the performance does not decrease significantly with the accuracy above 90% of the full performance (Morcos et al., 2018). Further studies quantify the number of non-zero units in response to objects and report a trend of decrease in the number of non-zero units along the hierarchy of DCNNs (Agrawal et al., 2014). These preliminary results suggest that DCNNs may adopt the sparse coding scheme, which likely evolves along hierarchy.

Here, we adopted a prevalent metric in neurophysiological studies on primates' brain, population sparseness index (PSI, Rolls and Tovee, 1995; Vinje and Gallant, 2000), to quantify the population sparseness along the hierarchy of two representative DCNNs, AlexNet (Krizhevsky, 2014) and VGG11 (Simonyan and Zisserman, 2015). Specifically, we first systematically evaluated the layer-wise sparseness in representing objects. Then, we characterized the functionality of sparseness by examining the relationship between sparseness and behavioral performance in each layer. Finally, we explored factors that may influence the coding scheme.

MATERIALS AND METHODS

Visual Images Datasets

ImageNet Dataset

The dataset from ImageNet Large Scale Visual Recognition Challenge 2012 (ILSVRC2012) (Russakovsky et al., 2015) contains 1,000 categories that are organized according to the hierarchy of WordNet (Miller, 1995). The 1,000 object categories consist of both internal nodes and leaf nodes of WordNet, but do not overlap with each other. The dataset contains 1.2 million images for model training, 50,000 images for model validation and 100,000 images for model test. In the present study, only the validation dataset (i.e., 1,000 categories × 50 images) was used to evaluate the coding scheme of DCNNs.

Caltech256 Dataset

The Caltech256 dataset consists of 30,607 images from 256 object categories with a minimum number of 80 images per category (Griffin et al., 2007). In the present study, 80 images per category were randomly chosen from the original dataset.

DCNNs and Activation Extraction

The well-known AlexNet and VGG11 that are pretrained for object classification were selected to explore the coding scheme of DCNNs. Besides the two pretrained models, corresponding weight-permuted models and ReLU-deactivated models were also examined to investigate the factors that may influence the coding scheme observed in the pretrained models.

Pretrained Models

AlexNet and VGG11 are pretrained on ILSVRC2012 dataset and were downloaded from PyTorch model Zoo[1]. Both DCNNs are purely feedforward: the input to each layer consists solely of the output from the previous layer. The AlexNet consists of 5 convolutional layers (Conv1 through Conv5) that contain a set of feature maps with linear spatial filters, and 3 fully-connected layers (FC1 through FC3). In between, a max (x, 0) rectifying non-linear unit (ReLU) is applied to all units after each convolutional and FC layer. In some convolutional layers, ReLU is followed by another max-pooling sublayer. VGG11 is similar to AlexNet in architecture except for two primary differences. First, VGG11 uses smaller receptive fields (3 × 3 with a stride of 1) than AlexNet (11 × 11 with a stride of 4). Second, VGG11 has more layers (8 convolutional layers) than AlexNet. When we refer to Conv#, we mean the outputs from the ReLU sublayer in the #th convolutional layer. Similarly, FC# means the outputs from the #th FC layer after ReLU. The DNNBrain toolbox[2] was used to extract the DCNN activation (Chen et al., 2020). For each unit (or channel), the activation map was averaged to produce a unit-wise (or channel-wise) activation for each exemplar, and the activation of the unit to an object category was then derived by averaging the unit-wise responses from all exemplars of the category.

[1] https://pytorch.org/

[2] https://github.com/BNUCNL/dnnbrain/

Weight-Permuted Models and Bias-Permuted Models

The weight-permuted models were derived by permuting weights of the pretrained models within each layer. That is, the structures of the original networks and the weight distribution of each layer were preserved while the exact feature filters obtained from the learning of the supervised task were disrupted. Weights in a given layer can be decomposed as channel x kernel, in which kernels are 3-D tensors (i.e., input channel x height x width). Three kinds of permutation strategies with various scales were performed: weights were permutated across all channels and kernels, across channels with all kernels intact, and across kernels with channel orders unaltered. The bias-permuted models were obtained by permuting biases in each layer with all weights and the network structure remaining unchanged.

ReLU-Deactivated Models

The ReLU-deactivated model was the same as the pretrained models with only ReLU being silenced in all layers by replacing it with an identity mapping. The ReLU-deactivated model disabled the non-linear operation after the feature extraction but still retained the same network architectures and the learned feature filters.

Population Sparseness Index

The PSI was calculated for each layer of DCNNs to quantify the peakedness of the distribution of population responses elicited by an object category, which is equivalent to the fraction of the units in the population that participated in coding objects in the case of binary responses (Vinje and Gallant, 2000).

$$\text{PSI} = \frac{1-a}{1-\frac{1}{N_u}}, \text{where } a = \frac{\left(\left(\sum r_u\right) / N_u\right)^2}{\sum \left({r_u}^2 / N_u\right)},$$

where r_u is the unit-wise activation of a unit u from a target layer in response to an object category, and N_u is the number of units in that layer. The unit-wise activation was z-scored across all categories for each unit, and then normalized across all units into a range from 0 to 1 to rescale the negative values to non-negative as required by the definition of PSI. Values of PSI near 0 indicate low sparseness that all units respond equally to the object category, and values near 1 indicate high sparseness that only a few units respond to the category.

Relationship Between Population Sparseness and Classification Performance

The relationship between sparseness and classification performance was first explored using correlation analyses. The Caltech256 classification task was used to estimate the classification performance of AlexNet and VGG11 on each category. Specifically, a logistic regression model was constructed using activation patterns from FC2 as features to perform a 256-class object classification. A 2-fold cross-validation procedure was used to evaluate the classification performance. Then, Pearson correlation coefficients between the PSI and the classification performance were calculated across all categories for each layer, respectively. Finally, to reveal how the sparse coding from different layers contribute to the classification performance, a stepwise multiple regression was conducted with the classification performance of each category as dependent variables and the PSI of the corresponding category from all layers as independent variables. The regressions were conducted for Conv layers and FC layers separately.

RESULTS

The coding scheme for object categorization in DCNN was characterized layer by layer in the pretrained AlexNet and VGG11 using PSI. The PSI was first evaluated on the ImageNet validation dataset, with the same categories on which these two DCNNs were trained. Similar findings were revealed in the two DCNNs. First, the values of the PSI were low for all object categories in all layers in general (median <0.4), with the maximum values no larger than 0.6 (**Figure 1**), suggesting that the sparse coding scheme was broadly adopted in all layers of the DCNNs to represent objects. Second, in each layer, the PSI of all categories exhibited a broad distribution (ranges >0.2), indicating great individual differences in sparseness among object categories. However, despite the large amount of inter-category differences, the median PSI of each layer showed a trend of increase along the hierarchy in both Conv and FC layers, respectively (AlexNet: Kendall's tau $= 0.40$, $p < 0.001$; VGG11: Kendall's tau $= 0.36$, $p < 0.001$). A similar result was found with the absolute value of activation before computing the PSI (AlexNet: Kendall's tau $= -0.44$, $p < 0.001$.; VGG11: Kendall's tau $= -0.52$, $p < 0.001$). Corroborative results were also observed by fitting the activation distribution of the neuron population with Norm distribution and Weibull functions (**Supplementary Figure 1**). Note that the increase in sparseness was not strictly monotonic, as the PSI of the first layer was slightly higher than the adjacent ones. More interestingly, although AlexNet and VGG11 have different numbers of Conv layers, the major increase occurred at the last Conv layer. Similar results have also been found in DCNNs (i.e., ResNet152 and GoogLeNet) whose architectures are significantly different from AlexNet and VGG11, suggesting that the hierarchical sparse coding scheme may be a general coding strategy in DCNNs (**Supplementary Figure 2**).

We replicated this finding with a new dataset, Caltech256, that is dissimilar to the ImageNet in object categories and is thus not in the training dataset. We found a similar pattern of the increase in sparseness along the hierarchy (AlexNet: Kendall's tau $= 0.35$, $p < 0.001$; VGG11: Kendall's tau $= 0.25$, $p < 0.001$; **Supplementary Figure 3**), suggesting that the increase in sparseness did not result from image dataset. Taken together, the hierarchically-increased sparseness suggested that there was a systematic shift from the distributed-like coding scheme in low layers to the local-like coding scheme in high layers.

Next, we examined the functionality of the sparse coding scheme observed in the DCNNs. To address this question, we tested the association between the population sparseness and the behavioral performance by performing correlation analyses within each layer of the DCNNs. In AlexNet, significant

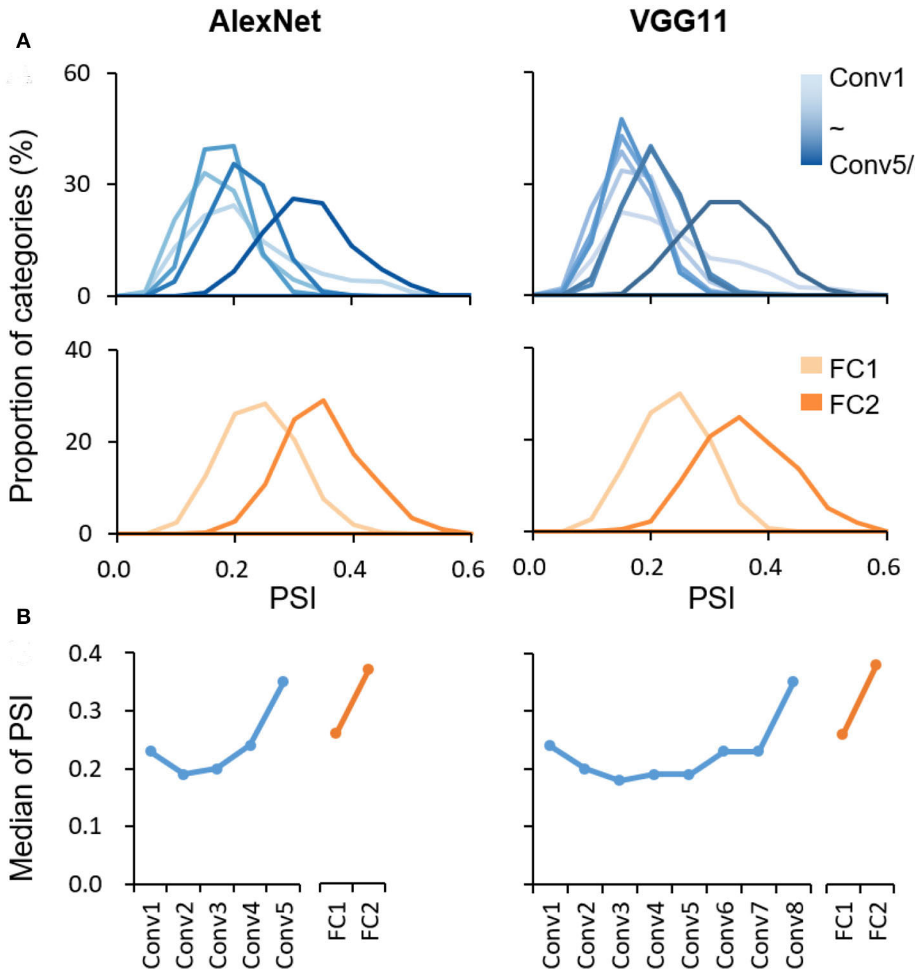


FIGURE 1 | Hierarchically sparse coding for object categories in DCNNs. **(A)** Layer-wise PSI distribution for object categories in DCNNs. The sparseness was evaluated using the PSI for each object category from the ImageNet dataset (1,000 categories) in each layer separately. The distribution of PSI right-shifted along hierarchy in general. X axis: the degree of sparseness, with higher PSI indicating a higher degree of sparseness; Y axis: the proportion of categories with a corresponding PSI value. **(B)** Median of PSI for each layer. In general, the median of PSI increased along hierarchy in Conv and FC layers, respectively. X axis: the name of layers along hierarchy; Y axis: the median of PSI.

correlations were found starting from Conv4 and beyond [*rs* (254) $>$ 0.19, *ps* $<$ 0.001, Bonferroni corrected; **Figure 2A**]. This result suggested that the degree of sparseness in coding object categories was predictive of performance accuracy. That is, the sparser an object category was represented, the better it was recognized and classified. Importantly, the correlation coefficients also increased along hierarchy (Kendall's tau = 0.90, $p = 0.003$), with the highest correlation coefficient observed at Conv5 (0.43) and FC2 (0.69), respectively (**Figure 2A**). This trend suggests a closer relationship between the population sparseness and the behavioral performance in higher layers. Indeed, with a stepwise multiple regression analysis in which PSI of all Conv/FC layers of certain categories were the independent variables and classification performance was the dependent variable, we confirmed that population sparseness was predictive of behavioral performance [Conv layers: $F_{(3,\,252)} = 22.54$, $p < 0.001$, adjusted $R^2 = 0.2$; FC layers: $F_{(2,\,253)} = 136.60$, $p < 0.001$, adjusted $R^2 = 0.52$]. Meanwhile, only PSI in higher layers starting from Conv3 remained in the regression models, further confirming that the coding scheme as a characteristic of representation became more essential with the increasing hierarchical level. Similar results were also found in VGG11 (**Figure 2B**), suggesting that the association between sparseness and performance may be universal in DCNNs.

Finally, we explored the factors that may affect the formation of such a hierarchical coding scheme in the DCNNs. The DCNNs consist of two subprocesses at the core of each layer (**Figure 3A**): one is the feature extraction process whose weights and biases are dynamically adjusted during learning, and the other is a gating process with a fixed non-linear function (i.e., ReLU) that silences units with negative activities. To examine whether the hierarchically-increased sparseness was constructed through learning, we measured the population sparseness of DCNNs with either the learned weights or biases randomly

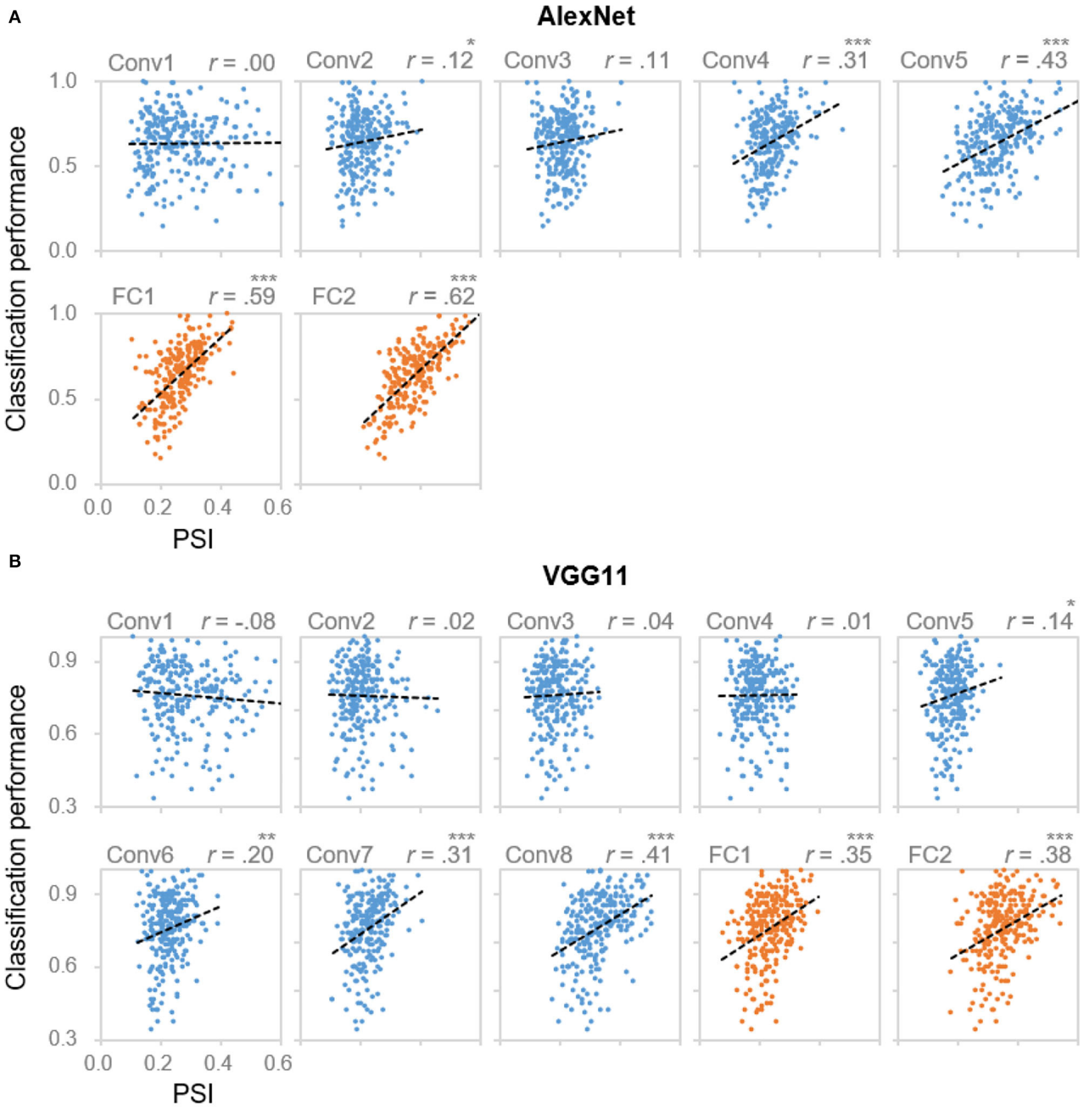


FIGURE 2 | Correlation between coding sparseness and behavioral performance. Layer-wise scatter plots of DCNNs' classification performance vs. PSI values from **(A)** AlexNet and **(B)** VGG11 for object categories from Caltech256. X axis: PSI value, the larger the value the sparser the coding; Y axis: DCNNs' classification performance for each object category. Each dot represents one category. *denotes $p < 0.05$, **denotes $p < 0.01$ and ***denotes $p < 0.001$. Categories with the best or the worst classification performances were listed in **Supplementary Figure 4**.

permuted. In the weight-permuted models where the weights were layer-wise permuted across all channels and kernels of the pretrained networks, we found that the degree of sparseness instead decreased along hierarchy (AlexNet: Kendall's tau = −0.53, $p < 0.001$; VGG11: Kendall's tau = −0.82, $p < 0.001$; **Figure 3B**), which was contradictory to the finding of the undisrupted one (**Figure 1**). This result was replicated when the weight permutation was performed across channels or kernels separately (AlexNet and VGG11: Kendall's taus <-0.53, *ps* < 0.001). Meanwhile, the population sparseness of the bias-permuted models in which all weights remained intact were also evaluated. We found that there was no increase in sparseness along hierarchy (AlexNet: Kendall's tau = 0.10, $p = 0.22$; VGG11: Kendall's tau = −0.15, $p = 0.03$; **Figure 3D**). In addition, when

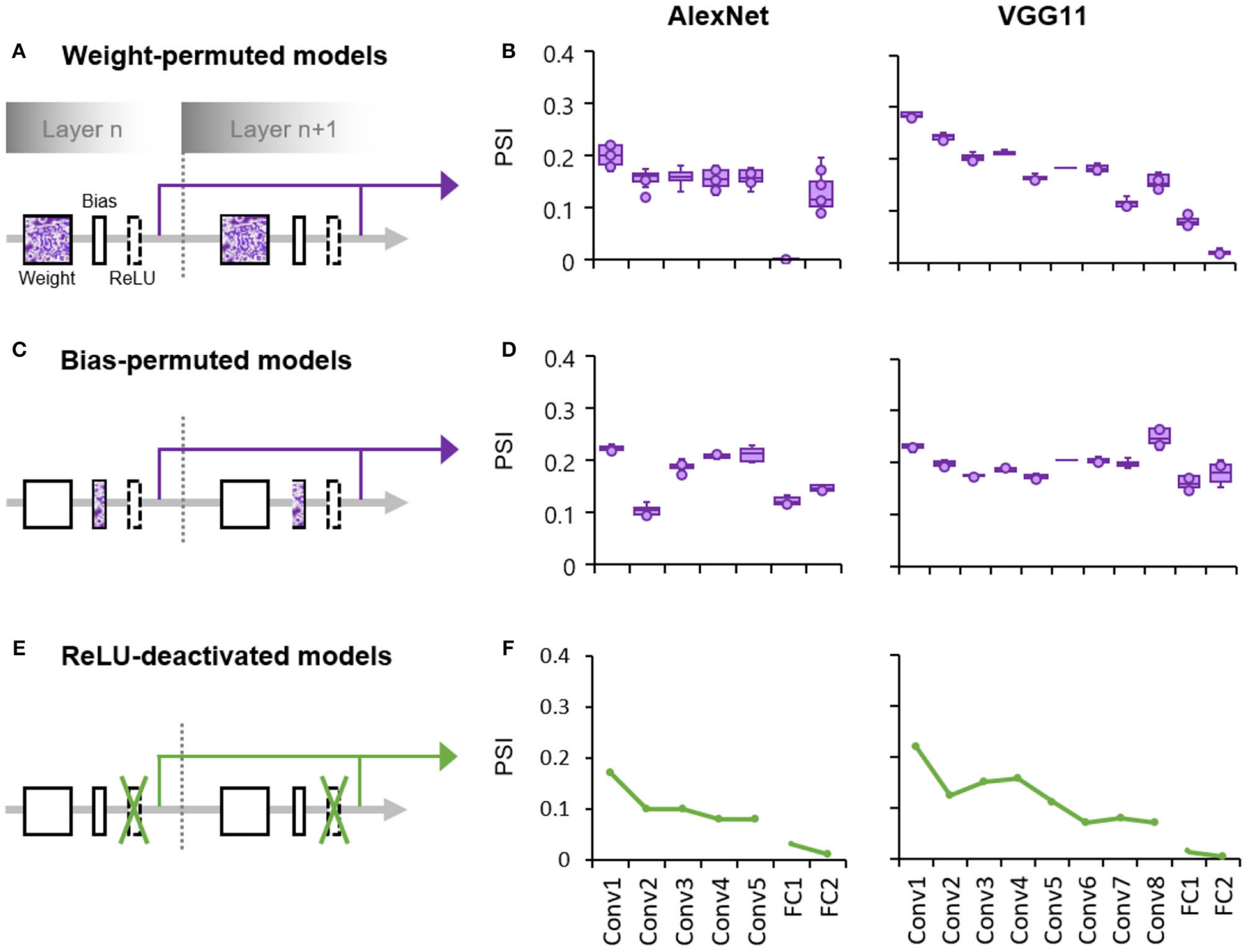


FIGURE 3 | Both the learning process and the gating process play an important role in the formation of the hierarchically-evolved coding scheme in the DCNNs. **(A)** A schematic diagram of the weight-permuted models. **(B)** Box plots of median PSI for objects across layers in the weight-permuted models, which represent the minimum, maximum, median, first quartile and third quartile of the distribution of the median PSI values. The PSI was measured in 10 permuted models using the same procedure as the intact one. **(C)** A schematic diagram of the bias-permuted models. **(D)** Box plots of median PSI for objects across layers in the bias-permuted models. **(E)** A schematic diagram of the ReLU-deactivated models. **(F)** Median PSI for objects across layers in the ReLU-deactivated models. X axis: the name of layers along hierarchy; Y axis: the median of PSI.

the ReLU sublayers were deactivated with the feature extraction sublayers intact (**Figure 3E**), we also observed a decreasing tendency of sparseness along the hierarchy (AlexNet: Kendall's tau = −0.21, $p < 0.001$; VGG11: Kendall's tau = −0.32, $p < 0.001$, **Figure 3F**), again in contrast to the AlexNet with functioning ReLU (**Figure 1**). Similar results were also found in VGG11, suggesting a general effect of learning and gating on the formation of the hierarchically-evolved coding scheme in DCNN.

DISCUSSION

In the present study, we systematically characterized the coding scheme in representing object categories at each layer of two typical DCNNs, AlexNet, and VGG11. We found that objects were in general sparsely encoded in the DCNNs, and the degree of sparseness increased along the hierarchy. Importantly, the hierarchically-evolved sparseness was able to predict the classification performance of the DCNNs, revealing the functionality of the sparse coding. Finally, lesion analyses of the weight-permuted models, the bias-permuted models, and the ReLU-deactivated models suggest that the learning experience and the built-in gating operation account for the hierarchically sparse coding scheme in the DCNNs. In short, our study provided one of the empirical evidence illustrating how object categories were represented in DCNNs for object recognition.

The finding that the degree of sparseness increased along the hierarchy in DCNNs is consistent with previous studies on DCNNs (Szegedy et al., 2013; Agrawal et al., 2014; Tripp, 2016; Wang et al., 2016; Morcos et al., 2018; Casper et al., 2019; Parde et al., 2020). Our study further extended these previous studies by conducting a layer-wise analysis throughout all hierarchical levels and calculating the degree of sparseness based on responses of the entire population of units ("neurons" in DCNN). Besides, our study tested two datasets of more than

1,000 object categories, and thus provided more comprehensive coverage of the object space. Finally, we also examined the functionality of sparse coding by showing that the sparser an object category was encoded, the higher accuracy of the object category was correctly recognized.

The fact that the hierarchically-increased coding sparseness coincides with a hierarchically-higher behavioral relevance in DCNNs suggests it as an organizing principle of representing a myriad of objects efficiently. That is, at the lower level of vision, representations recruit a larger number of generic neurons to process myriad natural objects with high fidelity. At the higher level, objects are decomposed into abstract features in the object space; therefore, only a smaller but highly-specialized group of neurons are recruited to construct the representation. Critically, a higher degree of sparseness makes representations more interpretable, because only at higher layers the degree of sparseness was able to read out for behavioral performance. One possibility is that distributed coding adopts more neuronal crosstalk that is difficult for readout, whereas sparser coding contains fewer higher-order relations and hence require less amount of computation for object recognition and memory storage/retrieval (Field, 1994; Froudarakis et al., 2014; Beyeler et al., 2019). That is, distributed coding is better at adapting and generalizing the variance across stimulus exemplars; sparse coding serves to explicit interpretation for goal-directed invariance (Földiák, 2009; Babadi and Sompolinsky, 2014; King et al., 2019). Taken together, the evolution of sparseness along the hierarchy likely mirrored the stages of objects being processed and the transformation of representation from stimulus-fidelity to goal-fidelity.

Interestingly, the sparseness was not accumulated gradually layer by layer. Instead, the sparseness was the highest at the last convolutional layer (i.e., Conv5 in AlexNet and Conv8 in VGG11) and fully-connected layer (i.e., FC2 in AlexNet and VGG11), much higher than that of their preceding ones regardless of the total number of layers in the DCNNs. This observation suggests a mechanism that the degree of sparseness dramatically increases at transitional layers either to the next processing stage (from Conv layers to FC layers) or to the generation of behavioral performance (from FC layers to the output layer). Further studies are needed to explore the functionality of the dramatic increase in sparseness. Note that the finding that the increase of sparseness was observed in two structurally-similar DCNNs (i.e., AlexNet and VGG11), and therefore it may not be applicable to other DCNNs.

As an intelligent system, DCNNs are products of the predesigned architecture by nature and learned features by nurture. Our lesion study revealed that both architecture and learning were critical for the formation of the hierarchically sparse coding scheme. As for the innate architecture, a critical built-in function is the non-linear gating sublayer, ReLU, that silences neurons with negative activity (Glorot et al., 2011; LeCun et al., 2015). Obviously, the gating function is bound to increase the sparseness of coding because it removes weak or irrelevant activations and thus leads objects to be represented by a smaller number of units. Our study confirmed this intuition by showing the disruption of hierarchically-increased sparseness when the gating function being disabled. Besides the commonly used gating operation ReLU, recently more approaches have been developed to directly serve the same purpose of sparsification (Liu et al., 2015; Kepner et al., 2018). On the other hand, the gating function was not sufficient for a proper sparse coding scheme, because after randomly permuting the weights of the learned filters in the feature sublayers, the sparseness was no longer properly constructed either. Further, the dependence of both external learning experiences and built-in non-linear operations implies that the sparse coding scheme may be also adopted in biological brains, because the gating function is the fundamental function of neurons (Lucas, 1909; Adrian, 1914) and the deprivation of visual experiences leads to deficits in a variety of visual functions (Wiesel and Hubel, 1963; Fine et al., 2003; Duffy and Livingstone, 2005). In short, the current study provides direct empirical evidence on the functionality and formation of hierarchy-dependent coding sparseness in DCNNs; However, the exact computational mechanisms underlying the evolution of sparse coding along hierarchy are needed for future work to unravel it.

Our findings with biologically-inspired DCNNs also lend insight into coding schemes in biological systems. Because the number of object categories, neurons, and sampling sites are largely limited by neurophysiological techniques, availability of subjects and ethical issues, it is difficult to characterize population sparseness along the visual pathway (Baddeley et al., 1997; Vinje and Gallant, 2000; Tolhurst et al., 2009). Several studies measured the population sparseness on certain single regions in mouse, ferret or macaque brain (Berkes et al., 2009; Froudarakis et al., 2014; Tang et al., 2018), but with diverse experimental setups, the evolution of population sparseness across brain regions is unclear. Lenky et al. did record both a group of V1 and the Inferotemporal neurons and found that the population sparseness increased from the V1 to Inferotemporal cortex (Lehky et al., 2005, 2011). In contrast, DCNNs can be used to examine not only coding schemes of a large number of objects (>1,000 object categories in our study) but also the degree of the sparseness of all units in all layers; therefore, DCNNs may serve as a quick-and-dirty model to pry open how visual information is represented in biological systems.

In sum, our study on the coding scheme of object categories in DCNNs invites future studies to understand how in DCNN objects are recognized accurately in particular, and how intelligence emerges under the interaction of internal architecture and external learning experiences in general. On one hand, approaches and findings from neurophysiological and fMRI studies help to transpire the black-box of DCNNs and enlighten the design of more effective DCNNs. For example, our study suggests new algorithms for better performance by increasing sparseness effectively possibly through learning or gating function built in the network. On the other hand, in contrast to the fact that neurophysiological studies on non-human primates and fMRI studies on human are limited either by the coverage of brain areas or by the spatial resolution, both architecture and units' activation in DCNNs are transparent. Therefore, DCNNs likely provides a perfect model to pry open mechanisms of object recognition at both micro- and macro-levels, which helps to understand how biological intelligent systems work.

DATA AVAILABILITY STATEMENT

Datasets analyzed in the present article were from two public datasets: (1) ImageNet: http://www.image-net.org/; (2) Caltech256: http://www.vision.caltech.edu/Image_Datasets/Caltech256. All codes for activation extraction and analyses are available on https://github.com/xingyu-liu/coding_sparseness.

AUTHOR CONTRIBUTIONS

XL, ZZ, and JL conceived the study and wrote the manuscript. XL developed the code and performed the research. All authors contributed to the article and approved the submitted version.

FUNDING

This study was funded by the National Natural Science Foundation of China (Grant No. 31861143039 and 31771251), and the National Basic Research Program of China (Grant No. 2018YFC0810602), the National Key R&D Program of China (Grant No. 2019YFA0709503).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fncom.2020.578158/full#supplementary-material

REFERENCES

Adrian, E. D. (1914). The all-or-none principle in nerve. *J. Physiol.* 47, 460–474. doi: 10.1113/jphysiol.1914.sp001637

Agrawal, P., Girshick, R., and Malik, J. (2014). "Analyzing the performance of multilayer neural networks for object recognition," in *Computer Vision – ECCV 2014*, eds D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars (Cham: Springer International Publishing), 329–344. doi: 10.1007/978-3-319-10584-0_22

Babadi, B., and Sompolinsky, H. (2014). Sparseness and expansion in sensory representations. *Neuron* 83, 1213–1226. doi: 10.1016/j.neuron.2014.07.035

Baddeley, R., Abbott, L. F., Booth, M. C. A., Sengpiel, F., Freeman, T., Wakeman, E. A., et al. (1997). Responses of neurons in primary and inferior temporal visual cortices to natural scenes. *Proc. R. Soc. Lond. B* 264, 1775–1783. doi: 10.1098/rspb.1997.0246

Barlow, H. B. (1972). Single units and sensation: a neuron doctrine for perceptual psychology? *Perception* 1, 371–394. doi: 10.1068/p010371

Barth, A. L., and Poulet, J. F. A. (2012). Experimental evidence for sparse firing in the neocortex. *Trends Neurosci.* 35, 345–355. doi: 10.1016/j.tins.2012.03.008

Berkes, P., White, B., and Fiser, J. (2009). "No evidence for active sparsification in the visual cortex," in *Advances in Neural Information Processing Systems*, eds Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta (Red Hook, NY: Curran Associates, Inc.), 108–116. Available online at: http://papers.nips.cc/paper/3774-no-evidence-for-active-sparsification-in-the-visual-cortex.pdf (accessed November 18, 2019).

Beyeler, M., Rounds, E. L., Carlson, K. D., Dutt, N., and Krichmar, J. L. (2019). Neural correlates of sparse coding and dimensionality reduction. *PLoS Comput. Biol.* 15:e1006908. doi: 10.1371/journal.pcbi.1006908

Casper, S., Boix, X., D'Amario, V., Guo, L., Schrimpf, M., Vinken, K., et al. (2019). Removable and/or repeated units emerge in overparametrized deep neural networks. *arXiv:1912.04783 [cs, stat]*. Available online at: http://arxiv.org/abs/1912.04783 (accessed June 26, 2020).

Chen, X., Zhou, M., Gong, Z., Xu, W., Liu, X., Huang, T., et al. (2020). DNNBrain: A Unifying Toolbox for Mapping Deep Neural Networks and Brains. *Front. Comput. Neurosci.* 14:580632. doi: 10.3389/fncom.2020.580632

Duffy, K. R., and Livingstone, M. S. (2005). Loss of neurofilament labeling in the primary visual cortex of monocularly deprived monkeys. *Cerebral Cortex* 15, 1146–1154. doi: 10.1093/cercor/bhh214

Field, D. J. (1994). What is the goal of sensory coding? *Neural Comput.* 6, 559–601. doi: 10.1162/neco.1994.6.4.559

Fine, I., Wade, A. R., Brewer, A. A., May, M. G., Goodman, D. F., Boynton, G. M., et al. (2003). Long-term deprivation affects visual perception and cortex. *Nat. Neurosci.* 6, 915–916. doi: 10.1038/nn1102

Földiák, P. (2009). Neural coding: non-local but explicit and conceptual. *Curr. Biol.* 19, R904–R906. doi: 10.1016/j.cub.2009.08.020

Froudarakis, E., Berens, P., Ecker, A. S., Cotton, R. J., Sinz, F. H., Yatsenko, D., et al. (2014). Population code in mouse V1 facilitates readout of natural scenes through increased sparseness. *Nat. Neurosci.* 17, 851–857. doi: 10.1038/nn.3707

Glorot, X., Bordes, A., and Bengio, Y. (2011). "Deep sparse rectifier neural networks," in *International Conference on Artificial Intelligence and Statistics* (Fort Lauderdale, FL), 315–323.

Griffin, G., Holub, A., and Perona, P. (2007). *Caltech-256 Object Category Dataset*. California Institute of Technology. Available online at: https://resolver.caltech.edu/CaltechAUTHORS:CNS-TR-2007-001

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. in *2015 IEEE International Conference on Computer Vision (ICCV)* (Santiago), 1026–1034. doi: 10.1109/ICCV.2015.123

Kepner, J., Gadepally, V., Jananthan, H., Milechin, L., and Samsi, S. (2018). "Sparse deep neural network exact solutions," in *2018 IEEE High Performance extreme Computing Conference (HPEC)* (Waltham, MA), 1–8. doi: 10.1109/HPEC.2018.8547742

King, M. L., Groen, I. I. A., Steel, A., Kravitz, D. J., and Baker, C. I. (2019). Similarity judgments and cortical visual responses reflect different properties of object and scene categories in naturalistic images. *NeuroImage* 197, 368–382. doi: 10.1016/j.neuroimage.2019.04.079

Krizhevsky, A. (2014). One weird trick for parallelizing convolutional neural networks. *arXiv:1404.5997 [cs]*. Available online at: http://arxiv.org/abs/1404.5997 (accessed June 26, 2020).

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

Lehky, S. R., Kiani, R., Esteky, H., and Tanaka, K. (2011). Statistics of visual responses in primate inferotemporal cortex to object stimuli. *J. Neurophysiol.* 106, 1097–1117. doi: 10.1152/jn.00990.2010

Lehky, S. R., Sejnowski, T. J., and Desimone, R. (2005). Selectivity and sparseness in the responses of striate complex cells. *Vis. Res.* 45, 57–73. doi: 10.1016/j.visres.2004.07.021

Li, Y., Yosinski, J., Clune, J., Lipson, H., and Hopcroft, J. (2016). Convergent learning: do different neural networks learn the same representations? *arXiv:1511.07543 [cs]*. Available online at: http://arxiv.org/abs/1511.07543 (accessed March 11, 2020).

Liu, B., Wang, M., Foroosh, H., Tappen, M., and Penksy, M. (2015). "Sparse convolutional neural networks," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Boston, MA: IEEE), 806–814.

Lucas, K. (1909). The "all or none" contraction of the amphibian skeletal muscle fibre. *J. Physiol.* 38, 113–133. doi: 10.1113/jphysiol.1909.sp001298

Miller, G. A. (1995). WordNet: a lexical database for English. *Commun. ACM* 38, 39–41. doi: 10.1145/219717.219748

Morcos, A. S., Barrett, D. G. T., Rabinowitz, N. C., and Botvinick, M. (2018). On the importance of single directions for generalization. *arXiv:1803.06959 [cs, stat]*. Available online at: http://arxiv.org/abs/1803.06959 (accessed December 20, 2019).

Olshausen, B. A., and Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* 381, 607–609. doi: 10.1038/381607a0

Parde, C. J., Colón, Y. I., Hill, M. Q., Castillo, C. D., Dhar, P., and O'Toole, A. J. (2020). Single unit status in deep convolutional neural network codes for face identification: sparseness redefined. *arXiv:2002.06274 [cs]*. Available online at: http://arxiv.org/abs/2002.06274 (accessed March 19, 2020).

Rolls, E. T. (2017). Cortical coding. *Lang. Cogn. Neurosci.* 32, 316–329. doi: 10.1080/23273798.2016.1203443

Rolls, E. T., and Tovee, M. J. (1995). Sparseness of the neuronal representation of stimuli in the primate temporal visual cortex. *J. Neurophysiol.* 73, 713–726. doi: 10.1152/jn.1995.73.2.713

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al. (2015). ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.* 115, 211–252. doi: 10.1007/s11263-015-0816-y

Simonyan, K., and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556 [cs]*. Available online at: http://arxiv.org/abs/1409.1556 (accessed June 26, 2020).

Szegedy, C., Wei, L, Yangqing, J, Sermanet, P., Reed, S., Anguelov, D., et al. (2015). "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Boston, MA), 1–9. doi: 10.1109/CVPR.2015.7298594

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., et al. (2013). Intriguing properties of neural networks. *arXiv:1312.6199 [cs]*. Available online at: http://arxiv.org/abs/1312.6199 (accessed September 1, 2019).

Tang, S., Zhang, Y., Li, Z., Li, M., Liu, F., Jiang, H., et al. (2018). Large-scale two-photon imaging revealed super-sparse population codes in the V1 superficial layer of awake monkeys. *eLife* 7:e33370. doi: 10.7554/eLife.33370.015

Thomas, E., and French, R. (2017). Grandmother cells: much ado about nothing. *Lang. Cogn. Neurosci.* 32, 342–349. doi: 10.1080/23273798.2016.1235279

Thorpe, S. (1989). Local vs. distributed coding. *Intellectica* 8, 3–40. doi: 10.3406/intel.1989.873

Thorpe, S., Fize, D., and Marlot, C. (1996). Speed of processing in the human visual system. *Nature* 381, 520–522. doi: 10.1038/381520a0

Tolhurst, D. J., Smyth, D., and Thompson, I. D. (2009). The sparseness of neuronal responses in ferret primary visual cortex. *J. Neurosci.* 29, 2355–2370. doi: 10.1523/JNEUROSCI.3869-08.2009

Tripp, B. (2016). Similarities and differences between stimulus tuning in the inferotemporal visual cortex and convolutional networks. *arXiv:1612.06975 [q-bio]*. Available online at: http://arxiv.org/abs/1612.06975 (accessed March 19, 2020).

Vinje, W. E., and Gallant, J. L. (2000). Sparse coding and decorrelation in primary visual cortex during natural vision. *Science* 287, 1273–1276. doi: 10.1126/science.287.5456.1273

Wang, J., Zhang, Z., Xie, C., Premachandran, V., and Yuille, A. (2016). Unsupervised learning of object semantic parts from internal states of CNNs by population encoding. *arXiv:1511.06855 [cs]*. Available online at: http://arxiv.org/abs/1511.06855 (accessed June 26, 2020).

Wiesel, T. N., and Hubel, D. H. (1963). Effects of visual deprivation on morphology and physiology of cells in the cat's lateral geniculate body. *J. Neurophysiol.* 26, 978–993. doi: 10.1152/jn.1963.26.6.978

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Computational Neuroscience

ORIGINAL RESEARCH
published: 26 January 2021
doi: 10.3389/fncom.2020.601314

Implementation-Independent Representation for Deep Convolutional Neural Networks and Humans in Processing Faces

Yiying Song [1*†], *Yukun Qu* [2†], *Shan Xu* [1] *and Jia Liu* [3*]

[1] Beijing Key Laboratory of Applied Experimental Psychology, Faculty of Psychology, Beijing Normal University, Beijing, China, [2] State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University, Beijing, China, [3] Department of Psychology & Tsinghua Laboratory of Brain and Intelligence, Tsinghua University, Beijing, China

Deep convolutional neural networks (DCNN) nowadays can match human performance in challenging complex tasks, but it remains unknown whether DCNNs achieve human-like performance through human-like processes. Here we applied a reverse-correlation method to make explicit representations of DCNNs and humans when performing face gender classification. We found that humans and a typical DCNN, VGG-Face, used similar critical information for this task, which mainly resided at low spatial frequencies. Importantly, the prior task experience, which the VGG-Face was pre-trained to process faces at the subordinate level (i.e., identification) as humans do, seemed necessary for such representational similarity, because AlexNet, a DCNN pre-trained to process objects at the basic level (i.e., categorization), succeeded in gender classification but relied on a completely different representation. In sum, although DCNNs and humans rely on different sets of hardware to process faces, they can use a similar and implementation-independent representation to achieve the same computation goal.

Keywords: deep convolutional neural network, face recognition, reverse correlation analysis, face representation, visual intelligence

****Correspondence:***
Yiying Song
songyiying@bnu.edu.cn
Jia Liu
liujiaTHU@tsinghua.edu.cn

† These authors have contributed equally to this work

Received: *31 August 2020*
Accepted: *30 December 2020*
Published: *26 January 2021*

Citation:
Song Y, Qu Y, Xu S and Liu J (2021) Implementation-Independent Representation for Deep Convolutional Neural Networks and Humans in Processing Faces. Front. Comput. Neurosci. 14:601314. doi: 10.3389/fncom.2020.601314

INTRODUCTION

In recent years, deep convolutional neural networks (DCNN) have made dramatic progresses to achieve human-level performances in a variety of challenging complex tasks, especially visual tasks. For example, DCNNs trained to classify over a million natural images can match human performance on object categorization tasks (Krizhevsky, 2014; Simonyan and Zisserman, 2015; Krizhevsky et al., 2017), and DCNNs trained with large-scale face datasets can approach human-level performance in face recognition (Taigman et al., 2014; Parkhi et al., 2015; Schroff et al., 2015; Ranjan et al., 2017). However, these highly complex networks have remained largely opaque, whose internal operations are poorly understood. Specifically, it remains unknown whether DCNNs achieve human-like performance through human-like processes. That is, do DCNNs use similar computations and inner representations to perform tasks as humans do?

To address this question, here we applied a reverse correlation approach (Ahumada and Lovell, 1971; Gold et al., 2000; Mangini and Biederman, 2004; Martin-Malivel et al., 2006), which has been widely used in psychophysical studies to infer internal representations of human observers that transform inputs (e.g., stimuli) to outputs (e.g., behavior performance). This data-driven method allows an unbiased estimate of what is in observers' "mind" when performing a task, rather than manipulating specific features that researchers *a priori* hypothesize to be critical for the task. Here we applied this approach to both DCNNs and human observers to investigate whether the DCNNs and humans utilized similar representations to perform the task of face gender classification.

Specifically, a gender-neutral template face midway between the average male and the average female faces was superimposed with random noises, which rendered the template face more male-like in some trials or more female-like in other trials. The noisy faces were then submitted to human observers and the VGG-Face, a typical DCNN pre-trained for face identification (Parkhi et al., 2015). Based on the output of an observer that a noisy face was classified as a male but not as a female, for example, we reasoned that the noise superimposed on the template face contained features matching the observer's internal male prototype. Therefore, the difference between noise patterns of trials classified as male and those as female revealed the facial features diagnostic for gender classification, and provided an explicit and unbiased estimate of the representation used by the observer for gender classification. Finally, we directly compared the similarity of the inner representations of human observers and the VGG-Face obtained from identical stimuli and procedures, and examined the hypothesis that different intelligent information-processing systems may use similar representations to achieve the same computation goal (Marr, 1982).

RESULTS

The VGG-Face and Humans Utilized Similar Information for Gender Classification

We used the reverse correlation approach to reconstruct the inner representations used by the DCNN and human observers for gender classification. Specifically, both the DCNN and human observers were asked to classify noisy faces from a gender-neutral template face embedded with random sinusoid noises as male or female (**Figure 1A**).

For the DCNN, we first trained the VGG-Face to classify gender using transfer learning with 21,458 face images of 52 identities (35 males) from the VGGFace2 dataset (see Methods), and the test accuracy of gender classification of the new network achieved 98.6% (see **Supplementary Tables 1, 2** for more details). The gender-neutral template face was roughly equally classified as male and female by the VGG-Face (female: 54%). The noise patterns were constructed from 4,092 sinusoids at five spatial scales, six orientations, and two phases. We presented the template face embedded in 20,000 noise patterns to the VGG-Face, of which 11,736 (58.7%) images were classified as male and 8,264 (41.3%) images as female. The noise patterns from trials classified as male or female were averaged separately (**Figure 1B**), and the difference between the two average noise patterns yielded a "classification image" (CI) that makes explicit the information used by the VGG-Face for gender classification (**Figure 1C**). A visual inspection of the CI showed that regions around the eyes, nose, and mouth were of high contrast in the CI, indicating the critical regions employed by the VGG-Face to classify male from female faces.

Then, we reconstructed the representation used by human observers in a similar way. In our study, 16 human observers performed the gender classification task, each presented with 1,000 noisy faces. Altogether, 16,000 images were presented to the human observers, of which 7,969 (49.8%) images were classified as males and 8,031 (50.2%) images as females. Similarly, the CI for human observers was obtained (**Figure 1C**). Visual inspections of the CIs for the VGG-Face and human observers revealed good agreement between them, and Pearson's correlation between the two CIs was high ($r = 0.73$). This result suggested that the VGG-Face and human observers utilized similar information to classifying gender.

Further, we reconstructed inner male and female prototypes by adding or subtracting the rescaled CI to or from the template face for the VGG-Face and humans, respectively (**Figure 1C**). As expected, the male and female prototype faces are perceptually male-like and female-like, and highly similar between the VGG-Face and human observers.

The Shared Representation Was Mainly Based on Low Spatial-Frequency Information

Having found that the VGG-Face and human observers utilized similar information for gender classification, next we asked whether the VGG-Face and human observers employed similar information in all spatial frequencies. In our study, the noise patterns were constructed from sinusoid components of five scales of spatial frequencies (2, 4, 8, 16, and 32 cycles/image), which enabled us to reconstruct the CIs for each scale separately (**Figure 2**) and examined the similarity at each scale. We found that the similarity was the highest at low spatial frequencies ($r = 0.87$ and 0.76 at 2 and 4 cycles/images), and then decreased sharply at high spatial frequencies ($r = 0.25$, 0.19, 0.11 at 8, 16, and 32 cycles/image). Consequently, male and female prototypes reconstructed with the noise patterns at low spatial frequencies (2 and 4 cycles/image) were more similar between human observers and the VGG-Face than those at high spatial frequencies (8, 16, and 32 cycles/images) (**Supplementary Analysis 1**). Therefore, the shared representation for gender classification was mainly based on information at low spatial frequencies, consistent with previous findings that face gender processing relies heavily on low spatial frequencies (Sergent, 1986; Valentin et al., 1994; Goffaux et al., 2003b; Mangini and Biederman, 2004; Khalid et al., 2013).

To further quantify the contribution of different spatial frequencies for gender classification, we calculated the contribution of each of the 4,092 parameters from all five spatial frequencies. For each parameter, we performed an independent sample *t*-test (two-sided) between the parameter values from the male trials and those from the female trials, and calculated the absolute value of Cohen's d as an index of the contribution of each parameter to gender classification. One hundred and four parameters in the VGG-Face and 12 in human observers contributed significantly for the classification (Bonferroni corrected for multiple comparisons, **Figures 3A,B**).

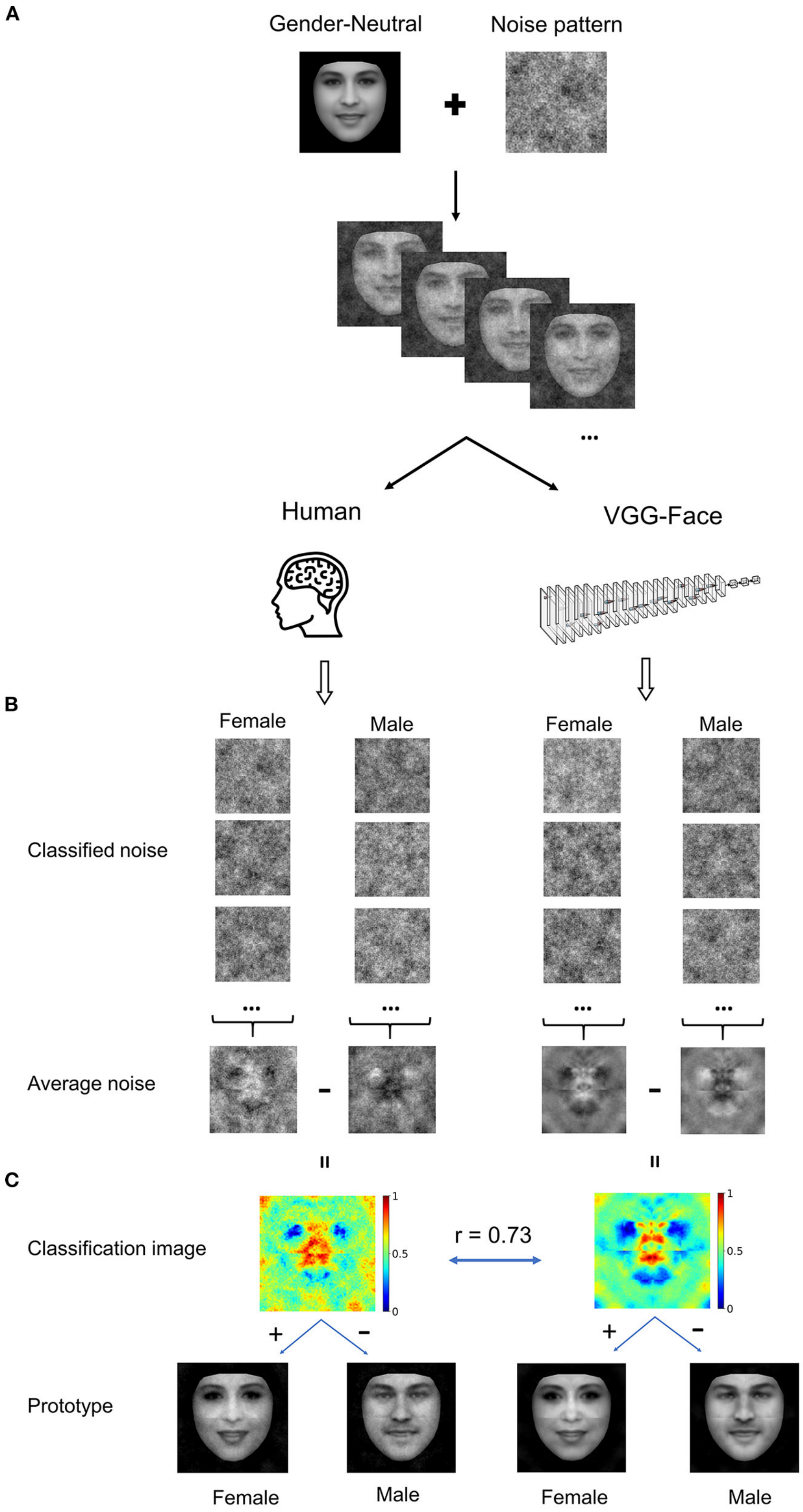


FIGURE 1 | (A) Experiment procedure. A gender-neutral template face was superimposed with noises to create a set of gender-ambiguous faces, which were submitted to the VGG-Face and human observers for gender classification. **(B)** Exemplars of noises extracted from noisy faces classified as either female or male,

(Continued)

FIGURE 1 | respectively. The noises were then averaged to reconstruct images that contained the critical information for classifying the noisy faces as male or as female. **(C)** Classification images (CI) were the difference of the average noise of female by that of male. For visualization, values in each CI were normalized separately to the range from 0 to 1, denoted by colors. By adding or subtracting the rescaled CI to or from the gender-neutral template face, female or male prototype of human observers (Left), and the VGG-Face (Right) were created. Brain icon made by Smashicons from www.flaticon.com.

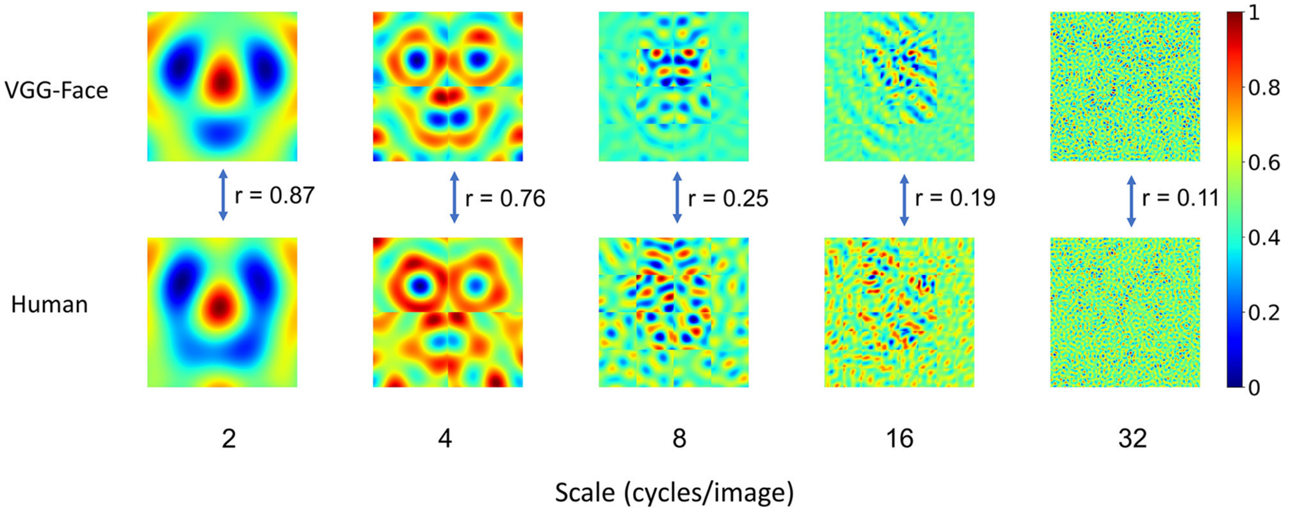


FIGURE 2 | Correspondence in representation at different scales of spatial frequencies. For visualization, values in each CI were normalized separately to the range from 0 to 1, denoted by colors. Note that the correspondence was the highest at the low-spatial frequencies, and then decreased sharply at the high-spatial frequencies. Scale number denotes cycles per image.

Of the 12 parameters in human observers, 9 were at the scales of 2 and 4 cycles/images. Similarly, most of the 104 parameters in the VGG-Face were also at low-frequency scales (7 at 2 cycles/images, 33 at 4 cycles/images, and 30 at 8 cycles/images), and the percentage of the significant parameters at low frequencies (58 and 69% at 2 and 4 cycles/images) were much higher than those at high frequencies (4 and 0% at 16 and 32 cycles/images). That is, both the VGG-Face and human observers mainly relied on information at low spatial frequencies for gender classification.

Another way is to select parameters that made the most contributions indexed by the absolute values of Cohen's d. We found that the 1,885 most contributing parameters of all 4,092 parameters already made up to 80% of the total contribution for the VGG-Face; importantly, these parameters also made up 48% of the contribution for human observers. Then, we examined the similarity of parameters' contribution by calculating the Spearman's correlation between Cohen's d of the VGG-Face and human observers for the highly-contributing parameters at each scale of spatial frequencies. We found that the correlation was high at low spatial frequencies (r = 0.79 and 0.74 at 2 and 4 cycles/images), and then declined sharply at high spatial frequencies (r = 0.21, 0.27, and 0.17 at 8, 16, and 32 cycles/images). In contrast, there were more parameters at high than low spatial frequencies that contributed differently between the VGG-Face and human observers (**Supplementary Analysis 2**). Taken together, at low spatial frequencies, not only were the representations more similar, but also the parameters underlying the representation contributed more significantly to the task.

Human-Like Representation Requires Prior Experience of Face Identification

Where did the representational similarity come from? One possibility is that information at low spatial frequencies is critical for face processing, and therefore both DCNN and human observers were forced to exact information at low spatial frequencies to successfully perform the task. An alternative hypothesis is that the VGG-Face and human observers share similar prior experiences of processing face at the subordinate level where faces are identified into different individuals. To test these two hypotheses, we examined another typical DCNN, the AlexNet, that also has abundant exposure to face images but is pre-trained to classify objects into 1,000 basic categories. We trained the AlexNet to perform the gender classification task with the same transfer learning procedure as that for the VGG-Face. The testing accuracy of gender classification of the AlexNet reached 89.3% (see **Supplementary Tables 1, 2** for more details), indicating that it was able to perform the task. However, the CIs obtained from the Alexnet (**Figures 4A,B**) were in sharp contrast to the CIs of human observers (**Figures 1C, 2**) as a whole (r = −0.04) and at different scales (r = −0.28, 0.03, 0.25, 0.10, and 0.03 at the scales of 2, 4, 8, 16, and 32). We also reconstructed the female and male prototype faces of AlexNet (**Figure 4A**),

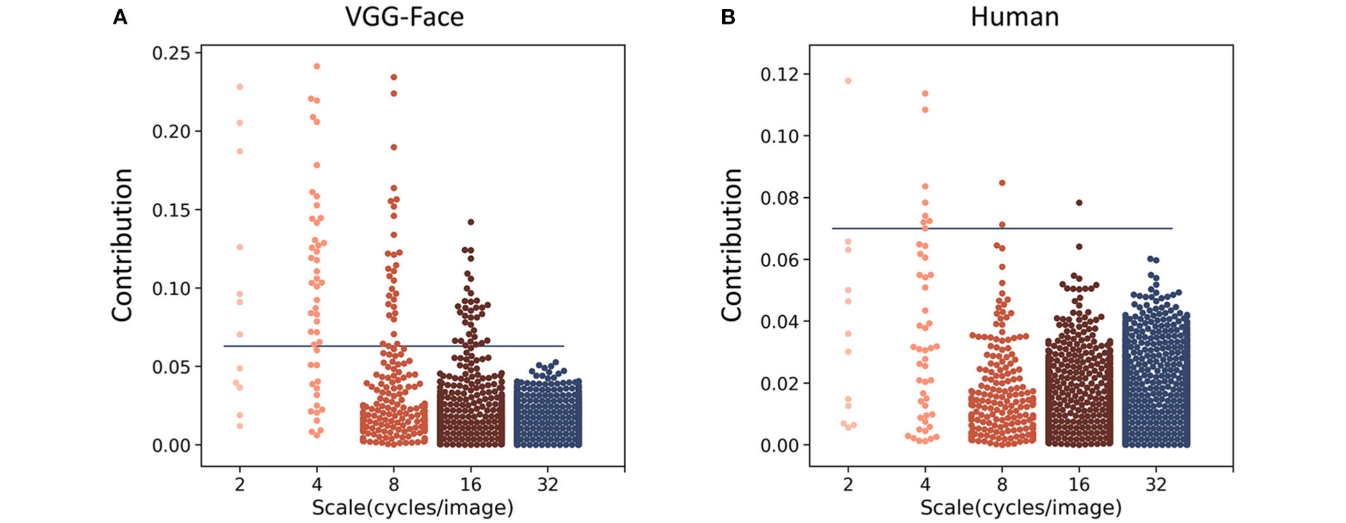


FIGURE 3 | Manhattan plot of the contribution (the absolute values of Cohen's d) of the parameters used to construct noises in the VGG-Face **(A)** and in human observers **(B)**. Each dot denotes a parameter, and the horizontal blue line indicates the significance level after Bonferroni correction.

and they appeared quite distinct from those of human observers and the VGG-Face (**Figure 1C**). This finding was unlikely due to the differences in architecture between the VGG-Face and the AlexNet, because the VGG-16, which has the same architecture as the VGG-Face but is pre-trained for object categorization as the AlexNet, showed a CI largely different from human observers (**Supplemental Analysis 3**). Therefore, although the AlexNet succeeded in performing the gender classification task, it relied on a set of information completely different from human observers to achieve the goal. Therefore, mere exposure to face stimuli or large categories of stimuli is not sufficient for the DCNNs to construct similar representations for gender classification as human observers; instead, the task requirement of face identification during prior experience was required.

Given that the training sample contained more male than female faces, we also trained the VGG-Face and AlexNet for face-gender classification with balanced training sample to exclude the possibility that our results was caused by unbalanced training sample (**Supplementary Analysis 4**).

In addition, to examine whether our results could transfer to other face databases, we trained the VGG-Face and AlexNet for face-gender classification using face images from another database FairFace, the Face Attribute Dataset for Balanced Race, Gender, and Age (Kärkkäinen and Joo, 2019), and the main findings were replicated with this new dataset (**Supplementary Analysis 5**).

Finally, to further illustrate that the CIs obtained here reflected representations for face gender classification, we built a simple new network that used the CIs to perform gender classification. Specifically, we first aligned each of 26,902 face images (13,738 females) to the neutral face template and convolved each aligned face with the CI to get an activation value (**Figures 5A,B**). This procedure is equivalent to using each aligned face as an input image and the CI as connected weights of a one-layer network with one output unit. If the CI does represent the differences between female and male faces, the activation distributions of male and female faces would be dissociated. As shown in **Figure 5C**, after convolving with the CI obtained from VGG-Face, the activation distribution of female faces dissociated from that of male faces (cohen's d = 1.62). A similar trend of dissociation was observed when using the CI obtained from humans (cohen's d = 1.58). As a baseline, we randomized the CI image and convolved each face with the randomized CI image, and the two activation distributions largely overlapped (cohen's d = 0.30). These results indicated that the CIs revealed the information used for face gender classification, and similar information was used by VGG-Face and humans. When using the CI obtained from AlexNet, the two activation distributions also largely overlapped (cohen's d = −0.35), and the difference between female and male activations was in an opposite direction to the results of VGG-Face and humans. Again, this result was consistent with our main finding that the CI of AlexNet differed from that of VGG-Face and humans.

DISCUSSION

Marr (1982) has proposed a three-level framework to understand an intelligent information-processing system. At the top is the computational level that defines the goal of the system, and in our study, the computation goal is face gender classification; at the bottom is the implementation level that is the physical substrate of the system, which are the DCNNs and human brain in our study. Most critically, in the middle is the representational and algorithmic level that establishes approaches through which the implementation achieves the computation goal. Despite dramatic differences in the physical implementations between the artificial and biological intelligent systems, similar representations may

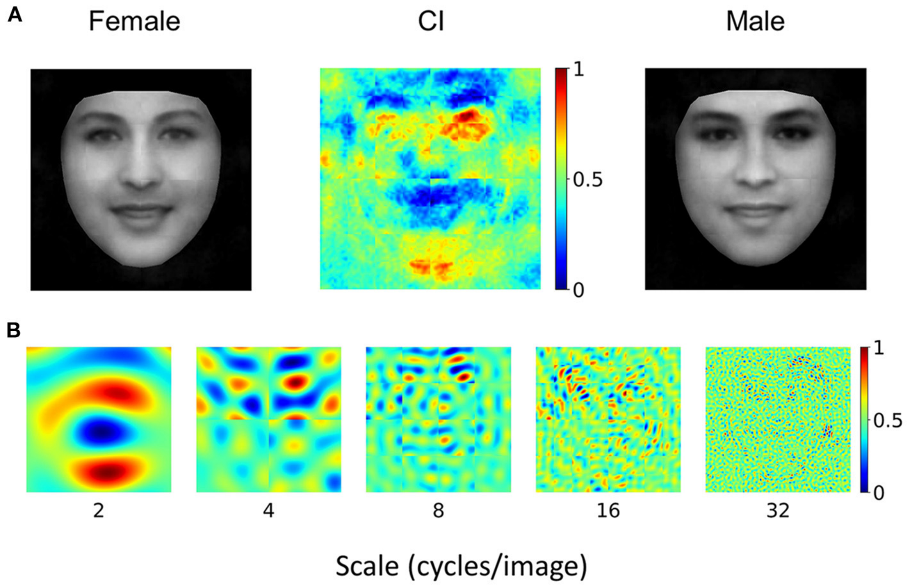


FIGURE 4 | (A) AlexNet's CI for gender classification. For visualization, values in each CI were normalized separately to the range from 0 to 1, denoted by colors. Note that the female prototype (left) and the male prototype (right) were not perceptually female-like and male-like, respectively. **(B)** Normalized CI at different scales of spatial frequencies. Note that they were significantly different from those of human observers.

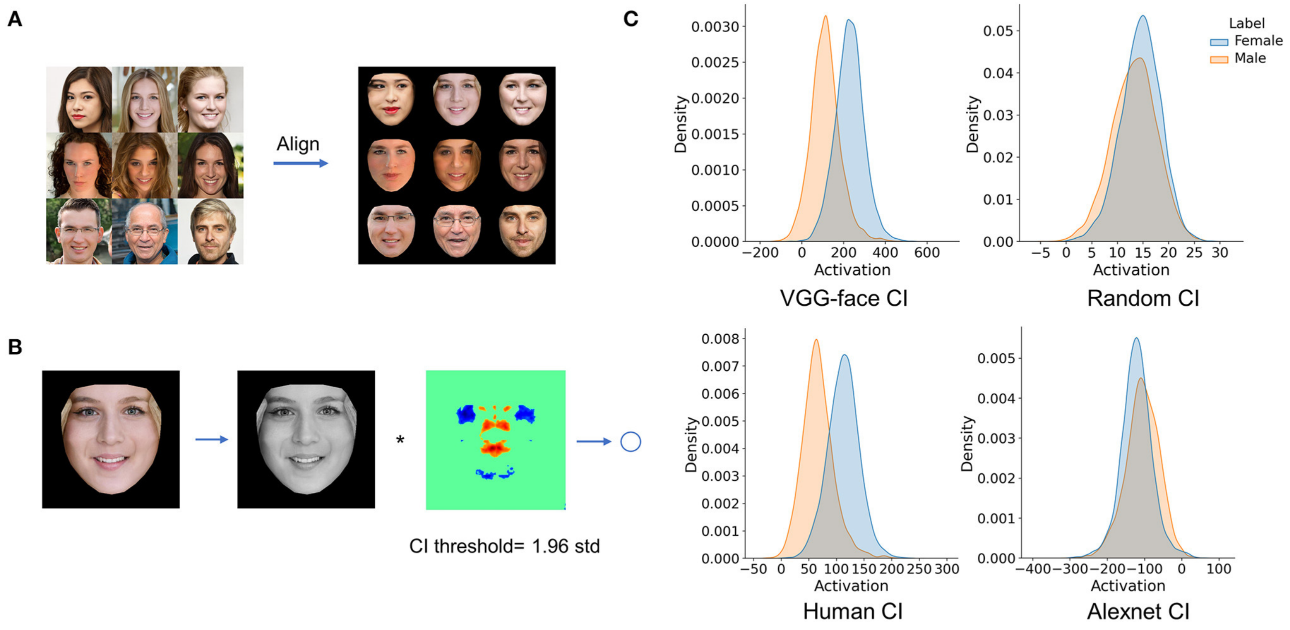


FIGURE 5 | Using CIs in a simple network. **(A)** Each of 26,902 face images was aligned to a neutral face template. **(B)** Each aligned face was convolved with the CI to get an activation value. **(C)** Activation distributions of the female and males faces after convolving with the CIs.

be used by different systems to achieve the same computation goal. Our study provides one of the first direct evidence to support this hypothesis by showing that the DCNNs and humans used similar representations to achieve the goal of face gender classification, which were revealed by highly similar CIs between the VGG-Face and humans. Admittedly, the present study examined face perception which is highly domain-specific in human visual cognition. Future study is needed to examine whether implementation-independent representation can also be observed in less specialized perceptual processes.

The shared representation, on one hand, may come from the critical stimulus information needed to achieve the computation goal. Previous human studies on gender classification suggest that the critical information humans used to solve the task is embedded mainly in low spatial frequencies (Sergent, 1986; Valentin et al., 1994; Goffaux et al., 2003b; Mangini and Biederman, 2004; Khalid et al., 2013). Here we found that the VGG-Face also relied heavily on low spatial frequencies of faces for gender classification. Further, it was the information only in this band that showed similarity to that of humans, but not in high spatial frequencies. In other words, one reason that the VGG-Face and humans established similar representations based on low spatial frequencies might be that this stimulus information is critical for the task of face gender classification.

On the other hand, the prior task experience before the gender classification task may also play a deterministic role for DCNNs to use a similar approach to achieve the goal as humans. Previous studies have shown that humans usually process faces at the subordinate level, that is, to recognize faces as individuals. Similar to humans, the VGG-Face is also pre-trained to recognize faces at the individual level, that is, to classify face images into different identities (e.g., John's face). Therefore, the similar task experience in the past likely led the similar approaches in achieving the new goal of gender classification.

In contrast, the AlexNet is pre-trained to recognize objects at the basic level, that is, to classify objects into categories (e.g., dogs) but not individuals (John's dog). Therefore, although the AlexNet experiences abundant exposure to face images during the pre-training, it processes faces as objects, different from humans and the VGG-Face. Previous studies on humans have shown that object recognition does not selectively rely on low- to middle- spatial frequencies as face recognition does (Biederman and Kalocsais, 1997; Goffaux et al., 2003a; Collin, 2006; Collin et al., 2012). Thus, it is not surprising that although the AlexNet also achieved a high performance (accuracy around 90%) in face gender classification, an approach significantly different from that of humans was adopted. Taken together, the similarity in representation between DCNNs and humans was not guaranteed by the common computational goal or by the passive experiences with stimuli; instead, it was constrained by the combination of experiences on the pre-training task in the past and critical stimulus information needed in performing the task in the present. The finding also suggests that DCNN can be used as a model of biological brains to experimentally investigate the effect of visual experience and task demands on human cognition.

The present study also brought insight from an engineering perspective. In history, two main approaches have been proposed to achieve and even excel human vision in artificial intelligence (Kriegeskorte and Douglas, 2018). The neuroscience approach adheres to biological fidelity at the implementation level, which simulates neural circuits of brains, whereas the cognitive approach emphasizes on cognitive fidelity, which focuses on goal-directed algorithms and disregards implementation. Our study suggests an intermediate approach lying in between these two. By simulating human intelligence at the representation level in Marr's framework, this approach provides an abstract description of how a system extracts critical features to construct representation for a specific task. Because the representation is relatively independent of implementation, the knowledge acquired in biological systems can be easily adopted by artificial systems with completely different substrates. Therefore, the simulation of representation may shed light on building new AI systems in a feasible way.

MATERIALS AND METHODS

Transfer Learning

We used the pre-trained VGG-Face network (Parkhi et al., 2015) that consists of 13 convolutional layers and 3 fully connected (FC) layers. Each convolutional layer and FC layer were followed by one or more non-linearities such as ReLU and max pooling. The VGG-Face network was pre-trained for face identification with the VGG-Face dataset containing over two million face images of 2,662 identities.

In our study, we trained the VGG-Face for face-gender classification using transfer learning. The final FC layer of the VGG-Face has 2,662 units, each for one identity. We replaced this layer with a two-unit FC layer for the binary gender classification. All weights of the network were frozen except the weights between the penultimate FC layer and the new final FC layer. The training sample contains 21,458 face images (male: 14,586) of 52 identities (male: 35) randomly selected from the VGGFace2 dataset (Cao et al., 2018). The validation sample contains other 666 face images (male: 429) from the same 52 identities. The testing sample contains 1,000 face images (male: 500) from 24 new identities from the VGGFace2 dataset. All face images were resized to 224×224 pixels to match the model input size. We used in-house python package DNNbrain (Chen et al., 2020) to train the network. The loss function was cross-entropy, and the optimizer was Adam. The learning rate was 0.03, and the network was trained for 25 epochs. After training, the accuracy of gender classification reached 100% on both the training and validation samples, and 98.6% on the testing sample.

The same training procedures were applied to AlexNet pre-trained for object categorization (Krizhevsky et al., 2017). The model consists of five convolutional layers and three FC layers. The AlexNet was pre-trained on ImageNet to classify 1.2 million images into 1,000 object categories. We also replaced the final layer of AlexNet with a two-unit FC layer for the binary gender classification. After transfer learning, the accuracy for gender classification reached 92.6% on the training sample, 93.2% on the validation sample, and 89.3% on the testing sample.

Reverse Correlation Approach

After the transfer learning on gender classification, we made the representation explicit with the reverse correlation approach on noisy faces. All stimuli consisted of a gender-neutral template face superimposed with sinusoid noise patterns. The template was a morphed face between a female average face and a male average face (**Figure 1A**). The female and male average faces were computed as a mathematical average of all female and all male faces of the training sample after they were aligned and wrapped into the same space with 68 landmarks using an open-access toolbox face_morpher (https://github.com/alyssaq/face_morpher). The average faces were 8-bit grayscale and 512 × 512 pixel images. We further created 500 morph faces that gradually changed from the female average face to the male average face using face_morpher. Then we presented 500 morphed faces evenly distributed between the female and the male average faces to the VGG-Face to find the face most equally classified as male and female in gender classification. The 250th morphed face, which was classified as female with a probability of 54% by the VGG-Face, was chosen as the gender-neutral template face in our study.

A random noise pattern was generated for each trial. Each noise pattern was composed of sinusoid patch layers of five different scales of spatial frequencies (2, 4, 8, 16, and 32 cycles/image), with each patch layer made up of 1, 4, 16, 64, and 256 sinusoid patches, respectively (Mangini and Biederman, 2004). For each sinusoid patch, sinusoids of six orientations (0, 30, 60, 90, 120, and 150 degrees) and two phases (0 and pi/2) were summed. The amplitude of each sinusoid came from a random sampling of a uniform distribution of values from −1 to 1. Therefore, each noise pattern was determined by 4,092 random amplitude parameters (12, 48, 192, 768, and 3,072 parameters for 2, 4, 8, 16, and 32 cycles/image). We use the R package rcicr to generate the sinusoid noises (Dotsch, 2017). We created 20,000 noise patterns for the DCNNs and 1,000 noise patterns for each human observer. Each noise pattern was then superimposed on the template face to create a different noisy face.

We resized the noisy face images to 224 × 224 pixels and submitted them to the VGG-Face and AlexNet, and obtained their classification prediction for each image. For VGG-Face, a noisy face was classified as male when the activation of the male unit was higher than the female unit. Note that the AlexNet showed a bias toward male faces when classifying the noisy faces; therefore, we modified the classification criterion for the AlexNet. That is, for AlexNet, a noisy face would be classified as male when the activation of the male unit to the to-be-classified face was higher than its average activation to all noisy faces. Note that the choice of criterion would not affect the results pattern of the VGG-Face and hence the dissociation between AlexNet and VGG-Face, because the two criteria lead to literally identical CIs for VGG-Face ($r = 0.99$).

To generate corresponding female or male prototype faces, each CI was separately rescaled to have the same maximum pixel value and then added or subtracted from the template face.

Participants

Sixteen college students (12 females, age 19–33 years, mean age 22 years) from Beijing Normal University, Beijing, China, participated in the gender classification task. All participants were right-handed and had normal or corrected-to-normal vision. The experiment protocol was approved by the Institutional Review Board of the Faculty of Psychology, Beijing Normal University. Written informed consent was obtained from all participants before the experiment.

Experimental Procedures

Before the experiment, participants were told that they would perform a difficult gender classification task because the faces were superimposed with heavy noises. The template image was not shown to participants in the experiment. The stimuli were 255-bit grayscale and 512 × 512 pixel images. PsychoPy (Peirce et al., 2019) was used to display the stimuli and record responses. The stimuli were presented on the screen of a Dell precision laptop at a distance of 70 cm. The stimuli subtended a visual angle of ∼8.2 degree. In each trial, a noisy face image was presented in the center of the screen for 1 s, and then the screen cleared until the participant made a response. The participants were instructed to provide one of four responses with a key press for each trial: probably female, possibly female, possibly male, or probably male. No feedback was provided. Each participant performed 1,000 trials. The participants could rest every 100 trials. The total experiment duration was about 1 h for each participant. In data analysis, the CI was calculated by subtracting the average noise patterns from all trials classified as male (probably male and possibly male) from those classified as female (probably female and possibly female).

DATA AVAILABILITY STATEMENT

All codes for analyses are available on https://github.com/YukunQu/cnnface. All face images used in the present study were from the public VGGFace2 dataset and FairFace dataset. The behavioral data are available on https://osf.io/6gqtx/.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Institutional Review Board of the Faculty of Psychology, Beijing Normal University. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

JL and YS conceived and designed the study. YQ collected and analyzed the data. YS interpreted the data and wrote the manuscript with input from JL, YQ, and SX. All authors contributed to the article and approved the submitted version.

FUNDING

This study was funded by the National Natural Science Foundation of China (31861143039, 31872786, and 31600925) and the National Basic Research Program of China (2018YFC0810602).

ACKNOWLEDGMENTS

This manuscript has been released as a pre-print at Song et al. (2020).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fncom.2020.601314/full#supplementary-material

REFERENCES

Ahumada, A. Jr., and Lovell, J. (1971). Stimulus features in signal detection. *J. Acoust. Soc. Am.* 49, 1751–1756. doi: 10.1121/1.1912577

Biederman, I., and Kalocsais, P. (1997). Neurocomputational bases of object and face recognition. *Philos. Trans. R. Soc. Lond B Biol. Sci.* 352, 1203–1219. doi: 10.1098/rstb.1997.0103

Cao, Q., Shen, L., Xie, W., Parkhi, O. M., and Zisserman, A. (2018). "VGGFace2: a dataset for recognising faces across pose and age," in *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)* (Xi'an: IEEE), 67–74. doi: 10.1109/FG.2018.00020

Chen, X., Zhou, M., Gong, Z., Xu, W., Liu, X., Huang, T., et al. (2020). DNNBrain: a unifying toolbox for mapping deep neural networks and brains. *Front. Comput. Neurosci*. 14:580632. doi: 10.3389/fncom.2020.580632

Collin, C. A. (2006). Spatial-frequency thresholds for object categorisation at basic and subordinate levels. *Perception* 35, 41–52. doi: 10.1068/p5445

Collin, C. A., Therrien, M. E., Campbell, K. B., and Hamm, J. P. (2012). Effects of band-pass spatial frequency filtering of face and object images on the amplitude of N170. *Perception* 41, 717–732. doi: 10.1068/p7056

Dotsch, R. (2017). *In Rcicr: Reverse-Correlation Image-Classification Toolbox (R Package Version 0.4.0)*.

Goffaux, V., Gauthier, I., and Rossion, B. (2003a). Spatial scale contribution to early visual differences between face and object processing. *Cogn. Brain Res.* 16, 416–424. doi: 10.1016/S0926-6410(03)00056-9

Goffaux, V., Jemel, B., Jacques, C., Rossion, B., and Schyns, P. G. (2003b). ERP evidence for task modulations on face perceptual processing at different spatial scales. *Cogn. Sci.* 27, 313–325. doi: 10.1207/s15516709cog2702_8

Gold, J. M., Murray, R. F., Bennett, P. J., and Sekuler, A. B. (2000). Deriving behavioural receptive fields for visually completed contours. *Curr. Biol.* 10, 663–666. doi: 10.1016/S0960-9822(00)00523-6

Kärkkäinen, K., and Joo, J. (2019). FairFace: face attribute dataset for balanced race, gender, and age. *arXiv [Preprint] arXiv:1908.04913*. Available online at: http://arxiv.org/abs/1908.04913

Khalid, S., Finkbeiner, M., König, P., and Ansorge, U. (2013). Subcortical human face processing? Evidence from masked priming. *J. Exp. Psychol. Hum. Percept. Perform.* 39, 989–1002. doi: 10.1037/a0030867

Kriegeskorte, N., and Douglas, P. K. (2018). Cognitive computational neuroscience. *Nat. Neurosci.* 21, 1148–1160. doi: 10.1038/s41593-018-0210-5

Krizhevsky, A. (2014). One weird trick for parallelizing convolutional neural networks. *arXiv* arXiv:1404.5997.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Commun. ACM*. 60, 84-90. doi: 10.1145/3065386

Mangini, M. C., and Biederman, I. (2004). Making the ineffable explicit: Estimating the information employed for face classifications. *Cogn. Sci.* 28, 209–226. doi: 10.1207/s15516709cog2802_4

Marr, D. (1982). *Vision: A Computational Investigation Into the Human Representation and Processing of Visual Information*. New York, NY: henry holt and co. Inc.

Martin-Malivel, J., Mangini, M. C., Fagot, J., and Biederman, I. (2006). Do humans and baboons use the same information when categorizing human and baboon faces? *Psychol. Sci.* 17, 599–607. doi: 10.1111/j.1467-9280.2006.01751.x

Parkhi, O. M., Vedaldi, A., and Zisserman, A. (2015). "Deep face recognition," in *Proceedings of the British Machine Vision Conference 2015* (Swansea: British Machine Vision Association), 41.1–41.12. doi: 10.5244/C.29.41

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., et al. (2019). PsychoPy2: experiments in behavior made easy. *Behav. Res. Methods* 51, 195–203. doi: 10.3758/s13428-018-01193-y

Ranjan, R., Sankaranarayanan, S., Castillo, C. D., and Chellappa, R. (2017). "An all-in-one convolutional neural network for face analysis," in *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)* (Washington, DC: IEEE), 17–24. doi: 10.1109/FG.2017.137

Schroff, F., Kalenichenko, D., and Philbin, J. (2015). "FaceNet: a unified embedding for face recognition and clustering," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Boston, MA: IEEE), 815–823. doi: 10.1109/CVPR.2015.7298682

Sergent, J. (1986). "Microgenesis of Face Perception," in *Aspects of Face Processing NATO ASI Series*, eds H. D. Ellis, M. A. Jeeves, F. Newcombe, and A. Young (Dordrecht: Springer Netherlands), 17–33. doi: 10.1007/978-94-009-4420-6_2

Simonyan, K., and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *arXiv [Preprint] arXiv:1409.1556*. Available online at: http://arxiv.org/abs/1409.1556

Song, Y., Qu, Y., Xu, S., and Liu, J. (2020). Implementation-independent representation for deep convolutional neural networks and humans in processing faces. *bioRXiv*, 2020.06.26.171298. doi: 10.1101/2020.06.26.171298

Taigman, Y., Yang, M., Ranzato, M., and Wolf, L. (2014). "DeepFace: closing the gap to human-level performance in face verification," in *2014 IEEE Conference on Computer Vision and Pattern Recognition* (Columbus, OH: IEEE), 1701–1708. doi: 10.1109/CVPR.2014.220

Valentin, D., Abdi, H., and O'toole, A. J. (1994). Categorization and identification of human face images by neural networks: a review of the linear autoassociative and principal component approaches. *J. Biol. Syst.* 2, 413–429. doi: 10.1142/S0218339094000258

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Computational Neuroscience

ORIGINAL RESEARCH
published: 16 February 2021
doi: 10.3389/fncom.2021.612041

Brain Inspired Sequences Production by Spiking Neural Networks With Reward-Modulated STDP

Hongjian Fang[1,2†‡], *Yi Zeng*[1,2,3,4*†‡] *and Feifei Zhao*[1]

[1] Research Center for Brain-Inspired Intelligence, Institute of Automation, Chinese Academy of Sciences, Beijing, China, [2] School of Future Technology, University of Chinese Academy of Sciences, Beijing, China, [3] Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Shanghai, China, [4] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China

Understanding and producing embedded sequences according to supra-regular grammars in language has always been considered a high-level cognitive function of human beings, named "syntax barrier" between humans and animals. However, some neurologists recently showed that macaques could be trained to produce embedded sequences involving supra-regular grammars through a well-designed experiment paradigm. Via comparing macaques and preschool children's experimental results, they claimed that human uniqueness might only lie in the speed and learning strategy resulting from the chunking mechanism. Inspired by their research, we proposed a Brain-inspired Sequence Production Spiking Neural Network (SP-SNN) to model the same production process, followed by memory and learning mechanisms of the multi-brain region cooperation. After experimental verification, we demonstrated that SP-SNN could also handle embedded sequence production tasks, striding over the "syntax barrier." SP-SNN used Population-Coding and STDP mechanism to realize working memory, Reward-Modulated STDP mechanism for acquiring supra-regular grammars. Therefore, SP-SNN needs to simultaneously coordinate short-term plasticity (STP) and long-term plasticity (LTP) mechanisms. Besides, we found that the chunking mechanism indeed makes a difference in improving our model's robustness. As far as we know, our work is the first one toward the "syntax barrier" in the SNN field, providing the computational foundation for further study of related underlying animals' neural mechanisms in the future.

Keywords: brain-inspired intelligence, spiking neural network, reward-medulated STDP, population coding, reinforcement learning

OPEN ACCESS

Edited by:
Si Wu,
Peking University, China

Reviewed by:
Jiang Wang,
Tianjin University, China
Thomas Wennekers,
University of Plymouth,
United Kingdom

Correspondence:
Yi Zeng
yi.zeng@ia.ac.cn

[†] These authors have contributed equally to this work

[‡] These authors share first authorship

Received: *30 September 2020*
Accepted: *19 January 2021*
Published: *16 February 2021*

Citation:
Fang H, Zeng Y and Zhao F (2021) Brain Inspired Sequences Production by Spiking Neural Networks With Reward-Modulated STDP. Front. Comput. Neurosci. 15:612041. doi: 10.3389/fncom.2021.612041

1. INTRODUCTION

The human capacity for language is unique on the earth: although most animals communicate, only humans show this unbounded expressive power (Fitch, 2018; Jiang et al., 2018). A significant topic for cognitive neuroscience is determining how human computational capacities differ from those of other animals (Deacon, 1998; Matsuzawa, 2013; Dehaene et al., 2015a; Yang, 2016; Jiang et al., 2018). Previously, the generative algorithms acquired by animals seem mainly restricted to the lowest level of the Chomsky hierarchy (Chomsky, 1957, 1965)—that is, regular languages (Fitch and Friederici, 2012; Fitch, 2014; Jiang et al., 2018). Thus, it has often been proposed that a pivotal gap lies between the levels of regular or "finite-state" grammars, which are accessible to nonhuman animals, and supra-regular grammars or "phrase-structure" grammars, which may only be available to humans (Hauser et al., 2002; Fitch, 2014; Jiang et al., 2018).

Some researchers attempt to teach animals understanding symbol sequences with nested or recursive structures, which are characteristic of human languages, have mostly been met with negative results (Miles, 1990; Pinker, 2003; Dehaene et al., 2015b).

So far, the generative algorithms acquired by animals seem mostly restricted to the lowest level of the Chomsky hierarchy (Chomsky, 1957, 1965)—that is, regular languages (Fitch, 2004; Fitch and Friederici, 2012). Thus, it has often been proposed that a "syntax barrier" lies between the levels of regular or "finite-state" grammars, which are accessible to nonhuman animals, and supra-regular grammars or "phrase-structure" grammars, which may only be available to humans (Hauser et al., 2002).

However, Jiang et al. (2018) designed the macaque monkeys supra-regular rule experimental paradigm, and they demonstrated that after extensive reinforcement training, macaque monkeys can master the supra-regular grammar, which breaks the barrier of syntax previously divided. Specifically, as Figure 1B in Fitch (2018) shown, Jiang and Wang designed a novel behavioral paradigm, delayed-sequence production task that required the animal to explicitly generate sequences according to the instructed grammars (Jiang et al., 2018). They compared two grammars:

(1) a "mirror" grammar of the form ABC|CBA, which in formal language theory involves recursive center embedding.

(2) a "repeat" grammar of the form ABC|ABC, i,e, repetition in serial order, as shown in Figure 1A in Jiang et al. (2018).

Like the grammars of all human languages, mirror grammars require a learner to possess supra-regular computational abilities, which requires specific computational machinery not needed at the lower sub-regular level Figure 1A in Fitch (2018). Besides, monkeys spontaneously generalized the learned grammar to novel sequences, including longer ones, and could generate hierarchical sequences formed by an embedding of two levels of abstract rules. Compared to monkeys, however, preschool children learned the grammars much faster using a chunking strategy (Jiang et al., 2018).

In fact, it is quite common for animals to complete the sequence by the "repeat" rule. The best example is that birds can imitate their parents' singing (Mooney, 2009). However, for the "mirror" sequence production, negative results are often obtained (Dehaene et al., 2015b). The essential reason is that most of the synapses (except electrical synapses) are unidirectional, and the reverse order production requires the agent to have the ordinal knowledge (Dehaene et al., 2015b). Even though when people are faced with the challenge of recalling a sequence of a phone number in reverse order, they often need to repeat the number sequence repeatedly to determine the position of a specific number in the sequence to complete the task. Therefore, the "mirror" sequence production task is a complex cognitive task that requires more advanced cognitive brain regions to participate in Fitch (2018). It is of great significance to reveal the cognitive process of reconstructing symbol sequence for understanding human language ability (Dehaene et al., 2015b; Jiang et al., 2018).

Their work inspired us to explore whether SNN can also break the "syntax barrier."

After experimental verification, we demonstrated SNN could indeed handle the same sequence production task. The innovative aspects of this work are as follows:

- As far as we know, we are the first one to demonstrated that SNN can break the "syntax barrier" with Population-Coding and Reward-Modulated STDP mechanism, coordinating STP and LTP mechanisms simultaneously.
- We demonstrated that the chunking mechanism, helping to improve the robustness and learning efficiency of the network.
- Our work provides the computational foundation for further study of underlying animal neural mechanisms in the future.

2. MODEL AND METHODS

2.1. Neuron Model and Synapse Learning Rule

There are various neuron models such as the famous H-H model (Hodgkin and Huxley, 1952), Leaky Integrate-and-Fire neuron (*LIF*) model (Miller, 2018), Izhikevich neuron model (Izhikevich, 2003), and so on.

In order to simplify the computational complexity of the model, we choose the Leaky Integrate-and-Fire neuron model as the building block of the Spiking Neural Network to complete the whole experiment. Standard LIF models are shown in Equations (1), (2), and (3).

$$C_m \frac{dV}{dt} = -g(V - V_s) + I \tag{1}$$

$$\tau_m \frac{dV}{dt} = -(V - V_s) + \frac{I}{g} \tag{2}$$

$$V \rightarrow V_{reset} \quad if(V \geq V_{threshold}). \tag{3}$$

C_m is the membrane capacitance of the neuron, V is the membrane potential of the neuron, g is the conductance of the membrane, V_s is the steady-state leaky potential, here we let $V_s = V_{reset}$ to simplify the model. I is the input current of the neuron. $\tau_m = \frac{C_m}{g}$ represents the voltage delay time, and different types of neurons have different values of τ_m.

$$I = \sum_j w_{j,i}\sigma_j(t-1) + I_s \tag{4}$$

$$\sigma_i(t) = \begin{cases} 0 & V < V_{threshold} \\ 1 & V \geq V_{threshold} \end{cases} \tag{5}$$

Equation (4) shows that the current of neurons consists of two parts: the current from other neurons and the external stimulating current I_s. $W_{j,i}$ is the weight of i-th neuron to j-th

TABLE 1 | Model parameters.

Model/Rule	Parameter	Value
LIF model	C_m	30nF
	τ_m	30 ms
	V_{reset}	−65 mv
	$V_{threshold}$	−35 mv
	τ_{ref}	10 ms
STDP rule	τ_s	15 ms
	τ_w	10 ms
	A_+	4
	A_-	0.95

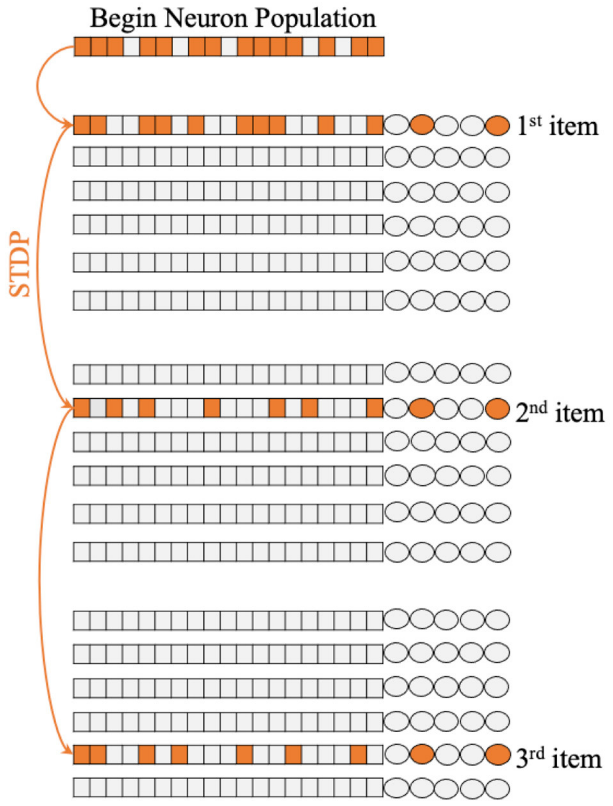


FIGURE 1 | The architecture of Working Memory Circuit (WMC), each row of neuron populations corresponding to the six symbol on the screen shown to macaques in biology experiment. And the synapses between populations update with STDP learning rule.

neuron. $\sigma_i(t)$ is the indicator to judge if the i-th neuron firing at the time of t in Equation (5). And external stimuli mainly corresponding to the appearance of a specific symbol.

As for the synapse learning rule, Spike Timing Dependent Plasticity (STDP) (Bi and Poo, 1998; Dan and Poo, 2004) is one of the most important learning principles for the biological brain. STDP postulates that the strength of the synapse is dependent on the spike timing difference of the pre- and post-neuron (Dan and Poo, 2006).

Here we use STDP to update synaptic weights according to the relative time between spikes of presynaptic and postsynaptic neurons. The modulation principle is that if the postsynaptic neuron fires a few milliseconds after the presynaptic neuron, the connection between the neurons will be strengthened, otherwise, the connection will be weakened (Wittenberg and Wang, 2006). The update function is shown in Equation (6), where A_+ and A_- are learning rates. τ_s and τ_w are STDP time constant, and Δt is the delay time from the presynaptic spike to the post-synaptic spike.

$$\Delta w_{j,i} = \begin{cases} A_+ e^{(\Delta t/\tau_+)} & -\tau_w < \Delta t < 0 \\ -A_- e^{(-\Delta t/\tau_-)} & 0 < \Delta t < \tau_w \end{cases} \tag{6}$$

All the parameters can be found in **Table 1**.

2.2. Working Memory Based on Population Coding

In the macaque monkeys' sequence producing experiment, researchers designed the paradigm where the macaque monkeys need to produce the sequence of the spatial symbols according to different rules, i.e., Repeat/Mirror.

Obviously, working memory is a necessary condition for sequence producing (Jiang et al., 2018). Just as the macaque monkeys must memorize the spatial symbols before producing process, our SNN should also include the corresponding circuit to accomplish working memory function. Therefore, we implemented the Working Memory Circuit (WMC) to realize related function, which will be covered in detail in this section.

Neurons can encode complicated temporal sequences such as the mating songs that songbirds learn, store, and replay (Quiroga, 2012; Yi et al., 2019). Inspired by the previous research work, an invariant, sparse, and explicit code, which might be important in the transformation of complex visual percepts into long-term and more abstract memories (Quiroga, 2012). It is reliable to assume when the macaques try to memorize the raw sequence, different populations of neurons are activated, i.e., they are bound to different light spots. Based on this assumption, we designed the Working Memory Circuit (WMC) to mimic the neuron activity of macaque monkeys.

Figure 1 shows the single unit of Working Memory Circuit, which includes six populations of neurons corresponding to six appear on the screen, the corresponding neuronal population will be stimulated in a short time window by an extra input current. Regarding the number of neurons of neuron population, we try different sizes in the experiment, which will be discussed in detail in the following chapters.

Inspired by biological discoveries we translate the appearance of single symbol in screen into the external input stimulation to corresponding spike neuron population. We choose Poisson Encoding as the method of input stimulation.

Due to the randomness of the Poisson Encoding, part of neurons in the population will fire at different times when the external stimulation window is given. The main function of inhibitory neurons is lateral inhibition. In order to make only one population of neurons fire among the six symbols, the inhibitory neurons in each population will inhibit remaining five populations of neurons.

Because symbols appear in sequential order, different populations of neurons will fire in turn. It is precisely because of different populations of neurons fire in a particular order, STDP rules can make a difference in the process of memory. **Figure 1** shows how the STDP rules influence the memorizing process with different temporal activation of neuron populations.

The whole memorizing process starts with the cross in the center of the screen in Figure 1B in Fitch (2018) lit, which corresponds to the "begin" neuron population in WMC. This population obtains extra current and part of the memorize will fire. Then, according to the examples of Figure 1A in Jiang et al. (2018), symbols 1, 2, and 5 appear in turn, and the corresponding neuron population obtains the extra current in turn and fire in turn. Due to the mechanism of STDP, new synapses are formed between the corresponding populations of neurons of symbols 1, 2, and 5, as shown in **Figure 1**, then completing the memory process. It is worth mentioning that the "125" sequence is just an example for convenience of understanding, WMC can memorize the sequence composed of any three symbols in the set of position symbols.

2.3. Motor Circuit

In the macaques' sequence producing experimental paradigm, macaques need to press the light spots in the screen by correct sequence to get the reward (Jiang et al., 2018). Neuroscientists have found that in the biological brain, action instructions are encoded by specific motor neurons (Wichterle et al., 2002). Correspondingly, in our model, we must define the triggers of pressing action.

The center part of **Figure 2** shows the concrete structure of the motor circuit. Motor circuit receives the projection from Working Memory Circuit and Reinforcement Learning Circuit.

In a nutshell the macaque monkeys perform a specific position keystroke operation once a corresponding motor neuron fire. In our SNN model, the network output a symbol once. (Six light spots, in this case, correspond to six motor neurons in each population).

2.4. Reinforcement Learning With Reward-Modulated STDP

Unlike the short-term plasticity (STP) (Markram and Tsodyks, 1996; Abbott et al., 1997; Zucker and Regehr, 2002) mechanism in the memory process, macaque monkeys use long-term plasticity (LTP) (Bi and Poo, 2001) mechanism as the mean of learning Repeat/Mirror rules (Jiang et al., 2018). That means macaque monkeys' memory about a particular sequence maintains short time, while the learning of producing rules are in the long term. According to the experimental paradigm in the references, macaque monkeys will be rewarded with food or water if they can complete the production of the sequence in the course of training, and punitive measures (i.e., blowing the monkey's eyes with air) will be launched if there are any symbolic errors in the production process (Jiang et al., 2018). Therefore, it is reasonable to assume that the learning of the Repeat/Mirror rules in macaque monkeys is based on reinforcement learning.

The Reinforcement Learning Circuit (RLC) on the right side of **Figure 2** is the core function circuit that enables the network to master different rules. The RLC consists of presynaptic and postsynaptic parts, each of which contains several populations of neurons. In **Figure 2**, on the right side of RLC are the presynaptic neuron populations, which receive external stimulation and affect postsynaptic neuron populations in RLC; on the left side are the postsynaptic neuron populations, which receive the projection from the presynaptic neuron populations and transmit the signal to the motor neurons, guiding the motor neurons to fire in a specific order.

In addition to the presynaptic "cue" neuron population and postsynaptic "end" neuron population, the remaining six populations of neurons correspond to each other, divided into three population by row, corresponding to the first, second, and third positions in the sequence, respectively. Dehaene have proposed that a taxonomy of five distinct cerebral mechanisms for sequence coding: transitions and timing knowledge, chunking, ordinal knowledge, algebraic patterns, and nested tree structures (Dehaene et al., 2015b), which inspired us that the ordinal knowledge should be encoded by different populations of neurons.

Therefore, the core of the so-called different "rules" lies in the connection mode of the reinforcement learning circuit. Just as macaque monkeys acquire rules in experiments, the acquisition of rule learning of our networks is also a reinforcement learning process.

Before the experiment was completed in the macaque monkeys (Jiang et al., 2018), these two rules were considered supra-regular rules that only human beings could master. **Figure 2** shows all the components of the network, including memory circuit, motor neurons population, and reinforcement learning circuit. In this case, for the convenience of introduction, we will describe the process to complete the sequences of length three production task.

How SNN can realize reinforcement learning is an open question hitherto, there is some leading research work in this field (Urbanczik and Senn, 2009; Frémaux and Gerstner, 2016; Wang et al., 2018). The main contradiction lies in the current learning rules of SNN synapses, such as STDP, Hebbian, etc., that the time of synaptic update was slightly later than the time of local neuron fire, however, in reinforcement learning, reward/punishment come after a trial. How to build a bridge between reward/punishment and synaptic learning rules such as STDP is where the crux lies (Izhikevich, 2007).

After full investigation, we adopt reward-modulated STDP (R-STDP) to implement the whole experiment due to the excellent biology plausibility (Frémaux and Gerstner, 2016).

The main idea of R-STDP is to modulate the outcome of "standard" STDP by a reward term (Friedrich et al., 2011).

Synaptic eligibility trace (right box in **Figure 3**) stores a temporary memory of the STDP outcome so that it is still available by the time a delayed reward signal is received (Frémaux and Gerstner, 2016). We regard the timing condition (or "learning window") of traditional STDP as $STDP(n_i, n_j)$, n_i and n_j denote the presynaptic and postsynaptic neuron in the network. The synaptic eligibility trace keeps a transient memory in the

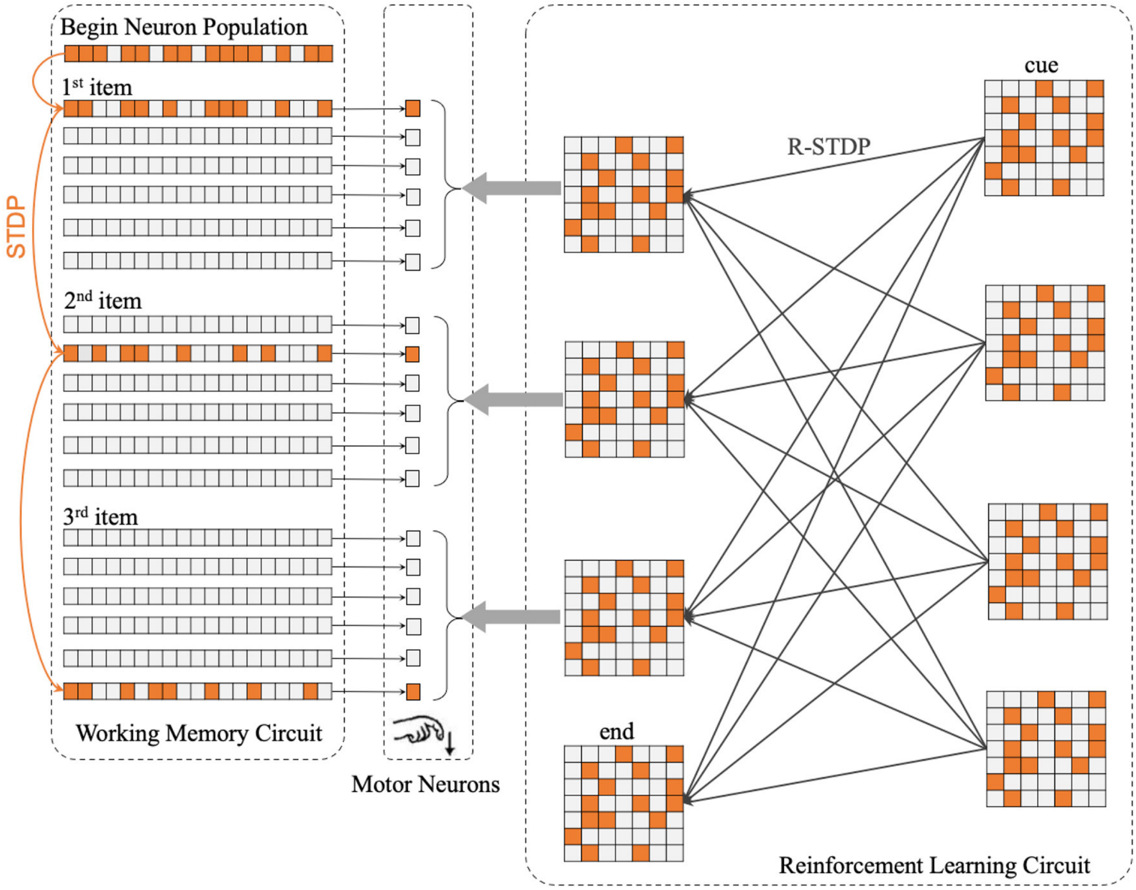


FIGURE 2 | The whole architecture of SP-SNN is divided into three neuron circuits. The orange lines in Working Memory Circuit mean the synapses inner WMC that is formulated by STDP rule. The thin black lines between WMC and Motor Neurons represent every population neurons in WMC project to the same motor neuron, which fire to trigger the output action. The gray arrows between Reinforcement Learning Circuit and Motor Neurons display each population neurons in RLC project to the same motor neuron as well. Moreover, the thick black lines in RLC show the synapses inner RLC. With reinforcement learning, the network will gradually learn different sequence reconstruction rules, which will be reflected in the weight distribution of synaptic connections in RLC.

form of a running average of recent spike-timing coincidences. Synaptic eligibility traces arise from theoretical considerations and effectively bridge the temporal gap between the neural activity and the reward signal.

$$\Delta e_{j,i} = -\frac{e_{j,i}}{\tau_e} + STDP(n_i, n_j) \tag{7}$$

$e_{j,i}$ is the eligibility traces between presynaptic neuron i and postsynaptic neuron j, τ_e is the time constant of the eligibility trace. The running average is equivalent to a low-pass filter. In R-STDP mechanism, the synaptic weight W changes when the neuromodulator M signals exist.

$$\Delta W = M * E \tag{8}$$

Considering the complexity of the network, we simply choose R-max policy i.e., $M = R$. R is the reward or punish signal toward network which is given by the experiment environment. Actually, R is the function of time t, Equation (9) shows how R changes through time.

$$R(t) = \begin{cases} C_r & t - t_r \leq T_R \\ C_p & t - t_p \leq T_R \\ 0 & t - t_r > T_R \\ 0 & t - t_p > T_R \end{cases} \tag{9}$$

C_r and C_p are the constants of reward and punish signal. t_r and t_p denote the latest time of reward and punish. And T_R is the size of time window of reward or punish signal. In the experiment, we set $C_r = 10$, $C_p = -10$, and $T_R = 5$.

Specifically, in the process of a sequence production, macaque monkeys need to memorize symbols sequence firstly, through STDP learning rules to complete the STP of WMC when the production process starts, under the guidance of the start signal, the neurons in WMC corresponding to the symbols in the original sequence fire in an ideal situation. Because there is a corresponding population to target connection between WMC and the motor neuron population, the membrane potential of

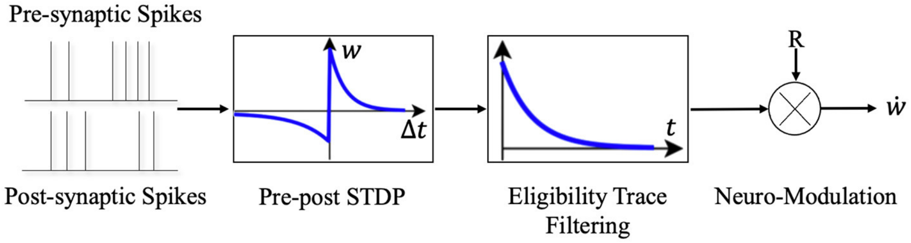


FIGURE 3 | The schematic diagram of Reward-modulated STDP. Different from the general STDP rules, when neurons implements the R-STDP rules the synaptic weights will not be updated once the pre- and post-synaptic neurons generate spike pair, but temporarily stores the variations of weights in the eligibility trace. Only when the reward or punishment signal comes, the corresponding synaptic weights will be updated according to the current value of the eligibility trace and the reward/punishment signal.

Algorithm 1: The learning process of SP-SNN

```
1. Initialize $N_{population} = 50, V_{threshold} = -35mv$, and other parameters of the network
2. Load Training Set(S)
3. Start training procedure
for every sequence in S do
    Memory stage:
    Increase $I_s$ of corresponding populations
    Update weights $W_{WMC}$ with STDP rule by Equation (6)                ▷ STDP rule

    Reinforcement learning stage:
    Increase $I_s$ of begin populations of RLC
    if output correct sequence then
        $R(t) \leftarrow C_r$                                                ▷ Give reward
    else
        $R(t) \leftarrow C_p$                                                ▷ Give punishment
    end if
    Update weights $W_{RLC}$ by Equation (8)                                ▷ R-STDP rule
end for
4. Start test procedure
```

motor neurons corresponding to the symbol increases. Although the membrane potential of these particular motor neurons has not reached the threshold of the action potential, it will be significantly increased compared with other neurons. Once the post-synapse neuron population in the RLC starts to fire frequently, the membrane potential of corresponding motor neurons population will rise quickly until fire. For a different rule, Repeat/Mirror, the network should produce the sequence by a different order, which means different firing order of post-synapse populations in RLC. It is where reinforcement learning makes a difference.

The more detailed learning process of SP-SNN is shown in Algorithm 1.

2.5. The Chunking Mechanism of SP-SNN

Through the design and verification of the macaque monkeys supra-regular rule experimental paradigm, it is found that after intensive training, macaque monkeys can master the supra-regular grammar, breaking the barrier of syntax previously divided (Jiang et al., 2018). Jiang et al. (2018) point out that whether there is a clear boundary between human and animal language competence needs to be discussed in detail again (Fitch, 2018).

Neuroscientists and psychologists have been exploring the Chunking Mechanism for a long time (Ellis, 1996; Gobet et al., 2001; Fujii and Graybiel, 2003). It is generally believed that this mechanism plays an essential role in human short-term working memory (Burtis, 1982), knowledge acquisition (Laird et al., 1984; Gobet, 2005), and even skill learning (Rosenbloom et al., 1989; Pammi et al., 2004). Bibbig et al. (1995) showed that after learning the hippocampus neurons form chunks that are special representations for co-occurrence of neural events in several association areas via computer simulations of a spiking neural network.

Decomposing a long sequence into several shorter sequences to improve the efficiency and accuracy of memory is the core component of chunking mechanism (Ellis, 1996). For example, it is difficult for one person to remember a whole sequence of mobile phone numbers. Instead, the mobile phone number sequence is decomposed into several shorter sequences to realize the memory. Therefore, inspired by Chunking Mechanism in the cognitive process of biological brain, we try to explore

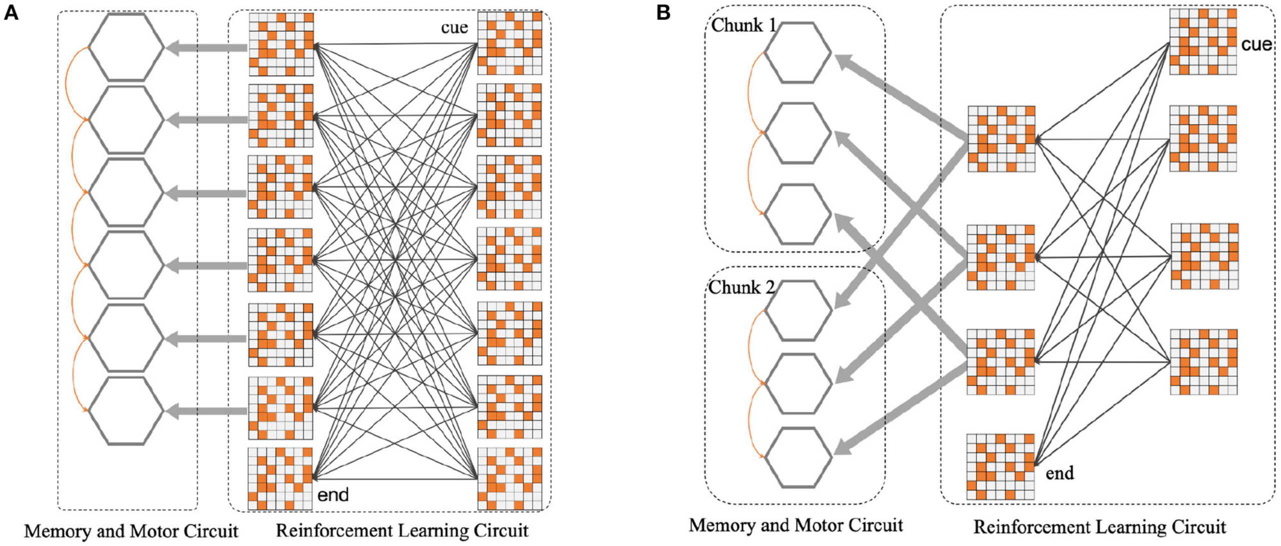


FIGURE 4 | (A) Before the introduction of chunking strategy, the neural network architecture diagram. **(B)** After the introduction of chunking strategy, the neural network architecture diagram. For the convenience of composition, WMC and motor circuit are merged into gray hexagons, each refers to six different position symbols.

the introduction of a biologically similar Chunking Mechanism into the network, and then observe the changes in network performance of sequence representation. Specifically, compared with the network without Chunking Mechanism, the main difference of the new network is the connection mode, i.e., after the introduction of Chunking Mechanism, the long sequence is segmented into several shorter chunks.

For instance, as shown in **Figure 4A**, several populations of neurons corresponding to six position symbols in the WMC fire sequentially according to the order in which symbols appear. The synaptic connections among six gray hexagons will be shaped by STDP rule, representing the memory for a 6-length sequence. **Figure 4B** shows how the network splitting a 6-length sequence into two 3-length chunks. Based on this instance, we completed the construction of spike neural network in the form of **Figures 4A,B** in the follow-up experiment, to explore the influence of chunking mechanism.

3. EXPERIMENT

3.1. Sequence Memory With Population Coding and STDP

First of all, we completed the construction of WMC, whose structure is shown in the **Figure 1**.

The experimental process is divided into two stages: memory stage and test stage. In the memory stage, the original sequence was repeatedly displayed to SP-SNN several times. Each time a position symbol appears, the corresponding population of neurons is stimulated by external current stimuli and fire. Here we use the Poisson encode to activate the neuron population. The sequential firing of different populations of neurons combined with the STDP rule formed specific synaptic connections, thus forming the memory of specific sequence. This process is shown in **Figure 5A** for 0–400 ms.

In the test stage, as shown in **Figure 5A** for 600–800 ms, we only give the network a start signal, i.e., activate the "begin" neuron population, and then let the network independent work without any external stimulation, and observe the firing state of the network. When more than half of the neurons in a neuron population fired, it is considered that the network outputs corresponding symbol. Only when all the symbols in the sequence are output in the correct order can we think that the sequence is correctly memorized. **Figure 5B** shows the synaptic connection between neurons after one trial of sequence memory experiment.

In Working Memory Circuit (WMC), all connections are initialized to a small random number very close to 0. Neural plasticity occurs between different populations of neurons to realize sequence memory, and weak synaptic connections are maintained within the neuron population.

In order to better understand whether the population coding strategy contributes to the robustness of the model, we tested the recall accuracy of the WMC toward 3-length sequence with and without population coding strategy for different intensities of background noise, which is widely present in the human brain (Hidaka et al., 2000; Mišic et al., 2010). We introduce Gaussian white noise as the background noise of the network.Each neuron receives a stimulus current of Gaussian white noise, that is, a random variable with a mean value of 0 and a variance of σ^2. According to the definition, the noise intensity of white Gaussian noise equals to σ^2. We test the accuracy of the network under different noise intensity.

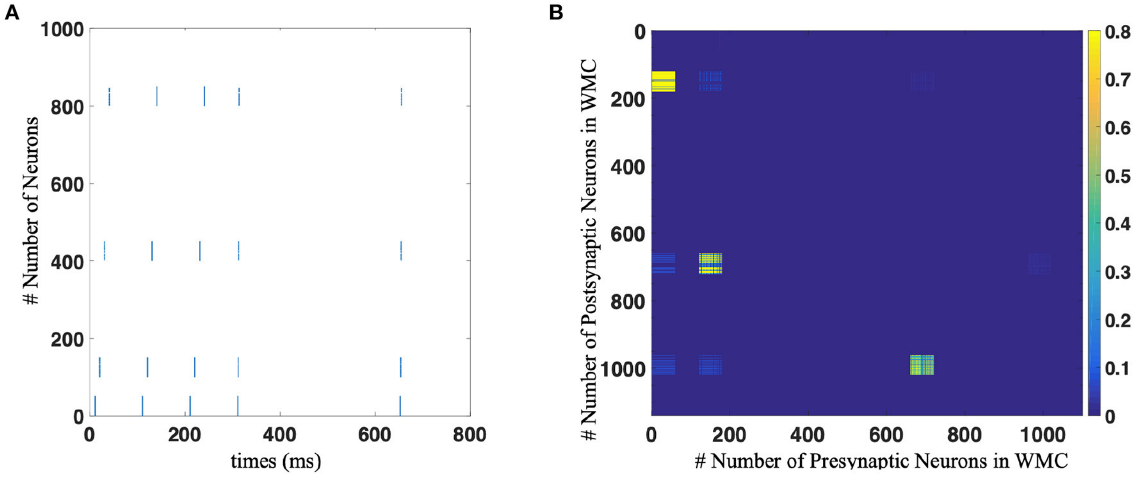


FIGURE 5 | (A) The spike trains of neural network of working memory task with population coding strategy. Each blue dot indicates that the neurons corresponding to the vertical axis discharge at the time node corresponding to the horizontal axis. The number of neuron in population is 60 in this implementation. **(B)** The weights distribution of neural network after implement STDP learning.

Intuitively, as **Figure 6** shown, we found with the increase of *C* (the number of single neurons population), the noise resistance of the model becomes more robust. When we increase *C* from 1 to 30 and then to 60, the accuracy of the model is greatly improved under the same noise conditions. Population coding strategy can indeed significantly improve the robustness of the model. However, when *C* continues to increase to 80 or even 100, the accuracy of the model increases slightly. It is interesting to explore the cause, we will discuss this phenomenon in the next chapter. Considering the computational complexity and experiment, we set $C = 60$ to carry out the follow-up experiments.

3.2. Sequence Production

Then, we completed the whole network test, including WMC, MC, and RLC, and demonstrated that SNN could reproduce the sequence according to different rules. For example, for the reverse production of a sequence of length 3, the spike trains and the strength of synaptic connections between different populations of neurons are shown in the **Figure 7**.

Compared with the result in Jiang et al. (2018), they found that the accuracy of macaque monkeys in the process of production sequences according to Mirror rules after the acquisition of rules is U-shaped as Figure 5A in Jiang et al. (2018) shown. Neuroscientists claim that the main reason causes this phenomenon are: the superposition of primacy and recency effect, which are considered essential in the process of evolution (Luchins, 1957; Jiang et al., 2018).

It is a well-established finding that the items at the beginning or at the end of the list are more likely to be recalled than the items in the middle of that list, which are termed the primacy and recency effects (Stewart et al., 2004). And both primacy and recency effects can be obtained in nonhuman primates (Castro and Larsen, 1992).

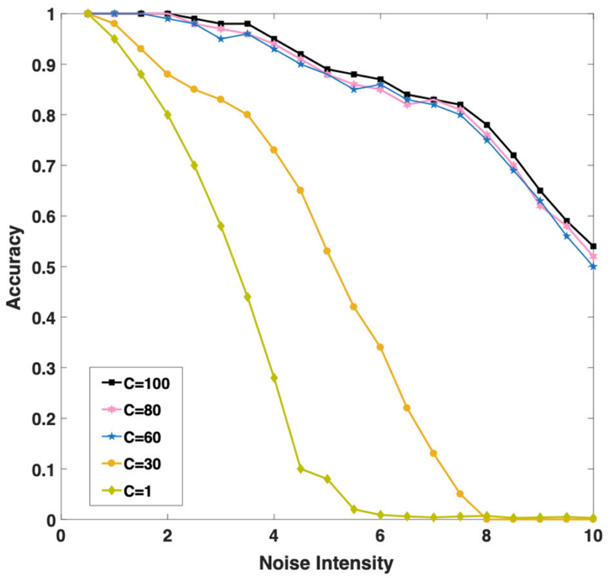


FIGURE 6 | Three different color curves show that the memory accuracy of the network for the sequence decreases with the background noise enhancement when *C* (the number of single neurons population) takes different values.

In our practical simulation experiment, we found that the accuracy of the production of different position symbols is close to 100% after the network acquires specific rules, which is difficult to reflect the difference of different positions in the sequence. Therefore, in order to compare with biological experiments, we pull-in the background noise based on the

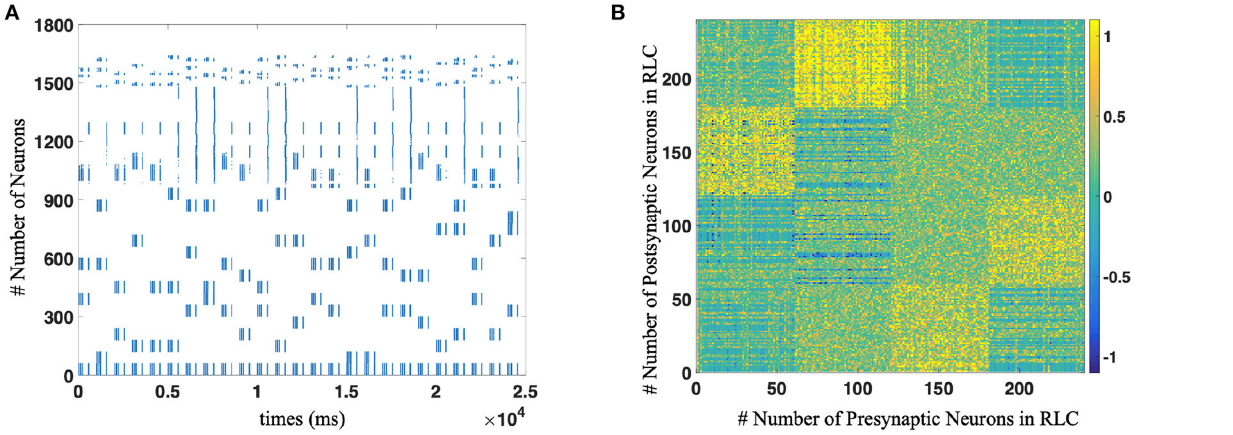


FIGURE 7 | (A) The spike trains of the whole neural network during R-STDP training process. Each blue dot indicates that the neurons corresponding to the vertical axis discharge at the time node corresponding to the horizontal axis. As for the number of neurons, WMC is composed of No. 1-979 neurons, No. 980–997 are motor neurons and RLC consists of No. 998 1639 neurons, respectively. **(B)** The weights distribution of Reinforcement Learning Circuit after implement R-STDP learning, which contains the supra-regular grammars (Mirror rule in this figure).

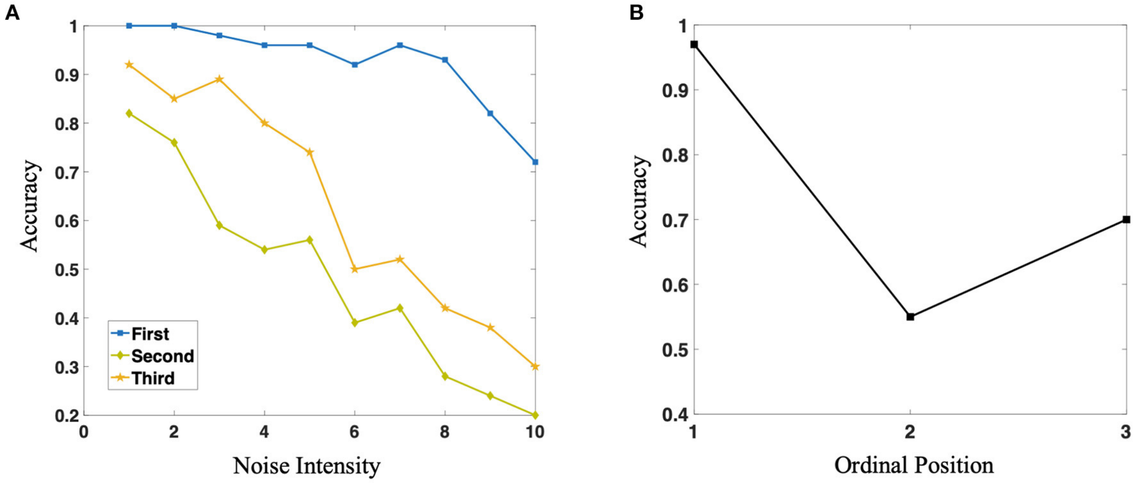


FIGURE 8 | (A) Three different color curves represent the reconstruction accuracy of the network for three different positions in the sequence, respectively. With the increase of noise intensity, the accuracy of the three positions show a downward trend, but they always maintain the relationship of U-shape. **(B)** The average value of production accuracy of neural network for three different positions under different noise intensity.

original network. What we are very excited about is: as shown in **Figure 8A**, with the increase of noise, the accuracy of different position symbols in the network production sequence is gradually decreasing, but the production accuracy of different position symbols in the sequence has always maintained this U-shaped structure, which shows that the network structure and connection structure we constructed is highly biologically interpretable. Furthermore, **Figure 8B** shows the average accuracy of three positions. It is of some enlightenment to further understand how macaque monkeys can complete the task of sequence production and break the grammatical barrier.

3.3. Sequence Production With Chunking Mechanism

For the network structure after the introduction of Chunking Mechanism, since the production rules are consistent within each chunk, different chunks can share a population of reinforcement learning postsynaptic neurons, as shown in **Figure 4B**.

Individually, in this case, in the sequence memory stage, six symbols appear sequentially and are cut into two chunks of 3+3. Two chunks form specific connections to complete local memory. In the sequence production stage, the RLC begins to work, and in conjunction with the WMC, correct motor neurons fire, completing the so-called sequence production. During

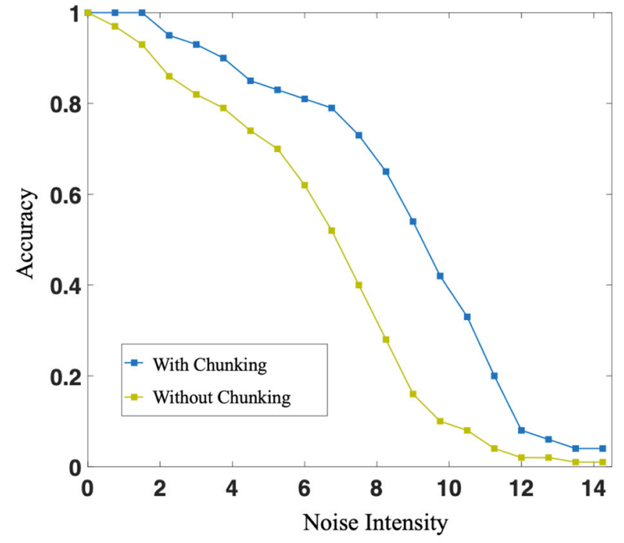


FIGURE 9 | The blue curve and the green curve represent the production accuracy of neural network with or without chunking strategy, respectively. With the increase of noise intensity, both accuracy decreases. However, the accuracy of networks with chunking strategy is always higher than that without chunking strategy.

the experiment, we trained the original network, as shown in **Figure 4A**, and the network after introducing Chunking Mechanism, as shown in **Figure 4B**, respectively. Then, we compared the difference in production accuracy between the two networks under different noise conditions, the result is shown in **Figure 9**. As the noise intensity increases, the accuracy of both networks decreases. However, the accuracy of Chunking Mechanism network is always higher than that of the original network, which fully reflects the vital role Chunking Mechanism plays in this task. We will discuss the reasons for Chunking Mechanism's role in detail in section 4.

4. DISCUSSION

We demonstrated that through the fusion of neuron population coding, STDP synaptic learning mechanism, and reinforcement learning mechanism (R-STDP), the SNN network could perform the same ability as macaque monkeys to construct the sequences according to super-regular rules, which was previously considered unique to humans. As far as we know, our work is the first to complete the research of sequence production based on SNN in accordance with supra-regular rules at the computational level.

Inspired by the research about "grandmother cells" in neuroscience (Bowers, 2009; Quiroga, 2012), we proposed to use the activation of a population of neurons to represent the emergence of a specific symbol in the brain, which may be caused by extra stimuli (correspond with the memory stage), or by the current from other populations of neurons within the network (correspond with the sequence production stage). In the process of experiment, we found that the population coding has stronger robustness and stability than the single neuron coding. However, when the number of neurons in the group reaches a certain degree, the robustness of the network tends to be stable and will not grow infinitely. How the brain sets the size of neuron populations to balance the robustness and consumption will be a very interesting topic in our future research.

Inspired by the training process toward macaque monkeys in the experimental paradigm of sequence production (Jiang et al., 2018), we introduced the reinforcement learning mechanism in super-regular rule learning, mainly using the R-STDP mechanism with eligibility traces method (Frémaux and Gerstner, 2016). Before defining the structure, we thought the difficulty of this work is that every time a network (or monkey) gets a reward or punishment, it is easy for the network (or monkey) to link the symbols with the reward and punishment signals, actually the production rules behind the symbols are related to the reward and punishment signals. In our experiments, we substantially helped the network to complete the transition from symbols to rules behind symbols, that was, reward and punishment signals are associated with rules, not just with symbols themselves. In the future work, how to let the network automatically complete this process, rather than directly tell the network in a priori way, is a very worthy of study.

In the experimental results of Jiang et al. (2018), the accuracy of symbol production for different positions in the sequence is different, and generally presents a U-shaped rule, i.e., the effect of sequence production in the middle of the sequence is weaker than that at the beginning and the end. This feature is considered to be an essential feature in the evolution process. Psychologists and cognitive scientists believe that this phenomenon is the result of the superposition of the primacy and recency effect (Luchins, 1957; Stewart et al., 2004).

During the experiment, our proposed network structure was consistent with macro-cognitive behavior of macaque monkeys, showing the accuracy of U-type production. The reason why the network can show U-type accuracy is that our proposed network structure combined the "primary effect" and "recency effect" simultaneously.

Respectively, as for the "primary effect," in WMC, the neuron populations corresponding to the first symbol in the sequence is stimulated by the "begin" neuron population, and most of the neurons in the "begin" neuron population are firing in the specific time window belong to "begin" population, which leads to the firing of the first symbol neuron population will be more intense, directly leading to the greater increase of membrane potential of the corresponding motor neuron of the first symbol. Therefore, the first symbol has better noise resistance, causing the so-called "primary effect."

About the "recency effect," because the building block of spike neural network is a LIF neuron model, and the LIF neuron model has a leakage mechanism (Miller, 2018), the membrane potential of the pre-activated motor neuron gradually decreases with the passage of time, and the last sign appears because of its shortest production time, and its membrane potential decreases the least, resulting in the "proximal effect."

Inspired by Chunking Mechanism in the biological brain, we implemented the Chunking Mechanism based on SNN in the experiment and verified that it plays a vital role in improving the accuracy of production. After analysis, we found that Chunking Mechanism can shorten the sequence length of the production process. In the example of **Figure 4**, the original network without Chunking Mechanism needs to recall the whole sequence of length six first, and then produce it with RLC. Because of the leakage characteristics of motor neurons, the membrane potential of motor neurons at corresponding locations decreases dramatically, which makes it easier to make mistakes. While Chunking Mechanism is introduced, hardly when every chunk is recalled, it will be produced immediately. Compared to the original network, the duration of the decline of the membrane potential of the new network motor neurons will be shorter, and the corresponding membrane potential will be higher, causing the higher accuracy.

However, we must admit that the chunking mechanism used in this work still has some limitations. The main limitation is that we implicitly help SP-SNN to divide a long sequence into several chunks, that is, to decompose a sequence of length six into two subsequences of length 3. However, in the actual human cognitive process, cutting long sequences into chunks has substantial autonomy and flexibility. How can SNN complete this process spontaneously? How do different segmentation methods, such as equal-length segmentation and unequal-length segmentation, affect the cognitive process's results? These are the problems worthy of exploration in the future. However, our work still completes the preliminary exploration of the chunking mechanism and demonstrates that the chunking mechanism is of great help to improve the model's robustness.

The following will discuss the difference between our work and the current popular artificial neural network or deep learning. The difference mainly consists of three parts.

First, almost all the current artificial neural networks set the weight between neurons to be fixed at the inference stage (only changes when the network is training), which is different from the real nervous system. In the real nervous system, the connection between neurons will be affected by the strength of input signal, time process and other factors, temporary change with neuron activity, which is called short-term plasticity of the synapse (STP), also known as dynamic connection (Markram and Tsodyks, 1996; Fung et al., 2015).

From the computational point of view, STP provides the biological neural network one more time dimension in information processing than the artificial neural network with a fixed weight, so it has more computational potential and can perform complex cognitive tasks. From this point of view, it is obvious that in our network WMC adopts short-term plasticity (STP). While the synapse in RLC changes in the long term, which can be looked like a particular kind of long-term plasticity (LTP). Compared with artificial neural network, our network integrates STP and LTP mechanism, makes full use of time dimension, and to a certain extent, expands the boundary of SNN's information processing capacity.

Second, although the neural network trained by the current deep learning technology can solve some specific problems, its plausibility is very poor (Castelvecchi, 2016), which leads to serious security problems (such as Adversarial Examples Problem) and becomes a black cloud over the head of deep learning (Goodfellow et al., 2014; Liu et al., 2016).

However, our model is totally different. In the experiment, we can check the firing state and weight distribution of network at any time, which can be used to judge the working memory of the current network, the learning process of rules, and so on. That is to say, the network we build is completely interpretable. Although the complexity of our model is not comparable to the current deep learning technology, our work may bring some inspiration for the construction of more interpretable artificial intelligence system in the future.

Finally, for researchers in the field of neuroscience and cognitive science, our work provides a new perspective to some extent, that is, how to use the neural network of connectionism to represent the symbol reasoning of symbolism. In this work, we demonstrate the feasibility of using spike neural network to complete the task of production sequence according to supra-regular rules, breaking the "syntax barrier" of animals. For further explore the representation of symbols in the animal brain, as well as how non-human primates such as macaque monkeys complete the task of sequence representation, our work lay the foundation of computing.

5. CONCLUSION

This paper proposed a Brain-inspired Sequence Production Spiking Neural Network (SP-SNN) to model the Sequence Production process, inspired by a biological experiment paradigm which showed that macaques could be trained to produce embedded sequences involved supra-regular grammars.

After experimental verification, we demonstrated that SP-SNN could also handle embedded sequence production tasks, striding over the "syntax barrier." SP-SNN coordinates STP and LTP mechanisms simultaneously. As for STP, Population-Coding and STDP mechanisms realize working memory. As for LTP, the R-STDP mechanism shapes Reinforcement Learning Circuit for different supra-regular grammars, whose synaptic weights do not change until a reward/punishment occurs. The U-shape accuracy of the results of SP-SNN and macaque, which is caused by the superposition of primacy and recency effect, further strengthened the biological plausibility of SP-SNN. Besides, we found the chunking mechanism, i.e., divides a long sequence into several subsequences, indeed makes a difference to improve our model's robustness.

As far as we know, our work is the first one toward the "syntax barrier" in the SNN field. In future research, we hope to compare the electrical activity of SP-SNN with the electrophysiological data of macaque in the sequence production task to expose more underlying animals' neural mechanisms in this cognitive process.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

HF and YZ designed the study and performed the experiments. HF, YZ, and FZ developed the algorithm, performed the analysis of the results, and wrote the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant No. XDB32070100),the new generation of artificial intelligence major project of the Ministry of Science and Technology of the People's Republic of China (Grant No. 2020AAA0104305), the Beijing Municipal Commission of Science and Technology (Grant No. Z181100001518006), and the Beijing Academy of Artificial Intelligence (BAAI).

ACKNOWLEDGMENTS

The authors appreciate Dongcheng Zhao, Qian Zhang, Cunqing Huangfu, and Qian Liang for valuable discussions. The authors would like to thank all the reviewers for their help in shaping and refining the paper.

REFERENCES

Abbott, L. F., Varela, J., Sen, K., and Nelson, S. (1997). Synaptic depression and cortical gain control. *Science* 275, 221–224. doi: 10.1126/science.275.5297.221

Bi, G.-Q. and Poo, M.-M. (1998). Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. *J. Neurosci*. 18, 10464–10472. doi: 10.1523/JNEUROSCI.18-24-10464.1998

Bi, G.-Q., and Poo, M.-M. (2001). Synaptic modification by correlated activity: Hebb's postulate revisited. *Annu. Rev. Neurosci*. 24, 139–166. doi: 10.1146/annurev.neuro.24.1.139

Bibbig, A., Wennekers, T., and Palm, G. (1995). A neural network model of the cortico-hippocampal interplay and the representation of contexts. *Behav. Brain Res*. 66, 169–175. doi: 10.1016/0166-4328(94)00137-5

Bowers, J. S. (2009). On the biological plausibility of grandmother cells: implications for neural network theories in psychology and neuroscience. *Psychol. Rev*. 116:220. doi: 10.1037/a0014462

Burtis, P. (1982). Capacity increase and chunking in the development of short-term memory. *J. Exp. Child Psychol*. 34, 387–413. doi: 10.1016/0022-0965(82)90068-6

Castelvecchi, D. (2016). Can we open the black box of AI? *Nat. News* 538:20. doi: 10.1038/538020a

Castro, C. A., and Larsen, T. (1992). Primacy and recency effects in nonhuman primates. *J. Exp. Psychol*. 18:335. doi: 10.1037/0097-7403.18.4.335

Chomsky, N. (1957). *Syntactic Structures*. The Hague: Mouton. doi: 10.1515/9783112316009

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press. doi: 10.21236/AD0616323

Dan, Y., and Poo, M.-M. (2004). Spike timing-dependent plasticity of neural circuits. *Neuron* 44, 23–30. doi: 10.1016/j.neuron.2004.09.007

Dan, Y., and Poo, M.-M. (2006). Spike timing-dependent plasticity: from synapse to perception. *Physiol. Rev*. 86, 1033–1048. doi: 10.1152/physrev.00030.2005

Deacon, T. W. (1998). *The Symbolic Species: The Co-Evolution of Language and the Brain*. New York, NY: WW Norton & Company.

Dehaene, S., Meyniel, F., Wacongne, C., Wang, L., and Pallier, C. (2015a). The neural representation of sequences: from transition probabilities to algebraic patterns and linguistic trees. *Neuron* 88, 2–19. doi: 10.1016/j.neuron.2015.09.019

Dehaene, S., Meyniel, F., Wacongne, C., Wang, L., and Pallier, C. (2015b). The neural representation of sequences: from transition probabilities to algebraic patterns and linguistic trees. *Neuron* 88, 2–19.

Ellis, N. C. (1996). Sequencing in SLA: phonological memory, chunking, and points of order. *Stud. Second Lang. Acquis*. 18, 91–126. doi: 10.1017/S0272263100014698

Fitch, T. (2004). Computational constraints on syntactic processing in a nonhuman primate. *Science* 303, 377–380. doi: 10.1126/science.1089401

Fitch, T. (2014). Toward a computational framework for cognitive biology: unifying approaches from cognitive neuroscience and comparative cognition. *Phys. Life Rev*. 11, 329–364. doi: 10.1016/j.plrev.2014.04.005

Fitch, T. (2018). Bio-linguistics: monkeys break through the syntax barrier. *Curr. Biol*. 28, R695–R697. doi: 10.1016/j.cub.2018.04.087

Fitch, T., and Friederici, A. D. (2012). Artificial grammar learning meets formal language theory: an overview. *Philos. Trans. R. Soc. B Biol. Sci*. 367, 1933–1955. doi: 10.1098/rstb.2012.0103

Frémaux, N., and Gerstner, W. (2016). Neuromodulated spike-timing-dependent plasticity, and theory of three-factor learning rules. *Front. Neural Circ*. 9:85. doi: 10.3389/fncir.2015.00085

Friedrich, J., Urbanczik, R., and Senn, W. (2011). Spatio-temporal credit assignment in neuronal population learning. *PLoS Comput. Biol*. 7:e1002092. doi: 10.1371/journal.pcbi.1002092

Fujii, N., and Graybiel, A. M. (2003). Representation of action sequence boundaries by macaque prefrontal cortical neurons. *Science* 301, 1246–1249. doi: 10.1126/science.1086872

Fung, C. A., Wong, K. M., Mao, H., Wu, S., et al. (2015). Fluctuation-response relation unifies dynamical behaviors in neural fields. *Phys. Rev. E* 92:022801. doi: 10.1103/PhysRevE.92.022801

Gobet, F. (2005). Chunking models of expertise: implications for education. *Appl. Cogn. Psychol*. 19, 183–204. doi: 10.1002/acp.1110

Gobet, F., Lane, P. C., Croker, S., Cheng, P. C., Jones, G., Oliver, I., et al. (2001). Chunking mechanisms in human learning. *Trends Cogn. Sci*. 5, 236–243. doi: 10.1016/S1364-6613(00)01662-4

Goodfellow, I. J., Shlens, J., and Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv [Preprint]. arXiv:1412.6572.*

Hauser, M. D., Chomsky, N., and Fitch, T. (2002). The faculty of language: what is it, who has it, and how did it evolve? *Science* 298, 1569–1579. doi: 10.1126/science.298.5598.1569

Hidaka, I., Nozaki, D., and Yamamoto, Y. (2000). Functional stochastic resonance in the human brain: noise induced sensitization of baroreflex system. *Phys. Rev. Lett*. 85:3740. doi: 10.1103/PhysRevLett.85.3740

Hodgkin, A. L., and Huxley, A. F. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. *Bull. Math. Biol*. 52, 25–71. doi: 10.1016/S0092-8240(05)80004-7

Izhikevich, E. M. (2003). Simple model of spiking neurons. *IEEE Trans. Neural Netw*. 14, 1569–1572. doi: 10.1109/TNN.2003.820440

Izhikevich, E. M. (2007). Solving the distal reward problem through linkage of stdp and dopamine signaling. *Cereb. Cortex* 17, 2443–2452. doi: 10.1093/cercor/bhl152

Jiang, X., Long, T., Cao, W., Li, J., Dehaene, S., and Wang, L. (2018). Production of supra-regular spatial sequences by macaque monkeys. *Curr. Biol*. 28, 1851–1859. doi: 10.1016/j.cub.2018.04.047

Laird, J. E., Rosenbloom, P. S., and Newell, A. (1984). "Towards chunking as a general learning mechanism," in *AAAI*, Pittsburgh, PA, 188–192.

Liu, Y., Chen, X., Liu, C., and Song, D. (2016). Delving into transferable adversarial examples and black-box attacks. *arXiv [Preprint]. arXiv:1611.02770.*

Luchins, A. S. (1957). "Primacy-recency in impression formation," in *Order of Presentation in Persuasion*, ed C. I. Hovland (New Haven, CT: Yale University Press), 33–61.

Markram, H., and Tsodyks, M. (1996). Redistribution of synaptic efficacy between neocortical pyramidal neurons. *Nature* 382, 807–810. doi: 10.1038/382807a0

Matsuzawa, T. (2013). Evolution of the brain and social behavior in chimpanzees. *Curr. Opin. Neurobiol*. 23, 443–449. doi: 10.1016/j.conb.2013.01.012

Miles, H. (1990). "The cognitive foundations for reference in a signing orangutan," in *"Language" and Intelligence in Monkeys and Apes: Comparative Developmental Perspectives*, eds S. T. Parker and K. R. Gibson (Cambridge: Cambridge University Press), 511–539.

Miller, P. (2018). *An Introductory Course in Computational Neuroscience*. Cambridge, MA: MIT Press. doi: 10.7551/mitpress/9780262533287.003.0011

Mišic, B., Mills, T., Taylor, M. J., and McIntosh, A. R. (2010). Brain noise is task dependent and region specific. *J. Neurophysiol*. 104, 2667–2676. doi: 10.1152/jn.00648.2010

Mooney, R. (2009). Neural mechanisms for learned birdsong. *Learn Mem*. 16, 655–669. doi: 10.1101/lm.1065209

Pammi, V. C., Miyapuram, K. P., Bapi, R. S., and Doya, K. (2004). "Chunking phenomenon in complex sequential skill learning in humans," in *International Conference on Neural Information Processing* (Berlin; Heidelberg: Springer), 294–299. doi: 10.1007/978-3-540-30499-9_44

Pinker, S. (2003). *The Language Instinct: How the Mind Creates Language*. City of Westminster; London: Penguin UK.

Quiroga, R. Q. (2012). Concept cells: the building blocks of declarative memory functions. *Nat. Rev. Neurosci*. 13, 587–597. doi: 10.1038/nrn3251

Rosenbloom, P. S., Laird, J. E., and Newell, A. (1989). "The chunking of skill and knowledge," in *Working Models of Human Perception*, eds B. A. G. Elsendoorn and H. Bouma (Einhoven: Academic Press Toronto, ON), 391–410. doi: 10.1016/B978-0-12-238050-1.50024-X

Stewart, D. D., Stewart, C. B., Tyson, C., Vinci, G., and Fioti, T. (2004). Serial position effects and the picture-superiority effect in the group recall of unshared information. *Group Dyn*. 8:166. doi: 10.1037/1089-2699.8.3.166

Urbanczik, R., and Senn, W. (2009). Reinforcement learning in populations of spiking neurons. *Nat. Neurosci*. 12, 250–252. doi: 10.1038/nn.2264

Wang, J. X., Kurth-Nelson, Z., Kumaran, D., Tirumala, D., Soyer, H., Leibo, J. Z., et al. (2018). Prefrontal cortex as a meta-reinforcement learning system. *Nat. Neurosci*. 21, 860–868. doi: 10.1038/s41593-018-0147-8

Wichterle, H., Lieberam, I., Porter, J. A., and Jessell, T. M. (2002). Directed differentiation of embryonic stem cells into motor neurons. *Cell* 110, 385–397. doi: 10.1016/S0092-8674(02)00835-8

Wittenberg, G. M., and Wang, S. S.-H. (2006). Malleability of spike-timing-dependent plasticity at the ca3-ca1 synapse. *J. Neurosci*. 26, 6610–6617. doi: 10.1523/JNEUROSCI.5388-05.2006

Yang, C. (2016). *The Price of Linguistic Productivity: How Children Learn to Break the Rules of Language*. Cambridge, MA: MIT Press. doi: 10.7551/mitpress/9780262035323.001.0001

Yi, H. G., Leonard, M. K., and Chang, E. F. (2019). The encoding of speech sounds in the superior temporal gyrus. *Neuron* 102, 1096–1110. doi: 10.1016/j.neuron.2019.04.023

Zucker, R. S., and Regehr, W. G. (2002). Short-term synaptic plasticity. *Annu. Rev. Physiol*. 64, 355–405. doi: 10.1146/annurev.physiol.64.092501.114547

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Computational Neuroscience

ORIGINAL RESEARCH
published: 19 February 2021
doi: 10.3389/fncom.2021.637144

Transfer of Learning in the Convolutional Neural Networks on Classifying Geometric Shapes Based on Local or Global Invariants

Yufeng Zheng [1]****, Jun Huang*** [2]***, Tianwen Chen*** [2]***, Yang Ou*** [2] ***and Wu Zhou*** [2]******

[1] Department of Data Science, University of Mississippi Medical Centre, Jackson, MS, United States, [2] Department of Otolaryngology-Head and Neck Surgery, University of Mississippi Medical Centre, Jackson, MS, United States

The convolutional neural networks (CNNs) are a powerful tool of image classification that has been widely adopted in applications of automated scene segmentation and identification. However, the mechanisms underlying CNN image classification remain to be elucidated. In this study, we developed a new approach to address this issue by investigating transfer of learning in representative CNNs (AlexNet, VGG, ResNet-101, and Inception-ResNet-v2) on classifying geometric shapes based on local/global features or invariants. While the local features are based on simple components, such as orientation of line segment or whether two lines are parallel, the global features are based on the whole object such as whether an object has a hole or whether an object is inside of another object. Six experiments were conducted to test two hypotheses on CNN shape classification. The first hypothesis is that transfer of learning based on local features is higher than transfer of learning based on global features. The second hypothesis is that the CNNs with more layers and advanced architectures have higher transfer of learning based global features. The first two experiments examined how the CNNs transferred learning of discriminating local features (square, rectangle, trapezoid, and parallelogram). The other four experiments examined how the CNNs transferred learning of discriminating global features (presence of a hole, connectivity, and inside/outside relationship). While the CNNs exhibited robust learning on classifying shapes, transfer of learning varied from task to task, and model to model. The results rejected both hypotheses. First, some CNNs exhibited lower transfer of learning based on local features than that based on global features. Second the advanced CNNs exhibited lower transfer of learning on global features than that of the earlier models. Among the tested geometric features, we found that learning of discriminating inside/outside relationship was the most difficult to be transferred, indicating an effective benchmark to develop future CNNs. In contrast to the "ImageNet" approach that employs natural images to train and analyze the CNNs, the results show proof of concept for the "ShapeNet" approach that employs well-defined geometric shapes to elucidate the strengths and limitations of the computation in CNN image classification. This "ShapeNet" approach will also provide insights into understanding visual information processing the primate visual systems.

Keywords: ShapeNet, topological perception, convolutional neural network (CNN), global feature, shape classification

OPEN ACCESS

Edited by:
Ke Zhou,
Beijing Normal University, China

Reviewed by:
Adel Said Elmaghraby,
University of Louisville, United States
Erik Blasch,
Air Force Research Laboratory,
United States

****Correspondence:***
Yufeng Zheng
yzheng@umc.edu
Wu Zhou
wzhou@umc.edu

Received: *02 December 2020*
Accepted: *01 February 2021*
Published: *19 February 2021*

Citation:
Zheng Y, Huang J, Chen T, Ou Y and Zhou W (2021) Transfer of Learning in the Convolutional Neural Networks on Classifying Geometric Shapes Based on Local or Global Invariants. Front. Comput. Neurosci. 15:637144. doi: 10.3389/fncom.2021.637144

INTRODUCTION

Over the past six decades, investigations of visual system anatomy, physiology, psychophysics and computation have resulted in a general model of vision, which begins from extracting the local features of the retinal images in the lower visual areas [e.g., Lateral Geniculate Nucleus (LGN), V1], then integrates the local features to extract the *global features* in the higher visual areas (e.g., V4 and IT) (Hubel and Wiesel, 1977; Marr, 1982). The convolutional neural networks (CNNs) are primarily inspired by this local-to-global hierarchical architecture of the visual pathways. Similar visual neurons that encode visual properties of a special region of the visual field (i.e., receptive fields), the CNN units perform computations using inputs from special regions of the image and the receptive fields of units at different CNN layers exhibit different properties. With roots in biology, math and computer science, the CNNs have been the most influential innovation in the field of computer vision and artificial intelligence (AI). The CNNs can be trained to classify natural images with accuracies comparable to or better than humans. It has become the core of top companies' services, such as Facebook's automatic tagging algorithms, Google's photo search, and Amazon's product recommendations.

Despite the commercial success of the CNNs, however, little is known about how the CNNs achieve image classification and whether there are inherent limitations. This knowledge is important for avoiding catastrophic errors of the CNN applications in critical areas. To provide insight into the CNNs limitations vs. advantages, we developed a new approach of training and testing the CNNs, which is an alternative to the popular ImageNet approach. The body of literatures (Liu et al., 2018; Hussain et al., 2018) reported the performance of CNNs transfer learning based on image classification. Instead of using natural images to train and test the CNNs, we employed geometric shapes as the training and testing datasets (Zheng et al., 2019). In addition to training the CNNs to perform shape classification tasks, we focused on assessing how the CNN learning in the training datasets is transferred to new datasets (i.e., transfer datasets), which have new shapes that share local/global features with the training datasets. By varying the train and transfer datasets, we will be able to determine whether a local/global feature is extracted by the CNNs during the learning process. The goal was to directly test two hypotheses on CNN image classification. The first hypothesis is that transfer of learning based on local features is higher than transfer of learning based on global features. The second hypothesis is that the CNNs with advanced architectures have higher transfer of learning based on global features. In this study, we analyzed transfer of learning in four representative CNN models, i.e., AlexNet, VGG-19, ResNet-101, and InceptionResNet-v2, which have been trained on the ImageNet and achieved high accuracies in classifying natural images. We found that the results rejected the two hypotheses. Although preliminary, the present study provided proof of concept for this new "ShapeNet" approach.

CONVOLUTIONAL NEURAL NETWORKS

Overview of Convolutional Neural Networks

In this study, four representative CNN models were tested, including the first deep-CNN (AlexNet), a significantly improved CNN model (VGG-19), and two milestones of the advanced CNNs (ResNet-101 and Inception-ResNet-v2). Their characteristics are summarized in **Table 1**. All the CNN models take color images as inputs, thus three-channel grayscale images are created for training. All shape images are scaled to proper size according to each model prior to training and testing. The four CNNs have been pretrained with the ImageNet.

AlexNet

AlexNet (Krizhevsky et al., 2012) is a deep CNN for image classification that won the ILSVRC (The ImageNet Large Scale Visual Recognition Challenge) 2012 competition (Russakovsky et al., 2015). It was the first model performed so well on the historically difficult ImageNet. AlexNet has eight layers with a total of 63 M parameters (**Table 1**). The first five layers are convolutional and the last three layers are fully connected. The AlexNet uses *Relu* instead of *Tanh* to add non-linearity and accelerates the speed by six times at the same accuracy. It uses *dropout* instead of *regularization* to deal with overfitting. AlexNet was trained using batch *stochastic gradient descent* (SGD), with specific values for *momentum* and weight decay.

VGG-19

Simonyan and Zisserman (2014) created a 19-layer (16 conv., 3 fully-connected) CNN that strictly used 3 × 3 filters with stride and pad of 1, along with 2 × 2 max-pooling layers with stride 2, called VGG-19 model[1] To reduce the number of parameters in such a deep network, it uses small 3 × 3 filters in all convolutional layers and best utilized with its 7.3% error rate. The VGG-19 has a total of 143.7M parameters. As the winner of ILSVRC 2015, it is one of the most influential models because it reinforced the notion that the CNNs need to have a deep network of layers for hierarchical representation of visual data.

ResNet-101 and Inception-ResNet-v2

As the winner of ILSVRC 2015, the ResNet-101 (He et al., 2016) has 101 layers, consisting 33 three-layer *residual* blocks plus input and output layers. *Identity connections* learn incremental, or residual, representations, which creates a path for back-propagation. The identity layers gradually transform from simple to complex. Such evolution occurs if the parameters for the $f(x)$ part begin at or near zero. The residual block helps overcome the hard training problem in DeepNet ($>$ 30 layers) due to vanishing gradients. The ResNet-101 model uses 3 × 3 filters with stride of 2, and 3 × 3 max-pooling layers with stride 2.

Inception-ResNet-v2 is a hybrid inception version with residual connections, which leads to dramatically improved recognition performance and training speed in contrast with

[1] Very Deep Convolutional Networks for Large-Scale Visual Recognition, http://www.robots.ox.ac.uk/\simvgg/research/very_deep/.

TABLE 1 | Summary of the four CNN models.

CNN Model	AlexNet	VGG-19	ResNet-101	Inception-ResNet-v2
Top-1 Accuracy (on ImageNet)	57.1%	71.3%	77.1%	80.0%
Number of Layers	8	19	101	164
Number of Parameters	63M	143.7M	44.5M	56M
Input Image Size	227 × 227 × 3	224 × 224 × 3	224 × 224 × 3	299 × 299 × 3

the inception architecture (Szegedy et al., 2017). The inception model uses variant kernel size (in v1) to capture the features from variant object size and location, introduces batch (weight) normalization (in v2) and factorizing convolutions (in v3), and uses bottleneck layers (1 × 1) to avoid a parameter explosion. The combination of the two most recent ideas: residual connections (Szegedy et al., 2016) and the latest revised version of the inception architecture (Szegedy et al., 2016). It is argued that residual connections are inherently important for training very deep architectures (He et al., 2016). Since inception networks tend to be very deep, it is natural to replace the filter concatenation stage of the inception architecture with residual connections. This would allow Inception to reap the benefits of the residual approach while retaining its computational efficiency.

EXPERIMENTAL RESULTS

Experimental Design and Datasets

As shown in **Table 2**, there were 24 categories of shapes (540 images per category), which were generated in MatLab with variations created with transforms including translation, rotation and scaling. For the learning tasks, 85% of the learning datasets were used for training and 15% of the learning datasets were used for measuring validation accuracies, which are reported as learning accuracies. For the transfer tasks, classification accuracies in the transfer datasets are reported as transfer accuracies. Notice that the training datasets and the transfer datasets were different and separated. For example, in Experiment A.1, the CNNs were trained with squares and trapezoids, but never with rectangles. The four CNNs, which were pretrained with the ImageNet, were retrained with the learning datasets for 20 epochs.

To quantitatively evaluate transfer of learning (**Figures 1–6**), we define *transfer index* (TFI) as

$$TFI = HAUC_{Transfer}/HAUC_{Learn} \times 100\%, \tag{1}$$

where the Half Area Under Curve (*HAUC*) is calculated using the (*Accuracy* 50) and *Epoch number*. In general, we are interested in classifiers with accuracies higher than 50%. Using (*Accuracy* 50) instead of *Accuracy* is also for normalization purpose. $HAUC_{Transfer}$ can be negative. The higher the *TFI*, the higher the transfer of learning of a CNN. A perfect transfer, $TFI =$ 100% can be achieved when both transfer accuracy and learning accuracy are 100%.

In the following discussion and all tables (**Tables 3–8**), we use the TFI values to measure the performance of transfer learning. The bold TFI values in each table indicate the best CNN model in that experiment.

Classification With Local Features: Different Shapes

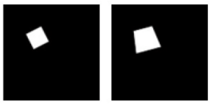
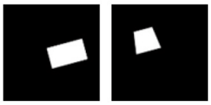
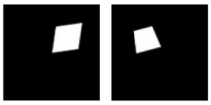
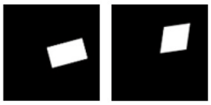
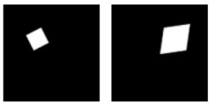
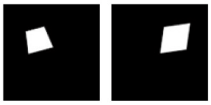
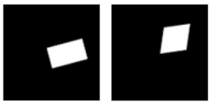
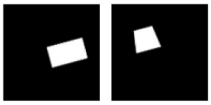
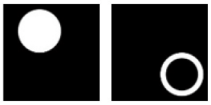
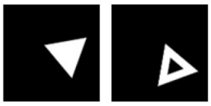
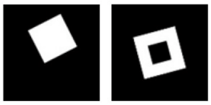
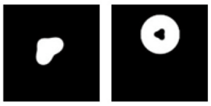
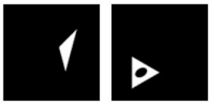
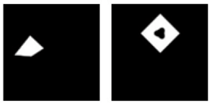
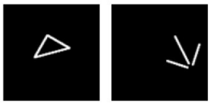
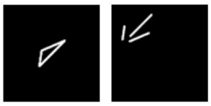
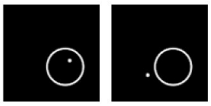
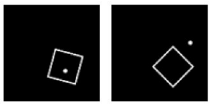
In Experiment A, the CNNs were trained to discriminate squares vs. trapezoids. If the classification was based on angles of neighboring sides or parallelism of opposing sides, we expect the models to classify rectangles vs. trapezoids as squares vs. trapezoids in Experiment A.1 (see the red curves labeled "Transferring 1" in **Figure 1**). In general, we expect that a trained model recognizes Column 1 images in the transfer dataset as Colum 1 images in the training dataset (**Table 2**) if the shared geometric invariants were used for classification. Note that the parallelograms appear in both columns, which partially explains that the accuracies of Transfer 2 and Transfer 3 are lower than that of Transfer 1 (**Figure 1** and **Table 3**). The Inception-ResNet-v2 model performed the best and achieved 62.94, 42.26, and 20.68% of transfer index on the three transfer tests (**Table 3**).

In Experiment B, squares vs. parallelograms were used to train the CNNs. Trapezoids were listed in both columns, which caused lower accuracies of Transfer 1 and Transfer 3 (shown in red and green curves in **Figure 2**). The VGG-19 model was the best and reached 47.4, 61.3, and 13.88% of transfer index (**Table 4**). The low transfer accuracies indicate that similarity extraction were not complete. Transfer of learning was the worst when classifying two unseen shapes (Transfer 3 in Experiment A and Experiment B, green curves in **Figures 1**, **2**). Note that the transfer curves were in "parallel" with the learning curves, indicating the CNNs did extract similarities between the training dataset and the transfer datasets. Linear regressions were performed to quantify the relationship between learning accuracy and transfer accuracy (**Supplemental Materials**). Slope and R of the regressions were used to assess the correlation between transfer accuracy and learning accuracy.

Classification With Global Features: No-Hole vs. One-Hole

The CNNs were trained to discriminate disks (no-hole) vs. rings (one-hole), and transfer of learning was tested on triangles vs. triangle-rings in Experiment C.1 and squares vs. square-rings in Experiment C.2, respectively. A perfect transfer would be expected if the presence of a hole was used to perform the classification. The Inception-ResNet-v2 performed the best (**Figure 3**) and achieved 77.2 and 81.48% of transfer index for the two transfer tasks, respectively (**Table 5**).

TABLE 2 | Examples of shapes for the learning and transfer tasks.

Exp. #	Sample images for learning tasks	Sample images for transfer tasks	Description
A.1			Learning: Square vs. Trapezoid Transfer: Rectangle vs. Trapezoid
A.2			Learning: Parallelogram vs. Trapezoid Transfer: Rectangle vs. Parallelogram
A.3			Learning: Square vs. Trapezoid Transfer: Rectangle vs. Trapezoid
B.1			Learning: Square vs. Parallelogram Transfer: Trapezoid vs. Parallelogram
B.2			Learning: Square vs. Parallelogram Transfer: Rectangle vs. Parallelogram
B.3			Learning: Square vs. Parallelogram Transfer: Rectangle vs. Trapezoid
C.1			Learning: Disk vs. Ring Transfer: Triangle vs. Triangle-ring
C.2			Learning: Disk vs. Ring Transfer: Square vs. Square-ring
D.1			Learning: Irregular-disk vs. Irregular-ring Transfer: Irregular-triangle vs. Irregular-triangle-ring
D.2			Learning: Irregular-disk vs. Irregular-ring Transfer: Irregular-square vs. Irregular-square-ring
E			Learning: Isosceles-triangle vs. Disassembled-Isosceles-triangle Transfer: Irregular-triangle vs. Disassembled-irregular-triangle
F			Learning: Dot-inside-circle vs. Dot-outside-circle Transfer: Dot-inside-square vs. Dot-outside-square

Similar tests with irregular shapes were conducted in Experiment D. The VGG-19 model had high transfer index of 94.68 and 97.22% for the two transfer tasks, respectively (**Figure 4** and **Table 6**). The high transfer performance were impressive when considering the fact of that the model had never been exposed to shapes in the transfer datasets, indicating that

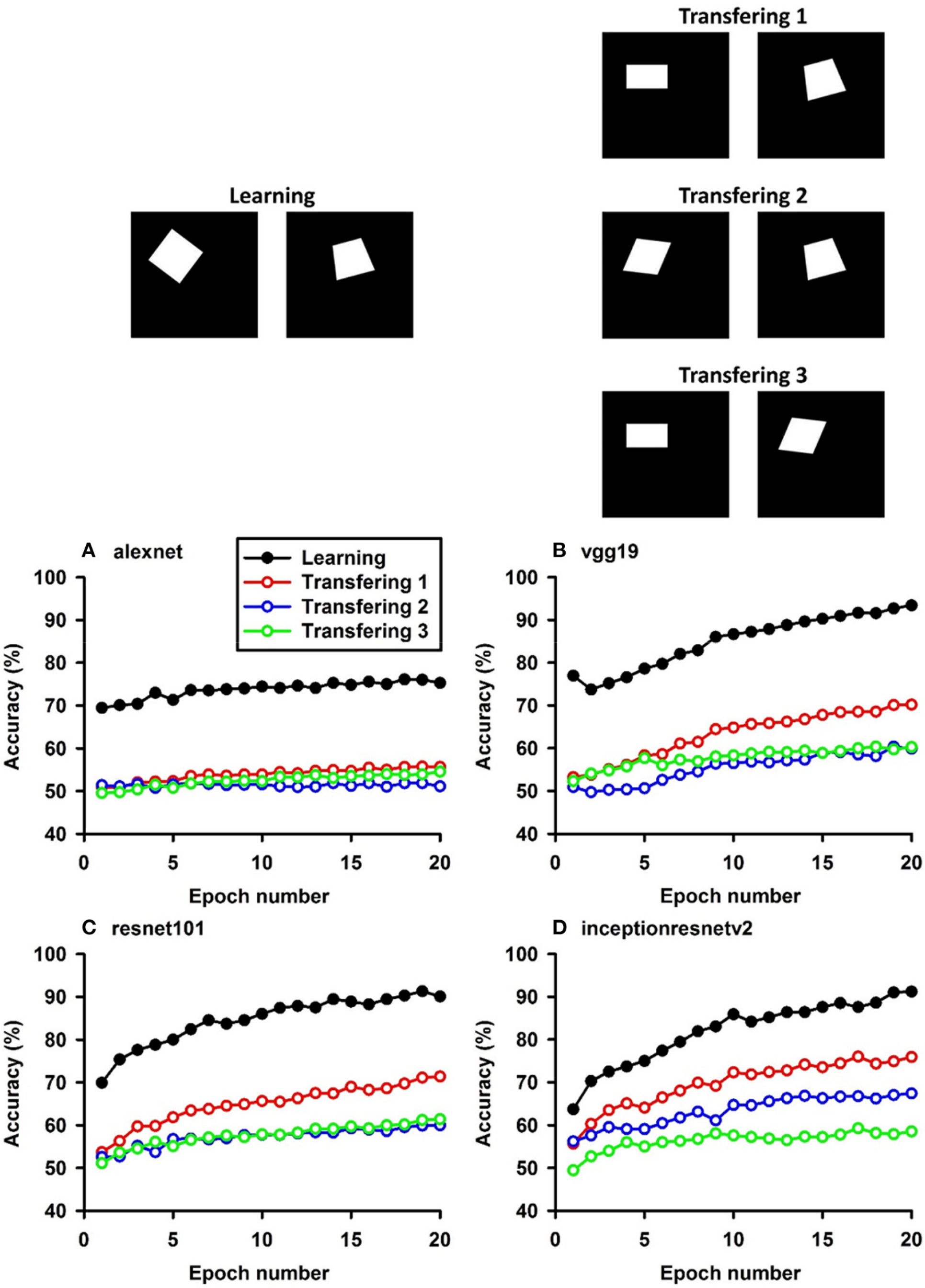


FIGURE 1 | Experiment A. The top panels show the images used in learning and transfer tasks. The lower panel is learning and transfer accuracies as a function of training epochs for the four CNN models.

presence of a hole (a topological invariant) was likely extracted and used for classification.

Note that the transfer index values of irregular shapes are higher than that of regular shapes. More experiments are needed to identify the underlying mechanisms.

Classification With Global Features: Connectivity

In Experiment E, the four CNNs were trained with isosceles-triangles (connected) vs. its three sides separated (not connected), and transfer of learning were tested

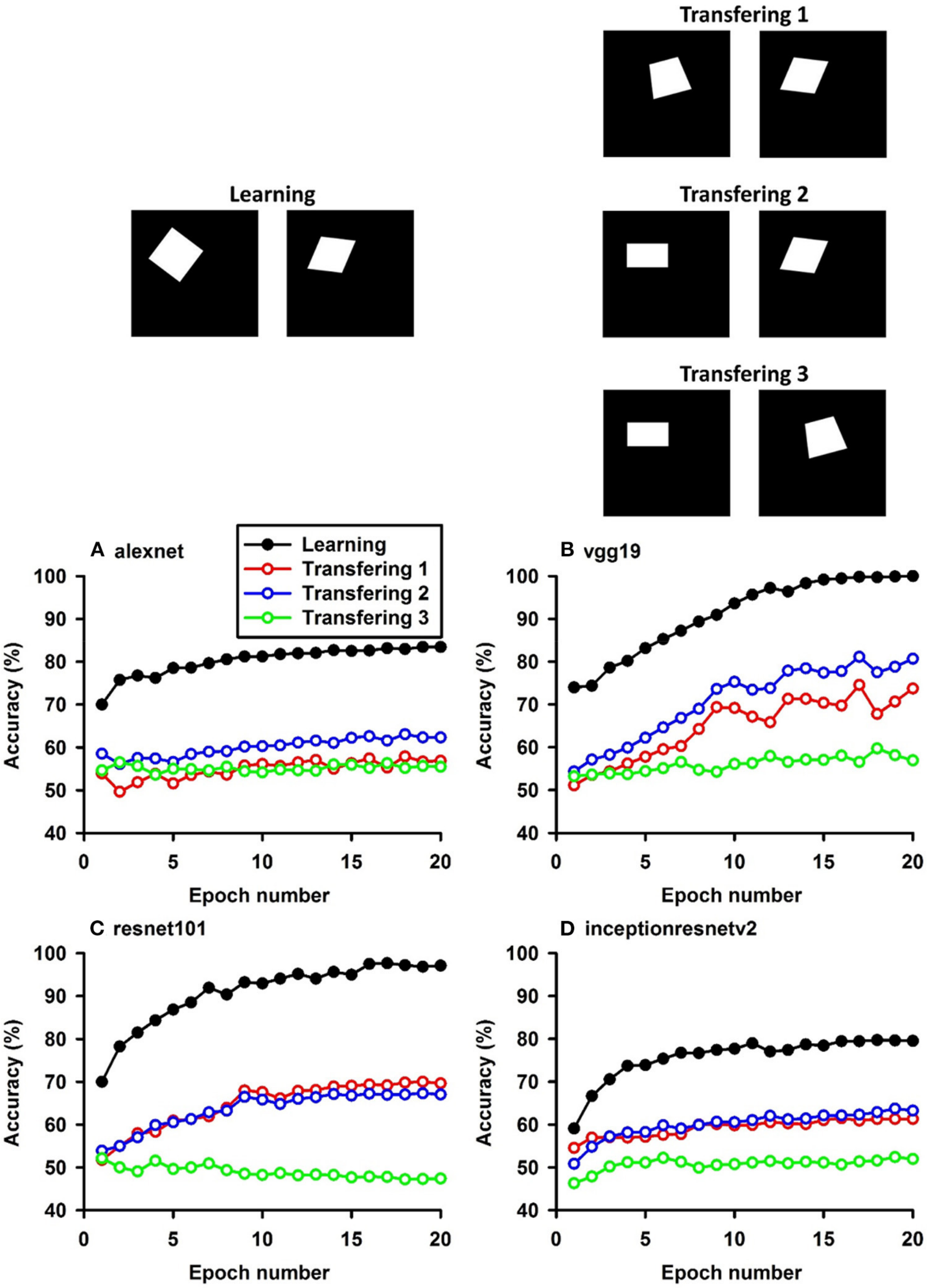


FIGURE 2 | Experiment B. The top panels show the images used in learning and transfer tasks. The lower panel is learning and transfer accuracies as a function of training epochs for the four CNN models.

on irregular-triangles vs. its three sides separated. If connectivity (a topological invariant) was extracted during the learning, we would expect high transfer accuracies in this task. Among the four models, VGG-19 exhibited the highest transfer index of 92.78% (**Figure 5** and **Table 7**).

Classification With Global Features: Inside/Outside Relationship

In Experiment F, the CNNs were trained to discriminate dot-inside-circle vs. dot-outside-circle, and transfer of learning was tested on dot-inside-square vs. dot-outside-square (**Figure 6**). While the VGG-19 achieved a moderate transfer index of 65.92%

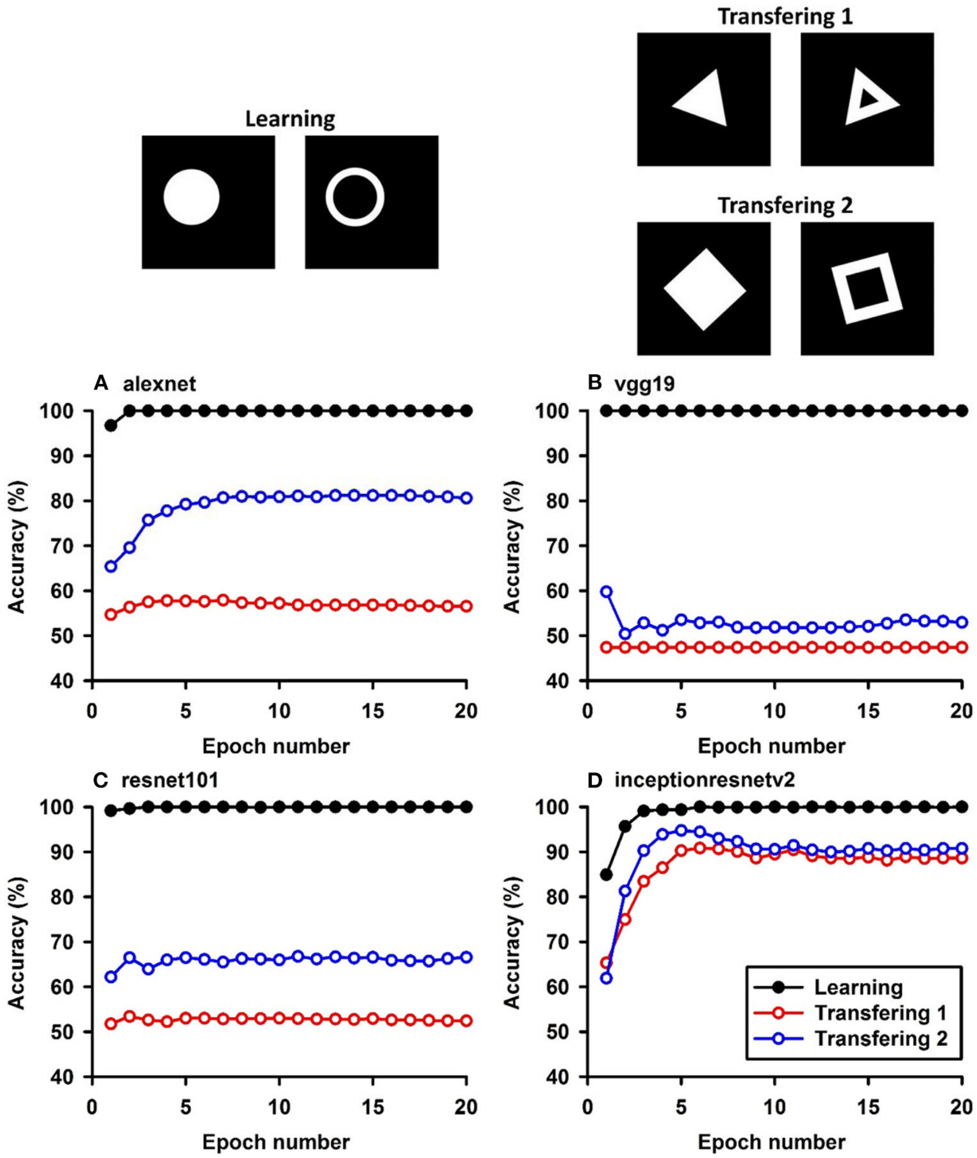


FIGURE 3 | Experiment C. The top panels show the images used in learning and transfer tasks. The lower panel is learning and transfer accuracies as a function of training epochs for the four CNN models.

(**Table 8**), the other three models, including the two advanced models, exhibited lower transfer index of 46.74 and 13.11%, respectively, indicating that inside/outside relationship was not extracted for classification during the learning phase.

SUMMARY AND DISCUSSION

In this study, we trained four CNNs to perform shape classification tasks based on local or global features and further examined how learning of classifying shapes in the training datasets was transferred to classifying shapes in the transfer datasets, which share local or global features with the training datasets. Experiments were designed to test two hypotheses on transfer of CNN learning. First, we wanted to test whether learning tasks based on local features have a higher transfer accuracy than that based on global features. This hypothesis was motivated by the local-to-global hierarchical organization of the CNN architecture, where local features are fully extracted and represented by the early layers. Second, we wanted to test whether the advanced CNNs have higher transfer accuracy for learning tasks based on global features than the early CNNs. This hypothesis was motivated by the fact that the advanced CNNs employ more layers and recurrent connections, which had advantages of extracting global features by integrating inputs from a large region. Although this is a pilot study using the ShapeNet approach, our results provide clear evidence that does not support the two hypotheses.

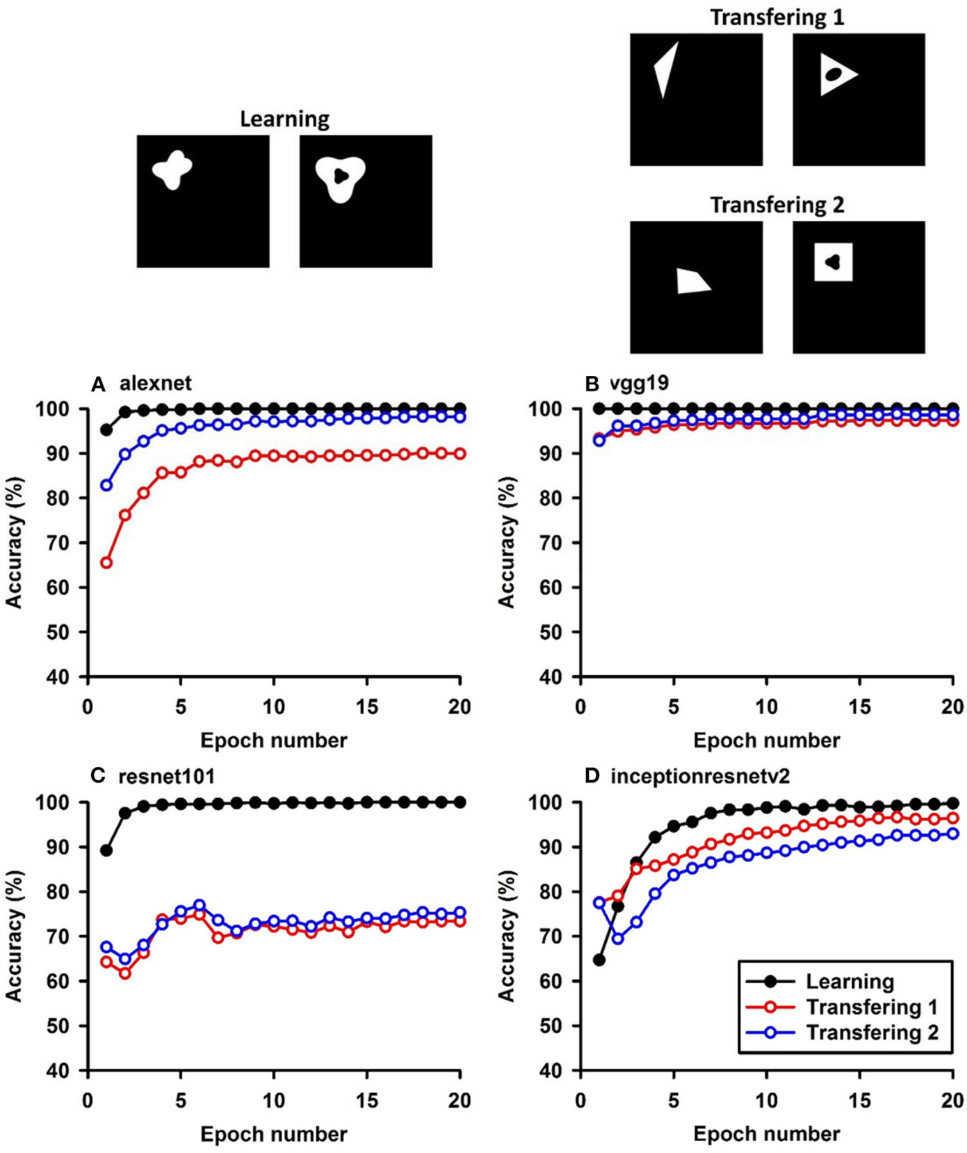


FIGURE 4 | Experiment D. The top panels show the images used in learning and transfer tasks. The lower panel is learning and transfer accuracies as a function of training epochs for the four CNN models.

As expected, the CNNs performed well in the learning tasks, regardless classifying shapes using local features (Experiments A and B) or global features (Experiments C–F). After 20 epochs of training, they classified the shapes in the training datasets at high accuracies (>95%), indicating feasibility of employing pre-trained CNNs to learn new tasks on a small dataset. However, their performance in the transfer experiments varied from task to task, and from model to model. Regarding the first hypothesis, we found that transfer accuracies for local features (Experiments A and B) were lower than those with global features (Experiments C–F). In the example of Resnet101, after it was trained to discriminate squares from trapezoids, they were tested to discriminate rectangles from trapezoids. The squares share many local features with squares, such as four angles of 90 degrees, two pairs of sides parallel to each other, etc. If the model learned to discriminate the pair of shapes based on these shared features, we should expect a perfect transfer, TFI = 100%. Contrary to this prediction, we found that Resnet101 only had 53.36% transfer to rectangles and 24.94% transfer to parallelograms. On the other hand, after Resnet-101 was trained to discriminate regular triangle from their separated sides, they were tested to discriminate irregular triangles from their separated sides. If Restnet101 learned to discriminate connected shape from disconnected shapes (i.e., connectivity, a topological invariant), we would expect to observe a high transfer accuracy. Indeed, it showed a transfer index

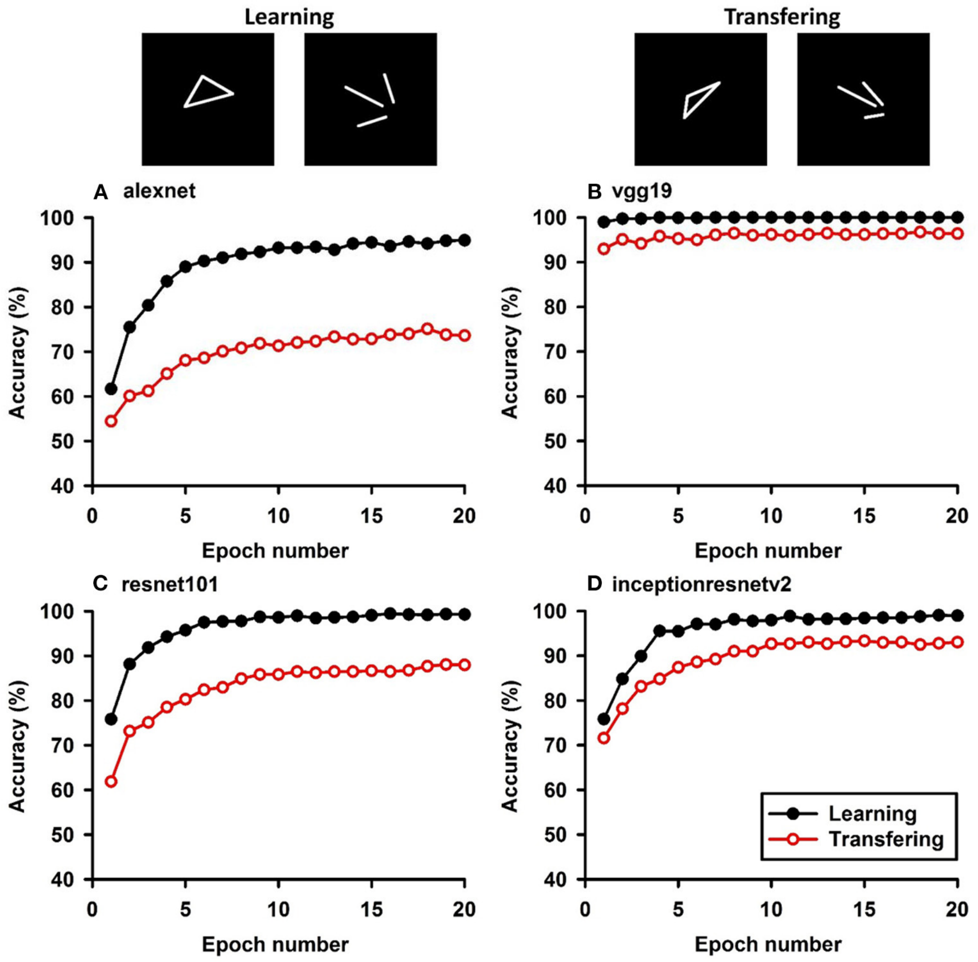


FIGURE 5 | Experiment E. The top panels show the images used in learning and transfer tasks. The lower panel is learning and transfer accuracies as a function of training epochs for the four CNN models.

of 77.06%, much higher than the transfer accuracy based on local features. This finding is counterintuitive, suggesting a lack of understanding of the mechanisms underlying CNN image classification. However, the ShapeNet analysis provides a quantitative approach to gain insight into this difficult problem. Future studies will systematically manipulate the differences between the learning datasets and the transfer datasets to tease out the features used in the learning tasks.

Regarding the second hypothesis, we found that the more advanced CNNs do not have higher transfer accuracies based on global features. For example, Inception-ResNet-v2 has 192 layers and VGG19 has 19 layers. However, VGG19 exhibited higher transfer accuracies on learning based on global features, such as connectivity (**Table 7**, 92.78 vs. 87.91%) and inside-outside relationship (**Table 8**, 65.92 vs. 13.11%). Among the three global (topological) invariants, we found that inside/outside relationship had the lowest transfer performance in the CNNs (**Figure 7**). In the training datasets of dot-inside-circle/dot-outside-circle, circle size, circle position and dot position with respect to the circle varied from image to image. The high learning accuracies (>99%) indicate that the CNNs successfully extracted the common features of the learning datasets. However, after only replacing circle by square, the models performed poorly in classifying dot-inside-square and dot-outside-square. The most advanced CNN model only had a transfer index of 13.11%. This counterintuitive result suggests that the CNNs achieved shape classification by adopting different strategies than extracting inside/outside relationship. Note that different from the other tasks, where the transfer curves are in parallel with the learning curves, indicating extracting shared properties between the training datasets and the transfer datasets (**Supplementary Figures 1, 2**), the transfer curve for the Inception-ResNet-v2 did not increase in parallel with the learning curve. In fact, the correlation coefficient between the transfer accuracy and learning accuracy was −0.17. Among the three tested global features, inside-outside relationship seems to be a limitation of the CNNs, which is not overcome by increasing depth and recurrent connections. This task may be an effective

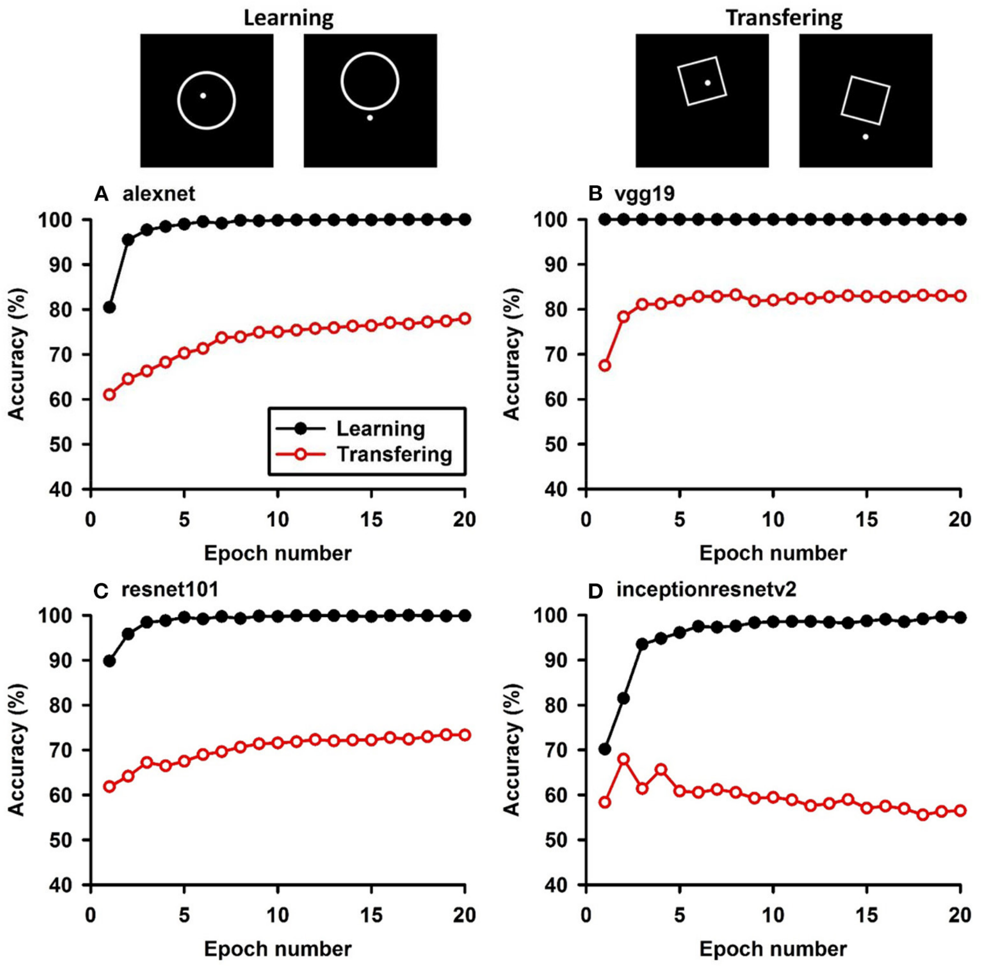


FIGURE 6 | Experiment F. The top panels show the images used in learning and transfer tasks. The lower panel is learning and transfer accuracies as a function of training epochs for the four CNN models.

TABLE 3 | Exp. A. Learning accuracies and transfer index (TFI, percentage) of the four CNNs (Epoch 1 and 20) and the slope and R of the regression.

Epoch#\CNN		AlexNet			VGG-19			ResNet-101			Inception-ResNet-v2	
1 (learning)		69.44			76.94			69.91			63.70	
20 (learning)		75.28			93.43			90.09			91.20	
Transfer Exp.	A.1	A.2	A.3	A.1	A.2	A.3	A.1	A.2	A.3	A.1	A.2	A.3
1 (TFI)	4.78	7.15	−2.37	12.03	3.45	8.57	18.58	13.01	5.58	41.24	45.26	−4.09
20 (TFI)	22.35	4.39	17.96	46.49	22.82	23.67	53.36	24.94	28.41	**62.94**	**42.26**	**20.68**
Slope of regression	0.738	0.029	0.708	0.868	0.548	0.320	0.809	0.366	0.443	0.733	0.462	0.271
R of regression	0.959	0.163	0.943	0.990	0.989	0.921	0.981	0.952	0.974	0.987	0.966	0.912

The bold values denote the best CNN model corresponding to the highest TFI values in each experiment.

benchmark for developing new CNNs that can extract global features under various conditions.

In summary, this pilot study presented a proof of concept of the "ShapeNet" approach that can be used to elucidate the mechanisms underlying CNN image classification. Rejecting the two intuitive hypotheses indicate clear knowledge gaps in our understanding of CNN image processing. Since the same stimuli and tasks can be used to study visual information processing in humans and monkeys, the "ShapeNet" approach may be an effective platform to compare CNN vision and biology vision. In fact, in addition to the well-known local-to-global approach, there are accumulating evidence for an alternative

TABLE 4 | Exp. B. Learning accuracies and transfer index (TFI, percentage) of the four CNNs (Epoch 1 and 20) and the slope and R of the regression.

Epoch#\CNN	AlexNet			VGG-19			ResNet-101			Inception-ResNet-v2		
1 (learning)	70.00			73.98			70.00			59.07		
20 (learning)	83.43			100.00			97.13			79.54		
Transfer Exp.	B.1	B.2	B.3	B.1	B.2	B.3	B.1	B.2	B.3	B.1	B.2	B.3
1 (TFI)	19.45	42.60	23.15	4.63	18.14	13.51	8.80	19.45	10.65	50.06	9.15	−40.79
20 (TFI)	20.49	36.82	16.33	**47.40**	**61.30**	**13.88**	41.65	36.16	−5.50	38.25	44.82	6.57
Slope of regression	0.435	0.483	0.048	0.766	0.941	0.175	0.723	0.566	−0.157	0.341	0.581	0.240
R of regression	0.692	0.780	0.223	0.955	0.983	0.876	0.970	0.974	0.822	0.880	0.970	0.873

The bold values denote the best CNN model corresponding to the highest TFI values in each experiment.

TABLE 5 | Experiment C. Learning accuracies and transfer index (TFI, percentage) of the four CNNs (Epoch 1 and 20) and the slope and R of the regression (N/A mean not applicable).

Epoch#\CNN	AlexNet		VGG-19		ResNet-101		Inception-ResNet-v2	
1 (learning)	96.76		100.00		99.17		84.91	
20 (learning)	100.00		100.00		100.00		100.00	
Transfer Exp.	C.1	C.2	C.1	C.2	C.1	C.2	C.1	C.2
1 (TFI)	10.01	32.93	−5.26	19.50	3.56	24.77	43.83	34.06
20 (TFI)	13.06	61.20	−5.26	5.84	4.88	33.14	**77.20**	**81.48**
Slope of regression	0.729	4.436	N/A	N/A	0.724	3.767	1.690	1.980
R of regression	0.754	0.758	N/A	N/A	0.412	0.707	0.940	0.967

The bold values denote the best CNN model corresponding to the highest TFI values in each experiment.

TABLE 6 | Experiment D. Learning accuracies and transfer index (TFI, percentage) of the four CNNs (Epoch 1 and 20) and the slope and R of the regression.

Epoch#\CNN	AlexNet		VGG-19		ResNet-101		Inception-ResNet-v2	
1 (learning)	95.27		100.00		89.20		64.68	
20 (learning)	100.00		100.00		100.00		99.72	
Transfer Exp.	D.1	D.2	D.1	D.2	D.1	D.2	D.1	D.2
1 (TFI)	34.26	72.61	86.58	85.64	36.33	44.87	87.67	86.85
20 (TFI)	79.86	96.30	**94.68**	**97.22**	46.76	50.70	93.34	86.36
Slope of regression	5.218	3.188	N/A	N/A	0.909	0.721	0.584	0.610
R of regression	0.907	0.909	N/A	N/A	0.641	0.584	0.911	0.813

The bold values denote the best CNN model corresponding to the highest TFI values in each experiment.

TABLE 7 | Experiment E. Learning accuracies and transfer index (TFI, percentage) of the four CNNs (Epoch 1 and 20) and the slope and R of the regression.

Epoch#\CNN	AlexNet	VGG-19	ResNet-101	Inception-ResNet-v2
1 (learning)	61.67	98.98	75.83	75.83
20 (learning)	94.91	100.00	99.26	98.98
1 (TFI)	38.05	87.71	45.88	83.51
20 (TFI)	52.57	**92.78**	77.06	87.91
Slope of regression	0.639	3.514	1.134	0.967
R of regression	0.967	0.897	0.976	0.967

The bold values denote the best CNN model corresponding to the highest TFI values in each experiment.

TABLE 8 | Experiment F. Learning accuracies and transfer index (TFI, percentage) of the four CNNs (Epoch 1 and 20) and the slope and R of the regression.

Epoch#\CNN	AlexNet	VGG-19	ResNet-101	Inception-ResNet-v2
1 (learning)	80.46	100.00	89.81	70.19
20 (learning)	100.00	100.00	99.91	99.44
1 (TFI)	36.18	35.00	29.77	41.26
20 (TFI)	55.92	**65.92**	46.74	13.11
Slope of regression	0.838	N/A	1.138	−0.167
R of regression	0.764	N/A	0.824	0.393

The bold values denote the best CNN model corresponding to the highest TFI values in each experiment.

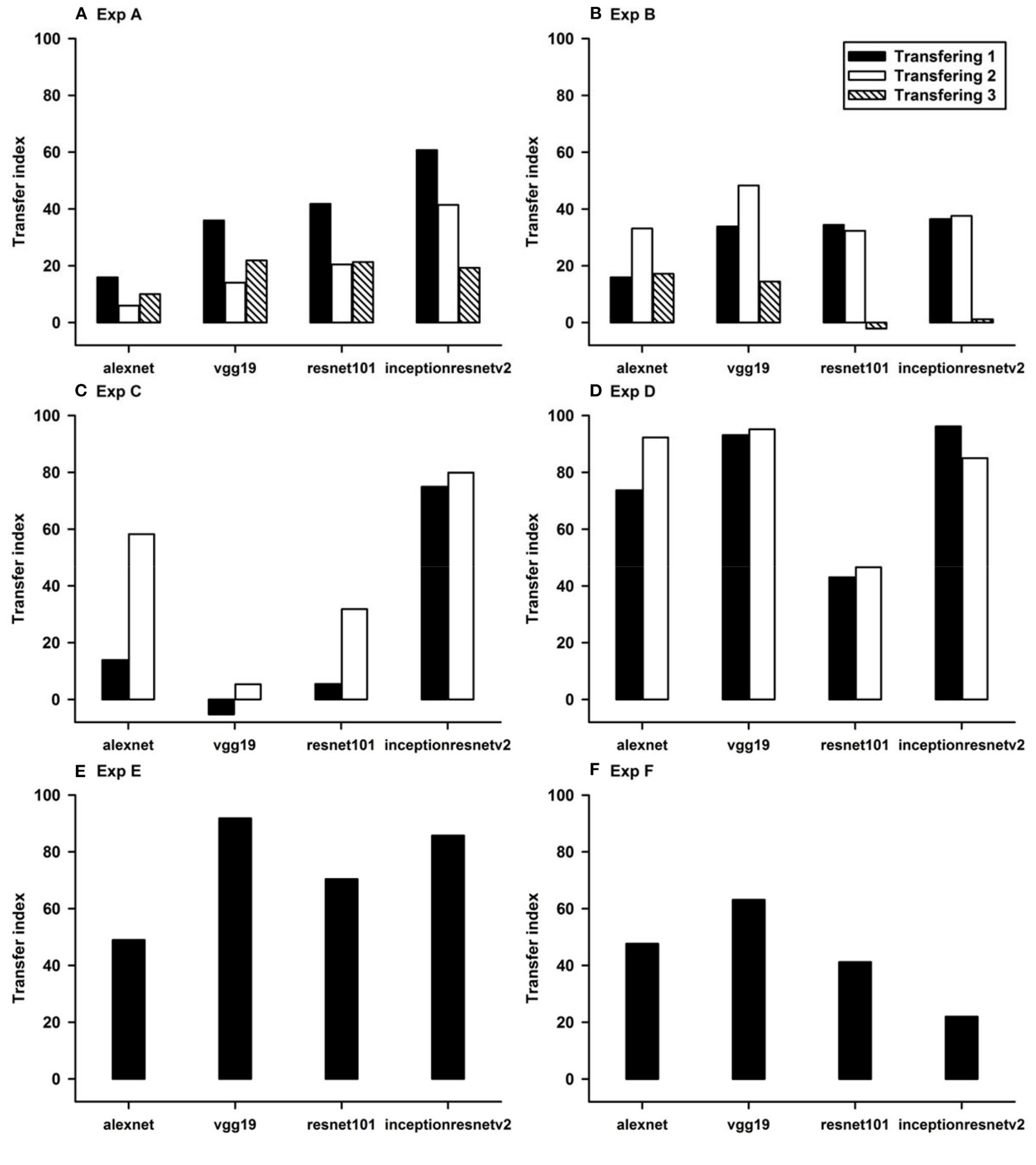


FIGURE 7 | Evaluation and comparison of transfer learning of the four CNN models using transfer index.

global-to-local approach, such as object-superiority (Weisstein and Harris, 1974), early detection of topological properties (Chen, 1982, 1990), and rapid processing of global features in non-human primates (Huang et al., 2017). While the exact underlying mechanisms and differences between CNN models and primate visual systems are unknown, the results suggested that the primate visual systems process local and global features in different ways than the CNNs. By recognizing the differences, future studies will be focused on extending the analysis to other local/global geometrical invariants to understand the CNNs and the biological visual functions. In particular, we will test how humans and monkeys transfer their learning based on inside/outside relationships (topological invariant). We believe comparison between the CNN vision and biological vision using the "ShapeNet" approach will provide insight into a better understanding of visual information processing in both systems.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation is available at http://r2image.com/transfer_learning/datasets.zip.

AUTHOR CONTRIBUTIONS

YZ analyzed all images with four CNN models and drafted the sections regarding CNN method and models. JH created all geometric images and figures. He also analyzed CNN results. TC and YO assisted in running all experiments. WZ designed all experiments and drafted the sections regarding hypotheses and result discussion. All authors contributed to the article and approved the submitted version.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fncom.2021.637144/full#supplementary-material

REFERENCES

Chen, L. (1982). Topological structure in visual perception. *Science* 218, 699–700. doi: 10.1126/science.7134969

Chen, L. (1990). Holes and wholes: a reply to Rubin and Kanwisher. *Percept. Psychophy.* 47, 47–53. doi: 10.3758/BF03208163

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition. 2016," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Las Vegas, NV), 770–778. doi: 10.1109/CVPR.2016.90

Huang, J., Yang, Y., Zhou, K., Zhao, X., Zhou, Q., Zhu, H., et al. (2017). Rapid processing of a global feature in the ON visual pathways of behaving monkeys. *Front. Neurosci.* 11:474. doi: 10.3389/fnins.2017.00474

Hubel, D. H., and Wiesel, T. N. (1977). Ferrier lecture. Functional architecture of macaque monkey visual cortex. *Proc. B. Soc. Lond. B.* 198, 1–59. doi: 10.1098/rspb.1977.0085

Hussain, M., Bird, J. J., and Faria, D. R. (2018). "A study on CNN transfer learning for image classification," in *Advances in Computational Intelligence Systems. UKCI 2018. Advances in Intelligent Systems and Computing, Vol. 840*, eds A. Lotfi, H. Bouchachia, A. Gegov, C. Langensiepen and M. McGinnity (Cham: Springer), 191–202.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). "ImageNet classification with deep convolutional neural networks," *Proceedings of the 25th International Conference on Neural Information Processing Systems—Volume 1* (Lake Tahoe, Nevada), 1097–1105.

Liu, S., John, V., Blasch, E., Liu, Z., and Huang, Y. (2018). "IR2VI: enhanced night environmental perception by unsupervised thermal image translation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops* (Salt Lake City, UT), 1153–1160. doi: 10.1109/CVPRW.2018.00160

Marr, D. (1982). *Vision: A Computational Investigation Into the Human Representation and Processing of Visual Information*. New York, NY: Freeman.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh S., Ma, S., et al. (2015). ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.* 115, 211–252. doi: 10.1007/s11263-015-0816-y

Simonyan, K., and Zisserman, A. (2014). *Very Deep Convolutional Networks for Large-Scale Image Recognition, arXiv Technical Report*, California Institute of Technology

Szegedy, C., loffes., Vanhoucke, V., and Alemi, A. A. (2017). "Inception-v4, Inception-ResNet and the impact of residual connections on learning," *AAAI* 2017, 4278–4284.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). "Rethinking the inception architecture for computer vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Boston, MA), 2818–2826. doi: 10.1109/CVPR.2016.308

Weisstein, N., and Harris, C. S. (1974). Visual detection of line segments: an object-superiority effect. *Science* 186:752. doi: 10.1126/science.186.4165.752

Zheng, Y., Huang, J., Chen, T., Ou, Y., and Zhou, W. (2019). "CNN classification based on global and local features," in *Proceedings of the. SPIE 10996, Real-Time Image Processing and Deep Learning 2019*, 109960G. doi: 10.1117/12.2519660

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Computational Neuroscience

ORIGINAL RESEARCH
published: 22 February 2021
doi: 10.3389/fncom.2021.625804

Semantic Relatedness Emerges in Deep Convolutional Neural Networks Designed for Object Recognition

Taicheng Huang[1], *Zonglei Zhen*[2]* and *Jia Liu*[3]*

[1] State Key Laboratory of Cognitive Neuroscience and Learning and IDG/McGovern Institute for Brain Research, Beijing Normal University, Beijing, China, [2] Beijing Key Laboratory of Applied Experimental Psychology, Faculty of Psychology, Beijing Normal University, Beijing, China, [3] Department of Psychology, Tsinghua University, Beijing, China

Human not only can effortlessly recognize objects, but also characterize object categories into semantic concepts with a nested hierarchical structure. One dominant view is that top-down conceptual guidance is necessary to form such hierarchy. Here we challenged this idea by examining whether deep convolutional neural networks (DCNNs) could learn relations among objects purely based on bottom-up perceptual experience of objects through training for object categorization. Specifically, we explored representational similarity among objects in a typical DCNN (e.g., AlexNet), and found that representations of object categories were organized in a hierarchical fashion, suggesting that the relatedness among objects emerged automatically when learning to recognize them. Critically, the emerged relatedness of objects in the DCNN was highly similar to the WordNet in human, implying that top-down conceptual guidance may not be a prerequisite for human learning the relatedness among objects. In addition, the developmental trajectory of the relatedness among objects during training revealed that the hierarchical structure was constructed in a coarse-to-fine fashion, and evolved into maturity before the establishment of object recognition ability. Finally, the fineness of the relatedness was greatly shaped by the demand of tasks that the DCNN performed, as the higher superordinate level of object classification was, the coarser the hierarchical structure of the relatedness emerged. Taken together, our study provides the first empirical evidence that semantic relatedness of objects emerged as a by-product of object recognition in DCNNs, implying that human may acquire semantic knowledge on objects without explicit top-down conceptual guidance.

Keywords: deep convolutional neural network, semantic relatedness, WordNet, perceptual experience, conceptual guidance

OPEN ACCESS

Edited by:
Florentin Wörgötter,
University of Göttingen, Germany

Reviewed by:
Minija Tamosiunaite,
University of Göttingen, Germany
Guillaume Lajoie,
Université de Montréal, Canada

***Correspondence:**
Zonglei Zhen
zhenzonglei@bnu.edu.cn
Jia Liu
liujiathu@tsinghua.edu.cn

Received: *04 November 2020*
Accepted: *01 February 2021*
Published: *22 February 2021*

Citation:
Huang T, Zhen Z and Liu J (2021) Semantic Relatedness Emerges in Deep Convolutional Neural Networks Designed for Object Recognition. Front. Comput. Neurosci. 15:625804. doi: 10.3389/fncom.2021.625804

SIGNIFICANCE

The origin of semantic concepts is in a long-standing debate, where top-down conceptual guidance is thought necessary to form the hierarchy structure of objects. However, an alternative hypothesis argues that semantic concepts derive from the perception of natural environments. Here, we addressed these hypotheses by examining whether deep convolutional neural networks (DCNNs), which only have abundant perceptual experience of objects, can emerge the semantic relatedness of objects with no conceptual relation information was provided. We found that in the DCNNs

representations of objects were organized in a hierarchical fashion, which was highly similar to WordNet in human. This finding suggests that top-down conceptual guidance may not be a prerequisite for human learning the relatedness among objects; rather, semantic relatedness of objects may emerge from the perception of visual experiences for object recognition.

INTRODUCTION

Objects in this world are complicated. Variations of objects (e.g., orientation, size, shape and color) create challenges for human to flexibly recognize and categorize them (Logothetis and Sheinberg, 1996). To survive in such difficult and diverse environments, humans learn to characterize objects into a rich and nested hierarchical structure, which finally evolves into semantic concepts (Tanaka, 1996; Yamins et al., 2014). However, how the hierarchically-structured semantic concepts are formed is still hotly debated.

Two hypotheses have been proposed. One hypothesis (Mahon and Caramazza, 2009; Leshinskaya and Caramazza, 2016) suggests that semantic concepts are only formed and accessed through abstract symbols that are independent of perceptual experiences. Supporting evidence comes from studies on congenitally blind people, whose core semantic retrieval system in the frontal-temporal cortex can still be activated for retrieving visually-experienced semantic information (Noppeney et al., 2003; Noppeney, 2007). In addition, functional brain imaging studies find that supramodal regions in the ventral temporal occipital cortex (e.g., superior occipital, inferior and superior parietal areas) are also involved in processing objects in blind individuals (Lambert et al., 2004; Ricciardi et al., 2014). Therefore, perceptual experiences seemed not necessary for the emergence of semantic concepts.

An alternative hypothesis argues that the development of semantic concepts derives from perception of natural environments (Sloutsky, 2003; Roy, 2005; Barsalou, 2008). For example, in a word/no word match-to-sample task, Imai et al. (1994) decouple taxonomic and perceptual similarity of words, and find that younger children rely on the visual property of objects, rather than taxonomic concepts, in response to novel words. More direct evidence comes from a study on 10-month-old infants who learn new words by the perceptual salience of an object rather than social cues provided by the caregivers (Pruden et al., 2006). That is, perceptual features are needed to form semantic concepts.

One inevitable limitation of these studies is that perceptual experiences and conceptual guidance are tightly intermingled during the development; therefore, it is impossible to examine one factor with the other controlled. In contrast, the advance of deep convolutional neural networks (DCNNs) provides a perfect model to examine how semantic relatedness is formed (Khaligh-Razavi and Kriegeskorte, 2014; Jozwik et al., 2017; Peterson et al., 2018). On one hand, DCNNs have abundant visual experiences on objects, as with the presence of millions of natural images, the DCNNs learn to extract critical visual features to classify objects into categories as perfectly as human. On the other hand, during the training, the relation among object categories is not provided in the training task or in the supervised feedback. Therefore, conceptual guidance is completely absent in the DCNNs. With such characteristics of the DCNNs, here we asked whether semantic relatedness among object categories was able to emerge with no top-down conceptual guidance.

To address this question, we used a typical DCNN, AlexNet, which is designed for classifying objects into 1,000 categories in ImageNet. Specifically, we first measured whether the representations of some object categories were more similar than their relation to others, which formed a hierarchical structure of object categories as a whole. We reasoned that if a stable and well-organized hierarchical structure was observed, the hypothesis of the necessity of conceptual guidance in forming the semantic relatedness was challenged.

MATERIALS AND METHODS

The ImageNet Dataset

We used the ILSVRC2012 dataset (Russakovsky et al., 2015) as the image stimulus (http://image-net.org/challenges/LSVRC/2012/). Both training and validation datasets were used in this study. The ILSVRC2012 training dataset contains about 1.2 million images with labels from 1,000 categories. The validation dataset contains 50,000 unduplicated images that belong to the same 1,000 categories as the training dataset.

Each object category from ILSVRC2012 dataset corresponds to one semantic concept in the WordNet (Deng et al., 2009). Semantic concepts are described with multiple words or phrases, coined as "synonym sets" or "synset" in abbreviation. The synsets used in the ILSVRC2012 are selected from WordNet, and none has a parent-child relation with others. All 1,000 synsets have the same ontology root (i.e., entity) and most of them are subsets of the superordinate synset of physical entity. Specifically, 3 synsets belong to abstract entity (e.g., bubble, street sign, and traffic light), 39 synsets belong to matter (e.g., menu), 9 synsets belong to geological formation (e.g., cliff), 517 synsets belong to artifact (e.g., abacus), 407 synsets belong to living things (e.g., tench), and 16 synsets belong to fruits (e.g., strawberry). As shown in **Figure 1A**, the 1,000 synsets are organized in a hierarchical structure based on the WordNet.

Deep Convolutional Neural Networks (DCNNs)

Six fully-pretrained DCNNs from three DCNN families were used to examine whether the emergence of semantic relatedness was a general feature of DCNNs. All DCNNs were pretrained on ImageNet with 1.2 million images for the classification of 1,000 object categories. The models were downloaded from PyTorch model Zoo (https://pytorch.org/docs/stable/torchvision/models.html).

AlexNet

AlexNet consists of 8 layers of computational units stacked into a hierarchical architecture, with the first 5 convolutional layers and the last 3 fully-connected layers for category classification.

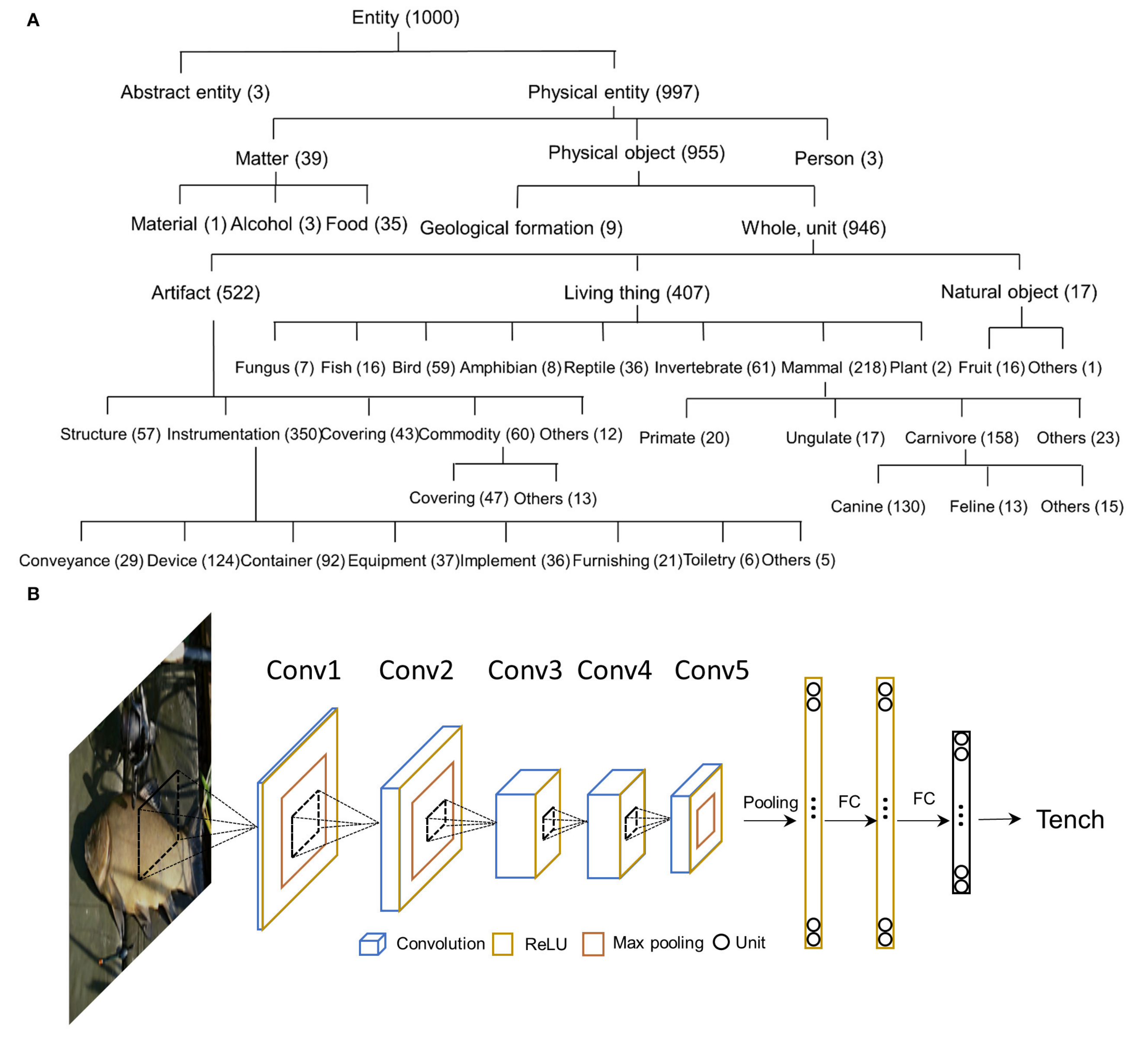


FIGURE 1 | (A) The hierarchical structure of 1,000 object categories in the WordNet. All categories were derived from an ontology root (e.g., entity), and most of them are the subsets of the physical entity. The 1,000 categories cover a wide range of physical objects, making it suitable to study the emerge of object relatedness. Numerals after each word are the number of categories belonging to this superordinate category. **(B)** The architecture of AlexNet. The AlexNet includes 8 layers of computational units stacked into a hierarchical architecture: the first 5 are convolutional layers, and the last 3 layers are fully connected for category classification.

Rectification (ReLU) non-linearity is applied after all layers except for the last fully-connected layer (**Figure 1B**).

VGG

Two VGG networks, including VGG11 and VGG19, were used to examine whether the number of layers was critical for the emergence of semantic relatedness. VGG11 and VGG19 include 11 and 19 weight layers, respectively, with the first 8 and 16 convolutional layers and the last 3 fully-connected layers. All hidden layers are equipped with the ReLU non-linearity.

ResNet

Three ResNet, including ResNet18, ResNet50, and ResNet101 were used to examine the effect of residue blocks on the emergence of semantic relatedness. ResNet18, ResNet50, and ResNet101 include 18, 50, and 101 weight layers, respectively, with all convolutional layers except for the last fully-connected layer. For every two convolutional layers, a residue block is constructed by inserting a shortcut connection. ReLU nonlinearity is applied within these residue blocks.

The Semantic Similarity of Category in WordNet

The semantic similarity of the 1,000 object categories was evaluated by the WordNet 3.0 (Miller, 1995), which is one of the most popularly-used and largest lexical databases of English. In WordNet, the lexical hierarchy is connected by several superordinate synsets in semantic relations, providing a hierarchical tree-like structure for the 1,000 synsets.

We measured the semantic similarity between each pair of the 1,000 synsets using Wu and Palmer's similarity (Wu and Palmer, 1994), which computed the similarities between concepts in an ontology restricted to taxonomic links. This measure is given by:

$$\mathrm{Sim}_{WP}(\mathrm{X,Y}) = \frac{2N}{N_1 + N_2}$$

Where N_1 and N_2 are the depth between the concepts X, Y and the ontology root (i.e., "entity" in WordNet) and N is the depth between the least common subsume (i.e., most specific ancestor node) and the ontology root.

Representation Similarity of Categories in DCNNs

Responses to each image were extracted from all of the convolutional layers and the last fully-connected layer using the ILSVRC2012 validation dataset with the DNNBrain toolbox (Chen et al., 2020) (https://github.com/BNUCNL/dnnbrain). No ReLU was performed for the responses. Responses of stimulus from the same category were averaged to make a response pattern for this category. The category similarity of a layer was measured as correlations of response patterns between each of two categories. In addition, correspondence between the category representational similarity from the DCNNs and the WordNet semantic similarity was calculated to measure the extent to which the relatedness of objects in the DCNNs was similar to that in humans.

The Development of the Relatedness in DCNNs

To investigate how the hierarchical structure of objects emerged in the AlexNet, we retrained it from scratch with about 1.2 million images that belong to the 1,000 categories from the ImageNet training dataset (Deng et al., 2009) using the PyTorch toolbox (Paszke et al., 2019). The network was trained for 50 epochs, with the initial learning rate as 0.01 and a step multiple of 0.1 every 15 epochs. The parameters of each model were optimized using stochastic gradient descent with the momentum and weight decay was fixed at 0.9 and 0.0005, respectively. Each input image was transformed by random crop, horizontal flip, and normalization to improve the training effect of the network. During the training progression, object classification accuracy was evaluated in predicting the category of 50,000 images from the ILSVRC2012 validation dataset in each epoch. In the end, the top-1 and top-5 accuracies for the AlexNet were 51.0% and 74.5%.

During the training progression, we input images from the ILSVRC2012 validation dataset by simply feedforwarding in each epoch to get the activation responses, and then averaged responses within each category and computed the similarity between each pair of categories for the category similarity. Correspondence between the category similarity from the AlexNet and the WordNet semantic similarity in each training stage was measured to evaluate how similar the relatedness of objects was between the AlexNet and human.

To reveal at which semantic level the category similarity from the AlexNet showed better correspondence to the WordNet semantic similarity, the category similarity from the AlexNet was measured at a coarse level and a fine-grained level, respectively. In particular, we first manually selected 19 superordinate concepts (i.e., food, fungus, fish, bird, amphibian, reptile, mammal, invertebrate, conveyance, device, container, equipment, implement, furnishing, toiletry, covering, commodity, structure, and geological formation) that covered most of the 1,000 categories by referring to the WordNet hierarchical relationship, then grouped categories into these superordinate concepts. The coarse-grained correspondence was measured as the correlation between the AlexNet category similarity and the WordNet semantic similarity in 19 superordinate concepts. In turn, the similarity among superordinate concepts was calculated by averaging the category representation similarities from each pair of superordinate concepts. The fine-grained correspondence was measured as the averaged correspondence between the AlexNet category similarity and the WordNet semantic similarity within each superordinate concept.

Effect of Object Co-occurrence to the Formation of Semantic Relatedness

We examined the effect of object co-occurrence in images on the emergence of semantic relatedness. To do this, annotations of object bounding boxes were collected from http://image-net.org/download-bboxes, which were annotated and verified through Amazon Mechanical Turk. To match results from the previous section, bounding boxes of the same 1,000 categories from the ILSVRC2012 dataset were selected, including 544,546 images and corresponding bounding boxes from the ILSVRC2012 training dataset, plus 50,000 images and corresponding bounding boxes from the ILSVRC2012 validation dataset.

Object bounding boxes provide information to distinguish objects from the background in each image. Pixels outside the object bounding boxes in each image were labeled as background, which was removed by setting to 255 (i.e., white color). In addition, for images containing multiple object bounding boxes (i.e., multiple objects), we randomly selected one of the object bounding boxes from these images, and retained the object within the box. Taken together, only one single object of an image remained, excluding the possibility of object co-occurrence as a source for the emergence of semantic relatedness.

We retrained an AlexNet with these single-object images using the Pytorch toolbox for 50 epochs. The top-1 and top-5 accuracies for the single-object AlexNet were 46.7% and 72.0%. Lower prediction accuracy was likely due to fewer images were used for training. Representational similarity of categories in the single-object AlexNet was measured with responses from the last fully-connected layer, and then compared with representation similarity of categories in the pre-trained AlexNet. The developmental trajectory of the single-object AlexNet was also evaluated in each training stage.

Effect of Task Demands on Semantic Relatedness

The effect of task demands on semantic relatedness was examined by re-training AlexNet to classify objects at superordinate levels (e.g., the living thing vs. artifact) as compared to the original AlexNet mainly at the basic level (e.g., traffic light, crane).

One superordinate classification occurred at the highest level of the WordNet: the living thing and the artifact, which consisted of 958 object categories from the ILSVRC2012 dataset. The other superordinate classification occurred at an intermediate level, which consisted of 19 superordinate categories (fungus, fish, bird, amphibian, reptile, canine, primate, feline, ungulate, invertebrate, conveyance, device, container, equipment, implement, furnishing, covering, commodity, and structure). They together consisted of 866 object categories, which were the subset of the 958 categories contained in the superordinate categories of living thing and the artifact. To match the number of object categories, here we used 866 object categories in both superordinate classification tasks, which included 1,108,643 images from the ILSVRC2012 training dataset and 43,301 images from the ILSVRC2012 validation dataset.

The AlexNet for superordinate classification shared the identical architecture as the original AlexNet, except that one extra FC layer was appended to the FC3 layer (i.e., the last FC layer of the original AlexNet). The extra FC layer was designated for different superordinate classification tasks, as the AlexNet for two superordinate categories (AlexNet-Cate2) had two output units, and the AlexNet for 19 intermediate categories (AlexNet-Cate19) had 19 output units. Besides, since

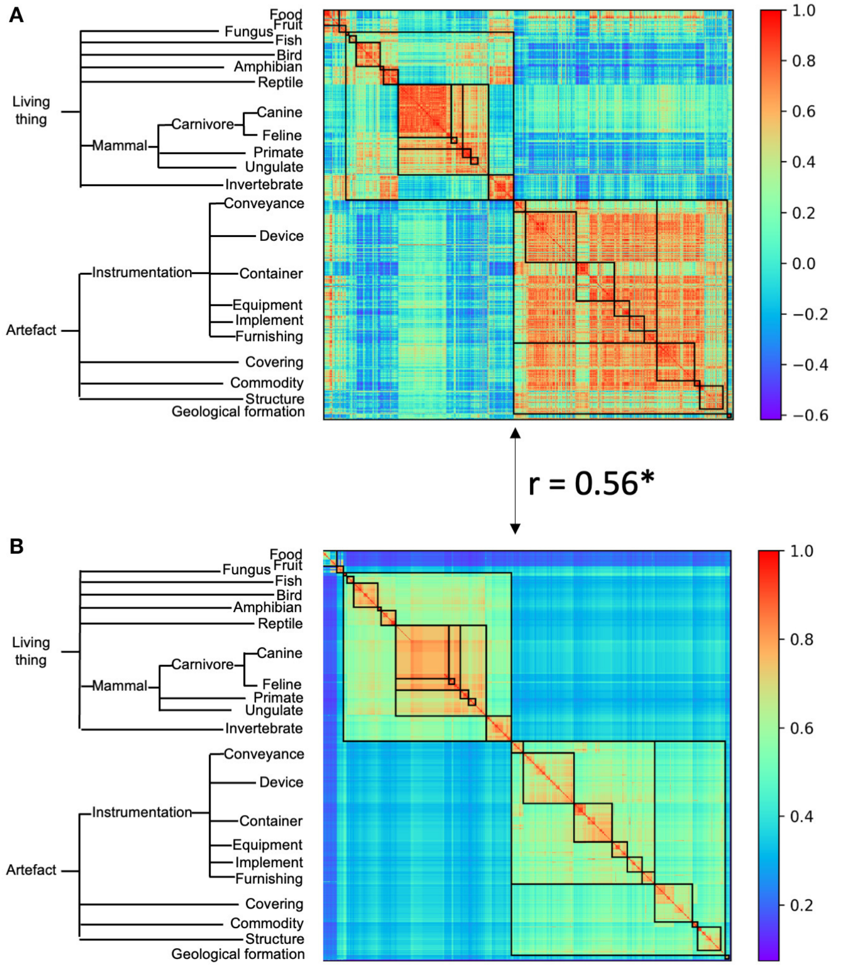


FIGURE 2 | The category representational similarity of the AlexNet **(A)** and the semantic similarity of WordNet hierarchy **(B)**. Categories were ordered according to the WordNet semantic hierarchy. A simplified hierarchical structure was shown as an indicator of superordinate categories in WordNet semantic similarity. For the ease of comparison between AlexNet's category similarity and WordNet semantic similarity, categories belong to the same superordinate category were marked with a black box. The AlexNet category similarity showed good correspondence to the WordNet semantic similarity. Asterisk denotes $p < 0.001$.

a new FC layer was appended to the original AlexNet that may change the dynamics of the network, we also built an AlexNet with an extra FC layer that included 1,000 output units (AlexNet-Cate1000) as the original one. The AlexNet-Cate1000 was designated for validation and for comparison with the AlexNet-Cate2 and AlexNet-Cate19.

The new AlexNets (i.e., AlexNet-Cate2/Cate19/Cate1000) were trained using the Pytorch toolbox for 50 epochs. The top-1 accuracy (top-5 accuracy) were 94.7% (100.0%), 68.7% (95.6%) and 49.0% (73.6%), respectively. Representational similarities of categories in the new AlexNets were measured with responses from all of their layers. Category similarity from the AlexNet-Cate1000 was compared with that of the original AlexNet to validate if they shared a similar hierarchy of semantic relatedness.

Data Availability

All data and code underlying our study and necessary to reproduce our results are available on Github: https://github.com/helloTC/SemanticRelation.

RESULTS

We first evaluated whether there was a hierarchical structure among object categories in the AlexNet, which was trained to classify object categories from the ImageNet containing no relation information among objects. For this, responses from the last fully-connected layer of the AlexNet (i.e., FC3) were averaged across images of each category as the response pattern for this category, and the similarity between two categories was calculated as the correlation between their response patterns. A great variance in similarity was observed, with the highest similarity between object toy poodle and object miniature poodle ($r = 0.99$), the lowest between object snail and object fur coat ($r = -0.62$), and the mean similarity of $r = 0.21$. The variance in similarity observed was significantly larger than variance from a randomized structure (permutation analysis, $p < 0.001$), suggesting that objects were structurally organized (**Figure 2A**, left). A close inspection of **Figure 2A** revealed two large clusters, one is living things and the other artifacts. Within each cluster, there are sub-clusters, as within-cluster variance was smaller than that of neighboring sub-clusters. The nested structure in similarity suggests that a

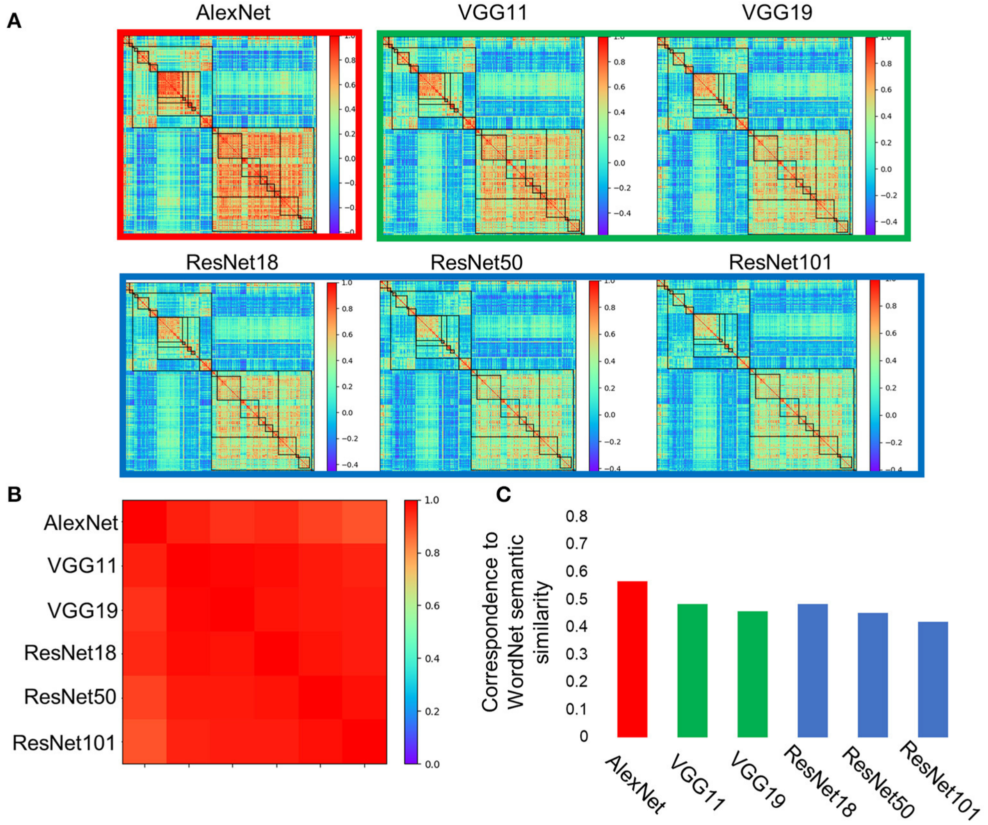


FIGURE 3 | Category representations in the DCNNs were stabled across architectures. **(A)** Categorical representations from AlexNet, two VGGs (i.e., VGG11 and VGG19), and three ResNets (i.e., ResNet18, ResNet50, and ResNet101) showed consistent hierarchical relation of object categories. **(B)** The hierarchical relations that emerged in these DCNNs were almost identical among each other. **(C)** Correspondences between the hierarchical relation among objects in the DCNNs and semantic similarity of WordNet in humans were significant.

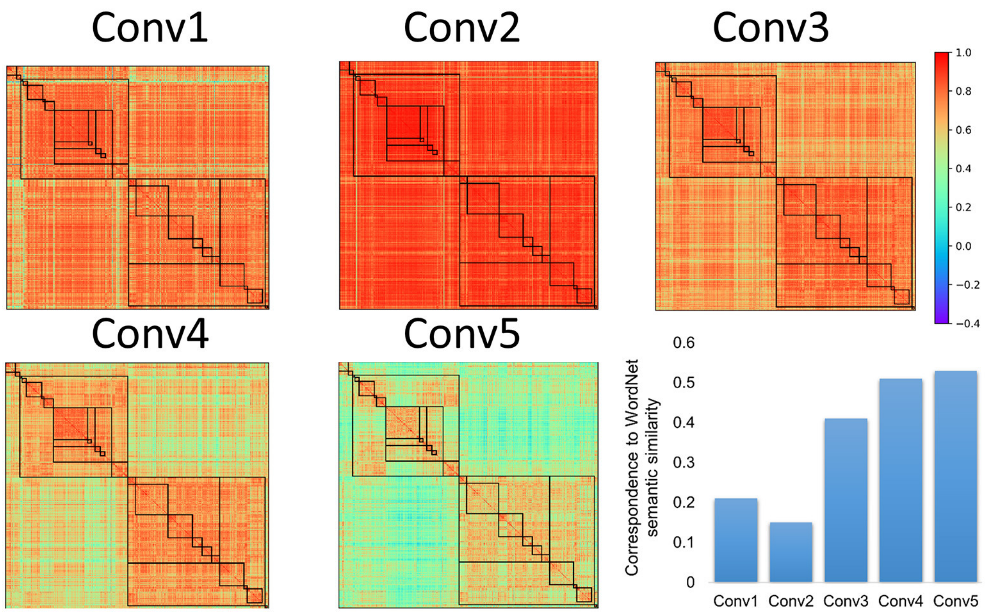


FIGURE 4 | The category representational similarity in different convolutional layers of the AlexNet. Hierarchical relations of objects in the AlexNet gradually emerged as a function of convolutional layers, so was the correspondence between the representational similarity in the AlexNet and WordNet semantic similarity in human. Coarse structure first emerged in lower layers, while the fine-grained structure was prominent only in higher layers.

hierarchical relation among objects emerged in the AlexNet without conceptual guidance.

Similar nested structures of objects were also observed in DCNNs with different architectures (e.g., layer number and kernel size), including two VGGs and three ResNets, which are designed for the same task (**Figure 3A**). Importantly, the hierarchical relations of the object categories that emerged from the VGG family and ResNet family were almost identical to that from the AlexNet ($r > 0.89$ for all DCNNs tested, **Figure 3B**), implying that the emerged hierarchical relation among object categories was invariant to implementations, but rather resulted from inherent properties of the stimulus and the task that DCNNs received and performed. Because human brains used images from the same physical world to perform the same task, one intuitive thought is that the hierarchical relation observed in the DCNNs may be similar to the semantic relatedness of objects in human.

To test this conjecture, the names of the object categories were put into WordNet derived from human, and their semantic similarity was calculated with the Wu and Palmer's similarity approach (**Figure 2B**). We found that there was a significant correlation between semantic similarity of WordNet in human and the hierarchical relation among objects in the AlexNet ($r = 0.56$, $p < 0.001$), and correlation also reached significance for both the living thing ($r = 0.70$, $p < 0.001$) and artifact ($r = 0.41$, $p < 0.001$). Similar correspondence to the semantic similarity of WordNet in human was also observed in DCNNs from the VGG family and ResNet family (**Figure 3C**). In addition, the correspondence of the AlexNet increased as a function of layers (**Figure 4**), with lower correlations observed in first two layers (layer 1: $r = 0.21$, layer 2: $r = 0.15$) and higher correlations in the third ($r = 0.41$), forth ($r = 0.51$), and fifth ($r = 0.53$) layers. A close inspection on the increases of hierarchy among layers revealed that coarse structure (e.g., the living thing vs. artifact) first emerged in lower layers, and a fine-grained structure was prominent only in higher layers.

How did the hierarchical relatedness of object categories emerge from unstructured image dataset in the DCNNs? To address this question, we explored the developmental trajectory of the relatedness when the AlexNet was trained to recognize objects. Two findings were observed. First, correspondence in the hierarchical relatedness of object categories between the AlexNet and the WordNet was established within the first epoch ($r = 0.60$, **Figure 5A**), whereas the performance for object recognition (top-1 accuracy: 8.9%) was far below that of the fully trained one (top 1 accuracy: 51%). Instead, at least 40 training epochs were needed to attain the matched performance to the fully trained model. The asynchronous development illuminated that the relatedness of object categories in the AlexNet was formed before it was capable of performing the task. Second, within the development of the hierarchical relatedness, there was a progression from a coarse structure to a fine-grained structure. That is, the coarse structure based on the 19 concepts (e.g., bird and device) merged from 1,000 object categories reached a plateau within the first epoch (**Figure 5B**), with a correlation of 0.65 to the WordNet. In contrast, the fine-grained structure within the 19 concepts (e.g., crane and flamingo in bird) did not approach a plateau until 40 epochs' training, with an averaged correlation of 0.38 to WordNet in humans. Therefore, the hierarchical relatedness of

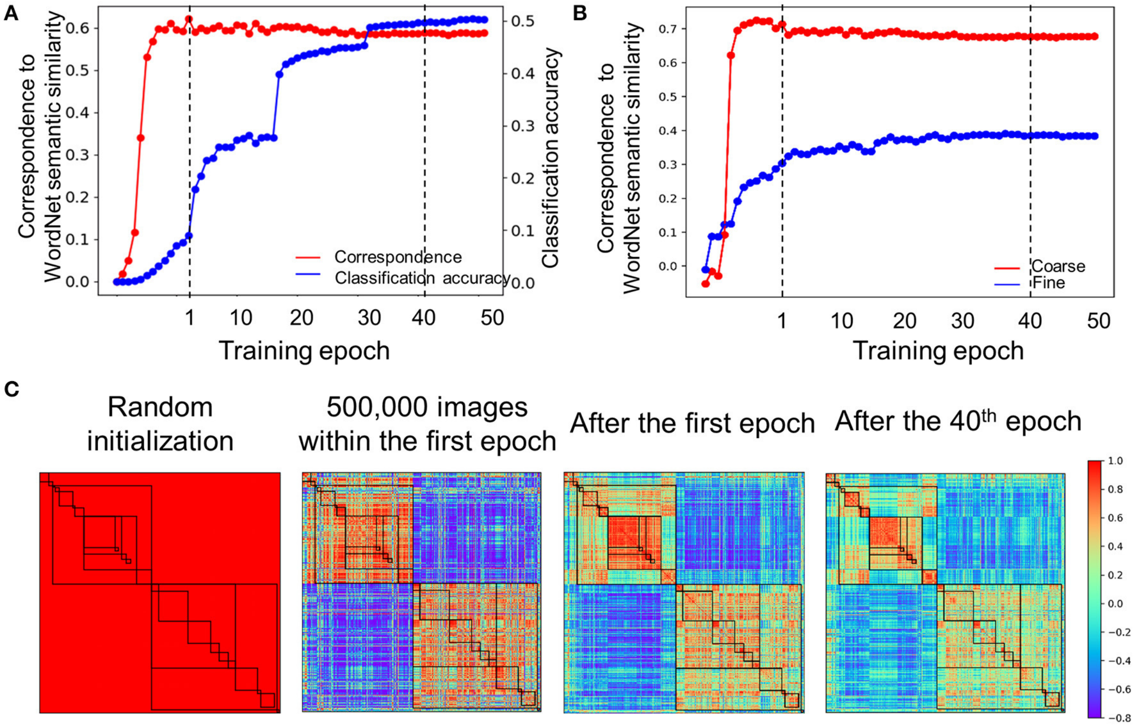


FIGURE 5 | Developmental trajectory of the relatedness. **(A)** The developmental trajectory of the correspondence in the hierarchical relatedness of object categories between the AlexNet and WordNet (red line). The classification accuracy of the AlexNet was shown in blue. The hierarchical structure evolved into maturity far before the establishment of object recognition ability. To illuminate results within the first epoch, correspondence to the WordNet semantic similarity for every 100,000 images was plotted. Dash line indicates epoch 1 and epoch 40, respectively. **(B)** A coarse to fine shift during training progression. The coarse structure based on the 19 superordinate categories reached a plateau within the first epoch (red line), while the fine-grained structure reached a plateau after 40 epochs' training (blue line). Dash line indicates epoch 1 and epoch 40, respectively. **(C)** The category similarities of the AlexNet in different training stages for comparison. From left to right, category similarities of the AlexNet without training, AlexNet trained with 500,000 images within the first epoch, AlexNet trained after the first epoch and AlexNet trained after the 40th epoch. Color bar indicates correlation coefficients.

object categories was formed in a coarse-to-fine fashion, with the coarse structure formed before the fine-grained structure (**Figure 5C**).

In natural environments, objects are seldom alone; further, semantically-related ones are often present together. This object co-occurrence may be preserved in images for training DCNNs, and thus contribute to the emergence of semantic relatedness in a DCNN. To rule out this possibility, we trained an AlexNet with images containing a single object without any background (i.e., the single-object AlexNet, see Materials and Methods) (**Figure 6A**). We found that the hierarchical relation of object categories from the single-object AlexNet was highly correlated with that in the pre-trained AlexNet ($r = 0.83$) (**Figure 6B**), suggesting that the object co-occurrence was not critical for the emergency of semantic relatedness in DCNNs. In addition, a similar developmental trajectory was also observed (**Figure 6C**).

Another probable factor that may shape the hierarchy is the task demand, as recent studies suggest behavior-related representations of DCNNs are largely shaped by tasks that DCNNs performed (Song et al., 2020), rather than the physical properties of stimuli (Xu et al., 2020). To test this possibility, we directly compared AlexNet-Cate2 and AlexNet-Cate19 that were designated to classify objects into 2 or 19 superordinate categories, respectively (**Figure 7A**). The newly added FC layer did not significantly change the internal dynamics of the original AlexNet, as the semantic hierarchy observed in the AlexNet-Cate1000 was almost identical to that of the original AlexNet ($r > 0.90$ for all layers).

We examined the semantic relatedness of the FC3 layer in AlexNet-Cate2 and AlexNet-Cate19, which corresponds to the last layer of the original AlexNet. First, the coarse structure was reserved, as the semantic relatedness emerged in the Alexnet-Cate2 ($r = 0.65$, $p < 0.001$) and AlexNet-Cate19 ($r = 0.89$, $p < 0.001$) was significantly correlated with that in the AlexNet-Cate1000 (**Figure 7B**). However, the degree of the fineness of the structures differed greatly, as the higher superordinate level of object classification was, the coarser the structure of the relatedness emerged. Importantly, such difference was prominent only at the later layers of the networks (**Figure 7C**). That is, the relatedness of object categories in the first four layers of AlexNet-Cate2 and AlexNet-Cate19 was similar to that in the AlexNet-Cate1000 ($rs > 0.89$), possibly driven by the physical properties of stimuli. Then, after the fourth layer, their correspondence to AlexNet-Cate1000 decreased gradually,

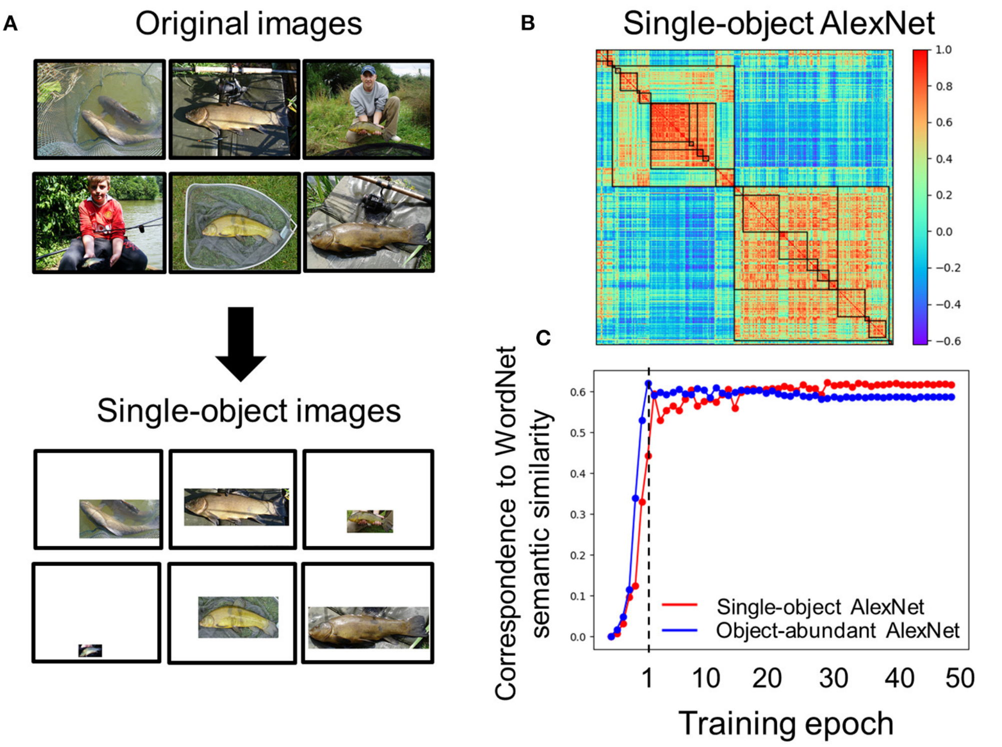


FIGURE 6 | Effect of object co-occurrence on the emergence of semantic relatedness in AlexNet. **(A)** Original images used for training AlexNet contain objects present in the background, which may contribute to the emergence of semantic relatedness in AlexNet. After removing the background, only one object remained. **(B)** Category similarity of the single-object AlexNet, which was trained with images containing only one object. Hierarchical relation is prominent. **(C)** The developmental trajectory of the single-object AlexNet (red) was drawn against that of the original AlexNet (blue). Note that to match the number of images used to train the single-object AlexNet, stages for training the original AlexNet with 600,000 to 1,200,000 images within the first epoch were not plotted.

with that of AlexNet-Cate2 decreasing more dramatically. The divergence in correspondence likely reflected the difference in task demands. In short, the stimulus-behavior dissociation that gradually formed along the hierarchy of the networks reflects the joint efforts of stimuli and tasks in shaping the semantic relatedness of object categories.

DISCUSSIONS

In this study, we used DCNNs as a model for human cognition to examine whether the semantic relatedness of object categories can automatically emerge without top-down conceptual guidance. First, we found that almost identical hierarchical structures of object categorizes emerged in AlexNet, VGG family, and ResNet family, which were highly similar to the WordNet derived in humans. This result suggests that the relation among object categories can be automatically formed without *a prior* conceptual relationship and independent of implementation hardware. Interestingly, the level of fineness of the semantic relatedness was attributed to the task demands of networks, as the stimulus-behavior dissociation was observed along the hierarchy of network layers. In sum, our study provided the first empirical evidence that even without top-down conceptual guidance, the semantic relatedness of objects can be formed from the joint effort of physical properties of stimuli and task demands of networks.

Unlike studies on humans where perceptual experiences are always intermingled with conceptual guidance, the DCNNs provide a perfect model to demonstrate how perceptual experiences contribute to the construction of relatedness among objects (Peterson et al., 2018). This finding is in line with developmental studies where children prefer to naming objects by referring to their perceptual features, suggesting that the perceptual property of objects play an important role in early accessing lexical knowledge (Imai et al., 1994; Gershkoffstowe and Smith, 2004; Samuelson and Smith, 2005). Further, the emerged semantic relatedness is likely independent of implementation, because the DCNNs and human brain, which differ significantly in hardware, show highly similar hierarchical structures of objects.

The similarity in the semantic structure may result from the similarity in architecture that DCNNs are designed with an architecture similar to the human sensory cortex. Accordingly, similar anatomy may lead to similar functions that give rise to similar structures of the relatedness among objects. For example, the top level of the hierarchy was the living things vs. artifacts, mirroring the axis of the mid-fusiform sulcus that separates the coding of animate objects and artifacts in the brain

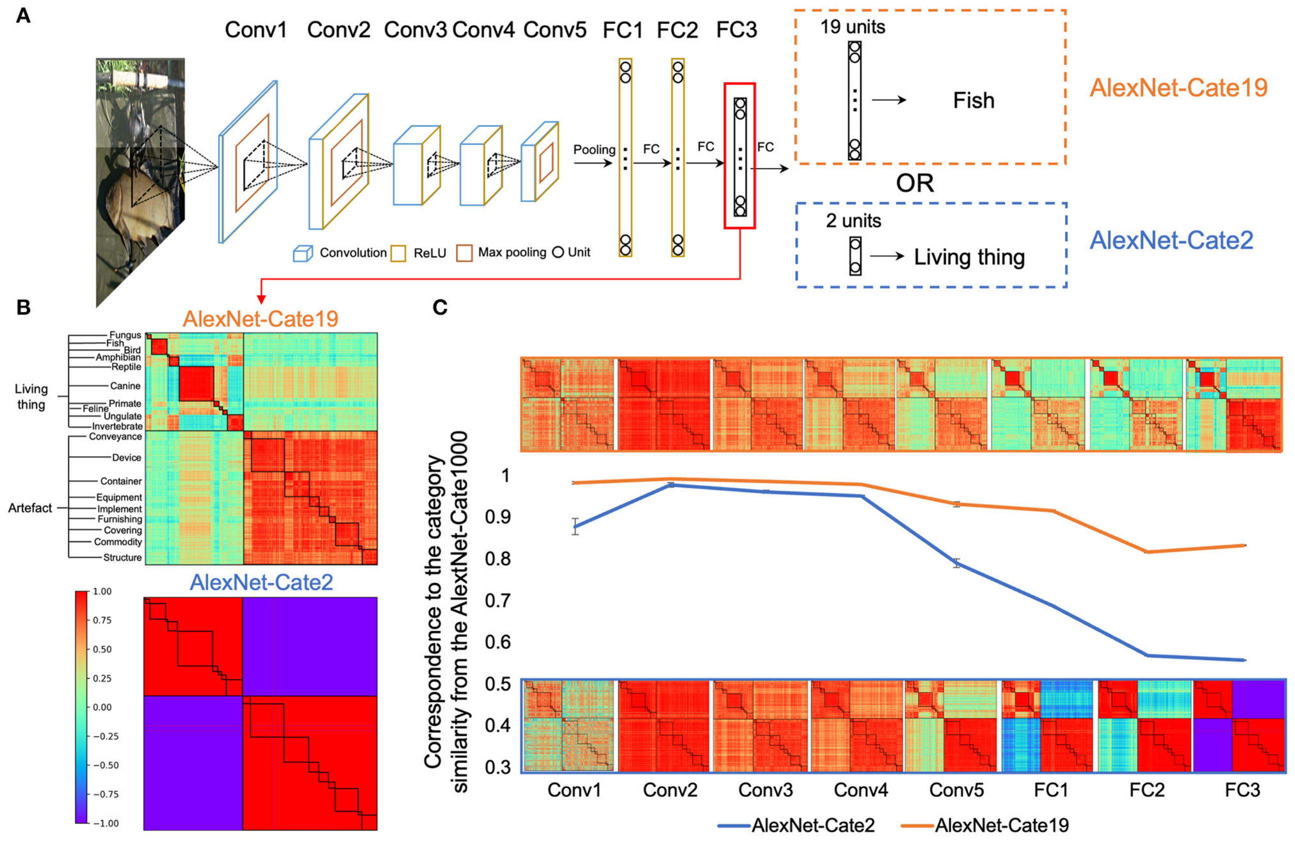


FIGURE 7 | Effect of task demands on semantic relatedness in AlexNet. **(A)** The architectures of AlexNet-Cate19 and AlexNet-Cate2, both of which inherited the same architecture as the original AlexNet, except that one extra FC layer was appended to the FC3 layer. **(B)** Category similarities of AlexNet-Cate19 and AlexNet-Cate2 from the FC3 layer. The hierarchical structures were less prominent in AlexNet-Cate19 as compared to the original AlexNet, and almost absent in AlexNet-Cate2. **(C)** Stimulus-behavior dissociation was formed along the hierarchy of the networks, with the similarity in representation diverging after the fourth convolutional layers. Error bars indicate the standard deviation of the AlexNet-Cate2 and AlexNet-Cate19 after the training was repeated eight times.

(Grill-Spector and Weiner, 2014), echoing the proposal that DCNNs are feasible models to understand visual cortex (Yamins et al., 2014; Yamins and DiCarlo, 2016). Indeed, Bao et al. (2020) have found that category-selective regions in the primate inferior temporal cortex are organized to encode the object space constructed by dimensions extracted from DCNNs.

Another and more plausible possibility may be the way by which objects are coded in representational space. In DCNNs, an object is firstly decomposed into multiple features, and mapped to a representational space (Xu et al., 2020). Then, the object is reconstructed from the feature repertoire of the representational space based on the demand of tasks (Xiang et al., 2019; Yang et al., 2019; Song et al., 2020). The representational space allows DCNNs to use the efficient coding scheme (Barlow, 1961; Liu et al., 2020) to reduce the redundancy of the natural stimuli, which is also widely observed in neuroscience studies (Dan et al., 1996; Kastner et al., 2015). Further, features of the representational space are distributedly represented by different units (Liu et al., 2020; Yang and Wang, 2020); therefore, if two objects are perceptually similar because of shared features, they are likely represented by the same set of units. In this way, the relation between two objects is then derived from the connections among units. This intuition is consistent with the hypothesis of parallel distributed processing (McClelland and Rogers, 2003; Saxe et al., 2019), where knowledge arises from the interactions of units through connections. Accordingly, the knowledge stored in the strengths of the connections finally becomes the building blocks of the hierarchical structure of object categories.

Importantly, such hierarchical structure emerged in a coarse-to-fine fashion. That is, at the initial stage of learning, DCNNs may encode global features to identify relations among objects when only a small number of exemplars are available. For example, dogs and cats are the same, but they are not trees based on general appearance. When more exemplars are learned, features in the repertoire are greatly enriched, and thus are capable of providing fine-grained representations for objects to establish the hierarchical structure of relationships among objects. This coarse-to-fine representation is also observed in infants, as infants are able to distinguish animals and vehicles at 7 months old, but fail to differentiate dogs from cats until 11 months old (Mandler and Mcdonough, 1993, 1998; Pauen, 2002).

Interestingly, we also found that the hierarchical structure evolved into maturity before the establishment of object recognition ability. This is not surprising because the enriched and structured feature repertoire is necessary for DCNNs to successfully recognize novel objects never seen before. For example, in a recent study where DCNN's experience on faces is selectively deprived, the DCNN is still capable of accomplishing a variety of face tasks behaviorally and evolving face-specific modules internally (Xu et al., 2020). Therefore, a mature representational space of objects will greatly benefit DCNNs' performance. This mechanism has already widely used in computer science, as transfer learning, for example, utilizes it to harness a pretrained network to work in another domain with a small number of exemplars but still with high accuracy (Olivas et al., 2009).

Besides the physical properties of stimuli, the demand of tasks also played an important role in shaping the representational space of objects especially when it needs to be read out for behavioral performance (Peterson et al., 2018; Turner et al., 2019). When the DCNN was designated to classify objects at superordinate levels rather than at the basic level, the representational space became coarser and the nested structure of the semantic relatedness was less prominent. However, at the earlier layers of the network, the representational space was less likely affected by task demands; rather it was mainly driven by the physical properties of stimuli. As the information flew into later layers, the stimulus-behavior dissociation was observed, as the representational space was mainly shaped by the demand of tasks. Therefore, it is possible that DCNNs extracted images' features based on image statistics into a repertoire to construct a representational space in lower layers, and then only selected features necessary for tasks that the network performed to constructed a new representational space in higher layers. Note that the demand of tasks did not provide any information on the hierachical stucture of objects, and therefore it only shaped the level of fineness of semantic relatedness. Given the similarity in anatomy between DCNNs and primates' systems, future studies are advocated to examine whether primates' visual cortex also follows similar rules to transfer sensation to perception and finally to concepts that lead to behaviors.

In sum, our study demonstrated that perceptual similarity among object categories and the demand of tasks jointly shaped the hierarchical structure among objects. However, there are several limitations to this study. First, this finding did not necessarily rule out the role of conceptual guidance in forming the semantic relatedness, which was clearly illustrated by a moderate correlation between the DCNNs and humans in the hierarchical structure among objects. In addition, the DCNNs used in this study are purely feedforward, and may not be suitable for studies on conceptual guidance. Therefore, other deep neural networks with feedback connections, such as Feedback-CNN or predictive coding network (Lotter et al., 2016; Cao et al., 2018), or networks directly trained with lexical and semantic relations (Bayer and Riccardi, 2016), shall be used to understand how relations between concepts modulate the semantic relatedness of objects without the influence of perceptual experiences. Second, it is counter-intuitive that the semantic relatedness was not derived from object co-occurrence in natural images. That is, it may result from features, rather than co-appearance frequencies, shared by objects. Further studies are needed to examine this hypothesis to unveil the bottom-up mechanism in forming the semantic relatedness of objects.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/supplementary material.

AUTHOR CONTRIBUTIONS

TH, ZZ, and JL designed research and wrote the paper. TH performed research and analyzed data. All authors contributed to the article and approved the submitted version.

FUNDING

This study was funded by the National Natural Science Foundation of China (Grant No. 31861143039, 31771251), the National Key R&D Program of China (Grant No. 2019YFA0709503), and the National Basic Research Program of China (Grant No. 2018YFC0810602).

REFERENCES

Bao, P., She, L., McGill, M., and Tsao, D. (2020). A map of object space in primate inferotemporal cortex. *Nature* 583, 103–108. doi: 10.1038/s41586-020-2350-5

Barlow, H. B. (1961). "Possible principles underlying the transformations of sensory messages," in *Sensory Communication* (Oxford, UK: Wiley), 217–234.

Barsalou, L. W. (2008). Grounded cognition. *Annu. Rev. Psychol* 59, 617–645. doi: 10.1146/annurev.psych.59.103006.093639

Bayer, A. O., and Riccardi, G. (2016). Semantic language models with deep neural networks. *Comput. Speech Lang.* 40, 1–22. doi: 10.1016/j.csl.2016.04.001

Cao, C., Huang, Y., Yang, Y., Wang, L., Wang, Z., and Tan, T. (2018). Feedback convolutional neural network for visual localization and segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* 41, 1627–1640. doi: 10.1109/TPAMI.2018.2843329

Chen, X., Zhou, M., Gong, Z., Xu, W., Liu, X., Huang, T., et al. (2020). DNNBrain: a unifying toolbox for mapping deep neural networks and brains. *Front. Comput. Neurosci.* 15:580632. doi: 10.3389/fncom.2020.580632

Dan, Y., Atick, J., and Reid, R. (1996). Efficient coding of natural scenes in the lateral geniculate nucleus: experimental test of a computational theory. *J. Neurosci.* 16, 3351–3362. doi: 10.1523/JNEUROSCI.16-10-03351.1996

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. in 2009 IEEE conference on computer vision and pattern recognition. *IEEE.* 52:248–255. doi: 10.1109/CVPR.2009.5206848

Gershkoffstowe, L., and Smith, L. B. (2004). Shape and the first hundred nouns. *Child Dev*. 75, 1098–1114. doi: 10.1111/j.1467-8624.2004.00728.x

Grill-Spector, K., and Weiner, K. S. (2014). The functional architecture of the ventral temporal cortex and its role in categorization. *Nat. Rev. Neurosci.* 15, 536–548. doi: 10.1038/nrn3747

Imai, M., Gentner, D., and Uchida, N. (1994). Children's theories of word meaning: the role of shape similarity in early acquisition. *Cogn. Dev*. 9, 45–75. doi: 10.1016/0885-2014(94)90019-1

Jozwik, K., Kriegeskorte, N., Storrs, K. R., and Mur, M. C. (2017). Deep convolutional neural networks outperform feature-based but not categorical models in explaining object similarity judgments. *Front. Psychol*. 8, 1726–1726. doi: 10.3389/fpsyg.2017.01726

Kastner, D., Baccus, S., and Sharpee, T. (2015). Critical and maximally informative encoding between neural populations in the retina. *Proc. Natl. Acad. Sci. U.S.A.* 112, 2533–2538. doi: 10.1073/pnas.1418092112

Khaligh-Razavi, S. M., and Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Comput. Biol*. 10:15. doi: 10.1371/journal.pcbi.1003915

Lambert, S., Sampaio, E., Mauss, Y., and Scheiber, C. (2004). Blindness and brain plasticity: contribution of mental imagery? An fMRI study. *Cogn. Brain Res*. 20, 1–11. doi: 10.1016/j.cogbrainres.2003.12.012

Leshinskaya, A., and Caramazza, A. (2016). For a cognitive neuroscience of concepts: moving beyond the grounding issue. *Psychon. Bull. Rev*. 23, 991–1001. doi: 10.3758/s13423-015-0870-z

Liu, X., Zhen, Z., and Liu, J. (2020). Hierarchical sparse coding of objects in deep convolutional neural networks. *Front. Comput. Neurosci*. 15:578158. doi: 10.3389/fncom.2020.578158

Logothetis, N. K., and Sheinberg, D. L. (1996). Visual object recognition. *Annu. Rev. Neurosci*. 19, 577–621. doi: 10.1146/annurev.ne.19.030196.003045

Lotter, W., Kreiman, G., and Cox, D. (2016). Deep predictive coding networks for video prediction and unsupervised learning. *arXiv [Preprint]. arXiv*:1605.08104.

Mahon, B. Z., and Caramazza, A. (2009). Concepts and categories: a cognitive neuropsychological perspective. *Annu. Rev. Psychol*. 60, 27–51. doi: 10.1146/annurev.psych.60.110707.163532

Mandler, J. M., and Mcdonough, L. (1993). Concept formation in infancy. *Cogn. Dev*. 8, 291–318. doi: 10.1016/S0885-2014(93)80003-C

Mandler, J. M., and Mcdonough, L. (1998). On developing a knowledge base in infancy. *Dev. Psychol*. 34, 1274–1288. doi: 10.1037/0012-1649.34.6.1274

McClelland, J. L., and Rogers, T. T. (2003). The parallel distributed processing approach to semantic cognition. *Nat. Rev. Neurosci*. 4, 310–322. doi: 10.1038/nrn1076

Miller, G. A. (1995). WordNet: a lexical database for English. *Commun.* 38, 39–41. doi: 10.1145/219717.219748

Noppeney, U. (2007). The effects of visual deprivation on functional and structural organization of the human brain. *Neurosci. Biobehav. Rev*. 31, 1169–1180. doi: 10.1016/j.neubiorev.2007.04.012

Noppeney, U., Friston, K. J., and Price, C. J. (2003). Effects of visual deprivation on the organization of the semantic system. *Brain* 126, 1620–1627. doi: 10.1093/brain/awg152

Olivas, E. S., Guerrero, J. D. M., Martinez-Sober, M., Magdalena-Benedito, J. R., and Serrano, L. (2009). *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*. IGI Global.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019). "Pytorch: an imperative style, high-performance deep learning library," *Advances in Neural Information Processing Systems* (Vancouver), 8026–8037.

Pauen, S. (2002). The global-to-basic level shift in infants' categorical thinking: first evidence from a longitudinal study. *Int. J. Behav. Dev*. 26, 492–499. doi: 10.1080/01650250143000445

Peterson, J. C., Abbott, J. T., and Griffiths, T. L. (2018). Evaluating (and improving) the correspondence between deep neural networks and human representations. *Cogn. Sci.* 42, 2648–2669. doi: 10.1111/cogs.12670

Pruden, S. M., Hirshpasek, K., Golinkoff, R. M., and Hennon, E. A. (2006). The birth of words: ten-month-olds learn words through perceptual salience. *Child Dev*. 77, 266–280. doi: 10.1111/j.1467-8624.2006.00869.x

Ricciardi, E., Bonino, D., Pellegrini, S., and Pietrini, P. (2014). Mind the blind brain to understand the sighted one! Is there a supramodal cortical functional architecture? *Neurosci. Biobehav. Rev*. 41, 64–77. doi: 10.1016/j.neubiorev.2013.10.006

Roy, D. (2005). Grounding words in perception and action: computational insights. *Trends Cogn. Sci*. 9, 389–396. doi: 10.1016/j.tics.2005.06.013

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al. (2015). ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis*. 115, 211–252. doi: 10.1007/s11263-015-0816-y

Samuelson, L. K., and Smith, L. B. (2005). They call it like they see it: spontaneous naming and attention to shape. *Dev. Sci*. 8, 182–198. doi: 10.1111/j.1467-7687.2005.00405.x

Saxe, A. M., Mcclelland, J. L., and Ganguli, S. (2019). A mathematical theory of semantic development in deep neural networks. *Proc. Natl. Acad. Sci. U.S.A*. 116, 11537–11546. doi: 10.1073/pnas.1820226116

Sloutsky, V. M. (2003). The role of similarity in the development of categorization. *Trends Cogn. Sci.* 7, 246–251. doi: 10.1016/S1364-6613(03)00109-8

Song, Y., Qu, Y., Xu, S., and Liu, J. (2020). Implementation-independent representation for deep convolutional neural networks and humans in processing faces. *bioRxiv*, 2020.06.26.171298. doi: 10.1101/2020.06.26.171298

Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annu. Rev. Neurosci.* 19, 109–139. doi: 10.1146/annurev.ne.19.030196.000545

Turner, M., Sanchez, G. L., Schwartz, O., and Rieke, F. (2019). Stimulus- and goal-oriented frameworks for understanding natural vision. *Nat. Neurosci* 22, 15–24. doi: 10.1038/s41593-018-0284-0

Wu, Z., and Palmer, M. (1994). Verb semantics and lexical selection. *ArXiv Prepr.* doi: 10.3115/981732.981751

Xiang, X., Xu, Y., and Huang, M. (2019). Task-driven common representation learning via bridge neural network. *Proc. AAAI Conf. Artif. Intell*. 33, 5573–5580. doi: 10.1609/aaai.v33i01.33015573

Xu, S., Zhang, Y., Zhen, Z., and Liu, J. (2020). The face module emerged in a deep convolutional neural network selectively deprived of face experience. *bioRxiv*, 2020.07.06.189407. doi: 10.1101/2020.07.06.189407

Yamins, D., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., and Dicarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci. U.S.A*. 111, 8619–8624. doi: 10.1073/pnas.1403112111

Yamins, D. L., and DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci*. 19, 356–365. doi: 10.1038/nn.4244

Yang, G. R., Joglekar, M. R., Song, H. F., Newsome, W. T., and Wang, X. (2019). Task representations in neural networks trained to perform many cognitive tasks. *Nat. Neurosci*. 22, 297–306. doi: 10.1038/s41593-018-0310-2

Yang, G. R., and Wang, X.-J. (2020). Artificial neural networks for neuroscientists: a primer. *Neuron* 107, 1048–1070. doi: 10.1016/j.neuron.2020.09.005

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Computational Neuroscience

ORIGINAL RESEARCH
published: 26 February 2021
doi: 10.3389/fncom.2021.619508

Target Detection Using Ternary Classification During a Rapid Serial Visual Presentation Task Using Magnetoencephalography Data

Chuncheng Zhang[1], *Shuang Qiu*[1], *Shengpei Wang*[1] *and Huiguang He*[1,2,3]*

[1] National Laboratory of Pattern Recognition and Research Center for Brain-Inspired Intelligence, Institute of Automation, Chinese Academy of Sciences, Beijing, China, [2] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China, [3] Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Beijing, China

Background: The rapid serial visual presentation (RSVP) paradigm is a high-speed paradigm of brain–computer interface (BCI) applications. The target stimuli evoke event-related potential (ERP) activity of odd-ball effect, which can be used to detect the onsets of targets. Thus, the neural control can be produced by identifying the target stimulus. However, the ERPs in single trials vary in latency and length, which makes it difficult to accurately discriminate the targets against their neighbors, the near-non-targets. Thus, it reduces the efficiency of the BCI paradigm.

Methods: To overcome the difficulty of ERP detection against their neighbors, we proposed a simple but novel ternary classification method to train the classifiers. The new method not only distinguished the target against all other samples but also further separated the target, near-non-target, and other, far-non-target samples. To verify the efficiency of the new method, we performed the RSVP experiment. The natural scene pictures with or without pedestrians were used; the ones with pedestrians were used as targets. Magnetoencephalography (MEG) data of 10 subjects were acquired during presentation. The SVM and CNN in EEGNet architecture classifiers were used to detect the onsets of target.

Results: We obtained fairly high target detection scores using SVM and EEGNet classifiers based on MEG data. The proposed ternary classification method showed that the near-non-target samples can be discriminated from others, and the separation significantly increased the ERP detection scores in the EEGNet classifier. Moreover, the visualization of the new method suggested the different underling of SVM and EEGNet classifiers in ERP detection of the RSVP experiment.

Conclusion: In the RSVP experiment, the near-non-target samples contain separable ERP activity. The ERP detection scores can be increased using classifiers of the EEGNet model, by separating the non-target into near- and far-targets based on their delay against targets.

Keywords: RSVP, ERP, MEG, CNN, SVM

OPEN ACCESS

Edited by:
Ke Zhou,
Beijing Normal University, China

Reviewed by:
Jing Jin,
East China University of Science and Technology, China
Wei Wei,
Xi'an Polytechnic University, China

****Correspondence:***
Huiguang He
huiguang.he@ia.ac.cn

Received: *20 October 2020*
Accepted: *20 January 2021*
Published: *26 February 2021*

Citation:
Zhang C, Qiu S, Wang S and He H (2021) Target Detection Using Ternary Classification During a Rapid Serial Visual Presentation Task Using Magnetoencephalography Data. Front. Comput. Neurosci. 15:619508. doi: 10.3389/fncom.2021.619508

INTRODUCTION

Rapid serial visual presentation (RSVP) is a high-speed brain–computer interface (BCI) experiment paradigm. In the rapid presented sequences, the odd-ball pictures can trigger the unique event-related potential (ERP) activity, known as P300 visual-evoked potentials in the brain (Won et al., 2019). This neural signal is generally chosen from a variety of well-studied non-invasive electroencephalography (EEG) and magnetoencephalography (MEG) signals (Lawhern et al., 2018). The detection of ERP onsets can be used to identify the pictures of interest in the sequence (Helfrich and Knight, 2019). As a result, the RSVP paradigm has been used in multiple BCI applications, e.g., picture identification, screen spellers, and other applications that require identifying target stimulus at high speed.

The applications of RSVP in BCI largely depend on the ERP detection accuracy. The machine learning methods have been widely used in ERP detection using the noisy single sample signals (Huang et al., 2011; Cecotti, 2016; Lin et al., 2017). Machine learning algorithm formulates the classifier to learn the ERP pattern in the high-dimensional neural signal, and automatically suppress the effect of noise. The xdawn algorithm was used to enhance ERP components in the EEG and MEG data. Support vector machine (SVM), linear discriminator (Cecotti, 2016), and convolution neural network (CNN) classifiers (Lawhern et al., 2018) have been applied to ERP detection tasks (Xiao et al., 2020). The weighted linear discriminant analysis has been used to reduce calibration time in the P300-based BCI paradigm; it not only reduces the computation request but also reduces the fatigue of subjects prior to BCI experiment (Jin et al., 2020b). Further, optimal feature selection method of common spatial pattern using L1-norm and Dempster–Shafer theory has been used in the EEG dataset to improve the robust against the non-stationary across time and subjects (Jin et al., 2020c). Despite the improvements in algorithm, it is still difficult to obtain the reliable ERP waveform from a single trial since the signal-to-noise rate is large in neural signal (Creel, 2019).

Besides the algorithm improvement, the paradigm of RSVP experiments also evolved. Jin et al. has developed a novel cheeks-stim paradigm for the P300 BCI experiment to substantially increase the efficiency and experience of BCI users (Lin et al., 2018; Jin et al., 2020a). Indeed, the reliable ERP can be obtained by averaging the waveform of several ERP trials, and there are RSVP paradigm improvements using the averaged multiple trials to increase the accuracy of ERP detection. Lin et al. developed a novel triple RSVP paradigm for the P300 BCI speller. It presented three single target character stimuli three times and uses the averaged signal to increase ERP detection accuracy (Lin et al., 2018). Cecotti et al. used the dual-RSVP paradigm. The sequence was presented synchronously with a fixed lag, and the succeeding two signals were used to increase the ERP detection accuracy (Cecotti, 2016). Additionally, the triple-RSVP paradigm has also been used to acquire higher accuracy (Mijani et al., 2019). It shows that the classifiers took the benefit from the dual sample combination and produced higher detection score. The new RSVP paradigm designs indeed improved the performance of the RSVP BCI application; however, it still left the difficulty of single sample ERP detection problem unsolved, which is important to common RSVP applications.

One of the main difficulties of ERP detection using a single trial is their complex dynamics (Barry and De Blasio, 2018), since they vary in latency and length across trials. The high-speed presented stimulus in the RSVP paradigm makes the stimulus closer with each other and the difference more ambiguous in temporal. Evenly, the presentation speed is becoming so fast that the ERP reaches its peak after the next stimuli onset, when the presentation rate is larger than 30 Hz. Thus, detecting the target samples against their neighbors is becoming more difficult and produces a higher error rate on the single-trial ERP detection.

In this study, we presented an RSVP experiment with MEG data acquired. The visual material is natural scene pictures with or without pedestrians, and the pictures with pedestrians were used as target pictures. We used a new training method to increase the ERP detection scores. In the new method, the samples were separated into three classes instead of two classes in the traditional method. They are target, near-non-target, and far-non-target samples. Thus, we used the classifier not only discriminating the target and other samples but also learning the difference between target samples and their neighbors. The SVM and CNN in EEGNet architecture classifiers were trained to detect ERP based on MEG data. The experiment results showed that the new training method improved ERP detection scores of the EEGNet classifier. The visualization results further explained the different underling of ERP detection of SVM and EEGNet classifiers.

MATERIALS AND METHODS

Visual Stimuli and Procedure

The participants were seated in the MEG scanner, and a screen was in front at 680 *mm*. During the MEG scanning, they were required to gaze on the center of the screen. The rapid visual stimuli were presented on the screen using a rapid flashed sequence of pictures. The picture size was 500 × 500 $pixels^2$ covering 150 × 150 mm^2 areas in the screen; thus, it subtended the area of 12.6 × 12.6 $degrees^2$ in visual angle. The flash rate of pictures was set as 10 *Hz*, and there were no gaps between two consecutive pictures.

All the pictures were selected from a dataset consisting of 1,400 colored street scene pictures. The pictures containing pedestrians were used as target pictures, and others were used as non-target pictures. There were 56 target pictures and 1,344 non-target pictures in the dataset.

During a block, 100 pictures were shown in random order. The ratio of target pictures was set to 4%, resulting in 100 pictures with 4 target pictures and 96 non-target pictures. In every block, the 100 pictures were randomly sampled from the dataset without replacement. As a result, one session contained 14 blocks. During a session, the participants were required to press a button in their right hand when they were ready to start a block and press the same button when they see a target picture as soon as possible. The aim of asking participants to press the button is to keep them focused on the screen, and the button-pressing events were also recorded to make sure that the participant saw the target pictures instead of missing them. All the participants finished 11 consecutive sessions during the RSVP experiment. The paradigm of the RSVP experiment can be found in **Figure 1**.

Participants

The experiment recruited 10 college students as participants in the RSVP experiment (seven males and three females, aged 23.79 ± 3.6) without previous training in the task. The participants practiced through a pseudo-RSVP block immediately before they entered the MEG scanner. The aim was only to make sure they had understood the rule of button pressing during the experiment. The participants exhibited normal or corrected-to-normal vision with no neurological problems and were financially compensated for their participation. The study was approved by the local ethics committee (Institute of Automation Chinese Academy of Sciences). All participants gave a written informed consent and received payment for their participation.

MEG Acquirement and Preprocessing

During MEG experiment, subjects performed RSVP experiment in a MEG scanner. MEG recordings were conducted in a magnetically shielded room with a whole-head CTF MEG system with 272 channels (MISL-CTF DSQ-3500, Vancouver, BC, Canada) at the MEG Center of Institute of Biophysics, Chinese Academy of Sciences. Prior to data acquisition, three coils were attached to the left and right pre-auricular points and nasion of each participant, and a head localization procedure was performed before and after each acquisition to locate the participant's head relative to the coordinate system fixed to the MEG system. Participants were asked to lie in a supine position, and a projection screen was used to present visual stimuli during recording.

MEG data were recorded at a sampling rate of $1,200$ *Hz*, filtered between 0 and 600 *Hz*. We preprocessed the data using MNE software (Gramfort et al., 2014). The artificial noise of eye moving was suppressed using ICA method (Dimigen, 2020). Since ICA is sensitive to low-frequency drifts, the 1-Hz high-pass filter was used to suppress low-frequency signal prior to ICA fitting. Then, the sources with large skewness, kurtosis, and variance scores were marked and zeroed out from raw data. Then, the raw data were down-sampled to the sample rate of 100 *Hz*. The down-sampled data were then filtered by a band-pass filter to fetch data in the frequency band of 0.115 *Hz*.

Data samples were then fetched from the filtered data. For every picture presented in the RSVP experiment, the time window ranging from -200 to $1,200$ *ms* from the onset was used to fetch the data sample. The samples also referred to the MEG epochs in some studies. The samples were baseline-corrected by the averaged value between -200 and 0 *ms* from the onset. The linear drifts were removed from the samples. As a result, the data sample could be represented by a matrix of 272 rows and 140 columns; 272 rows represented 272 channels and 140 columns represented 140 time points from −200 to 1,200 ms. The samples were then used to detect ERP activity. The averaged time series of the signals are plotted in **Figure 2**.

ERP-Based Target Detection

MEG Sample Labeling

The lag between samples was 100 ms since the presentation rate was 10 Hz. However, the length of the samples was 1,400 ms. Thus, the samples inevitably overlapped with their neighbors. The traditional ERP detection method used dual classification, which only separated target and non-target samples, e.g., labeling target epochs as label 1 and non-target epochs as label 2. As a result, they used the same label to represent the non-target samples with or without ERP components. It forced the classifier to distinguish the ERP signals against their neighbors, which might contain the same ERP with a small latency. Thus, the confusion will inevitably decrease the accuracy of ERP detection.

In this study, we used three classification methods to further separate the target signal from their neighbors. Three labels were used in the experiment: target label (noted as 1), far-non-target label (noted as 2), and near-non-target label (noted as 3). The far-non-target samples refer to the epochs whose onset was far from target stimulus, which means that there were no target stimuli occurring within a 0.5-s range. The other non-target epochs were labeled as near-non-target labels. Simply, the target samples were ERP samples, the near-non-target samples contained ERP but of incorrect latency, and the far-non-target samples did not contain ERP activity.

ERP Detection Using SVM

The SVM is a widely used statistical learning algorithm, especially for large datasets with high dimensionality (Vapnik, 1998). It has been reported that SVM outperforms other competing methods in many researches (Williams, 2003; Pohlmeyer et al., 2011). The SVM has also been used for ERP detection in the RSVP experiment (Huang et al., 2011). Since the SVM was originally designed for binary classification, the trinary classification method used the one-against-one method that was proposed by Chih-Wei and Chih-Jen (2002) in the "libsvm" software package.

The prior feature extraction was also necessary for SVM classifier. We used signal enhancement with xdawn algorithm (Rivet et al., 2009). The xdawn method was used as a supervised feature extraction method to enhance the ERP components in the MEG data by maximizing the signal-to-signal-plus-noise rate (Cecotti, 2016). The number of components was set to six in this study based on prior research and visualization results. Thus, the 272-sensor MEG data were converted into six-component feature data to fit the SVM classifier.

SVM uses RBF kernel to explore more flexible classification strategy for high-dimensional data. In this study, we set the prior parameter gamma as "scale" to automatically calculate the variance of the training data. Since non-target samples were dozen times outnumbered target samples, we set the class-weight option as "balanced" to increase the weight of target signal in loss function to obtain a meaningful classifier.

ERP Detection Using EEGNet

EEGNet is an outstanding CNN architecture to detect ERP signal in EEG data (Lawhern et al., 2018). In this study, we used EEGNet to detect the ERP signal in MEG data. The EEGNet

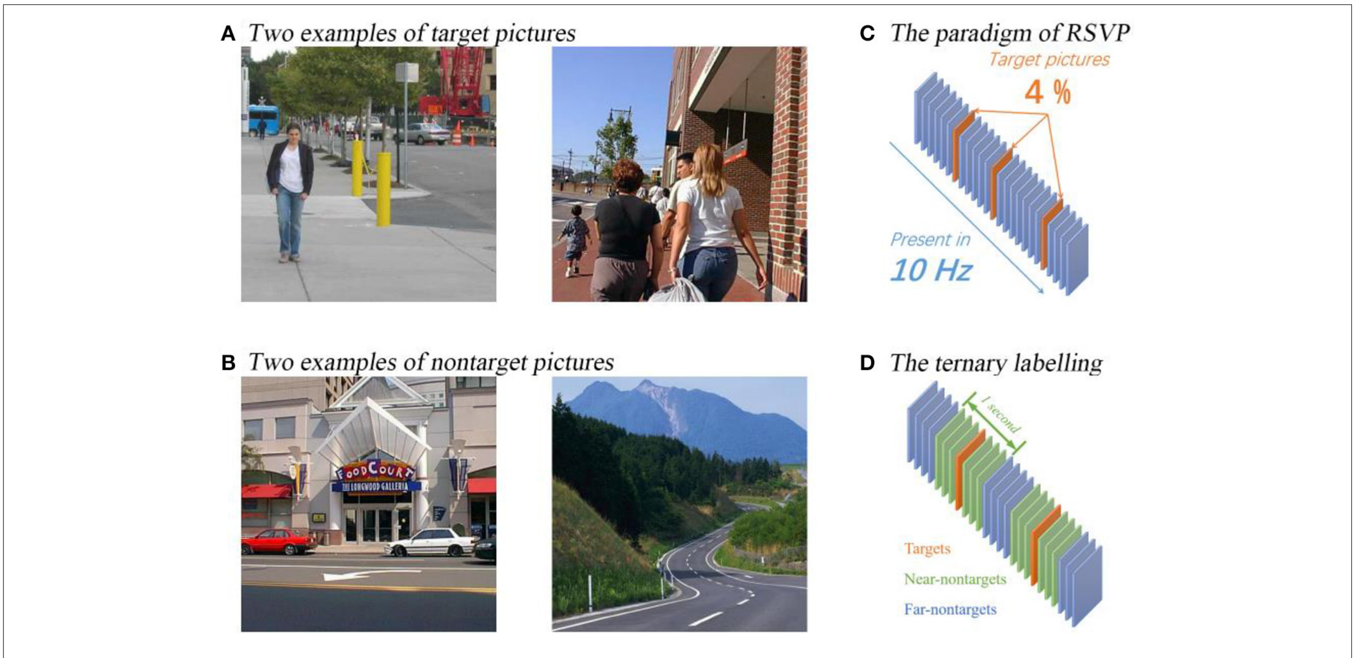


FIGURE 1 | The paradigm and examples of pictures used in the RSVP experiment. The examples of target and non-target pictures are plotted on **(A,B)**. The paradigm is plotted on **(C)** and the ternary classification labeling protocol is plotted on **(D)**.

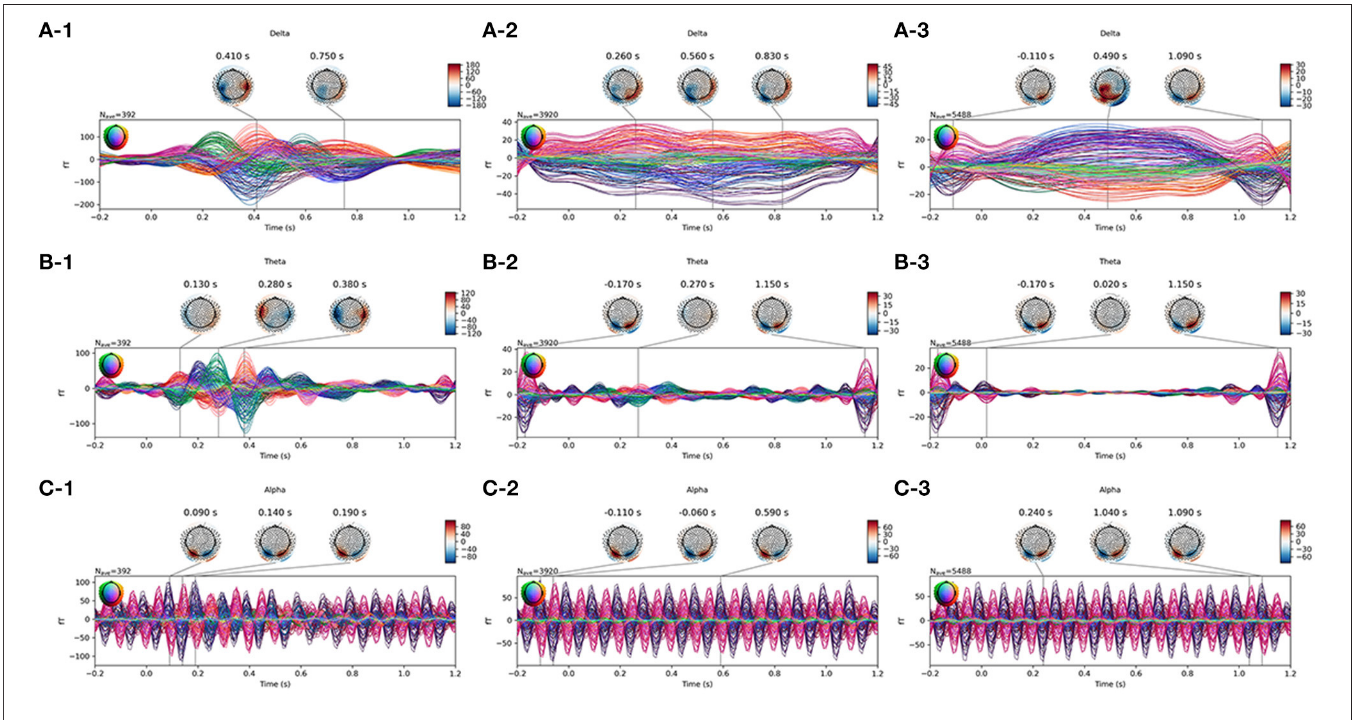


FIGURE 2 | Average waveform visualization of MEG samples. The A/B/C row plots the average waveform of the frequency of Delta/Theta/Alpha band. The 1/2/3 column plots the average waveform of target/near-/far-non-target samples.

classifier was built and trained using "pytorch" toolbox in the high-performance GPU server. Since there were 272 sensors in the MEG data other than the 64 sensors in the EEG data, we changed the input number to 272 accordingly. Additionally, we used softmax function in the output to match the ternary classification. The loss function was calculated using the output of EEGNet and one hot-coded sample label. The architecture was the same as the "DeepConvNet" model of EEGNet (Lawhern

TABLE 1 | Scoures table of ERP detection.

	Recall	Precision	F1 score	Accuracy
SVM (binary)	0.8206 ± 0.1304	0.8649 ± 0.0828	0.8364 ± 0.1027	0.9875 ± 0.0068
SVM (ternary)	0.8243 ± 0.1259	0.8610 ± 0.0823	0.8384 ± 0.1027	0.9876 ± 0.0070
Net (binary)	**0.8740** ± **0.0837**	0.7574 ± 0.1216	0.8085 ± 0.0987	0.9829 ± 0.0097
Net-3 (ternary)	0.8513 ± 0.0847	**0.8731** ± **0.0775**	**0.8608** ± **0.0749**	**0.9890** ± **0.0059**

The bold values refer the highest value of the column.

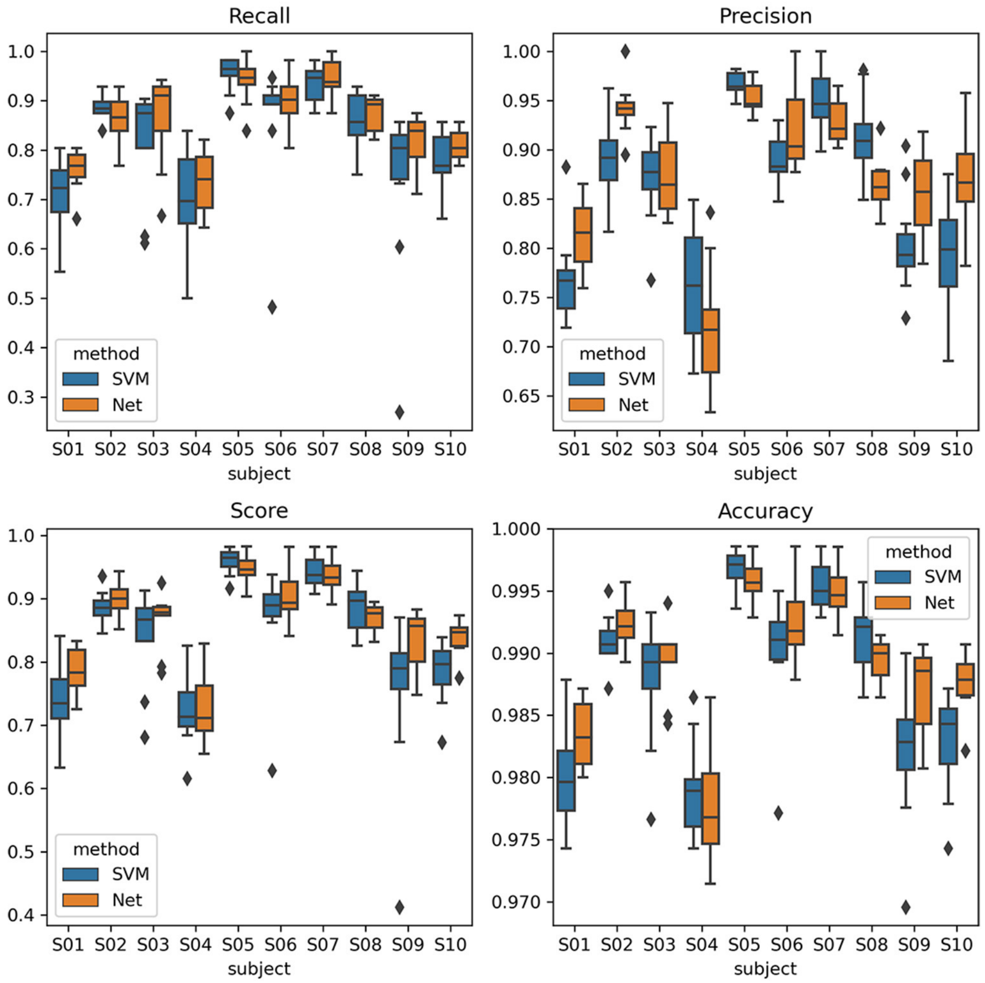


FIGURE 3 | ERP detection scores with ternary classification of all the 10 subjects in box-and-whisker plots. The SVM labels refer to the score of SVM classifier, and the Net labels refer to the score of EEGNet classifier.

et al., 2018). The parameters in the EEGNet were upgraded using the Adam optimizer. The learning rate was set as 0.001 for initiate and then the rate was set to shrink to 0.8 times every 50 epochs to avoid overfitting. The training process contained 500 epochs, and 300 training samples with equal class number were randomly selected in each epoch. Since the EEGNet classifier performed feature selection automatically using the first convolution layers, the band filtered MEG data were used directly without additional feature extraction process in prior.

Cross-Session Validation

We used the SVM and CNN model in EEGNet architecture classifier to detect ERP for identifying the target samples. To evaluate the reusability of the classifiers, we applied cross folder protocol to separate the MEG data into training and testing dataset recursively. The separation is based on the sessions of the experiment to keep the independency between the training and testing data. Since all the subjects finished 11 sessions of the RSVP experiment, we applied the 11-folder protocol. In each folder, the data of one session were used as testing dataset, and data of other sessions were concatenated to generate the training data.

In folders of 11 sessions, the following training and testing procedure were repeated. In the SVM part, the training dataset was used to train the xdawn spatial filter to perform feature extraction, and then the features were used to train an SVM classifier. The testing dataset was then applied by the trained xdawn spatial filter and SVM classifier to evaluate the detection scores. In the EEGNet part, the training dataset was used to train the parameter of the net without prior feature extraction and then the testing dataset was used to evaluate the detection scores.

As a result, we performed cross-session validation within subject to validate the discriminating power of the method. It was operated as the online experiment simulation. The model was fitted to samples in training sessions and then the test samples were transformed one by one to obtained the labels. Although the ternary classification gave labels of three class labels, we merged the near- and far-non-target labels as the non-target label. Thus, the ternary classification method was used to increase the discriminating power, and it was transparent to the experiment since it eventually produced binary labels.

Additionally, we also visualized the features to investigate the ERP detection underling of SVM and EEGNet classifiers. For the SVM classifier, the features extracted by the xdawn spatial filter were visualized. For EEGNet, the activity of the first convolution layer was visualized. We used the TSNE projection method to project the features into the two-dimensional manifold space. In the space, we showed the distribution of the target, near-, and far-non-target samples in a distance invariance manner.

RESULTS

ERP Detection Scores

The ERP detection scores were recorded and compared between SVM and EEGNet classifiers. The scores of interests are the recall rate, precision rate, and F1 score of the target samples, which was also the aim of the RSVP experiment. The average scores of all the subjects were shown in **Table 1**. The recall score was higher for the EEGNet classifier. Additionally, the ternary classification method increased the scores of the EEGNet classifier beyond the SVM. The scores of EEGNet and SVM using ternary classification of all the subjects were plotted in **Figure 3**. It showed that the scores of EEGNet was higher than SVM on more subjects. The variance among cross-session folders of the EEGNet method were smaller. Moreover, the EEGNet with the ternary classification method also produced the highest F1 scores.

To make sure the comparison was valid, we applied analysis of variance (ANOVA) (Rouder et al., 2016) and paired *t*-test (Xu et al., 2017) method to test the statistical level of the difference between the scores. Firstly, to settle the complicity of the experiment, we used ANOVA to testify if the difference between the scores was because of the usage of classifiers. As a result, we used three-factor ANOVA; the factors were subject factor, folder factor, and method factor. The results showed that the method factor had main effect, which suggested that the choice of classifiers affected the scores. Then, we used the *t*-test method to obtain the *p*-value of the difference. The results showed that the increase of the EEGNet was significant since the *p*-value was < 0.001 for recall score and F1 score, please see **Table 2** for the detail values.

TABLE 2 | ANOVA tables of scores.

	Df	sum_sq	mean_sq	*F*	*PR* (>*F*)
Recall					
Subject	9.0	1.1160	0.1240	21.4217	4.6338e−24
Method	1.0	0.0354	0.0354	6.1178	1.4347e−02
Folder	10.0	0.0956	0.0095	1.6530	9.5467e−02
Resibinary	173.0	1.0015	0.0057	NaN	NaN
Precision					
Subject	9.0	0.8201	0.0911	42.8921	1.3668e−39
Method	1.0	0.0071	0.0071	3.3651	6.8306e−02
Folder	10.0	0.0409	0.0040	1.9261	4.4532e−02
Resibinary	173.0	0.3675	0.002125	NaN	NaN
F1 Score					
Subject	9.0	0.9869	0.1096	37.6439	2.4480e−36
Method	1.0	0.0244	0.0244	8.3976	4.2433e−03
Folder	10.0	0.0620	0.0062	2.1310	2.4425e−02
Resibinary	173.0	0.5039	0.0029	NaN	NaN
Accuracy					
Subject	9.0	0.0058	0.0006	55.5035	1.9985e−46
Method	1.0	0.0000	0.0000	8.0118	5.1975e−03
Folder	10.0	0.0003	0.0000	2.7216	3.9466e−03
Resibinary	173.0	0.0020	0.0000	NaN	NaN

Figure 4 shows the confusion matrix of the classification. Firstly, it shows that the near- and far-non-target samples can be discriminated. The first row of the matrix had three columns, which showed the ratio of target samples being detected as target, near-non-target, and far-non-target samples. The second and third rows showed the ratio of near-non-target and far-non-target samples, respectively. As a result, the diagonal values were the ratio of the three classes of samples being correctly classified. The other values were the ratio of being incorrectly classified. The first row was used to calculate the scores of target samples classification. The value in the first column referred to the true-positive rate (TPR) (Albieri and Didelez, 2014). The value in the second and third columns referred to the false-negative rate (FNR) of target to near-non-target and far-non-target, respectively. The first column was used to calculate the scores of samples being classified to target samples. The false-positive rate (FPR) of near-non-target to target was the value of the second row and first column in the matrix.

The results showed that the TPR of target samples was higher in EEGNet, and the FNR of target to far-non-target was lower in EEGNet. According to the first column, the FPR of near-non-target to target is lower in EEGNet. According to the other elements in the matrixes, the discriminating power between target and near-non-target was also higher in EEGNet. It suggested that the higher scores of EEGNet were due to the fact that the new three classification method could increase the discriminating power between target and near-non-target of the EEGNet classifier.

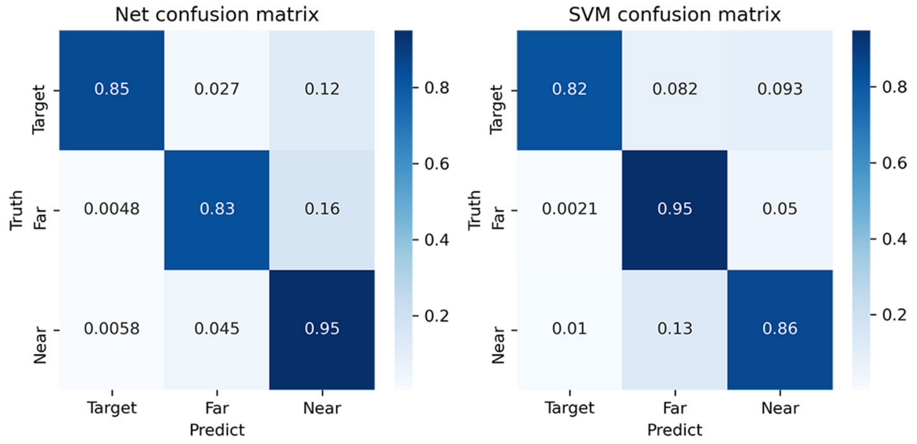


FIGURE 4 | The average confusion matrix of the SVM and EEGNet method using ternary classification. The float numbers on the grids are the average value of percentage.

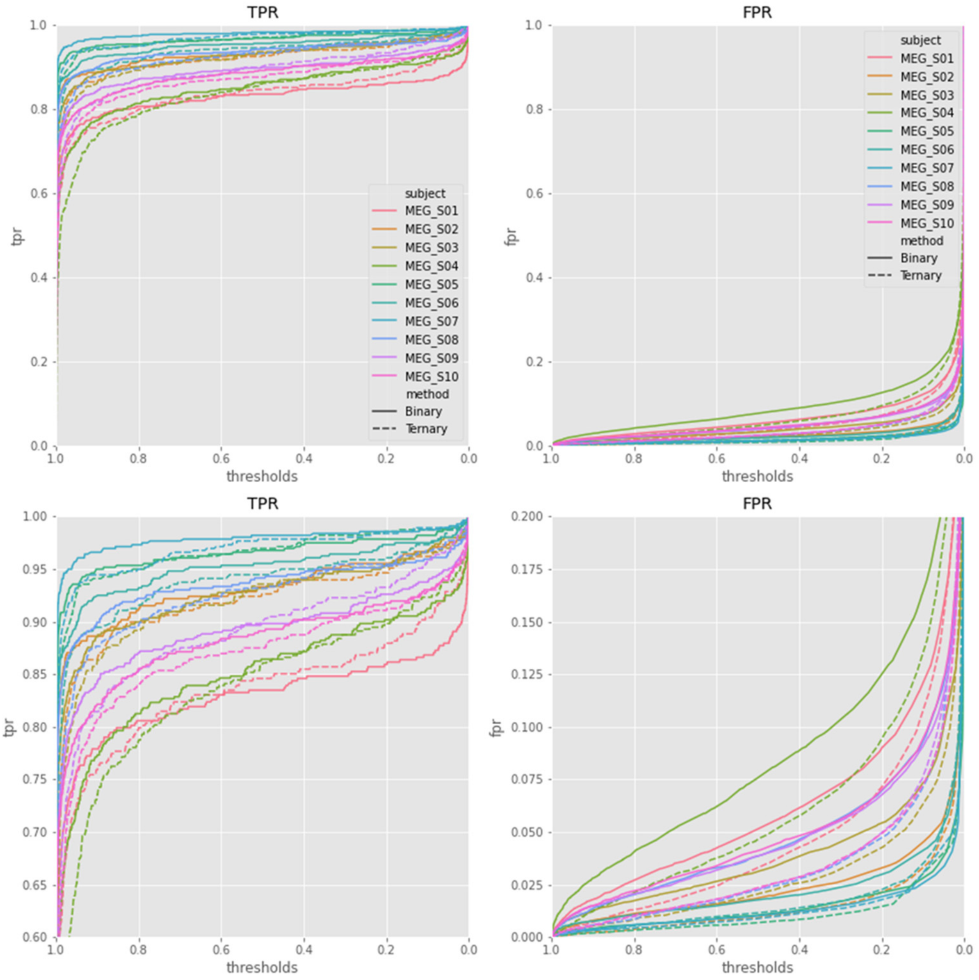


FIGURE 5 | The TPR and FPR curves of EEGNet with thresholds of all the 10 subjects. The two plots on the bottom were the same as the plots on the top, other than using a smaller value range.

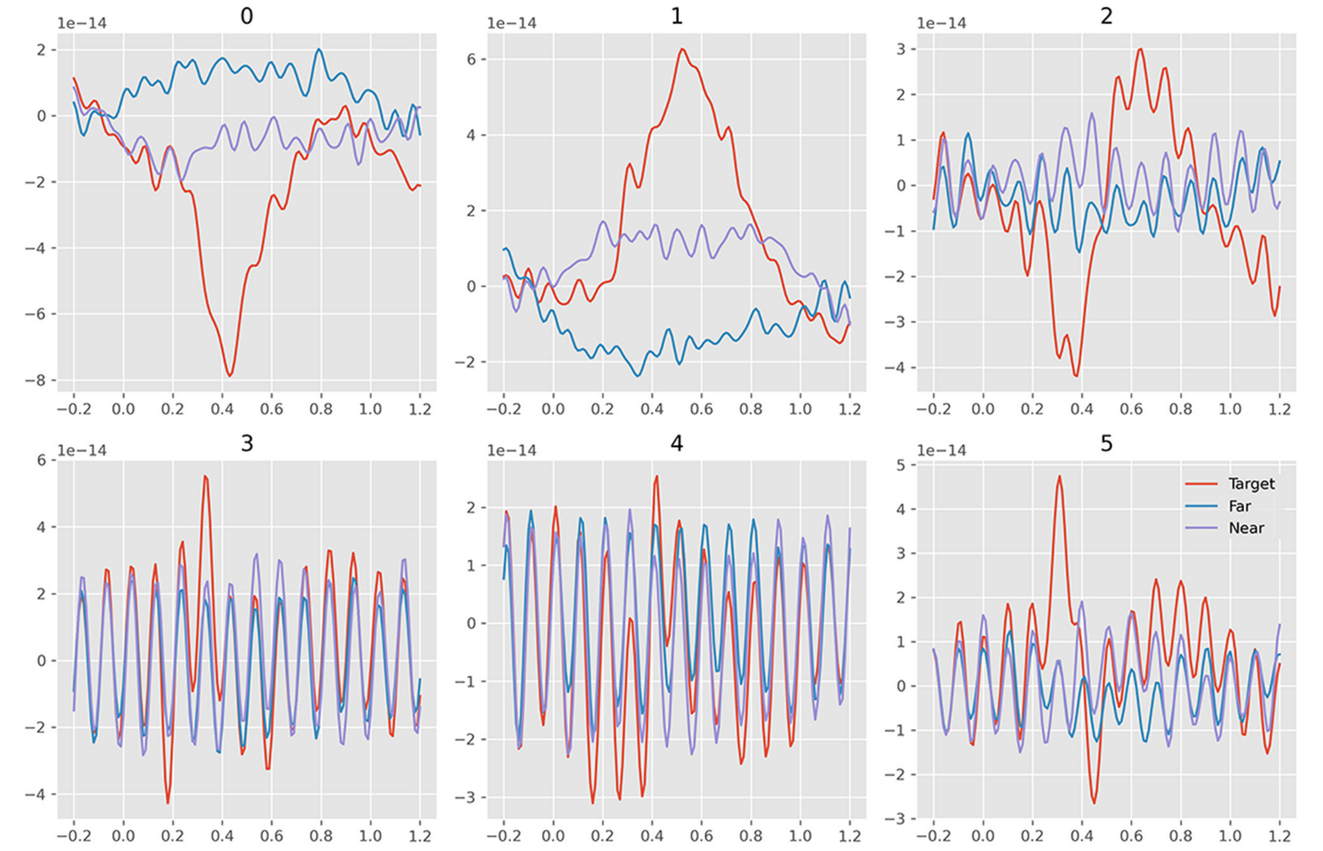


FIGURE 6 | Average waveform plots of six xdawn features. The six grids refer to the six features; the colors refer to ternary kinds of samples.

Additionally, since we used softmax function on the output layer of EEGNet, the probability of the sample as a target sample could be obtained. The TPR and FPR curves among different thresholds (Zhang et al., 2015) of target samples were plotted in **Figure 5** based on the output of EEGNet. The area under curve (AUC) values of EEGNet were 0.9808 ± 0.0197 of binary classification and 0.9858 ± 0.0136 of ternary classification. The results showed that the ternary classification produced higher AUC values and lower FPR values than traditional binary classification protocol. The results suggested that the ternary classification method can largely suppress the FPR of target samples.

Visualization

Figure 2 plots the waveform and topotactical activity of averaged samples of one subject on different frequency bands. The graphs used the joint plotting visualization method of MNE software, and the colors represented the 272 channels of the MEG set. The waveform of target samples on the Delta band clearly showed the ERP activity of the target pictures. The waveform on the Alpha band showed the SSVEP activity triggered by the 10-Hz presentation, and the SSVEP occurred in all the three kinds of samples. The differences between near- and far-non-target samples were mainly on the Delta band, and even their activities were both weak. It showed that the activity pattern of near-non-target samples was similar to target samples, and the far-non-target samples did not show similarity.

Figure 6 plots all the six averaged components of xdawn extraction. The order was set as decreasing order of explained variance. It turned out that the first three features cover the main differences between target and non-target signals. There was little difference between near- and far-non-target samples. The SSVEP components mainly existed in the latter three features, which suggested that they were less important to ERP detection. **Figure 7** plots samples in the two-dimensional manifold space. It showed a similar trend with the averaged plot. The first three features were more separated among the three kinds of samples.

The visualization of EEGNet features was done in the same way as the SVM features. **Figure 8** plots the waveform of the averaged 25 features. **Figure 9** plots the features in the two-dimensional manifold space. It showed that all the 25 features show difference between three kinds of samples. The difference between near- and far-non-target samples was also clear. Moreover, the features containing SSVEP also showed difference among three classes. The features of No. 11, 13, 14, 15, 16, 17, 20, 22, and 24 showed a large difference between target and non-target samples. The features of No. 3, 5, 20, and 19 showed moderate difference between near- and far-non-target

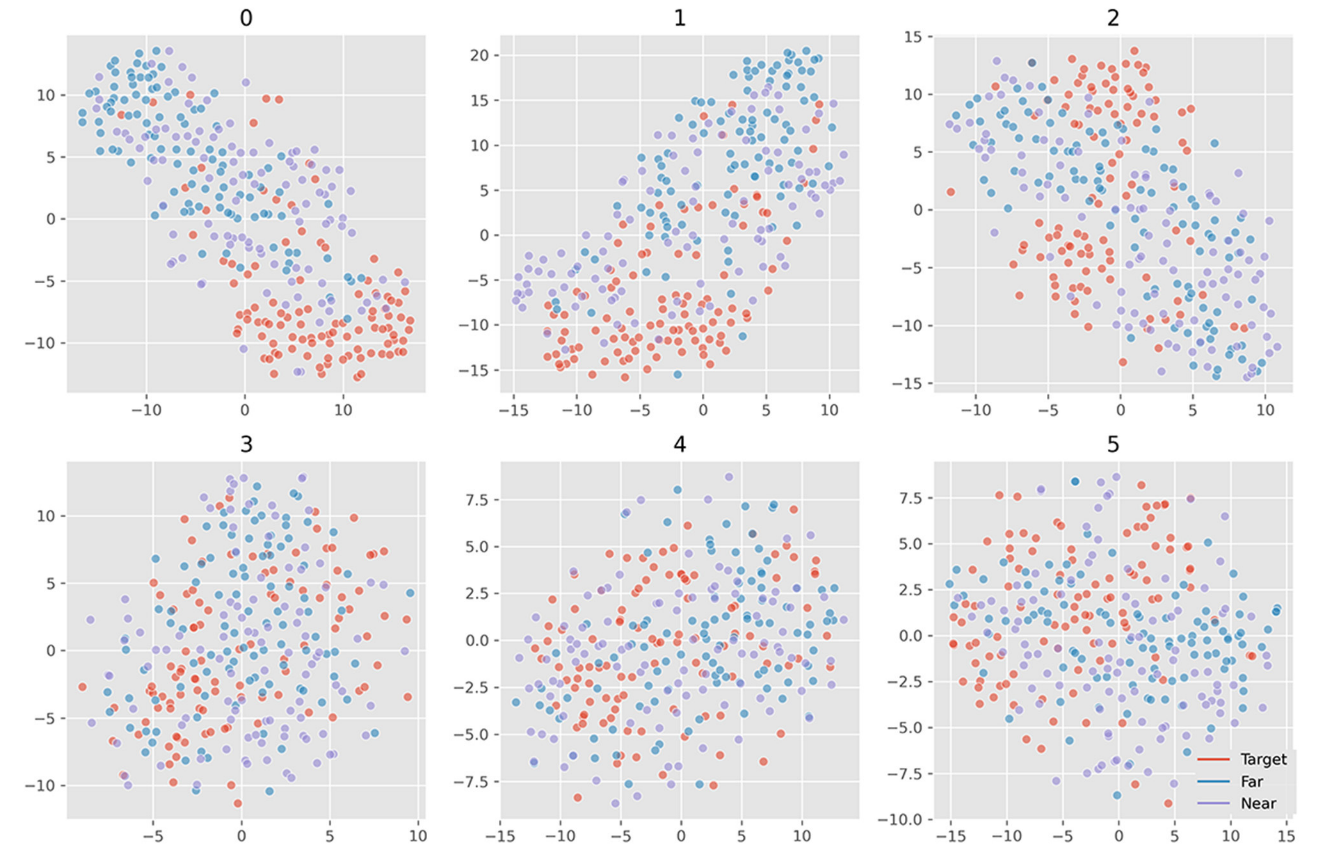


FIGURE 7 | Projection of samples of six xdawn features in two-dimensional manifold space. The six grids refer to the six features; the colors refer to ternary kinds of samples.

samples. The results were consistent with the confusion matrix of **Figure 4**, which showed large error rate between near- and far-non-target samples.

DISCUSSION

In this study, the MEG data were acquired during the RSVP experiment; the rapid presented pictures were natural scene pictures, and the pictures with pedestrians were used as target pictures. We presented the new ternary classification method to train the SVM and EEGNet classifiers to detect ERP signal to identify the onset of target stimulus. The new method has improved the detection scores using the EEGNet classifier.

The traditional machine learning method in the RSVP experiment only used binary classification to discriminate the target and other samples. The method ignores the similarity of target samples and their neighbors. The speed of RSVP in the experiment was 10 *Hz*. The latency between two samples was 0.1 s. However, the latency of a classic reliable ERP was about 0.3 s, which was widely known as the P300 feature signal (Mijani et al., 2019). The length of the ERP was not narrow either. As a result, the near-non-target samples inevitably contained the ERP the same as target samples (see the average waveform in **Figure 2**). The difference between them was only that the target samples contained the ERP with the "correct" latency, which was occasionally too small in some samples to distinguish them.

We separated the samples into three classes: target, near-, and far-non-target samples. The waveforms showed that the difference between them were mainly on the Delta band, and the near-non-target samples were more similar to the target samples (see **Figure 2**). The visualization of the features showed the difference between them either (see **Figures 6–9**). Thus, the non-target samples should be separated into two sets, the ones near a target sample (near-non-target) and others (far-non-target). The traditional methods did not separate the two kinds of non-target samples either. As a result, the classifiers had to solve the confusion by detecting some ERP signals and discarding others, which was bad to ERP detection.

The new ternary classification method trained the classifier to learn not only the difference between target and others but also the difference between target samples and their neighbors. It actually separated the ERP detection task into two folders. The first one was to detect ERP components in the samples to find target and near-non-target samples. The second one was to distinguish the two classes. The confusion matrix of EEGNet proved that the new method increased the TPR of target and near-target samples (see **Figure 4**).

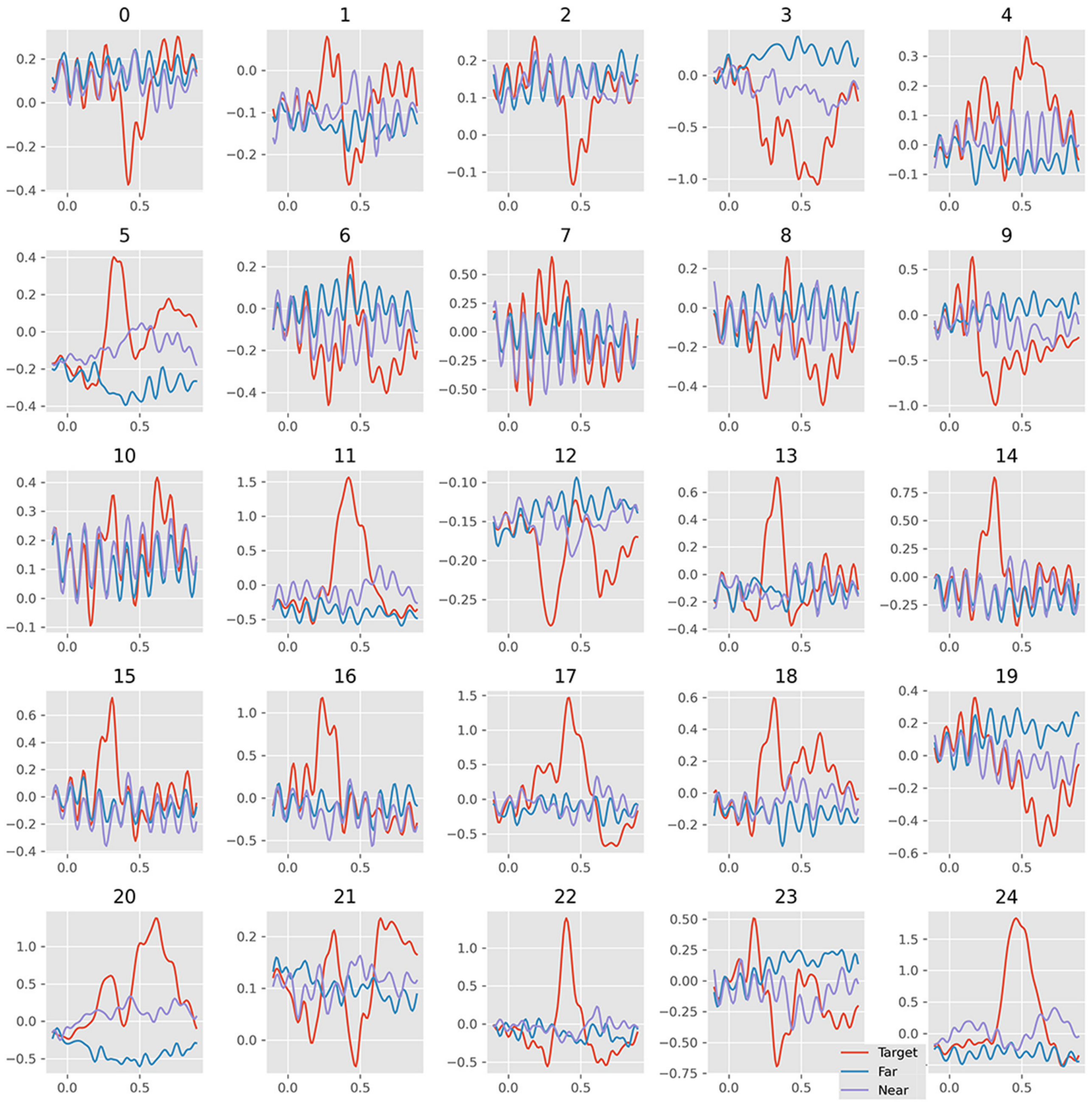


FIGURE 8 | Average waveform plots of 25 EEGNet features. The 25 grids refer to the 25 features; the colors refer to ternary kinds of samples.

Compared with the SVM classifier, the EEGNet provided higher TPR for ERP component detection. Although the TPR of far-non-target samples was lower than SVM, the incorrect samples were more likely to be classified as the near-non-target samples. Finally, the scores of target samples using EEGNet were overly higher than using the SVM classifier. The ROC plots of EEGNet showed the difference between the traditional binary and new ternary methods in detail (see **Figure 5**). The FPRs of target detection of the ternary method were largely lower than those of the binary method, while the TPRs of the two methods were similar. The results explained that the ternary method produced higher precision score than the binary method (see **Table 1**). It was shown that the TPR only reached 0.85 in confusion matrix (see **Figure 5**) and the overall accuracy reached 0.98 (see **Table 1**). The reason was the non-target samples largely outnumbered the target samples. Based on EEGNet classifier results, the TNR was extremely high (see the second and third row of the first column of the confusion matrix), which made the overall accuracy higher than the TRP value of target samples.

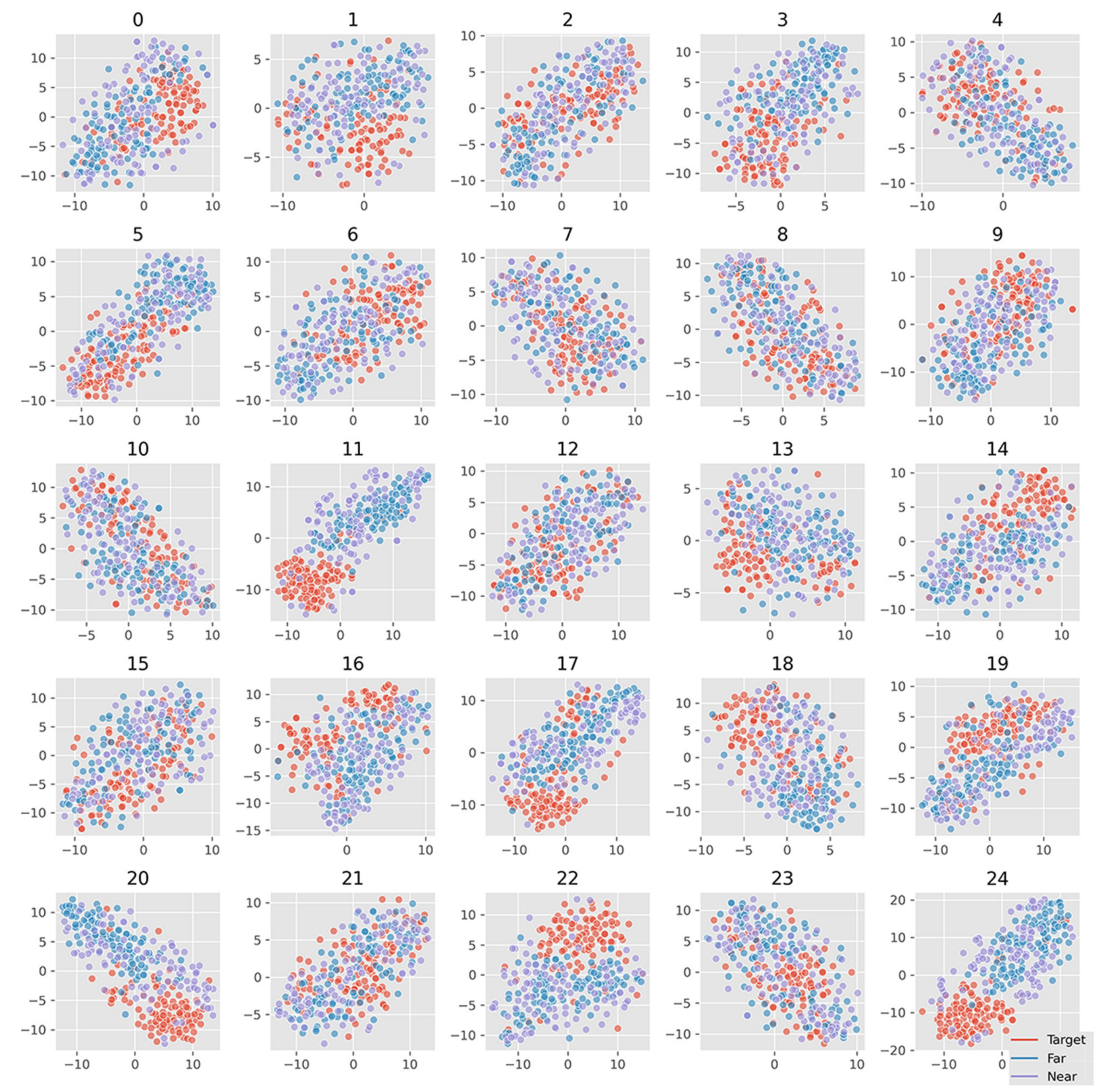


FIGURE 9 | Projection of samples of 25 EEGNet features in two-dimensional manifold space. The 25 grids refer to the 25 features; the colors refer to ternary kinds of samples.

Based on the results of the study, the separation increased the ERP detection scores. The results suggested that the reason EEGNet produced higher ERP detection scores was that it had learned the difference between the samples with ERP and other samples without ERP signals. Furthermore, the results also suggested that the CNN model was better at detecting ERP components despite their variance in latency, which were consistent with the translation invariance of the CNN model (Furukawa, 2017). The visualization of 25 features of samples also verified that the CNN model can effectively extract the useful features automatically in the RSVP experiment (see **Figure 9**). As a result, the xdawn spatial filter was not necessary for the EEGNet classifier. Meanwhile, it also hinted that the CNN model could benefit from the correct separation of the samples.

The SVM classifier did not benefit from the ternary method. It might be due to the fact that SVM used time points in the samples as independent feature dimensions. The shifts of ERP components in near-non-target samples converted the feature from dimensions. Thus, it was hard for the SVM classifier to track the dependence between the time points. The reason we used xdawn in SVM classification was the lack of automatically extracting features of the SVM classifier (Bascil et al., 2016). The results also suggested that the six components had fully covered the explainable variance, and the increase of the components was not necessary.

CONCLUSION

In this study, the MEG data were acquired during the RSVP experiment; the rapid presented pictures were natural scene pictures, and the pictures with pedestrians were used as target pictures. We also presented the new ternary classification method to train the SVM and EEGNet classifiers to detect ERP signal to identify the onset of target stimulus. We obtained a fair ERP detection accuracy using traditional SVM and EEGNet classifiers. The proposed ternary classification method showed the discrimination of the near- and far-non-targets in the RSVP experiment and increased accuracy in the EEGNet classifier. The visualization of the results also uncovered the different ERP detection underling between SVM and EEGNet classifiers.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Institute of Automation Chinese Academy of Sciences. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

CZ operated the experiment, analyzed the data, and wrote the manuscript. SQ operated the experiment. SW jointed in data analyzing. HH was in charge of the project. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported in part by the National Natural Science Foundation of China under Grant 61976209, Grant 62020106015, and Grant 61906188, in part by the Chinese Academy of Sciences (CAS) International Collaboration Key Project under Grant 173211KYSB20190024; and in part by the Strategic Priority Research Program of CAS under Grant XDA27000000, in part by the National Nature Science Foundation of China grant (31730039).

REFERENCES

Albieri, V., and Didelez, V. (2014). Comparison of statistical methods for finding network motifs. *Stat. Appl. Genet. Mol. Biol.* 13, 403–422. doi: 10.1515/sagmb-2013-0017

Barry, R. J., and De Blasio, F. M. (2018). EEG frequency PCA in EEG-ERP dynamics. *Psychophysiology* 55:e13042. doi: 10.1111/psyp.13042

Bascil, M. S., Tesneli, A. Y., and Temurtas, F. (2016). Spectral feature extraction of EEG signals and pattern recognition during mental tasks of 2-D cursor movements for BCI using SVM and ANN. *Australas. Phys. Eng. Sci. Med.* 39, 665–676. doi: 10.1007/s13246-016-0462-x

Cecotti, H. (2016). Single-trial detection with magnetoencephalography during a dual-rapid serial visual presentation task. *IEEE Trans. Biomed. Eng.* 63, 220–227. doi: 10.1109/tbme.2015.2478695

Chih-Wei, H., and Chih-Jen, L. (2002). A comparison of methods for multiclass support vector machines. *IEEE Trans. Neural Netw.* 13, 415–425. doi: 10.1109/72.991427

Creel, D. J. (2019). Visually evoked potentials. *Handb. Clin. Neurol.* 160, 501–522. doi: 10.1016/b978-0-444-64032-1.00034-5

Dimigen, O. (2020). Optimizing the ICA-based removal of ocular EEG artifacts from free viewing experiments. *Neuroimage* 207:116117. doi: 10.1016/j.neuroimage.2019.116117

Furukawa, H. (2017). Deep learning for target classification from SAR imagery: data augmentation and translation invariance.

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., et al. (2014). MNE software for processing MEG and EEG data. *Neuroimage* 86, 446–460. doi: 10.1016/j.neuroimage.2013.10.027

Helfrich, R. F., and Knight, R. T. (2019). Cognitive neurophysiology: event-related potentials. *Handb. Clin. Neurol.* 160, 543–558. doi: 10.1016/b978-0-444-64032-1.00036-9

Huang, Y., Erdogmus, D., Pavel, M., Mathan, S., and Hild, K. E. (2011). A framework for rapid visual image search using single-trial brain evoked responses. *Neurocomputing* 74, 2041–2051. doi: 10.1016/j.neucom.2010.12.025

Jin, J., Chen, Z., Xu, R., Miao, Y., Wang, X., and Jung, T.-P. (2020a). Developing a novel tactile p300 brain-computer interface with a cheeks-stim paradigm. *IEEE Trans. Biomed. Eng.* 67, 2585–2593. doi: 10.1109/tbme.2020.2965178

Jin, J., Li, S., Daly, I., Miao, Y., Liu, C., Wang, X., et al. (2020b). The study of generic model set for reducing calibration time in P300-based brain–computer interface. *IEEE Trans. Neural Syst. Rehabil. Eng.* 28, 3–12. doi: 10.1109/tnsre.2019.2956488

Jin, J., Xiao, R., Daly, I., Miao, Y., Wang, X., and Cichocki, A. (2020c). Internal feature selection method of CSP based on L1-Norm and Dempster-Shafer theory. *IEEE Trans. Neural Netw. Learn. Syst.* 1–12. doi: 10.1109/tnnls.2020.3015505

Lawhern, V. J., Solon, A. J., Waytowich, N. R., Gordon, S. M., Hung, C. P., and Lance, B. J. (2018). EEGNet: a compact convolutional neural network for EEG-based brain-computer interfaces. *J. Neural Eng.* 15:056013. doi: 10.1088/1741-2552/aace8c

Lin, Z., Zeng, Y., Gao, H., Tong, L., Zhang, C., Wang, X., et al. (2017). Multirapid serial visual presentation framework for EEG-based target detection. *Biomed. Res. Int.* 2017:2049094. doi: 10.1155/2017/2049094

Lin, Z., Zhang, C., Zeng, Y., Tong, L., and Yan, B. (2018). A novel P300 BCI speller based on the Triple RSVP paradigm. *Sci. Rep.* 8:3350. doi: 10.1038/s41598-018-21717-y

Mijani, A. M., Shamsollahi, M. B., and Sheikh Hassani, M. (2019). A novel dual and triple shifted RSVP paradigm for P300 speller. *J. Neurosci. Methods* 328:108420. doi: 10.1016/j.jneumeth.2019.108420

Pohlmeyer, E. A., Wang, J., Jangraw, D. C., Lou, B., Chang, S. F., and Sajda, P. (2011). Closing the loop in cortically-coupled computer vision: a brain-computer interface for searching image databases. *J. Neural Eng.* 8:036025. doi: 10.1088/1741-2560/8/3/036025

Rivet, B., Souloumiac, A., Attina, V., and Gibert, G. (2009). xDAWN algorithm to enhance evoked potentials: application to brain–computer interface. *IEEE Trans. Biomed. Eng.* 56, 2035–2043. doi: 10.1109/TBME.2009.2012869

Rouder, J. N., Engelhardt, C. R., McCabe, S., and Morey, R. D. (2016). Model comparison in ANOVA. *Psychon. Bull. Rev*. 23, 1779–1786. doi: 10.3758/s13423-016-1026-5

Vapnik, V. N. (1998). Statistical learning theory. *Encycl. Ences Learn.* 41, 3185.

Williams, C. K. I. (2003). Learning with kernels: support vector machines, regularization, optimization, and beyond. *J. Am. Stat. Assoc.* 98:489. https://www.researchgate.net/deref/http%3A%2F%2Fdx.doi.org%2F10.1198%2Fjasa.2003.s269 doi: 10.1198/jasa.2003.s269

Won, K., Kwon, M., Jang, S., Ahn, M., and Jun, S. C. (2019). P300 speller performance predictor based on RSVP multi-feature. *Front. Hum. Neurosci.* 13:261. doi: 10.3389/fnhum.2019.00261

Xiao, X., Xu, M., Jin, J., Wang, Y., Jung, T. P., and Ming, D. (2020). Discriminative canonical pattern matching for single-trial classification of ERP components. *IEEE Trans. Biomed. Eng.* 67, 2266–2275. doi: 10.1109/TBME.2019.2958641

Xu, M., Fralick, D., Zheng, J. Z., Wang, B., Tu, X. M., and Feng, C. (2017). The differences and similarities between two-sample T-test and paired T-test. *Shanghai Arch. Psychiatry* 29, 184–188. doi: 10.11919/j.issn.1002-0829.217070

Zhang, X., Li, X., Feng, Y., and Liu, Z. (2015). The use of ROC and AUC in the validation of objective image fusion evaluation metrics. *Signal Process.* 115, 38–48. doi: 10.1016/j.sigpro.2015.03.007

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Computational Neuroscience

ORIGINAL RESEARCH
published: 10 March 2021
doi: 10.3389/fncom.2021.620281

Multidimensional Face Representation in a Deep Convolutional Neural Network Reveals the Mechanism Underlying AI Racism

Jinhua Tian [1], Hailun Xie [1], Siyuan Hu [1] and Jia Liu [2]**

[1] Beijing Key Laboratory of Applied Experimental Psychology, Faculty of Psychology, Beijing Normal University, Beijing, China,
[2] Department of Psychology & Tsinghua Laboratory of Brain and Intelligence, Tsinghua University, Beijing, China

OPEN ACCESS

Edited by:
*Petia D. Koprinkova-Hristova,
Institute of Information and
Communication Technologies
(BAS), Bulgaria*

Reviewed by:
*Connor Parde,
The University of Texas at Dallas,
United States
Jiankang Deng,
Imperial College London,
United Kingdom*

****Correspondence:***
*Siyuan Hu
husiyuan@bnu.edu.cn
Jia Liu
liujiathu@tsinghua.edu.cn*

Received: *22 October 2020*
Accepted: *08 February 2021*
Published: *10 March 2021*

Citation:
*Tian J, Xie H, Hu S and Liu J (2021)
Multidimensional Face Representation
in a Deep Convolutional Neural
Network Reveals the Mechanism
Underlying AI Racism.
Front. Comput. Neurosci. 15:620281.
doi: 10.3389/fncom.2021.620281*

The increasingly popular application of AI runs the risk of amplifying social bias, such as classifying non-white faces as animals. Recent research has largely attributed this bias to the training data implemented. However, the underlying mechanism is poorly understood; therefore, strategies to rectify the bias are unresolved. Here, we examined a typical deep convolutional neural network (DCNN), VGG-Face, which was trained with a face dataset consisting of more white faces than black and Asian faces. The transfer learning result showed significantly better performance in identifying white faces, similar to the well-known social bias in humans, the other-race effect (ORE). To test whether the effect resulted from the imbalance of face images, we retrained the VGG-Face with a dataset containing more Asian faces, and found a reverse ORE that the newly-trained VGG-Face preferred Asian faces over white faces in identification accuracy. Additionally, when the number of Asian faces and white faces were matched in the dataset, the DCNN did not show any bias. To further examine how imbalanced image input led to the ORE, we performed a representational similarity analysis on VGG-Face's activation. We found that when the dataset contained more white faces, the representation of white faces was more distinct, indexed by smaller in-group similarity and larger representational Euclidean distance. That is, white faces were scattered more sparsely in the representational face space of the VGG-Face than the other faces. Importantly, the distinctiveness of faces was positively correlated with identification accuracy, which explained the ORE observed in the VGG-Face. In summary, our study revealed the mechanism underlying the ORE in DCNNs, which provides a novel approach to studying AI ethics. In addition, the face multidimensional representation theory discovered in humans was also applicable to DCNNs, advocating for future studies to apply more cognitive theories to understand DCNNs' behavior.

Keywords: deep convolutional neural network, faces, other race effect, multidimensional face race representation, contact theory

INTRODUCTION

With enormous progress in artificial intelligence (AI), deep convolutional neural networks (DCNN) have shown extraordinary performance in computer vision, natural language processing, and complex strategy video games. However, the application of DCNNs increases the risk of amplifying social bias (Zou and Schiebinger, 2018). For example, a word-embedding processing system may associate women with homemakers, or a face identification network may match non-white faces to inanimate objects, suggesting the existence of gender and race biases in DCNNs (Bolukbasi et al., 2016). Although the phenomenon of social bias has been widely recognized, the underlying mechanism of such bias is little understood (Caliskan et al., 2017; Garg et al., 2018). In this study, we explored how biased behaviors were generated in DCNNs.

Insight into human biases may help to understand DCNNs' biased responses. A classical race bias, the other race effect (ORE) (Malpass and Kravitz, 1969; Valentine, 1991), shows that people are better at identifying faces of their own race than those of other races (Meissner and Brigham, 2001). The reason underlying the ORE is that people usually have more experiences with faces of their own race (Valentine, 1991), which leads to a better capacity of recognizing faces of their own race. Accordingly, we reasoned that a similar biased response might also be present in DCNNs, as DCNNs tend to perform better on data that most closely resembles the training data. Note that the biased response in DCNNs is not identical to the ORE in humans; however, given the same underlying causes, here we borrowed the term "ORE" to index the biased responses in DCNNs for simplicity. On the other hand, one influential human recognition theory, the face multidimensional representation space (MDS) theory, proposes that ORE comes from the difference in representing faces in a multidimensional space, or simply "face space" (Valentine, 1991; Valentine et al., 2016; O'toole et al., 2018). According to this theory, face space is a Euclidean multidimensional space, with dimensions representing facial features. The distance between two faces in the space indexes their perceptual similarity. Under the frame of this theory, faces of one's own race are scattered widely in the face space (i.e., high distinctiveness) and faces of other races are clustered in a smaller space (i.e., low distinctiveness) (Valentine, 1991; Valentine et al., 2016). Therefore, the higher distinctiveness in representation leads to better recognition of own-race faces than that of other-race faces. In this study, we examine whether the ORE in DCNNs, if observed, may be accounted for by a similar mechanism.

To address the aforementioned question, the current study chose a typical DCNN, VGG-Face (**Figure 1A**), which is widely used for face recognition (Parkhi et al., 2015). We first examined whether there was a similar ORE in VGG-Face and explored its face representation space using MDS theory. First, we manipulated the ratio of face images of different races to examine whether the ORE in the VGG-Face changed as a function of the frequencies of encountered races (Chiroro and Valentine, 1995). Secondly, we examined whether frequent interaction with one race led to sparser distribution (i.e., high distinctiveness) in VGG-Face's representation space. Thirdly, we explored whether the difference in representation led to the ORE.

MATERIALS AND METHODS

Convolutional Neural Network Model

In this study, a well-known deep neural network, VGG-Face (available in http://www.robots.ox.ac.uk/~albanie/pytorch-models.html) was used for model testing, model retraining, and model activation extraction (Parkhi et al., 2015). An illustration of the VGG-Face architecture is shown in **Figure 1A**. This framework consists of five groups of convolutional layers and three fully connected layers, with 16 layers in total. Each convolutional layer comprises some convolution operators, followed by a non-linear rectification layer, such as ReLU and max pooling. The input images (for example, 3 × 224 × 224 pixels color image) are transferred into 2,622 representational units, each corresponding to a unit of the last fully connected layer (FC3), representing a certain identity.

Face Stimuli

The VGG-Face was originally trained for face identification tasks with the VGGFace dataset (including 2,622 identities in total, with 2,271 downloadable identities).

As shown in **Figure 1B**, to test the performance of the VGG-Face on three races, 300 different identities were selected from another face dataset, VGGFace2 (Cao et al., 2018). Face images that were present in both the VGGFace and VGGFace2 datasets were excluded (see https://github.com/JinhuaTian/DCNN_other_race_effect/tree/master/face_materials for details). We classified the remaining 8,250 identities into four groups: white (6,995 identities), black (518 identities), Asian (345 identities), and other races (392 identities). Three hundred identities were randomly selected from the first three groups (100 identities for each race) and separated into in-house transferring learning (300 identities, each containing 100 images), validating (300 identities, each containing 50 images), and testing (300 identities, each containing 50 images) datasets. These three datasets contained the same identities but with different face exemplars; therefore, biased responses were unlikely to be introduced at the phase of transferring learning. Note that the dataset for transferring learning, validating, and testing was not overlapped with the dataset used for pre-training the network. We performed the transfer learning on the VGG-Face with the transfer learning dataset, validated the model with the validating dataset, and finally used the testing dataset to measure the identification accuracy of three different races. To confirm the reproducibility of our results, we sampled the other two datasets for transfer learning (detailed information is provided in the **Supplementary Material 1.3**).

Transfer Learning

We tested the identification performance of VGG-Face with new identities using transfer learning (Yosinski et al., 2014), which trains a pre-trained network with another small set of related stimuli. Transfer learning was performed on the pre-trained VGG-Face with the in-house training set. We replaced the last FC

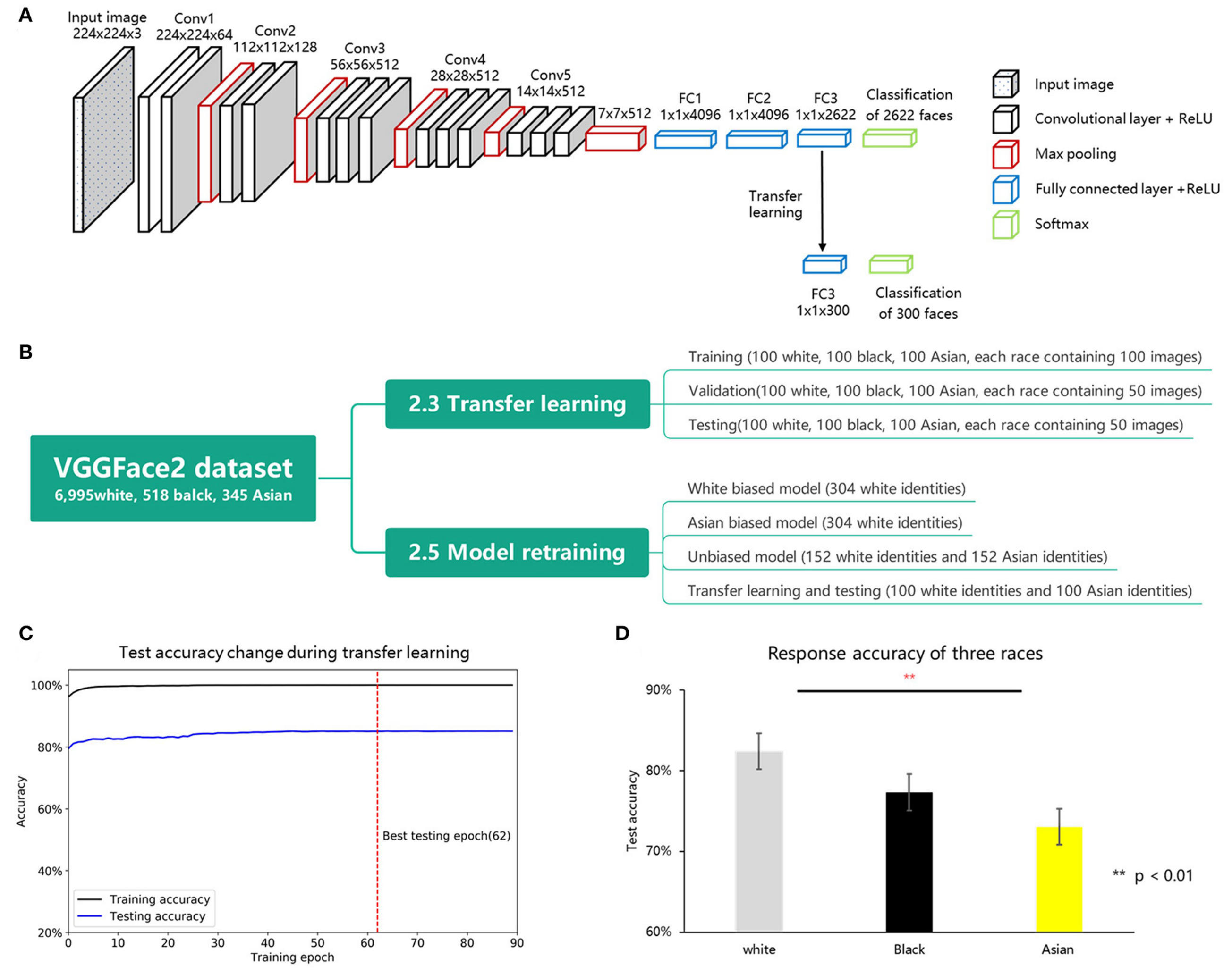

A
Input image 224x224x3
Conv1 224x224x64
Conv2 112x112x128
Conv3 56x56x512
Conv4 28x28x512
Conv5 14x14x512
7x7x512
FC1 1x1x4096
FC2 1x1x4096
FC3 1x1x2622
Classification of 2622 faces
Transfer learning
FC3 1x1x300
Classification of 300 faces
Input image
Convolutional layer + ReLU
Max pooling
Fully connected layer +ReLU
Softmax
B
VGGFace2 dataset
6,995white, 518 balck, 345 Asian
2.3 Transfer learning
Training (100 white, 100 black, 100 Asian, each race containing 100 images)
Validation(100 white, 100 black, 100 Asian, each race containing 50 images)
Testing(100 white, 100 black, 100 Asian, each race containing 50 images)
2.5 Model retraining
White biased model (304 white identities)
Asian biased model (304 white identities)
Unbiased model (152 white identities and 152 Asian identities)
Transfer learning and testing (100 white identities and 100 Asian identities)
C
Test accuracy change during transfer learning
Accuracy
100%
80%
60%
40%
20%
Training epoch
0
10
20
30
40
50
60
70
80
90
Best testing epoch(62)
Training accuracy
Testing accuracy
D
Response accuracy of three races
Test accuracy
90%
80%
70%
60%
**
white
Black
Asian
** p < 0.01

FIGURE 1 | (A) Illustration of VGG-Face's architecture used in this study. The model comprised five convolutional blocks (conv1-conv5) and three fully connected layers (FC1-FC3). **(B)** Data organization of transfer learning and model retraining. **(C)** The change of test accuracy during VGG-Face transfer learning. The x axis represents training accuracy, and the y axis represents training epochs. The black and blue line represent training and validation accuracy changes during model training separately. **(D)** Identification accuracy of the VGG-Face on white, black, and Asian faces.

layer (the third fully connected layer, FC3, containing 2,662 units) of the VGG-Face with another fully connected layer containing 300 units (each representing a unique face identity used in training and testing procedures). Subsequently, we froze the parameters prior to the classification layer (FC3) and trained the FC3 using the training dataset. Detailed training parameters were obtained from a previous study (Krizhevsky, 2014). All networks are trained for face identification using the cross-entropy loss function with a stochastic gradient descent (SGD) optimizer (initial learning rate $= 0.01$, momentum $= 0.9$). Images were normalized to the same luminance (mean $= [0.485, 0.456, 0.406]$, SD $= [0.229, 0.224, 0.225]$) and resized to the $3 \times 224 \times 224$ pixels. Data argumentation used 15° random rotation and a 50% chance of horizontal flip. All models were trained for 90 epochs, and the learning rate decayed $250^{-1/3}$ (≈ 0.159) after every 23 epochs (1/4 training epochs). To achieve optimal training accuracy and prevent overfitting, we saved the best model, which had the highest validating accuracy during training. The training procedure is shown in **Figure 1C**. After transfer learning, this network (the best model) was tested using the testing dataset. The performance difference between the three races was analyzed using a repeated-measures analysis of variance (ANOVA).

Model Retraining

According to human contact theory, low interracial interactions are the main cause of ORE. For a DCNN, biased training data may lead to biased performance. To examine this hypothesis, we further retrained the VGG-Face using two "biased" face sets and one matched face set, and then tested whether these models showed a face bias. The training face sets were composed of different numbers of Asian and white faces. The different composition of Asian and white faces simulates the "white biased," "Asian biased," and "unbiased" datasets.

Retraining Materials

All images used for model retraining and validating were selected from the VGGFace2 datasets. We selected 404 Asian identities and 404 white identities for model training and testing. For the white-biased model, we randomly selected 304 white identities out of 404 identities for model training. For the Asian-biased model, we randomly selected 304 Asian identities out of 404 identities for model training. For unbiased model training, we selected 152 Asian and 152 white identities. The training datasets were further separated into training and validation sets. We selected 30 of each identity (15,000 images in total) as the validation dataset, and the remaining faces (109,450 images for the Asian biased model, 103,745 images for the white biased model, and 105,781 images for the unbiased model) were used for model training. Two hundred other identities (100 identities for each race) were selected for transfer learning and testing.

Retraining Procedure

Recent studies have shown that the softmax loss function in VGG-Face lacks the power of discrimination (Cao et al., 2020), and therefore may result in the ORE observed in the network. To rule out this possibility, we re-trained VGG-Face with new loss functions, such as focal loss (Lin et al., 2017) and Arcface (Deng et al., 2019), which are designed to solve the simple hard example imbalance or long-tailed problem caused by imbalanced training data. We used the same VGG-Face framework as the pre-trained model. All networks were trained for face identification with a stochastic gradient descent (SGD) optimizer (initial learning rate $= 0.01$, momentum $= 0.9$). Images were normalized to the same luminance (mean $= [0.485, 0.456, 0.406]$, SD $= [0.229, 0.224, 0.225]$) and resized to $3 \times 224 \times 224$ pixels. Data argumentation used 15° random rotation and a 50% chance of horizontal flip. All models were trained for 90 epochs, and the learning rate decayed $250^{-1/3}$ (≈ 0.159) after every 23 epochs (1/4 training epochs). To achieve optimal training accuracy and prevent overfitting, we saved the best model, which had the highest validating accuracy during training. The saved model was used for further model testing using the testing dataset.

Face Representation Difference of Three Races in VGG

To explore the representation pattern of different races in VGG-Face, we further analyzed the face representation difference. It has been suggested that activation responses of the layer prior to the final classification layer (the second fully connected layer: FC2) is a typical representation of each face in DCNNs (O'toole et al., 2018). Thus, we extracted the activation responses in the FC2 layer for all the testing faces using an in-house Python package, namely, DNNBrain (Chen et al., 2020) with the PyTorch framework (Paszke et al., 2019).

To describe the distinctness of each race group, we used three measurements to describe the distribution of face space. First, we applied the representation similarity analysis to obtain the representational dissimilarity correlation matrix (RDM) of three race faces with FC2 activation. To further explore the representation difference between the three races, we used the in-group similarity to describe representation variance within a race group. The in-group similarity was calculated as the averaged Pearson correlation of a certain identity with other identities of the same race. Specifically, a face with larger in-group similarity indicated smaller representation distinctiveness. That is, the larger the distinctiveness, the better the performance in discriminating identities.

Next, we used FC2 activation to construct the face space describing the distribution of different faces. Valentine and Endo (1992) assume the face space to be an n-dimensional space; a face is represented as a point localized in the space. The axes of the space represent dimensions to discriminate faces. According to this hypothesis, we used the average activation of all faces as the possible center coordinates of this face space. Thus, we computed the Euclidean distance of the averaged activation from each face to all averaged face activations as a measurement of face distinctiveness. A face with a larger Euclidean distance indicated larger representation distinctiveness. The activation differences in the three races were also analyzed using a one-way ANOVA.

Face Representation Visualization

For a better visualization of the representation of the face space, we used the t-SNE (t-distributed stochastic neighbor embedding, t-SNE) method to reduce face representation dimensions and visualize the activation distribution. The t-SNE starts by converting the high-dimensional Euclidean distances between data points into conditional probabilities that represent similarities (Van Der Maaten and Hinton, 2008). We used the t-SNE to squeeze the activation vectors (2,622 units) of each face's activation into two dimensions and plotted these conditional probabilities on a two-dimensional coordinate for visualization. The t-SNE was performed using default parameters (learning rate $= 200$, iteration $= 1{,}000$).

Correlation Between Face Representation and Identification Performance

To explore whether VGG-Face activation and its performance were correlated, we computed the Spearman correlation as well as the Pearson correlation between the in-group similarity and Euclidean distance with face identification accuracy of the VGG-Face.

RESULTS

First, we used transfer learning to examine race bias in the VGG-Face. The average accuracy of all identities was 77.6%, significantly higher than the stochastic probability (0.33%), indicating the success of transfer learning. A one-way ANOVA showed a significant main effect of race ($F_{2,\,297} = 8.762$, $p < 0.001$, $\eta_p^2 = 0.056$), with white faces being identified significantly better than Asian faces ($p < 0.001$, $d^{'} = 0.545$) and marginally significantly better than black ($p = 0.071$, $d^{'} = 0.353$) faces (**Figure 1D**). No significant difference was found in accuracy between the identification of black and Asian faces ($p = 0.176$, $d' = 0.255$).

To verify face selection bias in VGG network training, we classified the available VGGFace dataset into four groups, namely, white (1,984 identities, 87.2%), black (211 identities, 9.7%), Asian (52 identities, 2.3%), and other races (brown or mixed race, 24 identities, 1.1%). As faces in the dataset were overwhelmingly white, the better identification accuracy for white faces suggested that the ORE also existed in the VGG-Face.

A direct test on whether the ORE observed in the VGG-Face resulted from the imbalance of races present in the dataset was to manipulate the ratio of the number of faces of each race. To do this, we retrained the VGG network using white-biased (white vs. Asian: 100 vs. 0%), Asian biased (0 vs. 100%), and unbiased (50 vs. 50%) datasets, respectively. As shown in **Figure 2**, the three DCNNs showed different patterns of ORE. For the DCNN trained with the white-biased dataset, white faces were identified significantly better than Asian faces (softmax: $t_{198} = 3.934$, $p < 0.001$, $d' = 0.562$; focal loss: $t_{198} = 4.203$, $p < 0.001$, $d' = 0.617$; Arcface: $t_{198} = 3.405$, $p < 0.001$, $d' = 0.486$). In contrast, in the Asian-biased DCNN, Asian faces were identified better than white faces (softmax: $t_{198} = 2.693$, $p = 0.008$, $d' = 0.381$; focal loss: $t_{198} = 2.689$, $p = 0.008$, $d' = 0.382$; Arcface: $t_{198} = 2.0880.$, $p = 0.038$, $d' = 0.296$). Finally, no ORE was found in the unbiased DCNN (softmax: $t_{198} = 1.135$, $p = 0.258$, $d' = 0.161$; Focal loss: $t_{198} = 0.905$, $p = 0.367$, $d' = 0.132$). Taken together, the ORE observed in the VGG-Face resulted from unbalanced experiences with different numbers of faces per race during model training.

How do unbalanced experiences shape the internal representation of faces in the VGG-Face? To address this question, we calculated the correlations between the representations of faces, which were indexed by the activations in the FC2 layer, and then constructed a correlation matrix consisting of Asian, white, and black faces (**Figure 3A**). A direct observation of **Figure 3A** revealed that faces of each race were grouped into one cluster; that is, the representations for faces were more similar within a race than between races, suggesting that faces from the same race were grouped together in the multidimensional space. Importantly, the representational similarity of white faces was smallest, compared with Asian ($p < 0.001$, $d' = 1.29$) and black ($p < 0.001$, $d' = 2.077$) faces, and that of Asian faces was smaller than that of black faces ($p < 0.001$, d'

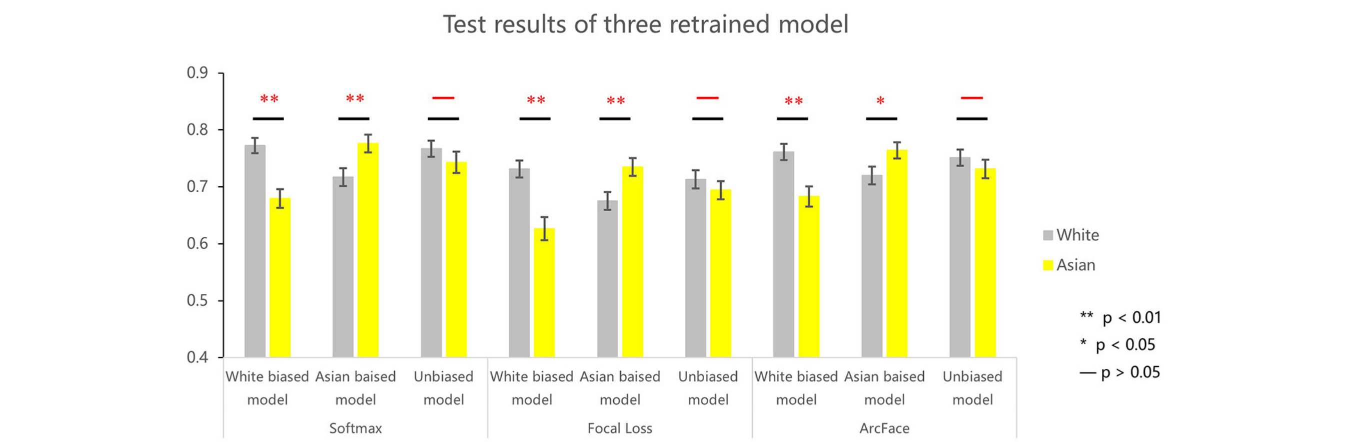


FIGURE 2 | Identification performance of three retrained VGG networks (i.e., white biased model, Asian biased model, and unbiased model) using softmax, focal loss, and Arcface. $^{**}p < 0.01$; $^{*}p < 0.05$; –, not significant.

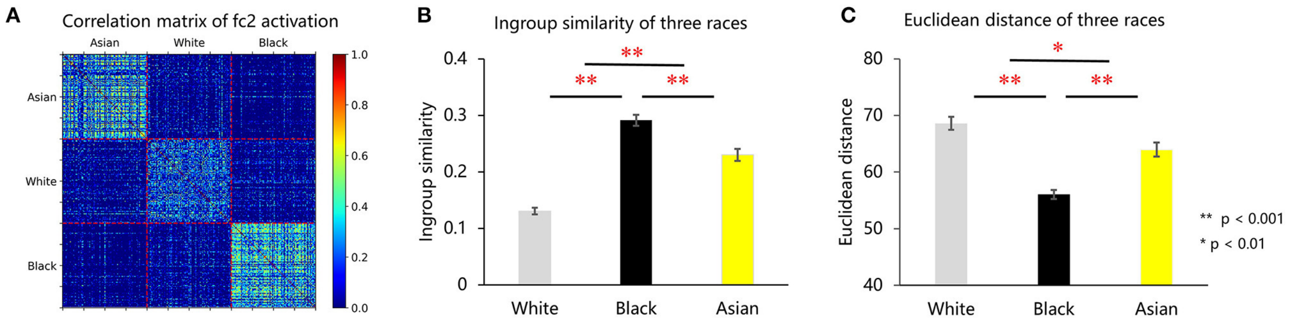


FIGURE 3 | (A) VGG-Face FC2 activation correlation matrix of Asian, white, and black faces. **(B)** Face distinctiveness of white, black, and Asian faces measured using in-group similarity. **(C)** Face distinctiveness of white, black, and Asian faces measured using face Euclidean distance.

= 0.4) (**Figure 3B**). That is, the representations for white faces were the sparsest in the face space. To quantify the sparseness of the representation, we calculated the Euclidean distance of the representation of individual faces to the averaged representation of all faces. As shown in **Figure 3C**, the representation of white faces was localized farther from the averaged representation than that of Asian ($p = 0.008$, $d' = 0.386$) and black ($p < 0.001$, $d' = 1.286$) faces, and that of Asian faces was farther than that of black faces ($p < 0.001$, $d' = 0.773$). The activation of faces in the last fully connected layer (FC3) was also extracted and analyzed, which showed a similar representational pattern as FC2 (detailed information is provided in the **Supplementary Materials**).

To visualize how race faces were represented in the face space, we used t-SNE to reduce multiple dimensions to two dimensions. As shown in **Figure 4A**, representations for each race were grouped into one cluster; however, the clusters for Asian and black faces were denser, whereas white faces were distributed more sparsely in the face space.

Finally, we explored whether the difference in sparseness of the representation was related to the ORE observed in VGG-Face. As shown in **Figure 4B**, the correlation analysis showed a significant negative correlation between in-group similarity and face identification accuracy (coefficient Pearson's correlation $R = -0.458$, $p < 0.001$, Spearman correlation $R = -0.499$, $p < 0.001$). As shown in **Figure 4C**, the correlation analysis showed a significant positive correlation between Euclidean distance and face identification accuracy (coefficient Pearson's correlation $R = 0.579$, $p < 0.001$, Spearman correlation $R = 0.621$, $p < 0.001$). That is, if a face was represented further from the average representation, it was more accurately identified by the VGG-Face. For the VGG-Face trained by a dataset dominated by white faces, white faces on average had the largest representational distance, and they were the most likely to be identified correctly, which therefore resulted in the ORE.

DISCUSSION

In this study, we examined the ORE in VGG-Face. By manipulating the ratio of faces of different races in the training dataset, the results demonstrated that unbalanced datasets led to the appearance of the ORE in VGG-Face, in line with studies on humans, which have reported that visual experiences affect the identification accuracy of a particular race's face (Chiroro and Valentine, 1995; Meissner and Brigham, 2001). Importantly, the representation similarity analysis revealed that if white faces dominated the dataset, they were distributed more sparsely in the multidimensional representational space of faces in VGG-Face, resulting in better behavioral performance. On the other hand, a similar phenomenon, called "long tailed problem," suggested that the model performs better on the head domains (i.e., high-frequency domain) than on the tail domains (i.e., low-frequency domain). The inter-class distance was usually used to distinguish the head domain from the tail domain. The head domain usually showed a larger inter-class indicator than that of the tail domain (Cao et al., 2020), which seems to be opposite to our result. In our study, we used intra-class distance (in-group similarity and in-group Euclidean distance), which was widely used to quantify the sparseness of the representation. We found the faces of the majority race were scattered more sparsely in the representational face space. This result is consistent with previous results in humans (Valentine, 1991; Valentine et al., 2016), which implied a similar mechanism. In sum, with the MDS theory in human, we provided a novel approach to understand race biases in DCNNs.

The AI ethical problem has attracted broad attention to the field of AI (Zemel et al., 2013; Zou and Schiebinger, 2018). However, the mechanism underlying AI biases is poorly understood. Our study confirmed that the ORE bias might be derived from an unbalanced training dataset. This is consistent with the contact theory (Chiroro and Valentine, 1995) in humans, according to which high-contact faces are recognized more accurately than low-contact ones. Previous studies in humans suggest that high in-group interaction leads to sparser representation (high distinctiveness) of in-group faces in face space, whereas low interaction leads to denser representation (low distinctiveness) of out-group faces (Valentine, 1991; Valentine et al., 2016). In the current study, we also found that in the representational space of VGG-Face, "own-race" faces (i.e., white faces) showed larger distinctiveness than that of "other-race" faces (i.e., Asian and black faces). Furthermore, the distinctiveness was indexed by the representational similarity of faces, which may serve as a more sensitive index than the ratio of faces in the unbalanced dataset. Therefore, before formal training, an examination of representational similarity in MDS

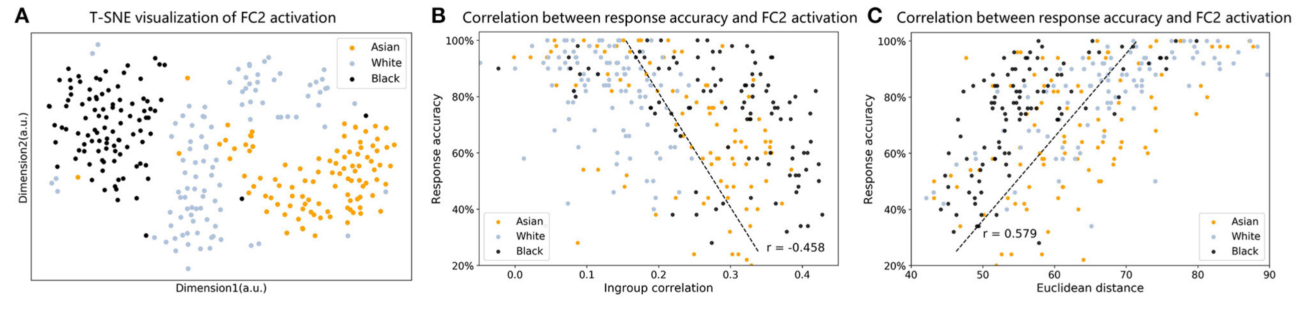


FIGURE 4 | (A) T-SNE visualization of FC2 activation of Asian, white, and black faces. **(B)** Correlation between in-group similarity and face identification accuracy. **(C)** Correlation between face Euclidean distance to averaged face activation and face identification accuracy.

with a portion of the training dataset may provide an estimate of the skewness of the datasets and the biased performance under current task demands.

Therefore, a more effective way of controlling AI biases may come from new algorithms that can modulate the internal representations of DCNNs. Currently, most efforts have been focused on the construction of balanced datasets and the approaches of training DCNNs, and guidelines have been advised (Gebru et al., 2018; Mitchell et al., 2019). However, it is laborious to balance datasets not only in terms of data collection, but also in terms of task demands. It might be more efficient if a revised back-propagation algorithm could minimize errors between outputs and goals and rectify differences in distinctiveness of the representation of interests. For example, in the field of natural language processing, Beutel et al. (2017) and Zhang et al. (2018) proposed a multi-task adversarial learning method to manipulate the biased representational subspace and thus mitigate the gender bias of model performance. They built a multi-head DCNN where one head was for target classification and another was for removing information about unfair attributes learned from the data. Similarly, in the field of computer vision, further studies could also explore ways to manipulate the face representational space to reduce social bias in DCNNs.

In conclusion, our study used a well-known phenomenon, the ORE, to investigate the mechanism inside DCNNs that leads to biased performance. In addition, we found a human-like multidimensional face representation in DCNN, suggesting that paradigms and theories discovered in human studies may also be helpful in identifying the underlying mechanisms of DCNNs. There are many other types of biases in AI, such as gender bias and age bias; therefore, our study invites broad investigation on these ethical problems in AI.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are publicly available. All face image materials and model training codes used in this article are provided on git-hub: https://github.com/JinhuaTian/DCNN_other_race_effect.

AUTHOR CONTRIBUTIONS

JL and SH designed the research. JT and HX collected and analyzed the data. JT wrote the manuscript with input from JL and SH. All authors reviewed and commented on this manuscript.

FUNDING

This study was funded by the National Natural Science Foundation of China (31861143039) and the National Basic Research Program of China (2018YFC0810602).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fncom.2021.620281/full#supplementary-material

REFERENCES

Beutel, A., Chen, J., Zhao, Z., and Chi, E. H. (2017). Data decisions and theoretical implications when adversarially learning fair representations. *arXiv [Preprint]. arXiv:*1707.00075.

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., and Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *arXiv [Preprint]. arXiv:*1607.06520.

Caliskan, A., Bryson, J. J., and Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 183–186. doi: 10.1126/science.aal4230

Cao, D., Zhu, X., Huang, X., Guo, J., and Lei, Z. (2020). "Domain balancing: face recognition on long-tailed domains," in: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Seattle, WA: IEEE). doi: 10.1109/CVPR42600.2020.00571

Cao, Q., Shen, L., Xie, W., Parkhi, O. M., and Zisserman, A. (2018). "VGGFace2: a dataset for recognising faces across pose and age," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)* (Xi'an: IEEE). doi: 10.1109/FG.2018.00020

Chen, X., Zhou, M., Gong, Z., Xu, W., Liu, X., Huang, T., et al. (2020). DNNBrain: a unifying toolbox for mapping deep neural networks and brains. *Front. Comput. Neurosci.* 14:580632. doi: 10.3389/fncom.2020.580632

Chiroro, P., and Valentine, T. (1995). An investigation of the contact hypothesis of the own-race bias in face recognition. *Quart. J. Experi. Psychol. Section A* 48, 879–894. doi: 10.1080/14640749508401421

Deng, J., Guo, J., Xue, N., and Zafeiriou, S. (2019). "ArcFace: additive angular margin loss for deep face recognition," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Long Beach, CA: IEEE). doi: 10.1109/CVPR.2019.00482

Garg, N., Schiebinger, L., Jurafsky, D., and Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proc. Natl. Acad. Sci. U.S.A.* 115, E3635–E3644. doi: 10.1073/pnas.1720347115

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumeé Iii, H., et al. (2018). Datasheets for datasets. *arXiv [Preprint]. arXiv:*1803.09010.

Krizhevsky, A. (2014). One weird trick for parallelizing convolutional neural networks. *arXiv [Preprint]. arXiv:*1404.5997.

Lin, T., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017). "Focal loss for dense object detection," in *2017 IEEE International Conference on Computer Vision (ICCV)* (Venice: IEEE). doi: 10.1109/TPAMI.2018.2858826

Malpass, R. S., and Kravitz, J. (1969). Recognition for faces of own and other race. *J. Pers. Soc. Psychol.* 13, 330–334. doi: 10.1037/h0028434

Meissner, C. A., and Brigham, J. C. (2001). Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychol. Pub. Policy Law* 7, 3–35. doi: 10.1037/1076-8971.7.1.3

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., et al. (2019). "Model cards for model reporting," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*: *Association for Computing Machinery* (New York, NY). doi: 10.1145/3287560.3287596

O'toole, A. J., Castillo, C. D., Parde, C. J., Hill, M. Q., and Chellappa, R. (2018). Face space representations in deep convolutional neural networks. *Trends Cogn. Sci.* 22, 794–809. doi: 10.1016/j.tics.2018.06.006

Parkhi, O., Vedaldi, A., and Zisserman, A. (2015). "Deep face recognition," in *Proceedings of the British Machine Vision Conference 2015* (Swansea: British Machine Vision Association), 1–12. doi: 10.5244/C.29.41

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019). Pytorch: an imperative style, high-performance deep learning library. *arXiv [Preprint]. arXiv:*1912.01703.

Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion, and race in face recognition. *Q. J. Exp. Psychol. A* 43, 161–204. doi: 10.1080/14640749108400966

Valentine, T., and Endo, M. (1992). Towards an exemplar model of face processing: the effects of race and distinctiveness. *Q. J. Exp. Psychol. A* 44, 671–703. doi: 10.1080/14640749208401305

Valentine, T., Lewis, M. B., and Hills, P. J. (2016). Face-space: a unifying concept in face recognition research. *Q. J. Exp. Psychol.* 69, 1996–2019. doi: 10.1080/17470218.2014.990392

Van Der Maaten, L., and Hinton, G. (2008). Visualizing data using t-SNE. *J. Machine Learn. Res.* 9, 2579–2605. doi: 10.1007/s10846-008-9235-4

Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. (2014). How transferable are features in deep neural networks? *arXiv [Preprint]. arXiv*:1411.1792.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. (2013). "Learning fair representations," in *International Conference on Machine Learning*: *PMLR* (Scottsdale, AZ). doi: 10.5555/3042817.3042973

Zhang, B. H., Lemoine, B., and Mitchell, M. (2018). "Mitigating unwanted biases with adversarial learning," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*: *Association for Computing Machinery* (New Orleans, LA). doi: 10.1145/3278721.3278779

Zou, J., and Schiebinger, L. (2018). AI can be sexist and racist - it's time to make it fair. *Nature* 559, 324–326. doi: 10.1038/d41586-018-05707-8

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Computational Neuroscience

ORIGINAL RESEARCH
published: 31 March 2021
doi: 10.3389/fncom.2021.641335

Optimal Organization of Functional Connectivity Networks for Segregation and Integration With Large-Scale Critical Dynamics in Human Brains

Xinchun Zhou[1†], Ningning Ma[2†], Benseng Song[1], Zhixi Wu[1], Guangyao Liu[3], Liwei Liu[4], Lianchun Yu[1] and Jianfeng Feng[2,5*]*

[1] Key Laboratory of Theoretical Physics of Gansu Province, Lanzhou Center for Theoretical Physics, Lanzhou University, Lanzhou, China, [2] School of Mathematical Sciences and Centre for Computational Systems Biology, Fudan University, Shanghai, China, [3] Department of Magnetic Resonance, Lanzhou University Second Hospital, Lanzhou, China, [4] College of Electrical Engineering, Northwest University for Nationalities, Lanzhou, China, [5] School of Mathematical Sciences, School of Life Science and the Collaborative Innovation Center for Brain Science, Fudan University, Shanghai, China

OPEN ACCESS

Edited by:
Si Wu,
Peking University, China

Reviewed by:
Jingxin Nie,
South China Normal University, China
Daqing Guo,
University of Electronic Science and Technology of China, China

***Correspondence:**
Lianchun Yu
yulch@lzu.edu.cn
Jianfeng Feng
jianfeng64@gmail.com

[†] These authors have contributed equally to this work

Received: *14 December 2020*
Accepted: *23 February 2021*
Published: *31 March 2021*

Citation:
Zhou X, Ma N, Song B, Wu Z, Liu G, Liu L, Yu L and Feng J (2021) Optimal Organization of Functional Connectivity Networks for Segregation and Integration With Large-Scale Critical Dynamics in Human Brains. Front. Comput. Neurosci. 15:641335. doi: 10.3389/fncom.2021.641335

The optimal organization for functional segregation and integration in brain is made evident by the "small-world" feature of functional connectivity (FC) networks and is further supported by the loss of this feature that has been described in many types of brain disease. However, it remains unknown how such optimally organized FC networks arise from the brain's structural constrains. On the other hand, an emerging literature suggests that brain function may be supported by critical neural dynamics, which is believed to facilitate information processing in brain. Though previous investigations have shown that the critical dynamics plays an important role in understanding the relation between whole brain structural connectivity and functional connectivity, it is not clear if the critical dynamics could be responsible for the optimal FC network configuration in human brains. Here, we show that the long-range temporal correlations (LRTCs) in the resting state fMRI blood-oxygen-level-dependent (BOLD) signals are significantly correlated with the topological matrices of the FC brain network. Using structure-dynamics-function modeling approach that incorporates diffusion tensor imaging (DTI) data and simple cellular automata dynamics, we showed that the critical dynamics could optimize the whole brain FC network organization by, e.g., maximizing the clustering coefficient while minimizing the characteristic path length. We also demonstrated with a more detailed excitation-inhibition neuronal network model that loss of local excitation-inhibition (E/I) balance causes failure of critical dynamics, therefore disrupting the optimal FC network organization. The results highlighted the crucial role of the critical dynamics in forming an optimal organization of FC networks in the brain and have potential application to the understanding and modeling of abnormal FC configurations in neuropsychiatric disorders.

Keywords: fMRI, functional connection networks, criticality, DTI, Greenberg-Hasting model, E/I ratio

INTRODUCTION

Functional connectivity (FC) analysis of resting state human brain allows to understand how the functional networks are organized, how this organization is related to behavior, and how it changes in case of pathology (van den Heuvel and Hulshoff Pol, 2010; Lee et al., 2013; Yu et al., 2016). Recent studies have identified the so-called resting-state networks which consist of anatomically separated, but functionally linked brain regions that show a high level of ongoing FC during rest (Heine et al., 2012; Raichle, 2015). The graph theoretical analysis of resting-state functional magnetic resonance imaging (fMRI) has revealed the "small-world" feature of the whole brain functional connectivity network (Rubinov and Sporns, 2010). Compared with random networks, small-world networks exhibit shorter characteristic path length but higher clustering coefficients (Watts and Strogatz, 1998). This specific organization of functional network is believed to benefit the higher-level cognitive functions requiring the integration of information from different brain regions (Watts and Strogatz, 1998), maximize efficiency at a minimal cost for effective information processing between multiple brain regions (Achard and Bullmore, 2007), and promote low wiring costs (Bassett and Bullmore, 2006). The small-world organization of brain functional network is likely to be related to human intellectual performance (van den Heuvel et al., 2009) and disrupted with normal aging (Wang et al., 2010). Extensive studies also showed that this small-world properties of functional network are altered by diseases such as schizophrenia (Liu et al., 2008), AD (Sanz-Arigita et al., 2010), autism (Rudie et al., 2013), etc. Specifically, the alterations are normally characterized by increased characteristic path length, as well as decreased clustering coefficient and efficiency [for an example, pleases see Ref. (Liu et al., 2008) for details], implying the disrupted organization of FC networks for integration and segregation. However, little is known about the underlying dynamics based on which this optimal FC network is established, and how its disruption induced by disease is associated with changes in brain dynamics.

The theory from statistical physics has predicted that resting state brain dynamics operates close to a critical point, hallmarked by power-law distributions of spatiotemporal cascades of activity-termed neuronal avalanche. Scale-free avalanches have been observed in different scales of neural systems with different methods (Beggs and Plenz, 2003; Gireesh and Plenz, 2008; Ribeiro et al., 2010), including local field potentials (Thiagarajan et al., 2010; Plenz, 2012), human electroencephalography (EEG) (Meisel et al., 2013), magnetoencephalogram (MEG) (Palva et al., 2013; Shriki et al., 2013), and fMRI (He, 2011; Tagliazucchi et al., 2012). It is suggested that there are many computational advantages for the neural systems being poised around this critical point. It maximize the number of meta-stable states (Haldeman and Beggs, 2005), the dynamic range to the input stimuli (Shew et al., 2009; Gautam et al., 2015), as well as the information capacity and transmission (Shew et al., 2011) of the cortical neural networks. Furthermore, cortical EI balance are found to be crucial for the forming of critical behavior at multiple levels of neuronal organization (Poil et al., 2012), perhaps achieved through self-organization with synaptic plasticity (Stepp et al., 2015). On the other hand, a leading theory, proposed over a decade ago as a model for autism (Rubenstein and Merzenich, 2003), holds that brain disorders arise from imbalanced EI in brain circuitry. This concept has since been applied to many other brain disorders, such as schizophrenia, tuberous sclerosis, and Angelman syndrome (O'Donnell et al., 2017). These studies have led to the conjecture that criticality may be a signature of healthy neural systems, and conversely excursion from such an optimal point may be responsible for a diversity of brain disorder (Massobrio et al., 2015; Cocchi et al., 2017).

Recent modeling studies have also revealed crucial role of critical dynamics in understanding the relation between large scale brain architecture and function. For example, the spatial organization of resting state networks observed in the resting state fMRI data, such as default mode network, emerge at the critical point in the dynamic network derived from human brain neuroanatomical connections (Haimovici et al., 2013). The structure-function coupling is maximal when the global network dynamics operate at a critical point (Deco et al., 2014a), and the decoupling of functional connectivity from anatomical constraints is found in the brains losing consciousness, accompanied with fading signature of criticality (Tagliazucchi et al., 2016). In addition, the local excitation/inhibition ratio (E/I ratio) significantly improves the model's prediction of the empirical human functional connectivity at the large-scale brain level (Deco et al., 2014b). The loss of small-world organization of FC networks and failure of critical dynamics in diseased brain implies the potential relationship between them. However, it is still not clear how the organization of the FC network depends on the large-scale critical dynamics in brains.

In this work, we answered this question by investigating: (i) the correlation between topological metrics of FC network and the long-range temporal correlations (LRTCs) of BOLD signals in fMRI data of healthy subjects; (ii) the dependence of these metrics on the control parameter (excitation threshold) in a toy model which combines the structural diffusion tensor imaging (DTI) and Greenberg-Hasting (GH) dynamics around the critical point; (iii) the impact of local E/I ratio on the critical dynamics and thus the functional network metrics in a biological plausible whole brain model. We showed that with the critical dynamics, the brain FC network exhibited optimized organization, characterized by maximized efficiency and clustering coefficient, but shortest characteristic path length. We also showed that local E/I ratio have a great influence on this large-scale critical dynamics and organization of FC networks. We discussed the potential application of our findings to the understanding and modeling of abnormal FC configurations in brain disorders.

RESULTS

Correlation of Network Metrics With LRTCs in Resting-State fMRI Data

We first assessed LRTCs in BOLD signals from the resting-state fMRI data of 95 healthy subjects by computing the Hurst exponent in the temporal domain using classical rescaled range

(RS) method (Blythe and Nikulin, 2017) ("**METHOD AND MATERIALS**," "**Hurst Exponent H**"). A Hurst exponent in the range $0.5 < H < 1$ indicates the presence of LRTCs, i.e., a high value in the series will probably be followed by another high value. An exponent of $0 < H < 0.5$ is obtained when the time series is anticorrelated (switching between high and low values in consecutive time steps). The uncorrelated temporal activity with exponential decay of the autocorrelation function yields an exponent of $H = 0.5$. After preprocessing with the standard preprocessing procedure ("**METHOD AND MATERIALS**," "**fMRI Data Acquisition and Preprocessing**"), the automated anatomical labeling atlas (AAL) (Tzourio-Mazoyer et al., 2002) was used to parcellate the brain into 90 brain regions, and the mean BOLD signals were extracted in each brain region by averaging the signals of all voxels within the region. For each subject, the Hurst exponents were calculated for each mean BOLD time series, and the mean Hurst exponent (H) from 90 brain regions was taken as a measure of LRTCs at the whole brain level for this subject. The Hurst exponent reflects the temporal correlations of a signal. The group averaged Hurst exponent in our study is 0.8628 ± 0.0188, indicating the long-range memory of the BOLD signals in human brain (He, 2011). However, the variance in the Hurst exponent among subjects should not be ascribed to noise sources, such as physiological noise. On the contrary, considering criticality as a theory of long-range fluctuation in the human brain, it reflects the different intrinsic brain dynamics among subjects that can be described by a departure from the criticality (Blythe and Nikulin, 2017), as we demonstrated below.

For each subject, we applied a binarizing threshold T_d to the absolute values of the correlation coefficients among mean BOLD signals from 90 brain regions to construct the FC network. Then six typical topological metrics, namely global and local efficiency (E_{global} and E_{local}), characteristic path length (L), clustering coefficient (C_{global}), mean connection strength (E_{corr}), and sparsity (S) of the FC networks for each subject were calculated ("**METHOD AND MATERIALS**," "**Network Metrics**"). We found there existed significant correlations between these metrics and the Hurst exponents (**Figure 1**). The longer temporal memory in BOLD signals yields higher global efficiency (**Figure 1A**), local efficiency (**Figure 1B**), clustering coefficient (**Figure 1D**), mean connection strength (**Figure 1E**), and sparsity (**Figure 1F**), but shortest characteristic path length (**Figure 1C**).

We then investigated the dependence of topological metrics and their correlations with Hurst exponents on the binarized threshold T_d. We first determined the small-world regime of the FC networks for T_d (Liu et al., 2008). The upper criteria for T_d are so set to make sure there is no isolated node in the network (red vertical lines in **Figure 2**). To determine the lower criteria for T_d, we compared the global efficiency of brain FC networks with that estimated in a random graph with the same degree distribution over a range of network sparsity (**Supplementary Figure 1A**). Then the lower criteria are set as the smallest value of the threshold T_d (blue vertical line in **Figure 2**) with which the global efficiency curve for the brain networks is below the global efficiency curve for the random networks. In this range of threshold T_d, the Hurst exponents and the topological metrics are significantly correlated (**Figure 2**, the threshold values with correlation coefficients $R > 0.26$ are marked with open circles and $R < -0.26$ with filled circles. The corresponding p-values are indicated with triangles if $p < 0.01$). It was also noticed in **Figure 2** that as the threshold T_d increases, the global efficiency (**Figure 2A**), local efficiency (**Figure 2B**), clustering coefficient (**Figure 2D**), and sparsity (**Figure 2F**) decrease, whereas the characteristic path length (**Figure 2C**) and the mean connection strength (**Figure 2E**) increase. The binarizing threshold dependent changes of these topological metrics are in line with previous study, e.g., Ref. (Liu et al., 2008).

The Critical Dynamics in the DTI+GH Brain Network Model

We built a toy brain network model which combines the DTI structural connection data among the 90 brain regions and GH excitable cellular automatons to simulate the BOLD signals from 90 brain regions (**Figure 3**, see details in the "**METHOD AND MATERIALS**," "**DTI+GH Brain Network Model**"). In this model, the DTI connection data provides the number of fibers connecting every two brain regions, which is taken as the connection weights among the regions. The regional dynamics is given by simple rules that describe the excitation of the active media. Previous work has demonstrated that such simple dynamical brain model is sufficient to replicate fundamental features of spontaneous brain activity observed in fMRI data. For example, the resting state networks, such as default mode network, emerge in such kind of whole brain models with the critical dynamics (Haimovici et al., 2013).

The criticality refers to a balanced state between ordered and disordered and is characterized by power law distribution of avalanche activity (i.e., the avalanche size distribution shows no characteristic scale). The supercritical state refers to the ordered states that are characterized by avalanche with large size, whereas the subcritical state refers to the disordered states that are characterized by avalanches with small size (Beggs and Plenz, 2003; Tagliazucchi et al., 2012; Shriki et al., 2013). In our model, we calculated the avalanche size distribution from the spatiotemporal patterns of excited nodes for different excitation threshold T_{m} ("**METHOD AND MATERIALS**," "**Avalanche Detection**"). When T_{m} is low, the nodes in the model are excited easily, and their activities are highly synchronized to result in a rather ordered state (**Figure 4D**). Thus, the activities tend to form avalanches with large size to have a power-law slope with a heavier tail in the distribution (**Figure 4A**), indicating the supercritical dynamics. Whereas, when T_{m} is high, the nodes in the network are less excitable, and their activities are random and less synchronized (**Figure 4F**). Thus, the groups of activity are small and die out quickly, unlikely to form avalanches with large size, which is termed as subcritical regime (**Figure 4C**). In both cases, the size distribution of avalanches demonstrates a characteristic scale. However, with moderate T_{m} ($T_m \approx 0.52$) the scale-free avalanche distribution emerges with an exponent of

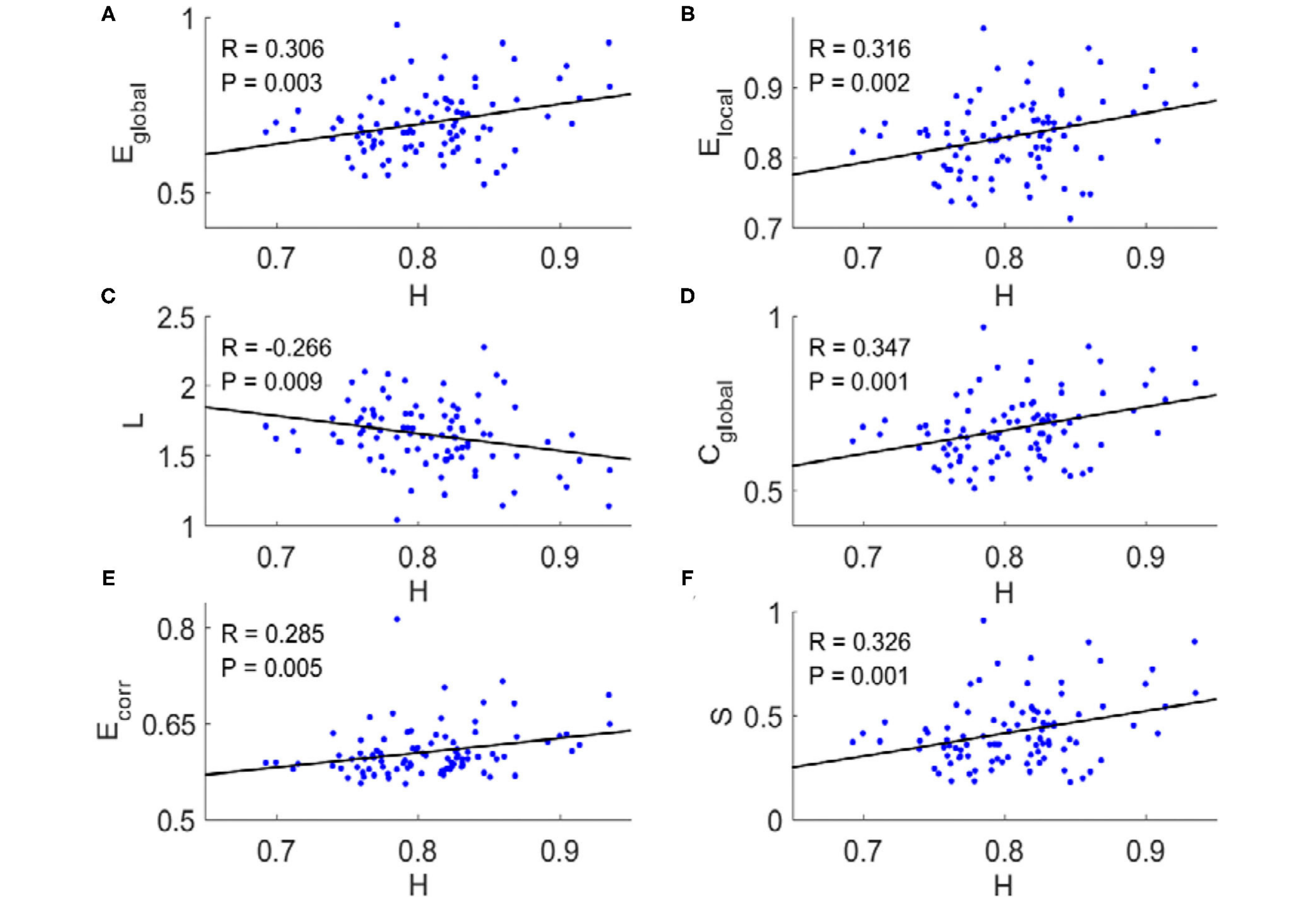


FIGURE 1 | Scatter plots with trend line showing the dependence of topological metrics of FC network on Hurst exponents. **(A)** Global efficiency. **(B)** Local efficiency. **(C)** Characteristic path length. **(D)** Clustering coefficient. **(E)** Mean connection strength. **(F)** Sparsity. The topological metrics of FC networks were calculated with threshold $T_d = 0.4$. Pearson correlation coefficient (R) for all six topological metrics were significant ($p < 0.01$).

−1.5 (**Figure 4B**), and the system is perched between order and random (**Figure 4E**).

We then convolved the activities of each node in the model with the hemodynamic response function to generate 90 simulated BOLD time series ("**METHOD AND MATERIALS**," "**DTI+GH Brain Network Model**"). Typical BOLD signals of arbitrarily chosen brain regions for supercritical, critical, and subcritical regimes are demonstrated in **Figures 4G–I**, respectively. The Hurst exponent calculated from these simulated BOLD signals yields its largest value at the critical point (**Supplementary Figure 2A**). The FC matrices corresponding to different regimes were then obtained by calculating the correlation coefficients among 90 simulated BOLD time series as before. We compared the similarity between simulated FC matrices and experimental FC networks and found the maximal similarity occurs around the critical point (**Supplementary Figure 3A**). It is also seen that when the system is poised at the critical point, the FC matrix exhibits patterns which is similar to the DTI structural connections (**Figure 4K**), whereas the supercritical and subcritical dynamics fail to replicate the DTI structural connections (**Figures 4J,L**). This phenomenon has been systematically studied with computer modeling and experiment on propofol-induced departure from critical dynamics (Tagliazucchi et al., 2016). It was argued that the functional organization of the brain is constrained and enabled by the unique structural organization (Tagliazucchi et al., 2016), and the spontaneous brain activity can be understood as an ever-transient exploration of the repertoire of paths offered by structural connections (Deco et al., 2014a; Tagliazucchi et al., 2016). The critical dynamics of the system would allow a more widespread exploration of all possibilities offered by the structural connections, makes FCs better reproduce its structural connections [see Ref. (Tagliazucchi et al., 2016) for a vivid explanation].

Optimal Organization of the FC Network at Criticality

We then investigated the changes of FC network metrics across the transition from supercritical to subcritical regime in the DTI+GH model. The simulated FC matrices were binarized with threshold T_d and the corresponding metrics were calculated in the small-world regime as before (**Supplementary Figure 1B**).

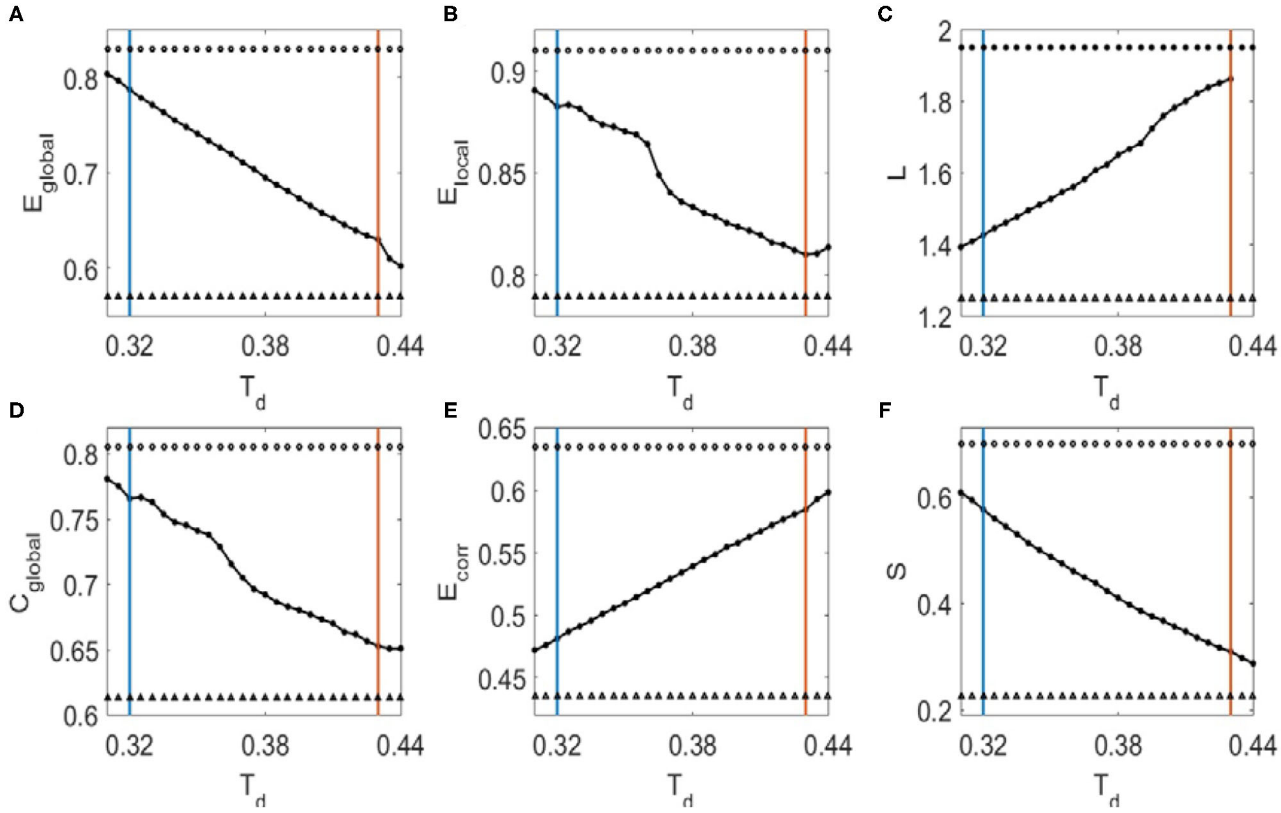


FIGURE 2 | The dependence of topological metrics of the FC networks and their correlations with Hurst exponents on the threshold T_d. **(A)** Global efficiency. **(B)** Local efficiency. **(C)** Characteristic path length. **(D)** Clustering coefficient. **(E)** Mean connection strength. **(F)** Sparsity. The red vertical lines mark the upper criteria above which there is no isolated node in the network, and the blue vertical lines mark the lower criteria below which the global efficiency curve for the brain networks is less than the global efficiency curve for the random networks. Open circles indicate that the correlation coefficients between the Hurst exponent and the corresponding topological metrics are larger than 0.26, where filled circles mark the correlation coefficients that is smaller than −0.26. Triangles mark the corresponding p-values of the correlation analysis that is significant ($p < 0.01$).

It is seen from **Figure 5** that around the critical point ($T_m = 0.52$), all the network metrics are maximized except for the characteristic path length which is minimized (**Figure 5C**). These results imply that critical dynamics can optimize brain FC network organization and the departure from criticality will cause the disruption of this optimal balance between integration and segregation.

It was also seen from **Figure 5** that with the increase of binarizing threshold T_d, global efficiency (**Figure 5A**), local efficiency (**Figure 5B**), clustering coefficient (**Figure 5D**), and sparsity (**Figure 5F**) decrease, whereas the characteristic path length (**Figure 5C**) and the mean connection strength (**Figure 5E**) increase. The dependence of these network metrics on the binarizing threshold T_d predicted by our model is in line with that obtained from the fMRI data (**Figure 2**), and that reported by other researchers (Liu et al., 2008).

Local *E/I* Ratio Tunes Critical Dynamics in the DTI+EI Network Model

Next, we built a large-scale brain functional model based on DTI structural connection data and EI neuronal networks (**Figure 6**). In this model, the neural activity in each region is modeled with a neuronal network composing 100 excitatory (*E*) and 25 inhibitory (*I*) neurons so that the ratio of number of excitatory neurons to that of inhibitory ones is 4:1. The single neuron dynamics is modeled with Izhikevich cortical spiking neuron model, which is computationally efficient and biologically plausible (Izhikevich, 2004). The neurons in each EI networks are connected with a probability of 0.5. The excitatory neurons send out only excitatory synaptic connections to other neurons and inhibitory neurons send out only inhibitory ones. In the simulation, we systematically change the *E–E* connection strength but fixed other ones and define the ratio of *E–E* to *I–I* synaptic connection strength as the local *E*/*I* ratio. The number of excitatory neurons that establish inter-regional connections is proportional to the number of fibers connecting corresponding brain regions (see "**METHOD AND MATERIALS**," "DTI+EI Whole Brain Model" for details).

Through adjusting the local *E*/*I* ratio in each region simultaneously, we observed the power law distribution of avalanche activities with exponent of −1.5 within each brain region when the *E*/*I* ratio is around 2.025 (**Figure 7B**), indicating the critical dynamics of the system. Whereas, the system is supercritical when the *E*/*I* ratio is high (**Figure 7A**) but

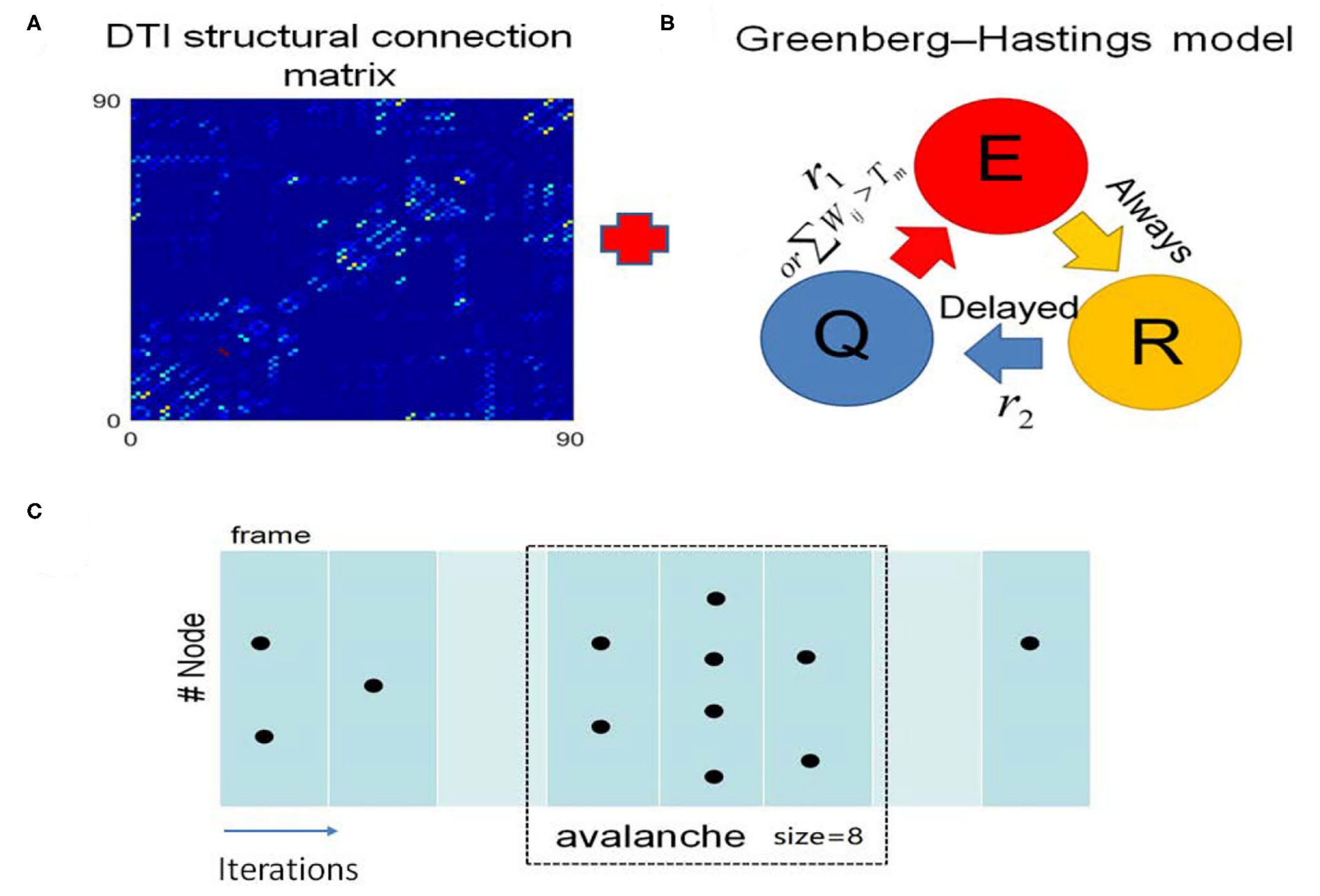


FIGURE 3 | The DTI+GH whole brain model. **(A)** The DTI structural connection matrix. **(B)** The Greenberg-Hastings (GH) cellular automaton model for dynamics of each brain region. The GH model has three states: quiescent (*Q*), excitation (*E*), and refractory (*R*). The colored arrows indicated the transition between these states. The transition from *Q* to *E* can happen with a small probability r_1, or if the sum of the connection weights w_{ij} with the active neighbors *j* is higher than a threshold T_m. Once the system is excited, it always goes to *R*. Then it transits from *R* to *Q* with a probability r_2 after several steps of delaying. **(C)** Demonstration of the method to extract avalanches from simulation of the whole brain model. The spatial activity in several simulation step is assigned as a frame (consecutive frames are divided by white lines). An avalanche is defined as the consecutive frames that are preceded by a blank frame (in which no activation occurs, marked with light cyan) and ended by a blank frame. The avalanche size is the total number of excited nodes in this avalanche. Black dots represent the excited nodes that are in the state *E*.

subcritical when the *E*/*I* ratio is low (**Figure 7C**). The spikes of the neurons are quite synchronized when the system is supercritical, especially for the excitatory neurons because of the strong excitatory connections among them (**Figure 7D**), but the firings are rather random when the system is subcritical with decreased excitatory connections (**Figure 7F**). The critical dynamics of the system exhibits moderate synchrony where synchronous firing occurs occasionally among excitatory neurons (**Figure 7E**). After, taking spike rate in each region as the input, we used the Balloon-Windkessel hemodynamic model to generate simulated BOLD signal for each region (see "**METHOD AND MATERIALS**" for details). **Figures 7G–I** demonstrates the examples of simulated BOLD signals from arbitrarily chosen brain regions for each case. Again, the Hurst exponent of these simulated BOLD signals exhibits its maximal values at the critical point (**Supplementary Figure 2B**). We then obtained the 90 × 90 FC matrices from the 90 simulated BOLD time series for each regime. We observed FC patterns emerge when the system is critical (**Figure 7K**). However, the pattern vanishes if the system is poised in the supercritical (**Figure 7J**) or subcritical regimes (**Figure 7L**). Again, the simulated FC matrices are most close to experimental FC network when the model is at its critical point (**Supplementary Figure 3B**).

Dependence of FC Network Metrics on Local *E*/*I* Ratio in the DTI+EI Model

We then calculated the dependence of simulated FC network metrics on the *E*/*I* ratio and the binarizing threshold T_d. The range of T_d for small-world regime was determined in the same way as before (**Supplementary Figure 1C**). It is seen from **Figure 8** that the *E*/*I* ratio, at where the critical dynamics emerges, maximizes the global efficiency (**Figure 8A**), local efficiency (**Figure 8B**), clustering coefficient (**Figure 8D**), mean

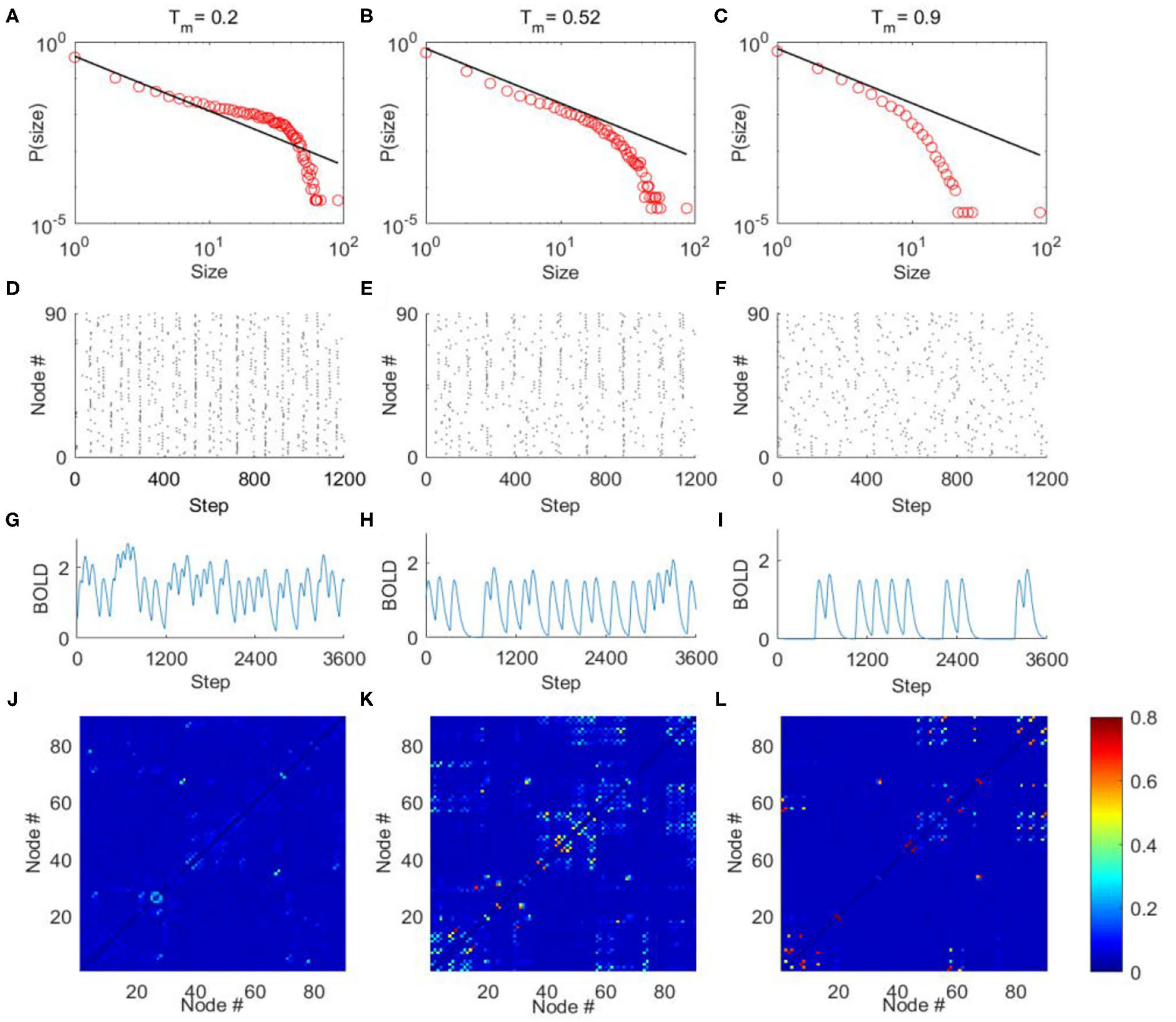


FIGURE 4 | The avalanche activity in the DTI+GH brain network model, the simulated BOLD signals, and FC matrices. Through tuning the excitation threshold T_m, the mode can exhibit typical avalanche distribution corresponding to supercritical **(A)**, critical **(B)**, and subcritical **(C)** dynamics. The horizontal axes are the size of avalanches, and vertical axes are the corresponding probability. The black lines in **(A–C)** indicate power law with exponent = −1.5. **(D–F)** The raster plots of spatial-temporal excitation distributions corresponding to **(A–C)**. The dots in the raster plots represent the excitation of the nodes (i.e., in state E). **(G–L)** The typical simulated BOLD signals of an arbitrarily chosen nodes and simulated FC matrices from the DTI+GH brain model in the supercritical **(G,J)**, critical **(H,K)**, and subcritical **(I,L)** regimes. Scale bar indicates the FC strength among the nodes in the model. The parameters of GH model in simulations are $r_1 = 0.005$, $r_2 = 0.98$, and $n_{delay} = 55$.

connection strength (**Figure 8E**), and sparsity (**Figure 8F**) but minimizes the characteristic path length (**Figure 8C**). These results further suggest that the local E/I ratio could adjust the global brain dynamics to the critical state, so as to achieve the balance between segregation and integration in FC networks. On the other hand, this optimal organization of FC networks could be damaged when the optimal E/I ratio is altered. As the binarizing threshold T_d increases, global efficiency, local efficiency, clustering coefficient, and sparsity decrease, but the characteristic path length and the mean connection strength increase. The dependence of these metrics on the binarizing threshold is in line with our above results from fMRI data analysis (**Figure 2**), the DTI+GH brain model (**Figure 5**) and that reported by other researchers (Liu et al., 2008).

DISCUSSION AND CONCLUSION

It has been shown that a network with shorter characteristic path length benefit the global efficiency, while a network with densely local connectivity benefit the local efficiency. Only

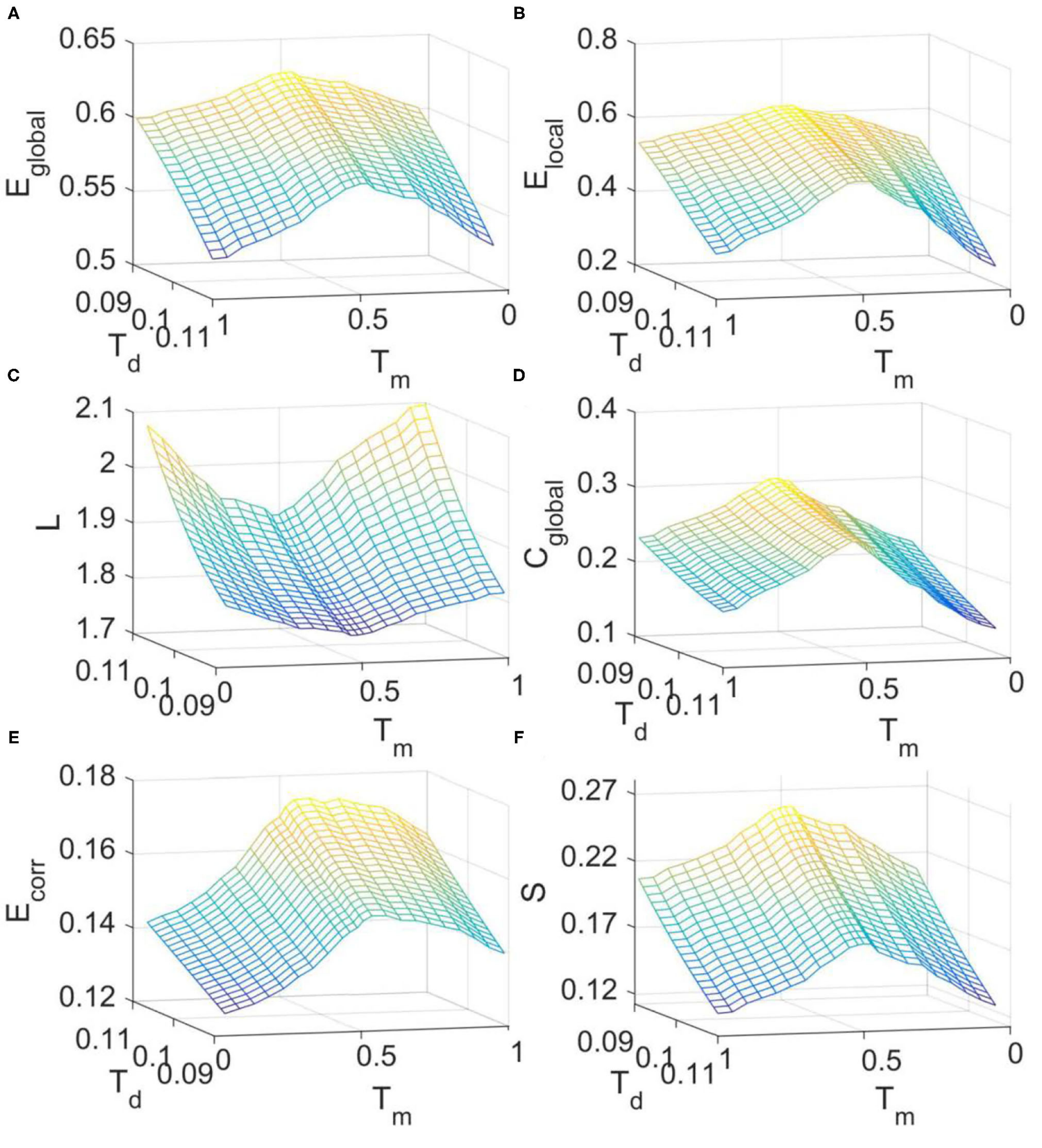


FIGURE 5 | The dependence of FC network topological metrics simulated with DTI+GH model on the excitation threshold T_m and binarizing threshold T_d. **(A)** Global efficiency. **(B)** Local efficiency. **(C)** Characteristic path length. **(D)** Clustering coefficient. **(E)** Mean connection strength. **(F)** Sparsity. The results were obtained by averaging results from 1,000 times of simulation. In each simulation, the obtained raw BOLD signals were sampled every 140 iteration steps to achieve the simulated BOLD time series of 200 time points.

in the small-world region, i.e., low characteristic path length combined with large clustering coefficient, does the network display globally and locally efficient at the same time (Latora and Marchiori, 2001). Recent analysis of human brain functional networks derived from EEG/MEG and fMRI experiments showed that these networks exhibit prominent small-world

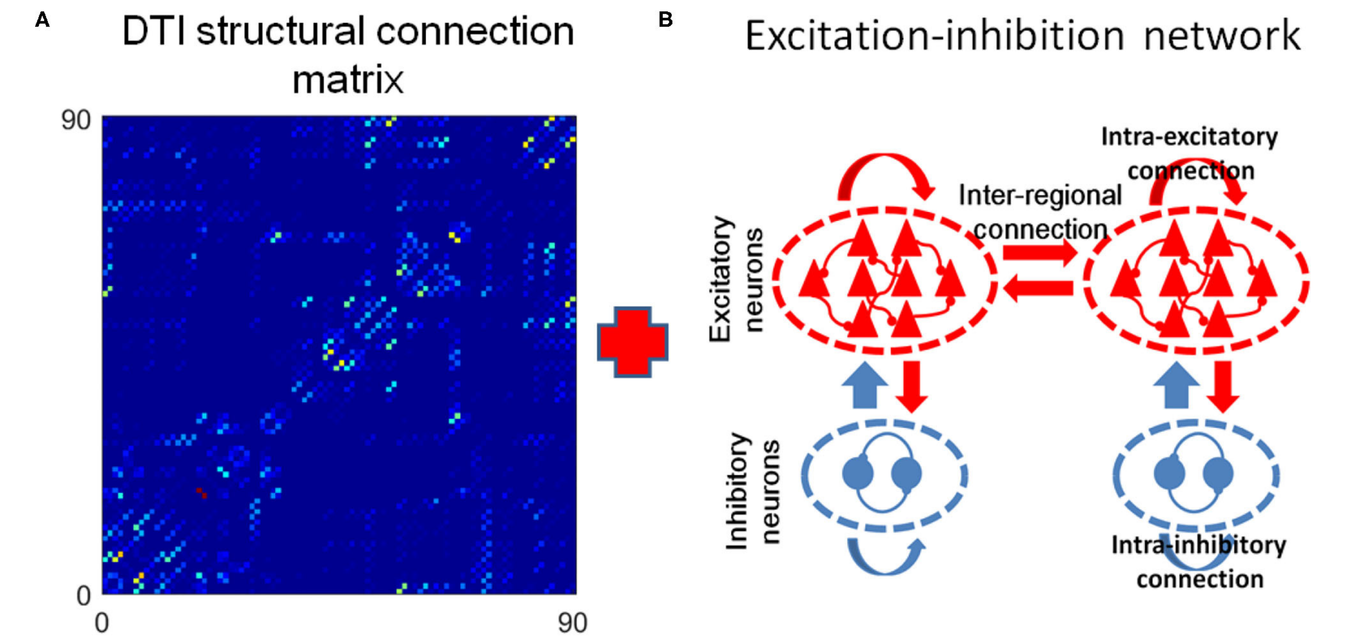

A
DTI structural connection matrix
90
0
0
90
B
Excitation-inhibition network
Intra-excitatory connection
Inter-regional connection
Excitatory neurons
Inhibitory neurons
Intra-inhibitory connection

FIGURE 6 | The DTI+EI whole brain model. **(A)** The DTI structural connection matrix. **(B)** Example of two excitation-inhibition (EI) neuronal networks that represent two brain regions. Each regional neuron network consists of one excitation neuron pool and one inhibitory neuron pool. They are 80% excitatory neurons and 20% inhibitory neurons in the network. The excitatory neurons send our only excitatory synapses to other neurons and the inhibitory neurons send out only inhibitory synapses. The two EI networks are coupled only by excitatory inter-regional connections.

organization. Through forming intrinsically densely connected and strongly coupled local network communities, the small-world topology facilitates functional segregation. Meanwhile, by enabling global communication between communities through network hubs, it also promotes functional integration (Bassett and Bullmore, 2006; Sporns, 2013). In this work, through resting-state fMRI data analysis and computational model of whole brain dynamics, we demonstrated that the critical dynamics favors this optimal organization of FC networks, and failure of critical dynamics causes the collapse of balance between segregation and integration in the network by increasing the characteristic path length and decreasing the cluster coefficient.

Critical dynamics in brains has been observed at brains at different levels, from single neuron to the whole brain levels, with different recoding techniques (Shew et al., 2009; Gal and Marom, 2013; Gollo et al., 2013; Mora et al., 2015). Recent work with resting-state fMRI data analysis demonstrated the existence of large-scale critical dynamics, hallmarked by scale-free avalanche activity, in the human cortex (Tagliazucchi et al., 2012). Beside these observations, the critical brain hypothesis argued that criticality benefits neural information processing in many ways, e.g., the maximal information transmission and storage capabilities (Shew et al., 2011; Timme et al., 2016). However, these arguments usually defined the advantages of criticality in the general framework of information-theoretic [e.g., mutual information entropy (Shew et al., 2011)], neural dynamics [e.g., maximal dynamic range (Shew et al., 2009; Gautam et al., 2015), or the number of the metastable states in the energy landscape (Shew et al., 2009)], but their direct relations to brain functions are unclear. The FC network metrics have been related to many factors that affect brain functional performance, e.g., intellectual performance (van den Heuvel et al., 2009), aging (Wang et al., 2010), and a variety of brain diseases (Stam et al., 2007; Liu et al., 2008; Wang et al., 2009; Sanz-Arigita et al., 2010; Zhang et al., 2011; Rudie et al., 2013). Furthermore, it is believed that both segregated and integrated information processing are facilitated by the small-world topology of FC networks. The information transmission efficiency is maximized with this small-world topology, with their high clustering coefficient for segregated processing and short characteristic path length for integrated processing. Meanwhile, the disrupted network organization was found in neuropsychiatric disorders, usually characterized by increased characteristic path length and decreased cluster coefficient, and these changes were correlated with symptom severity in clinical-scale examinations (Stam et al., 2007; Liu et al., 2008; Wang et al., 2009; Zhang et al., 2011; Rudie et al., 2013). In our work, we found the critical dynamics maximizes clustering coefficient but minimizes the characteristic path length and yields both maximal local and global efficiency of the FC network. So our findings presented in this work not only uncovered the possible underlying dynamics from which the small-world FC network organization emerges but also revealed the advantage of large-scale critical dynamic in information processing at the whole brain level.

It is well-established that the EI balanced is critical for the forming of critical dynamics in healthy brains (Poil et al.,

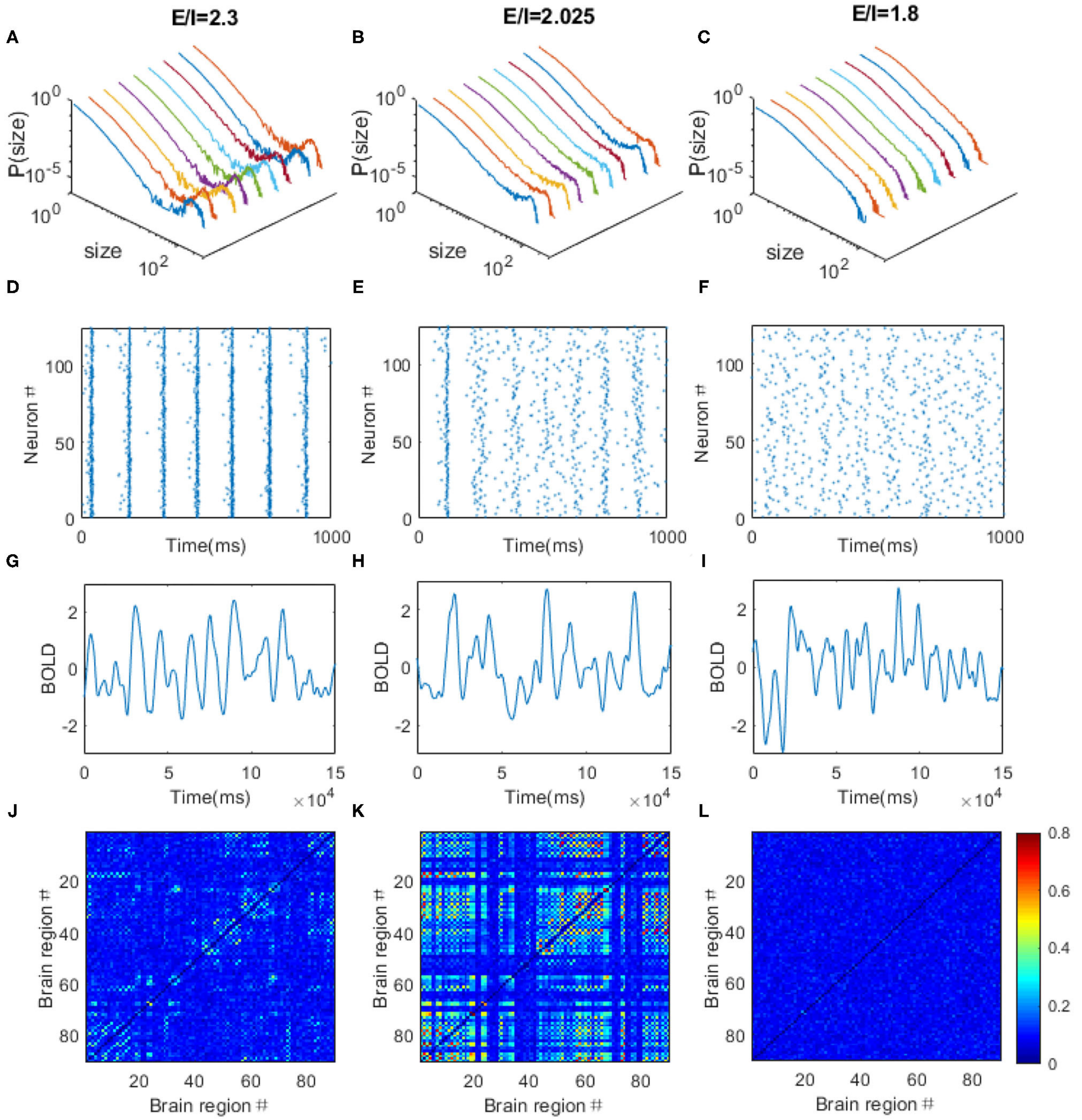


FIGURE 7 | The dynamical behaviors of the DTI+EI brain model. Through tuning the *E/I* ratio, the model can exhibits supercritical **(A)**, critical **(B)**, and subcritical **(C)** avalanche dynamics in each regions. Colored lines are corresponding to different brain regions chosen arbitrarily. **(D–F)** The raster plots of spatial-temporal firing distributions of an arbitrarily chosen brain regions corresponding to **(A–C)**. The firings are more synchronized in the supercritical regions **(D)**, but the firings are rather random when the system is subcritical with decreased excitatory connections **(F)**. The critical dynamics is characterized with moderate synchrony **(E)**. The dots in the raster plots indicate the firing of the neurons. **(G–L)** The typical simulated BOLD signals of an arbitrarily chosen brain region and simulated FC matrices from the DTI+EI brain model in the supercritical **(G,J)**, critical **(H,K)** and subcritical **(I,L)** regimes. Scale bar indicates the FC strength among the nodes in the model.

2012; Yang et al., 2012), and the neural systems may achieved this balanced state through synaptic plasticity (de Arcangelis et al., 2006; Stepp et al., 2015). On the contrary, the EI imbalance hypothesis has been postulated to underlie brain dysfunction across neurodevelopmental and neuropsychiatric disorders (Canitano and Pallagrosi, 2017; Foss-Feig et al., 2017). It was recently demonstrated that regulating the local *E/I* ratio crucially changes not only the characteristics of the emergent

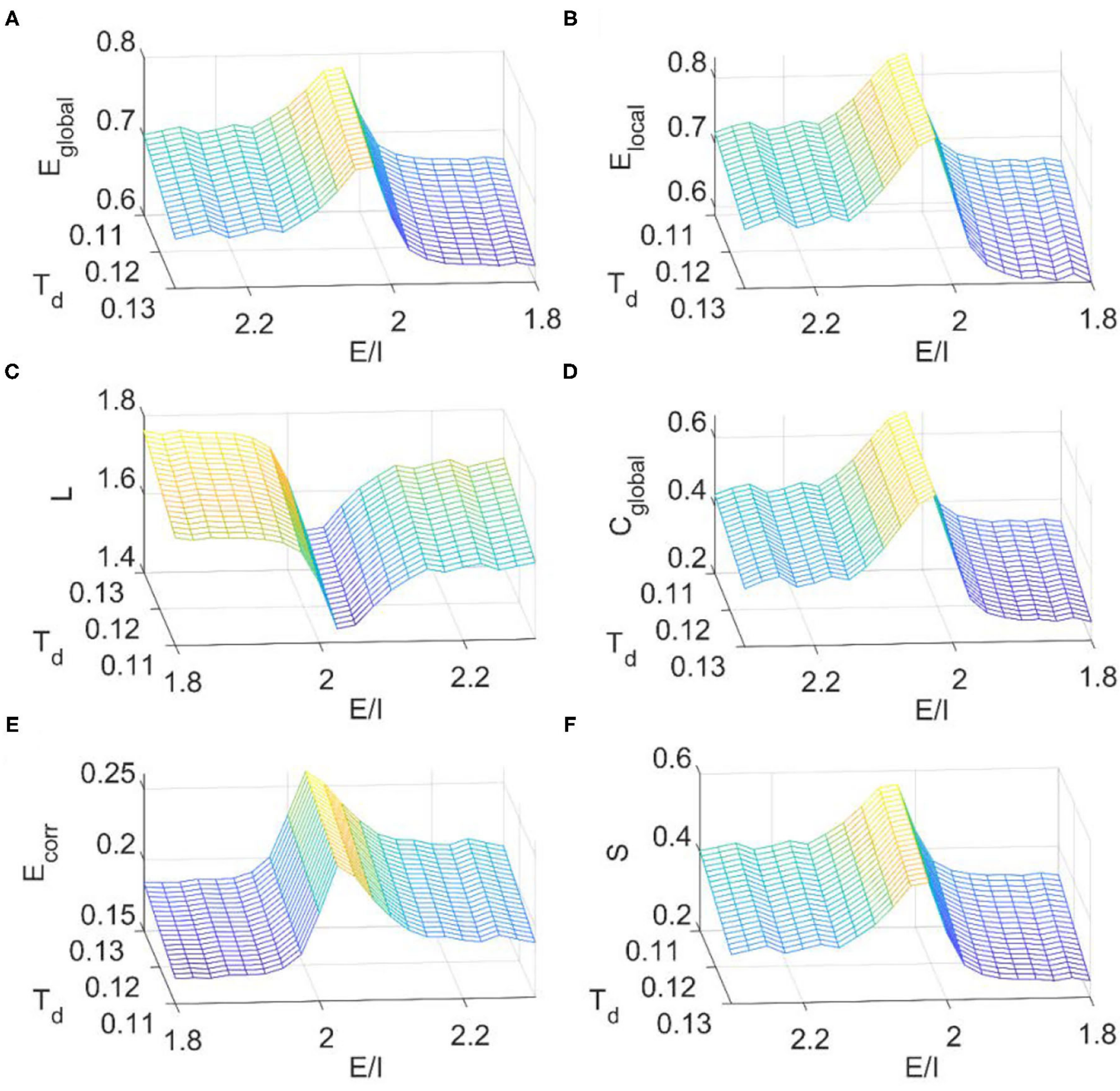


FIGURE 8 | The dependence of topological metrics of the FC network on the thresholding value T_d and *E/I* ratio in the DTI+EI whole brain model. **(A)** Global efficiency. **(B)** Local efficiency. **(C)** Characteristic path length. **(D)** Clustering coefficient. **(E)** Mean connection strength. **(F)** Sparsity. The results were obtained by averaging results from 10 times of simulation. Each simulation last for 480 s with a time step of 1 ms and the first 180 s was removed for stability. The obtained raw BOLD signals were then normalized and sampled at a rate of 0.5 Hz.

resting activity but also evoked activity. It also gives a more robust prediction of resting state FCs. Furthermore, it enhances the information capacity and the discrimination accuracy in the global networks (Deco et al., 2014b). These arguments have led to another hypothesis that criticality is a signature of healthy neural systems (Massobrio et al., 2015). In this study, we demonstrated that through tuning the *E*/*I* ratio of the brain model, the system could be poised at the critical point, and at this critical point, the functional integration and segregation of brain FC network is optimized. Considering the well-reported disruption of FC network in brain diseases, our modeling work with EI networks not only revealed the crucial role of the local *E*/*I* ratio in the forming of the optimal organization of whole brain FC networks but also provided supportive evidence for the hypothesis of EI imbalance by linking it with disruption of FC organization at the whole brains level, which has been observed in many brain diseases.

One attractive point and also the limitation of EI imbalance hypothesis is that brain disorders can be arranged in an imaginary line around the optimal point that balances excitation and inhibition. The limitation for unidimensionality of the EI imbalance has been discussed recently and it was argued that the higher dimensional models can better capture the multidimensional computational functions of neural circuits (O'Donnell et al., 2017). Therefore, EI balance may be not the only factor that is responsible for aberrant neural activity and

FC network organization in diseased brains. Our results from DTI+GH model suggested that the general conclusion in this work still holds even in this case, since optimal organization of FC networks can emerge from critical dynamics without EI connections. These results implies the possibility of utilizing criticality to bridge the gap between altered FC organization caused by diseases at the whole brain level and aberrant neural activity described by higher dimensional models at the circuit level, rather than one-dimensional EI model.

However, there are several limitations in the current study. In the fMRI data analysis, the Hurst exponent was used as an indicator of criticality. However, it cannot distinguish the super- or subcritical state of the system. The full solution for this problem requires the calculation of avalanche size distribution, branching ratio, as well as mean synchronization, as had done in EEG (Meisel et al., 2013). However, though the scale-free distribution of avalanche has been observed with fMRI (Tagliazucchi et al., 2012), the applicability of this method alone to identify super- or subcritical dynamics is still questionable. The major concern is that unlike EEG, fMRI does not measure neural activity directly but *via* the changes of BOLD signals. Therefore, future investigations that combines EEG and fMRI are necessary to validate the conclusions drawn from this study (Fagerholm et al., 2015).

It is also noticed that after the critical dynamics in our models is established, there is quite a few parameters that must be determined to obtain the simulated BOLD signals. Due to the simplifications made in the models, the neural activities produced by models are not exactly in the same time scale as in the real brains. Therefore, though we used the standard parameters for hemodynamic response function in models [it is also noticed that though these function and model were used widely in simulation of BOLD signals (Deco et al., 2011; Haimovici et al., 2013; Tagliazucchi et al., 2016), they were actually proposed for task-related hemodynamic response, not for resting state], simulation parameters (such as fMRI sampling rate, duration for scanning session) are not exactly the same as these in the experiment. Therefore, our results in this work requires further test with more detailed simulations of whole brain neural dynamics, as well as more detailed simulation of hemodynamic response in the resting state fMRI (Rangaprakash et al., 2017).

In this study, we tested the hypothesis that critical dynamics is responsible for optimal organization of brain FC networks which is usually featured with "small worldness." We found that the LRTCs of the BOLD signals measured with Hurst exponent is significantly correlated with the topological metrics of the FC networks, suggesting there exists an optimal dynamics for the brain FC network organization. Based on the inter-regional structural connection provided by DTI data, we built two kinds of whole brain dynamics model, using either simple cellular automaton, or more biological plausible neuronal networks with EI synaptic connections. In these models, we demonstrated that the critical dynamics could optimize the brain FC network organization through maximizing its cluster coefficient, while minimizing the shortest characteristic path length, so to achieve highest efficiency information transmission in the brain. We further showed that the local *E*/*I* ratio would have a great impact on critical dynamics and the organization of whole brain FC networks, suggesting imbalanced EI in brain circuitry may be responsible for the loss of small worldness in FC networks of brain disorder.

In conclusion, we demonstrated that the critical dynamics could optimize the brain FC network organization through maximizing its cluster coefficient, while minimizing the characteristic path length, so to achieve highest efficiency information transmission in the brain. Furthermore, imbalanced EI in brain circuitry may be responsible for the loss of the optimal organization in FC networks observed in brain disorder. Our findings revealed the crucial role of large scale critical dynamics in the forming of optimal FC network organization for efficient information processing, and potential relationship between local EI imbalance and the disrupted small-world organization. We hope that in the future these findings could not only lead to fundamental understanding on human brain function in health and its alterations in disease, but also help to develop whole brain computer models that could account for these alterations in brain disorder.

METHODS AND MATERIALS

fMRI Data Acquisition and Preprocessing

One hundred right-handed healthy subjects (mean age: 31.2 ± 8.8 years, range: 15–70 years, 63 males) participated in the study. The degree of education is from 0 to 23 years (mean: 8.5 years). All participants were screened to ensure they were free of neurological or psychiatric disorders. The data was acquired using a Siemens Trio 3.0 Tesla MRI scanner at the Second Hospital of Lanzhou University. All subjects provided written informed consent prior to the study which was approved by the medical ethics committee of the Second Hospital of Lanzhou University in accordance with the 1964 Declaration of Helsinki and its later amendments or comparable ethical standards. Participants were instructed to relax and keep their eyes closed, remain as motionless as possible, and not to think of anything in particular. Both functional and high-solution structural MRI were applied to all participants. T2*-weighted resting-state fMRI data were acquired using a gradient-echo EPI sequence, TR = 2 s, TE = 30 ms, slice thickness = 3 mm, gap = 0.99 mm, FOV = 240 mm, matrix size = 64 × 64. The scans lasted 360 s (180 volumes). High-resolution T1-weighted images were acquired with a magnetization prepared rapid gradient echo sequence, TR = 2 s, TE = 2.67 ms, inversion time = 900 ms, slice thickness = 1 mm, gap = 1 mm, FOV = 220 × 220 mm, matrix size = 256 × 224.

Preprocessing of fMRI data was performed using Statistical Parametric Mapping (SPM) 8 (http://www.fil.ion.ucl.ac.uk/spm) and the Data Processing Assistant for Resting-State fMRI (DPARSF) within the Data Processing and Analysis for Brain Imaging (DPABI) (Yan and Zang, 2010). Volumes were corrected for slice timing and head movements, and five subjects were excluded for excessive head movement (>3 mm or >3°) during the scan. After spatial normalization (Montreal Neurological Institute space), resampling (3 mm isotropic voxels), and spatial smoothing (4 mm, full-width, half-maximum Gaussian kernel), volumes were preprocessed using linear trend subtraction and temporal filtering (0.01–0.08 Hz). In addition, using the general linear regression, nuisance regressors including head motions, global mean signals, white matter signals, and cerebrospinal fluid signals were regressed out from the fMRI time series.

The DTI Data Acquisition and Processing

In this study, the DTI data was obtained from IMAGEN consortium, which included 142 healthy participants (76 females, age: 14.5 ± 0.2 years). The detailed information of data acquisition could be found in Ref. (Schumann et al., 2010). The DTI data were corrected for motion and eddy current distortion using FMRIB Software Library v5.0 (FSL, http://www.fmrib.ox.ac.uk/fsl) (Jenkinson et al., 2012). In addition, we extracted the brain mask from the B0 image. We used the TrackVis (Wang et al., 2007) to perform the fiber tractography with the deterministic tracking method. Maps of fractional anisotropy (FA) were computed from the DTI data. The regions of interest (ROIs) were determined by the AAL atlas-based T1 image from each subject (Tzourio-Mazoyer et al., 2002), using the PANDA suite (Cui et al., 2013). Finally, between each pair of ROIs, we assessed the fiber number to construct the DTI structural connection matrix.

Hurst Exponent

We use the Hurst exponent to measure the extent of long-range memory of the BOLD time series, either from the fMRI data or from the simulation with both brain models. The Hurst exponent is estimated using the method of classical rescaled range (RS) method (Blythe and Nikulin, 2017):

1. Divide the time series $\{y(t)\}_{t=1}^{T}$ into M subseries by choosing an appropriate number n, and each subseries has a window length of n.
2. For each subseries ($m = 1, 2, M$), calculate the local statistic $LS_{n,m} = \frac{R_{n,m}}{S_{n,m}}$. The range of mth subseries $R_{n,m} = max(Z_1, Z_2, \ldots, Z_n) - \min(Z_1, Z_2, \ldots, Z_n)$, where $Z_k = \sum_{t=1}^{k}(y_{t,m} - y_{n,m})$. $S_{n,m}$ is the standard deviation of mth subseries, which is calculated as $S_{n,m} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_{t,m} - y_{n,m})^2}$. Then by averaging over all subseries, we obtain the global statistic, i.e., $SS_n = \frac{1}{M}\sum_{m=1}^{M} LS_{n,m}$.
3. Through changing n and repeating the previous steps, we obtain a series of SS_n corresponding to a different choice of n.
4. The Hurst exponent is estimated by fitting the power law $SS_n \approx Cn^H$ to the data. This can be done by running a double logarithm regression for a series of SS_n corresponding to different values of n.

In the calculation, the global Hurst exponent of the whole brain level was obtained by average the local Hurst exponent across 90 brain regions in both fMRI data analysis and the DTI+EI model. Whereas, in the DTI+GH models, to obtain the stable estimation of Hurst exponent of the systems, we first averaged the 90 simulated BOLD time series and then calculated its Hurst exponent.

Network Metrics

First, we used the AAL template to extract from 90 brain regions 90 time series, each of which is the averaged BOLD signals across all the voxels in each region. The correlation coefficients for each pair of the time series was then calculated to build the FC matrix $z(i, j)$ ($i, j = 1, 2, 90$), in which each off-diagonal element is the correlation coefficient between a pair of brain regions. The FC network was constructed by setting a threshold T_d to each element in the absolute FC matrix:

$$a_{ij} = \begin{cases} 1 \text{ if } |z(i,j)| \geq T_\mathrm{d} \\ 0 \text{ otherwise} \end{cases} \quad (1)$$

The network metrics of a FC network with n nodes was calculated as follows:

The degree of a node i is defined as the number of its direct neighbors:

$$k_i = \sum\nolimits_{j \in N} {}^{a_{ij}} \quad (2)$$

where a_{ij} is the connection between nodes i and j, $a_{ij} = 1$ when they are directly linked, $a_{ij} = 0$ if not.

The connectivity strength of the node i is:

$$E_{i\_\mathrm{corr}} = \frac{1}{k_i}\sum_{j \in N} |z(i,j)| \cdot a_{ij}, \quad (3)$$

which is a measure to evaluate the strength of the connectivity between node i and the nodes connected to it. The connectivity strength of a network is:

$$E_\mathrm{corr} = \frac{1}{n}\sum_{i \in N} E_{i\_\mathrm{corr}}. \quad (4)$$

The ratio of the number of existing edges to the number of maximum possible number:

$$S = \frac{1}{n(n-1)}\sum_{i \in N} k_i, \quad (5)$$

is defined as the sparsity of the network.

Characteristic path length measures the extent of average connectivity or overall routing efficiency of the network (Sanz-Arigita et al., 2010), which is defined as

$$L = \frac{1}{n}\sum_{i \in N} L_i = \frac{1}{n}\sum_{i \in N}\frac{\sum_{j \in N, j \neq i} d_{ij}}{n-1}, \quad (6)$$

in which $d_{ij} = \sum_{a_{uv} \in g_{i \leftrightarrow j}} a_{uv}$ is the shortest path length between nodes i and j with the shortest way $g_{i \leftrightarrow j}$ and L_i is the mean shortest path length of node i.

Global efficiency is a measure of the efficiency of parallel information transfer in the network at the global level:

$$E_{global} = \frac{1}{n}\sum_{i \in N} E_i = \frac{1}{n}\sum_{i \in N}\frac{\sum_{j \in N, j \neq i} d_{ij}^{-1}}{n-1}. \quad (7)$$

Local efficiency, which measures the efficiency at the local level, is defined as:

$$E_{local} = \frac{1}{n}\sum_{i\in N} E_{loc,i} = \frac{1}{n}\sum_{i\in N} \frac{\sum_{j,h\in N, j\neq i} a_{ij}a_{ih}\left[d_{jh}(N_i)\right]^{-1}}{k_i(k_i-1)}, \quad (8)$$

where $d_{jh}(N_i)$ is the shortest path length between nodes j and h which should be the nodes directly connected to node i.

Clustering coefficient measures the possibility that any two neighbors of one node are also connected, i.e., the extent of the local density of the network:

$$C_{global} = \frac{1}{n}\sum_{i\in N} C_i = \frac{1}{n}\sum_{i\in N} \frac{2t_i}{k_i(k_i-1)}, \quad (9)$$

where $t_i = \frac{1}{2}\sum_{j,h\in N} a_{ij}a_{ih}a_{jh}$ is the number of triangles around node i.

The network metrics were calculated in a range of threshold that is small enough to assure the mean number of connected nodes within each group was > 89, yet large enough so that the small worldness of network still holds, i.e., the global efficiency of the normal networks is less than the global efficiency of the random networks.

DTI+GH Brain Network Model

The DTI+GH brain network model used in this study was adapted from the model proposed by Haimovici et al. (2013). The coupling strength w_{ij} between any two nodes i and j was given by the corresponding element in the DTI structural connection matrix multiplied by 0.01. Each node was modeled with the Greenberg-Hastings (GH) dynamics. The detailed description of GH dynamics could be found in Ref. (Haimovici et al., 2013). We binarized the time series of each node by assigning state $E = 1$ and the rest of the states into 0 s. To model the brain neurometabolic coupling, we then convolved the binarized time series with a hemodynamic response function (Henson and Friston, 2007):

$$f(t) = \left(\frac{t-o}{d}\right)^{p-1}\left(\frac{\exp(-(t-o)/d)}{d(p-1)!}\right)^{p-1}, \quad (10)$$

where $d = 0.6$ is the time-scaling, $o = 0$ is the onset delay, and $p = 3$ is an integer phase-delay (the peak delay is given by pd, and the dispersion by pd^2). The obtained raw BOLD signals were sampled every 140 iteration steps to have the simulated BOLD time series of 200 time points.

DTI+EI Whole Brain Model

In this model, each brain region is modeled by an EI neuronal network comprising 100 excitatory and 25 inhibitory neurons. As in the mammalian neocortex, the ratio of excitatory to inhibitory cells is 4 to 1 (DeFelipe et al., 2002). The connection probability between these neurons is set to 0.5. Then we use the Izhikevich model to produce single neuron dynamics (Izhikevich, 2004):

$$\frac{\mathrm{dv^i}}{\mathrm{dt}} = 0.04(\mathrm{v^i})^2 + 5\mathrm{v^i} + 140 - \mathrm{u^i} + \mathrm{I_{synapse}} + \xi(\mathrm{t}), \quad (11)$$

$$\frac{\mathrm{du^i}}{\mathrm{dt}} = \mathrm{a}\left(\mathrm{bv^i} - \mathrm{u^i}\right), \quad (12)$$

$$\text{If } \mathrm{v^i} \geq 30\,\mathrm{mV}, \text{ then} \begin{cases} v^i \leftarrow c \\ u^i \leftarrow u^i + d \end{cases} \quad (13)$$

where v^i and u^i represent the ith neuron's membrane potential and recovery, respectively. The parameters a, b, c, and d are set to model either excitatory (0.02, 0.2, $-65 + 15r^2$, $2.8{-}6r^2$) or inhibitory ($0.02 + 0.08r$, $0.2{-}0.05r$, -65, 2) neurons. To introduce some variability in the neuronal population, the variable r is drawn from a uniform distribution $U(0,1)$. $I_{synapse}$ represents synaptic currents this neuron receives from other neurons. $\xi(t)$ is the background Gaussian white noise with $\langle\xi(t)\rangle$=0 and $\langle\xi(t)\xi(t')\rangle$=D$\delta$(t-t'), where the noise intensity $D = 25$ for excitatory neurons and $D = 6.25$ for inhibitory neurons. Equation (13) models the after-spike reset behavior when the membrane potential v^i exceeds a threshold. This model is widely used in large-scale neuronal network modeling because of its computational efficiency and biological plausibility (Izhikevich, 2004).

In our model, the synaptic current received by one neuron ($I_{synapse}$) can be divided into two parts: $I^i_{intra<uscore>synapse}$ is the synaptic current the ith neuron receives from other neurons in this brain region, which is written as:

$$\mathrm{I^i_{intra<uscore>synapse}} = \sum_{\mathrm{j\neq i}} \mathrm{g^{ij}_{E\to E,E\to I,I\to E,I\to I}}\,\delta\left(\mathrm{t} - \mathrm{t^j_{spike}}\right), \quad (14)$$

where t^j_{spike} is the time instant when the presynaptic neuron j that exerts synaptic connection to the neuron i fires a spike. The summation runs across all the neurons that exert synaptic connection to the neuron i. The intra-regional synaptic connecting strength is set as follow: $g^{ij}_{E\to I} = 1$ if the jth neuron is excitatory and ith neuron is inhibitory; $g^{ij}_{I\to E,I\to I} = -1$ if the jth neuron is inhibitory no matter if the ith neuron is excitatory or inhibitory. In the simulation, we systematically varied the connections among excitatory neurons $g^{ij}_{E\to E}$ to change the local E/I ratio, which was defined as the ratio of $g^{ij}_{E\to E}$ to $g^{ij}_{I\to\ I}$.

The inter-regional connections are set only for excitatory neurons among the different brain regions. Therefore, $I^{ij}_{inter<uscore>synapse} = 0$ if neurons i and j belong to different regions and at least one of them is an inhibitory neuron. The inter-regional connection probability of excitatory neurons in each pair of brain regions is proportional to their corresponding DTI structural connection strength and the maximum is set to be 0.5. Specifically, if the DTI connection between region m and n is q, then the excitatory neurons in these two brain regions have an inter-regional connection probability of $0.5q_{\mathrm{mn}}/q_{\mathrm{max}}$, where q_{max} is the highest value in the DTI matrix. For example, for neuron *i* in one region, it receives inter-regional synaptic currents from neuron *j* in another region is written in the following form:

$$\mathrm{I}^{\mathrm{ij}}_{\mathrm{inter<uscore>synapse}} = \sum_{\substack{i \in M \\ j \in N}} \mathrm{g}^{\mathrm{ij}}_{\mathrm{E}\leftrightarrow\mathrm{E}}\, \delta(\mathrm{t} - \mathrm{t}^{\mathrm{j}}_{\mathrm{spike}}), \tag{15}$$

where $E \leftrightarrow E$ represents the inter-regional excitatory synaptic coupling. The inter-regional synaptic connection $g^{ij}_{E\leftrightarrow E} = 0.15$ in the simulation.

For DTI+EI model, the fMRI BOLD signals are computed with Balloon-Windkessel hemodynamic model (Friston et al., 2003). The regional BOLD signal is driven by the collective neuronal activity of both excitatory and inhibitory neurons. For region *i*, we define neuronal activity z_i as the ratio of number of spikes to the number of neurons in the region within a time window of 1 ms. We assume z_i causes an increase in a vasodilatory signal s_i that increases the flow f_i. The inflow f_i then causes changes in blood volume v_i and deoxyhemoglobin content q_i:

$$\frac{\mathrm{ds_i}\,(\mathrm{t})}{\mathrm{dt}} = \epsilon_{\mathrm{i}}\mathrm{z_i} - \mathrm{k_i}\mathrm{s_i} - \gamma_{\mathrm{i}}\left(\mathrm{f_i} - 1\right), \tag{16}$$

$$\frac{\mathrm{df_i}\,(\mathrm{t})}{\mathrm{dt}} = \mathrm{s_i}, \tag{17}$$

$$\tau_{\mathrm{i}}\frac{\mathrm{dv_i}\,(\mathrm{t})}{\mathrm{dt}} = \mathrm{f_i} - \mathrm{v_i}^{1/\alpha}, \tag{18}$$

$$\tau_{\mathrm{i}}\frac{\mathrm{dq_i}\,(\mathrm{t})}{\mathrm{dt}} = \frac{\mathrm{f_i}(1-(1-\rho_{\mathrm{i}})^{\mathrm{f_i}})}{\rho_{\mathrm{i}}} - \frac{\mathrm{q_i}\mathrm{v_i}^{1/\alpha}}{\mathrm{v_i}}, \tag{19}$$

where ρ is the resting oxygen extraction fraction. Taken as a static non-linear function of volume and deoxyhemoglobin that comprises a volume-weighted sum of extra- and intravascular signals, the BOLD signal is then calculated as:

$$\mathrm{y_i} = \mathrm{V_0}(7\rho_{\mathrm{i}}\left(1-\mathrm{q_i}\right) + 2\left(1 - \frac{\mathrm{q_i}}{\mathrm{v_i}}\right) + (2\rho_{\mathrm{i}} - 0.2)(1-\mathrm{v_i})),(20)$$

where $V_0 = 0.02$ is the resting blood volume fraction. The biophysical parameters in the simulation were set as $\epsilon_i = 0.2$, $k_i = 0.65$, $\gamma_i = 0.41$, $\tau_i = 0.98$, $\alpha_i = 0.32$, and $\rho_i = 0.34$. The simulation last for 480 s with a time step of 1 ms, and the first 180 s was removed for stability. The obtained raw BOLD signals were then normalized and sampled every 2 s (TR).

Avalanche Detection

For the DTI+GH model, the simulated time series were subsequently binarized by assigning the active state to 1 and the other two to 0. Then the raster plot of the activations was divided into many consecutive frames. We calculated the number of activated nodes N_i for frame *i*. In addition, this frame is blank if $N_i = 0$. If the consecutive frames contain activated nodes, proceeding with blank frame, and ended with blank frame, then the activities in these consecutive frames is defined as an avalanche. The number of total activated nodes in this avalanche is defined as its size. The frame length of DTI+GH model was set to two iteration steps so to obtain avalanche size distribution with power law distribution of −1.5 (Beggs and Plenz, 2003).

For the DTI+EI model, the detection of avalanche in each region is the same as before, except that the activation of nodes is replaced with the firing events of the neurons. The frame length is chosen to be 2 ms so as to produce power law avalanche distributions with exponents closest to −1.5.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Second Hospital of Lanzhou University. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR'S NOTE

The hypothesis that the brain might operate at or near-phase transitions because criticality facilitates information processing capabilities and health. This hypothesis was strongly driven by theoretical concepts and supported by many experimental studies. Recent structure-dynamics-function modeling studies combining the structural and functional imaging data at whole brain level demonstrated the functional connectivity (FC) emerges from structural connectivity when the brain dynamics is poised at the criticality. It is therefore conjectured that criticality may facilitate the optimal organization of FC networks, usually characterized by "small worldness" which are corrupted in disordered brains. There are several arguments for this conjecture: First, criticality has been argued to optimize the neural systems for computation, whereas the "small worldness" FC network has been considered an efficient way for inter-regional communication in brains. Second, it has been shown in experiments and simulations that a proper excitation-inhibition (*E*/*I*) balance is required to maintain critical dynamics in cortical networks. Accordingly, *E*/*I* imbalances have been implicated in various brain disorders, such as autism, schizophrenia, etc. In this study, we demonstrated that the FC network organization is optimized by critical dynamics by maximizing the cluster coefficient while minimizing the characteristic path length, so to yield maximal global and local efficiency in information transmission. We also demonstrated with whole brain model that the local *E*/*I* ratio can be optimized to produce critical dynamics in the system, thereby yielding optimal organization of FC networks at the whole brain level.

AUTHOR CONTRIBUTIONS

LY and JF designed research. LY, XZ, NM, BS, ZW, and GL performed research. LY, JF, LL, and ZW wrote the paper. All authors reviewed the manuscript.

FUNDING

This study was funded by the National Natural Science Foundation of China (Grants Nos. 11564034 and 11105062) and the Fundamental Research Funds for the Central Universities (Grant Nos. lzujbky-2015-119 and 31920130008). JF and NM also received funding from the 111 Project (Grant No. B18015), the key project of Shanghai Science & Technology (Grant No. 16JC1420402), the Shanghai Municipal Science and Technology Major Project (Grant No. 2018SHZDZX01), and the Shanghai Sailing Program (Grant No. 17YF1426400).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fncom.2021.641335/full#supplementary-material

REFERENCES

Achard, S., and Bullmore, E. (2007). Efficiency and cost of economical brain functional networks. *PLoS Comput. Biol.* 3:e17. doi: 10.1371/journal.pcbi.0030017

Bassett, D. S., and Bullmore, E. (2006). Small-world brain networks. *Neuroscientist* 12, 512–523. doi: 10.1177/1073858406293182

Beggs, J. M., and Plenz, D. (2003). Neuronal avalanches in neocortical circuits. *J. Neurosci.* 23, 11167–11177. doi: 10.1523/JNEUROSCI.23-35-11167.2003

Blythe, D. A. J., and Nikulin, V. V. (2017). Long-range temporal correlations in neural narrowband time-series arise due to critical dynamics. *PLoS ONE* 12:e0175628. doi: 10.1371/journal.pone.0175628

Canitano, R., and Pallagrosi, M. (2017). Autism spectrum disorders and schizophrenia spectrum disorders: excitation/inhibition imbalance and developmental trajectories. *Front. Psychiatry.* 8:69. doi: 10.3389/fpsyt.2017.00069

Cocchi, L., Gollo, L. L., Zalesky, A., and Breakspear, M. (2017). Criticality in the brain: a synthesis of neurobiology, models and cognition. *Progr. Neurobiol.* 158, 132–152. doi: 10.1016/j.pneurobio.2017.07.002

Cui, Z., Zhong, S., Xu, P., Gong, G., and He, Y. (2013). PANDA: a pipeline toolbox for analyzing brain diffusion images. *Front. Hum. Neurosci.* 7:42. doi: 10.3389/fnhum.2013.00042

de Arcangelis, L., Perrone-Capano, C., and Herrmann, H. J. (2006). Self-organized criticality model for brain plasticity. *Phys. Rev. Lett.* 96:028107. doi: 10.1103/PhysRevLett.96.028107

Deco, G., Jirsa, V. K., and McIntosh, A. R. (2011). Emerging concepts for the dynamical organization of resting-state activity in the brain. *Nat. Rev. Neurosci.* 12, 43–56. doi: 10.1038/nrn2961

Deco, G., McIntosh, A. R., Shen, K., Hutchison, R. M., Menon, R. S., Everling, S., et al. (2014a). Identification of optimal structural connectivity using functional connectivity and neural modeling. *J. Neurosci.* 34, 7910–7916. doi: 10.1523/JNEUROSCI.4423-13.2014

Deco, G., Ponce-Alvarez, A., Hagmann, P., Romani, G. L., Mantini, D., and Corbetta, M. (2014b). How local excitation–inhibition ratio impacts the whole brain dynamics. *J. Neurosci.* 34, 7886–7898. doi: 10.1523/JNEUROSCI.5068-13.2014

DeFelipe, J., Alonso-Nanclares, L., and Arellano, J. I. (2002). Microstructure of the neocortex: comparative aspects. *J. Neurocytol.* 31, 299–316. doi: 10.1023/a:1024130211265

Fagerholm, E. D., Lorenz, R., Scott, G., Dinov, M., Hellyer, P. J., Mirzaei, N., et al. (2015). Cascades and cognitive state: focused attention incurs subcritical dynamics. *J. Neurosci.* 35, 4626–4634. doi: 10.1523/JNEUROSCI.3694-14.2015

Foss-Feig, J. H., Adkinson, B. D., Ji, J. L., Yang, G., Srihari, V. H., McPartland, J. C., et al. (2017). Searching for cross-diagnostic convergence: neural mechanisms governing excitation and inhibition balance in schizophrenia and autism spectrum disorders. *Biol. Psychiatry* 81, 848–861. doi: 10.1016/j.biopsych.2017.03.005

Friston, K. J., Harrison, L., and Penny, W. (2003). Dynamic causal modelling. *Neuroimage* 19, 1273–1302. doi: 10.1016/S1053-8119(03)00202-7

Gal, A., and Marom, S. (2013). Self-organized criticality in single-neuron excitability. *Phys. Rev. E Statist. Nonlinear Soft Matter Phys.* 88:062717. doi: 10.1103/PhysRevE.88.062717

Gautam, S. H., Hoang, T. T., McClanahan, K., Grady, S. K., and Shew, W. L. (2015). Maximizing sensory dynamic range by tuning the cortical state to criticality. *PLoS Comput. Biol.* 11:e1004576. doi: 10.1371/journal.pcbi.1004576

Gireesh, E. D., and Plenz, D. (2008). Neuronal avalanches organize as nested theta- and beta/gamma-oscillations during development of cortical layer 2/3. *Proc. Natl. Acad. Sci. U.S.A.* 105, 7576–7581. doi: 10.1073/pnas.0800537105

Gollo, L. L., Kinouchi, O., and Copelli, M. (2013). Single-neuron criticality optimizes analog dendritic computation. *Sci. Rep.* 3:3222. doi: 10.1038/srep03222

Haimovici, A., Tagliazucchi, E., Balenzuela, P., and Chialvo, D. R. (2013). Brain organization into resting state networks emerges at criticality on a model of the human connectome. *Phys. Rev. Lett.* 110:178101. doi: 10.1103/PhysRevLett.110.178101

Haldeman, C., and Beggs, J. M. (2005). Critical branching captures activity in living neural networks and maximizes the number of metastable states. *Phys. Rev. Lett.* 94:058101. doi: 10.1103/PhysRevLett.94.058101

He, B. J. (2011). Scale-free properties of the fMRI signal during rest and task. *J. Neurosci.* 31, 13786–13795. doi: 10.1523/JNEUROSCI.2111-11.2011

Heine, L., Soddu, A., Gómez, F., Vanhaudenhuyse, A., Tshibanda, L., Thonnard, M., et al. (2012). Resting state networks and consciousness: alterations of multiple resting state network connectivity in physiological, pharmacological, and pathological consciousness states. *Front. Psychol.* 3:295. doi: 10.3389/fpsyg.2012.00295

Henson, R., and Friston, K. (2007). "Convolution models for fMRI," in *Statistical Parametric Mapping: The Analysis of Functional Brain Images,* eds K. Friston, J. Ashburner, S. Kiebel, T. Nichols, and W. Penny (London: Academic Press), 178–192. doi: 10.1016/B978-012372560-8.X5000-1

Izhikevich, E. M. (2004). Which model to use for cortical spiking neurons? *IEEE Transact. Neural Netw.* 15, 1063–1070. doi: 10.1109/TNN.2004.832719

Jenkinson, M., Beckmann, C. F., Behrens, T. E. J., Woolrich, M. W., and Smith, S. M. (2012). FSL. *Neuroimage* 62, 782–790. doi: 10.1016/j.neuroimage.2011.09.015

Latora, V., and Marchiori, M. (2001). Efficient behavior of small-world networks. *Phys. Rev. Lett.* 87:198701. doi: 10.1103/PhysRevLett.87.198701

Lee, M. H., Smyser, C. D., and Shimony, J. S. (2013). Resting state fMRI: a review of methods and clinical applications. *AJNR Am. J. Neuroradiol.* 34, 1866–1872. doi: 10.3174/ajnr.A3263

Liu, Y., Liang, M., Zhou, Y., He, Y., Hao, Y., Song, M., et al. (2008). Disrupted small-world networks in schizophrenia. *Brain* 131, 945–961. doi: 10.1093/brain/awn018

Massobrio, P., de Arcangelis, L., Pasquale, V., Jensen, H. J., and Plenz, D. (2015). Criticality as a signature of healthy neural systems. *Front. Syst. Neurosci.* 9:22. doi: 10.3389/978-2-88919-503-9

Meisel, C., Olbrich, E., Shriki, O., and Achermann, P. (2013). Fading signatures of critical brain dynamics during sustained wakefulness in humans. *J. Neurosci.* 33, 17363–17372. doi: 10.1523/JNEUROSCI.1516-13.2013

Mora, T., Deny, S., and Marre, O. (2015). Dynamical criticality in the collective activity of a population of retinal neurons. *Phys. Rev. Lett.* 114:078105. doi: 10.1103/PhysRevLett.114.078105

O'Donnell, C., Gonçalves, J. T., Portera-Cailliau, C., and Sejnowski, T. J. (2017). Beyond excitation/inhibition imbalance in multidimensional models of neural circuit changes in brain disorders. *Elife* 6:e26724. doi: 10.7554/eLife.26724

Palva, J. M., Zhigalov, A., Hirvonen, J., Korhonen, O., Linkenkaer-Hansen, K., and Palva, S. (2013). Neuronal long-range temporal correlations and avalanche dynamics are correlated with behavioral scaling laws. *Proc. Natl. Acad. Sci. U.S.A*. 110, 3585–3590. doi: 10.1073/pnas.1216855110

Plenz, D. (2012). Neuronal avalanches and coherence potentials. *Eur. Phys. J. Special Topics* 205, 259–301. doi: 10.1140/epjst/e2012-01575-5

Poil, S.-S., Hardstone, R., Mansvelder, H. D., and Linkenkaer-Hansen, K. (2012). Critical-state dynamics of avalanches and oscillations jointly emerge from balanced excitation/inhibition in neuronal networks. *J. Neurosci*. 32, 9817–9823. doi: 10.1523/JNEUROSCI.5990-11.2012

Raichle, M. E. (2015). The brain's default mode network. *Annu. Rev. Neurosci*. 38, 433–447. doi: 10.1146/annurev-neuro-071013-014030

Rangaprakash, D., Dretsch, M. N., Yan, W., Katz, J. S., Denney, T. S., and Deshpande, G. (2017). Hemodynamic response function parameters obtained from resting-state functional MRI data in soldiers with trauma. *Data Brief*. 14, 558–562. doi: 10.1016/j.dib.2017.07.072

Ribeiro, T. L., Copelli, M., Caixeta, F., Belchior, H., Chialvo, D. R., Nicolelis, M. A. L., et al. (2010). Spike avalanches exhibit universal dynamics across the sleep-wake cycle. *PLoS ONE* 5:e14129. doi: 10.1371/journal.pone.0014129

Rubenstein, J. L. R., and Merzenich, M. M. (2003). Model of autism: increased ratio of excitation/inhibition in key neural systems. *Genes Brain Behav*. 2, 255–267. doi: 10.1034/j.1601-183X.2003.00037.x

Rubinov, M., and Sporns, O. (2010). Complex network measures of brain connectivity: uses and interpretations. *Neuroimage* 52, 1059–1069. doi: 10.1016/j.neuroimage.2009.10.003

Rudie, J. D., Brown, J. A., Beck-Pancer, D., Hernandez, L. M., Dennis, E. L., Thompson, P. M., et al. (2013). Altered functional and structural brain network organization in autism. *NeuroImage Clin*. 2, 79–94. doi: 10.1016/j.nicl.2012.11.006

Sanz-Arigita, E. J., Schoonheim, M. M., Damoiseaux, J. S., Rombouts, S. A. R. B., Maris, E., Barkhof, F., et al. (2010). Loss of small-world networks in Alzheimer's Disease: graph analysis of fMRI resting-state functional connectivity. *PLoS ONE* 5:e13788. doi: 10.1371/journal.pone.0013788

Schumann, G., Loth, E., Banaschewski, T., Barbot, A., Barker, G., Büchel, C., et al. (2010). The IMAGEN study: reinforcement-related behaviour in normal brain function and psychopathology. *Mol. Psychiatry* 15:1128. doi: 10.1038/mp.2010.4

Shew, W. L., Yang, H., Petermann, T., Roy, R., and Plenz, D. (2009). Neuronal avalanches imply maximum dynamic range in cortical networks at criticality. *J. Neurosci*. 29, 15595–15600. doi: 10.1523/JNEUROSCI.3864-09.2009

Shew, W. L., Yang, H., Yu, S., Roy, R., and Plenz, D. (2011). Information capacity and transmission are maximized in balanced cortical networks with neuronal avalanches. *J. Neurosci*. 31, 55–63. doi: 10.1523/JNEUROSCI.4637-10.2011

Shriki, O., Alstott, J., Carver, F., Holroyd, T., Henson, R. N. A., Smith, M. L., et al. (2013). Neuronal avalanches in the resting MEG of the human brain. *J. Neurosci*. 33, 7079–7090. doi: 10.1523/JNEUROSCI.4286-12.2013

Sporns, O. (2013). Network attributes for segregation and integration in the human brain. *Curr. Opin. Neurobiol*. 23, 162–171. doi: 10.1016/j.conb.2012.11.015

Stam, C., Jones, B., Nolte, G., Breakspear, M., and Scheltens, P. (2007). Smallworld networks and functional connectivity in Alzheimer's disease. *Cerebral Cortex*. 17, 92–99. doi: 10.1093/cercor/bhj127

Stepp, N., Plenz, D., and Srinivasa, N. (2015). Synaptic plasticity enables adaptive self-tuning critical networks. *PLoS Comput. Biol*. 11:e1004043. doi: 10.1371/journal.pcbi.1004043

Tagliazucchi, E., Balenzuela, P., Fraiman, D., and Chialvo, D. (2012). Criticality in large-scale brain fMRI dynamics unveiled by a novel point process analysis. *Front. Physiol*. 3:15. doi: 10.3389/fphys.2012.00015

Tagliazucchi, E., Chialvo, D. R., Siniatchkin, M., Amico, E., Brichant, J.-F., Bonhomme, V., et al. (2016). Large-scale signatures of unconsciousness are consistent with a departure from critical dynamics. *J. R. Soc. Interface* 13:20151027. doi: 10.1098/rsif.2015.1027

Thiagarajan, T. C., Lebedev, M. A., Nicolelis, M. A., and Plenz, D. (2010). Coherence potentials: loss-less, all-or-none network events in the cortex. *PLoS Biol*. 8:e1000278. doi: 10.1371/journal.pbio.1000278

Timme, N. M., Marshall, N. J., Bennett, N., Ripp, M., Lautzenhiser, E., and Beggs, J. M. (2016). Criticality maximizes complexity in neural tissue. *Front. Physiol*. 7:425. doi: 10.3389/fphys.2016.00425

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., et al. (2002). Automated anatomical labeling of activations in SPM Using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage* 15, 273–289. doi: 10.1006/nimg.2001.0978

van den Heuvel, M. P., and Hulshoff Pol, H. E. (2010). Exploring the brain network: a review on resting-state fMRI functional connectivity. *Eur. Neuropsychopharmacol*. 20, 519–534. doi: 10.1016/j.euroneuro.2010.03.008

van den Heuvel, M. P., Stam, C. J., Kahn, R. S., and Hulshoff Pol, H. E. (2009). Efficiency of functional brain networks and intellectual performance. *J. Neurosci*. 29, 7619–7624. doi: 10.1523/JNEUROSCI.1443-09.2009

Wang, L., Li, Y., Metzak, P., He, Y., and Woodward, T. S. (2010). Age-related changes in topological patterns of large-scale brain functional networks during memory encoding and recognition. *Neuroimage* 50, 862–872. doi: 10.1016/j.neuroimage.2010.01.044

Wang, L., Zhu, C., He, Y., Zang, Y., Cao, Q., Zhang, H., et al. (2009). Altered small-world brain functional networks in children with attention-deficit/hyperactivity disorder. *Hum. Brain Mapp*. 30, 638–649. doi: 10.1002/hbm.20530

Wang, R., Benner, T., Sorensen, A. G., and Wedeen, V. J. (eds.). (2007). Diffusion toolkit: a software package for diffusion imaging data processing and tractography. *Proc. Intl. Soc. Mag. Reson. Med*. 15:3720.

Watts, D. J., and Strogatz, S. H. (1998). Collective dynamics of "small-world" networks. *Nature* 393, 440–442. doi: 10.1038/30918

Yan, C., and Zang, Y. (2010). DPARSF: a MATLAB toolbox for "pipeline" data analysis of resting-state fMRI. *Front. Syst. Neurosci*. 4:13. doi: 10.3389/fnsys.2010.00013

Yang, H., Shew, W. L., Roy, R., and Plenz, D. (2012). Maximal variability of phase synchrony in cortical networks with neuronal avalanches. *J. Neurosci*. 32, 1061–1072. doi: 10.1523/JNEUROSCI.2771-11.2012

Yu, L., De Mazancourt, M., Hess, A., Ashadi, F. R., Klein, I., Mal, H., et al. (2016). Functional connectivity and information flow of the respiratory neural network in chronic obstructive pulmonary disease. *Hum. Brain Mapp*. 37, 2736–2754. doi: 10.1002/hbm.23205

Zhang, J., Wang, J., Wu, Q., Kuang, W., Huang, X., He, Y., et al. (2011). Disrupted brain connectivity networks in drug-naive, first-episode major depressive disorder. *Biol. Psychiatry*. 70, 334–342. doi: 10.1016/j.biopsych.2011.05.018

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Computational Neuroscience

ORIGINAL RESEARCH
published: 20 May 2021
doi: 10.3389/fncom.2021.626259

The Face Module Emerged in a Deep Convolutional Neural Network Selectively Deprived of Face Experience

Shan Xu [1*†], *Yiyuan Zhang* [1†], *Zonglei Zhen* [1] *and Jia Liu* [2*]

[1] *Beijing Key Laboratory of Applied Experimental Psychology, Faculty of Psychology, Beijing Normal University, Beijing, China,*
[2] *Department of Psychology & Tsinghua Laboratory of Brain and Intelligence, Tsinghua University, Beijing, China*

Can we recognize faces with zero experience on faces? This question is critical because it examines the role of experiences in the formation of domain-specific modules in the brain. Investigation with humans and non-human animals on this issue cannot easily dissociate the effect of the visual experience from that of the hardwired domain-specificity. Therefore, the present study built a model of selective deprivation of the experience on faces with a representative deep convolutional neural network, AlexNet, by removing all images containing faces from its training stimuli. This model did not show significant deficits in face categorization and discrimination, and face-selective modules automatically emerged. However, the deprivation reduced the domain-specificity of the face module. In sum, our study provides empirical evidence on the role of nature vs. nurture in developing the domain-specific modules that domain-specificity may evolve from non-specific experience without genetic predisposition, and is further fine-tuned by domain-specific experience.

Keywords: face perception, face domain, deep convolutional neural network, visual deprivation, experience

OPEN ACCESS

Edited by:
Dezhong Yao,
University of Electronic Science and Technology of China, China

Reviewed by:
Peiyang Li,
Chongqing University of Posts and Telecommunications, China
Mingming Chen,
Zhengzhou University, China

****Correspondence:***
Jia Liu
liujiaTHU@tsinghua.edu.cn
Shan Xu
shan.xu@bnu.edu.cn

†These authors have contributed equally to this work

Received: *05 November 2020*
Accepted: *14 April 2021*
Published: *20 May 2021*

Citation:
Xu S, Zhang Y, Zhen Z and Liu J (2021) The Face Module Emerged in a Deep Convolutional Neural Network Selectively Deprived of Face Experience.
Front. Comput. Neurosci. 15:626259.
doi: 10.3389/fncom.2021.626259

INTRODUCTION

A fundamental question in cognitive neuroscience is how nature and nurture form our cognitive modules. In the center of the debate is the origin of face recognition ability. Numerous studies have revealed both behavioral and neural signatures of face-specific processing, indicating a face module in the brain (for reviews, see Kanwisher and Yovel, 2006; Freiwald et al., 2016). Further studies from behavioral genetics revealed the contribution of genetics on the development of the face-specific recognition ability in humans (Wilmer et al., 2010; Zhu et al., 2010). Collectively, these studies suggest an innate domain-specific module for face cognition. However, it is unclear whether the visual experience is also necessary for the development of the face module.

A direct approach to address this question is visual deprivation. Two studies on monkeys selectively deprived the visual experience of faces since birth, while leaving the rest of experiences untouched (Sugita, 2008; Arcaro et al., 2017). They report that face-deprived monkeys are still capable of categorizing and discriminating faces (Sugita, 2008), though less prominent in selective looking preference to faces over non-face objects (Arcaro et al., 2017). Further examination of the brain of the experience-deprived monkeys fails to localize typical face-selective cortical regions with the standard criterion; however, in the inferior temporal cortex where face-selective regions are normally localized, weak and variable face-selective activation (i.e., neural responses to faces larger than non-face objects) is observed (Arcaro et al., 2017). Taken together, without visual experiences of faces, rudimental functions to process faces may still evolve to some extent.

Two related but independent hypotheses may explain the emergence of the face module without face experiences. An intuitive answer is that the rudimental functions are hardwired in the brain by genetic predisposition (Wilmer et al., 2010; McKone et al., 2012). Alternatively, we argue that the face module may emerge from experiences on non-face objects and related general-purpose processes, because representations for faces may be constructed by abundant features derived from non-face objects. Unfortunately, studies on humans and monkeys are unable to thoroughly decouple the effect of nature and nurture to test these two hypotheses.

Recent advances in deep convolutional neural network (DCNN) provide an ideal test platform to examine the impact of visual experiences on face modules without genetic predisposition. DCNNs are found similar to human visual cortex both structurally and functionally (Kriegeskorte, 2015), but free of any predisposition on functional modules. Therefore, with DCNNs we can manipulate experiences without considering interactions from genetic predisposition. In this study, we asked whether DCNNs can achieve face-specific recognition ability when visual experiences on faces were selectively deprived.

To do this, we trained a representative DCNN, AlexNet (Krizhevsky et al., 2012), to categorize non-face objects with face images carefully removed from the training dataset. Once this face-deprived DCNN (d-AlexNet) was trained, we compared its behavioral performance to that of a normal AlexNet of the same architecture but with faces present during training. Specifically, we examined their performance in both face categorization (i.e., differentiating faces from non-face objects) and discrimination (i.e., discriminating faces among different individuals) tasks. We predicted that the d-AlexNet, though without predisposition and experiences of faces, may still develop face selectivity through its visual experiences of non-face objects.

MATERIALS AND METHODS

Stimuli

Deprivation Dataset

The deprivation dataset was constructed to train the d-AlexNet. It was based on the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) 2012 dataset (Deng et al., 2009), which contains 1,281,167 images for training and 50,000 images for validation, in 1,000 categories. These images were first subjected to automated screening with an in-house face-detection toolbox based on VGG-Face (Parkhi et al., 2015), and then further screened by two human raters, who separately judged whether a given image contains faces of humans or non-human primates regardless of the orientation and intactness of the face, or anthropopathic artwork, cartoons, and artifacts. We removed images judged by either rater as containing any above-mentioned contents. Finally, we removed categories whose remaining images were <640 images (approximately half of the original number of images in a category). The resultant dataset consists of 736 categories, with 662,619 images for training and 33,897 for testing the performance.

Classification Dataset

To train a classifier that can classify faces, we constructed a classification dataset consisting of 204 categories of non-face objects and one face category, each of 80 exemplars. For the non-face categories, we manually screened Caltech-256 (Griffin et al., 2007) to remove images containing human, primate, or cartoon faces, and then removed categories whose remaining images were <80. In each of the 204 remaining non-face categories, we randomly chose 70 images for training and another 10 for calculating classification accuracy. The face category was constructed by randomly selecting 1,000 faces images from Faces in the Wild (FITW) dataset (Berg et al., 2005). Among them, 70 were used as training data and another 10 for classification accuracy. In addition, to characterize DCNN's ability in differentiating faces from object categories, we compiled a second dataset consisting of all images in the face category except those used in training.

Discrimination Dataset

To train a classifier that can discriminate faces at individual level, we constructed a discrimination dataset consisting of face images of 133 individuals, 300 images each, selected from the Casia-WebFace database (Yi et al., 2014). For each individual in the dataset, 250 were randomly chosen for training and another 50 for calculating discrimination accuracy.

Representation Dataset

To examine representational similarity of faces and non-face images between the d-AlexNet and the normal one, we constructed a representation dataset with two categories, faces and bowling pins as an "unseen" non-face object category that was not presented to the DCNNs during training. Each category consisted of 80 images. The face images were a random subset of FITW, and images of bowling pins were randomly chosen from the corresponding category in Caltech-256.

Movies Clips for DCNN-Brain Correspondence Analysis

We examined the correspondence between the face-selective response of the DCNNs and brain activity using a set of 18 clips of 8-min natural color videos from the Internet that are diverse yet representative of real-life visual experiences (Wen et al., 2017).

The Deep Convolutional Neural Network

Our model of selective deprivation, the d-AlexNet, was built with the architecture of the well-known DCNN "AlexNet" (Krizhevsky et al., 2012, see **Figure 1A** for illustration). AlexNet is a feed-forward hierarchical convolutional neural network consisting of five convolutional layers (denoted as Conv1–Conv5, respectively) and three fully connected layers denoted as FC1–FC3. Each convolutional layer consists of a convolutional sublayer, followed by a ReLU sublayer, and Conv1, 2, and 5 are further followed by a pooling sublayer. Each convolutional sublayer consists of a set of distinct channels. Each channel convolves the input with a distinct linear filter (kernel) which extracts filtered outputs from all locations within the input with a particular stride size. FC1–FC3 are fully connected

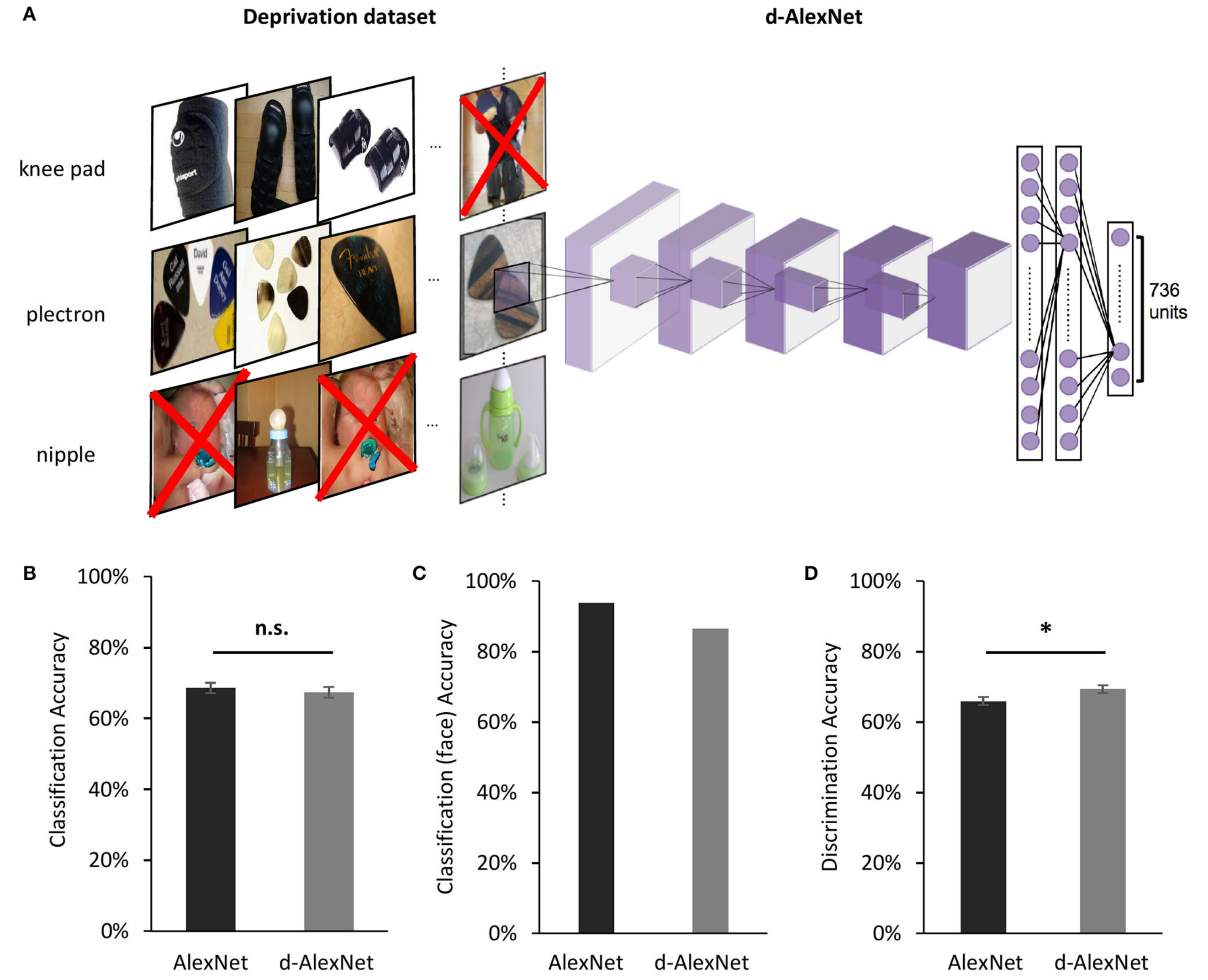


FIGURE 1 | (A) An illustration of the screening to remove images containing faces for the d-AlexNet. The "faces" shown in the figure were AI-generated for illustration purpose only, and therefore have no relation to real person. In the experiment, face images were from the ImageNet, with real persons' faces. **(B)** The classification performance across categories of the two DCNNs was comparable. **(C)** Both DCNNs achieved high accuracy in categorizing faces from other images. **(D)** Both DCNNs' performance in discriminating faces was above the chance level, and the d-AlexNet's accuracy was significantly higher than that of the AlexNet. The error bars in **(B)** denote the standard error of the mean across the 205 categories in the Classification dataset. The error bars in **(D)** denote the standard error of the mean across the 133 identities in the Discrimination dataset. The asterisk denotes statistical significance ($\alpha = 0.05$). n.s. denotes no significance.

layers. FC3 is followed by a sublayer using a softmax function to output a vector that represents the probability of the visual input containing the corresponding object category (Krizhevsky et al., 2012).

The d-AlexNet used the architecture of AlexNet but changed the number of units in FC3 to 736 and changed the following softmax function accordingly to match the number of categories in the deprivation dataset. The d-AlexNet was initialized with values drawn from a uniform distribution, and was then trained on the deprivation dataset following the approach specified in Krizhevsky (2014). We used the pre-trained AlexNet from pytorch 1.2.0 as the normal DCNN, referred to as the AlexNet in this paper for brevity.

The present study referred to channels in the convolutional sublayers by the layer they belong to and a channel index, following the convention of pytorch 1.2.0. For instance, Layer 5-Ch256 refers to the 256th convolutional channel of Layer 5.

To test the generalizability of the main findings of the present study, we also applied the same deprivation on another well-known DCNN, "ResNet-18" (He et al., 2016). ResNet-18 introduces residual learning blocks in a DCNN to overcome the degradation problem in the training of DCNNs, and achieves even better performance than AlexNet in object categorization task with a deeper architecture. The d-ResNet used the architecture of ResNet-18 but changed the number of units in the FC layer to 736 and changed the following softmax function accordingly to match the number of categories in the deprivation dataset. The d-ResNet was trained on the deprivation dataset following the same approach specified above. For comparison, we used the pre-trained ResNet-18 from pytorch 1.2.0 as the normal DCNN, referred to as the ResNet in this study for brevity.

Transfer Learning for Classification and Discrimination

To examine to what extent our manipulation of the visual experience affected the categorical processing of faces, we replaced the fully-connected layers of each DCNN with a two-layer face-classification classifier. The first layer was a fully connected layer with 43,264 units as inputs and 4,096 units as outputs with sigmoid activation function, and the second was a fully connected layer with 4,096 units as inputs and 205 units as outputs, each of which corresponded to one category of the classification dataset. This classifier, therefore, classified each image into one category of the classification dataset. The face-classification classifier was trained for each DCNN with the training images in the classification dataset for 90 epochs.

To examine to what extent our manipulation of the visual experience affected face discrimination, we similarly replaced the fully connected layers of each DCNN with a discrimination classifier. The discrimination classifier differed from the classification classifier only in its second layer, which had 133 units instead as outputs, each corresponding to one individual in the discrimination dataset. The face-discrimination classifier was trained for each DCNN with the training images in the discrimination dataset for 90 epochs. The same transfer learning was applied to the d-ResNet and the pre-trained ResNet-18.

The Face Selective Channels in DCNNs

To identify the channels selectively responsive to faces, we submitted images in the classification dataset to each DCNN, recorded the average activation in each channel of Conv5 after ReLU in response to each image, and then averaged the channel-wise activation within each category. We selected channels where the face category evoked the highest activation, and used the Mann-Whitney U test to examine the activation difference between faces and objects that had the second-highest activation in these channels ($p < 0.05$, Bonferroni corrected). The selectivity of each face channel thus identified was indexed by the selective ratio. The selective ratio was calculated by dividing the face activation by the second-highest activation. In addition, we measured the lifetime sparseness of each face-selective channel as an index for selectivity of faces among all non-face objects. We first normalized the mean activations of a face channel in Layer5 to all the categories to the range of 0–1, and then calculated lifetime sparseness with the formula:

$$S = \frac{\left(\sum_{i=1,n} r_i/n\right)^2}{\sum_{i=1,n} \left(r_i{}^2/n\right)}$$

where r_i is the normalized activations to the ith object category. The smaller this value is, the higher the selectivity is.

To confirm the face selectivity of the selected channels, we also tested their categorical selectivity with the fMRI localizer stimuli typically used to identify face-selective regions. More specifically, we recorded each channels' responses to the localizer stimuli from the face and the tool condition of the Human Connectome Project dataset (Van Essen et al., 2013), and examined the significance of face selectivity of each face channel by comparing the activation in the face condition and that of the tool condition in this channel using the Mann-Whitney U test described above.

Since we found face-selective channels in the d-AlexNet and reduced face selectivity of these channels comparing with face-selective channels in the AlexNet, we proceeded to test the robustness of these findings. Another five instances of face-deprived AlexNet were each independently trained in the same way as the d-AlexNet. In these instances, we searched for face-selective channels, computed their face selectivity, and examined the significance of their face selectivity by the Mann-Whitney U test on their responses to the classification dataset as well as on the fMRI localizer stimuli, in the same way as we did in the d-AlexNet and the AlexNet. The same procedure of channel identification was also applied to the d-ResNet and the pre-trained ResNet-18.

DCNN-Brain Correspondence

We submitted the movie clips to the DCNNs. Following Wen et al. (2017)'s approach, we extracted and log-transformed the channel-wise output (the average activation after ReLU) of each face-selective channel using the toolbox DNNBrain (Chen et al., 2020), and then convolved it with a canonical hemodynamic response function (HRF) with a positive peak at 4 s. The HRF convolved channel-wise activity was then down-sampled to match the sampling rate of functional magnetic resonance imaging (fMRI) and the resultant timeseries was standardized before further analysis.

Neural activation in the brain was derived from the preprocessed data in Wen et al. (2017). The fMRI data were recorded while human participants viewed each movie clips twice. We averaged the standardized time series across repetition and across subjects for each clip. Then, for each DCNN, we conducted multiple regression for each clip, with the activation time series of each brain vertex as the dependent variable and that of face-selective channels in this network as independent variables. For the d-AlexNet, all face-selective channels were included. For the AlexNet, we included the same number of face-selective channels with the highest face selectivity to match the complexity of the regression model. We used the R^2 of each vertex as the index of the overall Goodness of fit of the regression in that vertex. The R^2 values were then averaged across clips. The larger the R^2 value, the higher correspondence between the DCNN and the brain in response to movie clips.

To test whether the correspondence changes between networks reflected an overall increase in the correspondence between fMRI signal and the activation of the face channels of the AlexNet comparing with the d-AlexNet (in contrast to an increase selectively within the face-selective regions), we delineated the face-selective regions and the object-selective regions and compared the correspondence between the top two face channels of each network and the face- and the object-selective regions. The face- and the object- selective regions were defined by functional localizer data of Human Connectome Project (Van Essen et al., 2013). Two hundred vertexes of the highest Z value in the tool-avg contrast were delineated as the object-selective ROIs, and two hundred vertexes of the highest Z value in the face-tool contrast were delineated as the face-selective ROIs. The channel-brain correspondence of each vertex with the ROIs was indexed by R^2 of the regression with the fMRI time series of this vertex as the dependent variable and the time series of the top-two face channels as the independent variables. A two-way ANOVA with visual experiences (d-AlexNet vs. AlexNet) and categorical selectivity (the object-selective regions vs. the face-selective regions) as independent variables was conducted to examine the difference between the channel-brain correspondence between the categorical-selective regions and the face-selective channels of the d-AlexNet and the AlexNet.

To examine whether the channel-brain correspondence changed in different face-selective regions equally, we delineated the bilateral fusiform face areas (FFA) and the occipital face area (OFA) with the maximum-probability atlas of face-selective regions (Zhen et al., 2015). Two hundred of vertexes of the highest probability of the left FFA and 200 of the right FFA were included in the ROI of FFA, and the ROI of OFA was delineated in the same way. The correspondence with brain activation in each ROI and the impact of the visual experience was examined by submitting the vertex-wise R^2 into a two-way ANOVA with visual experience (d-AlexNet vs. AlexNet) as within-subject factor and regional correspondence (OFA and FFA) as between-subject factor.

Face Inversion Effect in DCNNs

The average activation amplitude of the top two face-selective channels of each DCNN in response to upright and inverted version of 20 faces from the Reconstructing Faces dataset (VanRullen and Reddy, 2019) was measured. The inverted faces were generated by vertically flipping the upright ones. The face inversion effect in the d-AlexNet was measured with paired sample *t*-tests (two-tailed) and the impact of the experience on the face inversion effect was examined by two-way ANOVAs with visual experience (d-AlexNet vs. AlexNet) and inversion (upright vs. inverted) as within-subject factors.

Representational Similarity Analysis

To examine whether faces in the d-AlexNet were processed in an object-like fashion, we compared the within-category representational similarity of faces to that of bowling pins, an "unseen" non-face object category never exposed to either DCNN. Specifically, for each image in the representation dataset, we arranged the average activations of each channel of Conv5 after ReLU into vectors, and then for each pair of images we calculated and then Fisher-z transformed the correlation between their vectors, which served as an index of pairwise representational similarity. Within-category similarity between pairs of face images and that between pairs of object images were calculated separately. A 2 × 2 ANOVA was conducted with visual experience (d-AlexNet vs. AlexNet) and category (face vs. object) as independent factors. In addition, cross-category similarity between faces and bowling pins was also calculated for each DCNN, and a paired sample *t*-test (two-tailed) on two DCNNs was conducted.

Sparse Coding and Empirical Receptive Field

To quantify the degree of sparseness of the face-selective channels in representing faces, we submitted the same set of 20 natural images containing faces from FITW to each DCNN, and measured the number of activated units (i.e., the units showing above-zero activation) in the face-selective channels. The more non-zero units observed in the face-selective channels, the less sparse the representation for faces is. The coding sparseness of the two DCNNs was compared with a paired-sample *t*-test.

We also calculated the size of the empirical receptive field of the face-selective channels. Specifically, we obtained the activation maps of 1,000 images randomly chosen from FITW. Using the toolbox DNNBrain (Chen et al., 2020), we up-sampled each activation map to the same size of the input. For each image, we averaged the up-sampled activation within the theoretical receptive field of each unit (the part of the image covered by the convolution of this unit and the preceding computation, decided by the network architecture), and selected the unit with the highest average activation. We then cropped the up-sampled activation map by the theoretical receptive field of this unit, to locate the image part that activated this channel most across all the units. Then, we averaged corresponding cropped activation maps across all the face images, and the resultant map denotes the empirical receptive field of this channel, delineating the part of the theoretical receptive field that causes this channel to respond strongly in viewing its preferred stimuli.

RESULTS

The d-AlexNet was trained with a dataset of 662,619 non-face images consisting of 736 non-face categories, generated by removing images containing faces from the ILSVRC 2012 dataset (**Figure 1A**). The d-AlexNet was initialized and trained in the same way as the AlexNet. Both networks were trained following the approach specified in Krizhevsky (2014). The resultant top-1 accuracy (57.29%) and the top-5 accuracy (80.11%) were comparable with the pre-trained AlexNet.

We first examined the performance of the d-AlexNet in two representative tasks of face processing, face categorization (i.e., differentiating faces from non-face objects) and face discrimination (i.e., identifying different individuals). The output of Conv5 after ReLU of the d-AlexNet was used to classify objects in the classification dataset (see Materials and Methods). The averaged categorization accuracy of the d-AlexNet (67.40%) was well above the chance level (0.49%), and comparable to that in the AlexNet [68.60%, $t_{(204)} = 1.26$, $p = 0.209$, Cohen's $d = 0.007$, **Figure 1B**]. Critically, the d-AlexNet, although with no experience on faces, succeeded in the face categorization task, with an accuracy of 86.50% in categorizing faces from non-face objects. Note that the accuracy was numerically smaller than the AlexNet's accuracy in categorizing faces (93.90%) though (**Figure 1C**).

A similar pattern was observed in the face discrimination task. In this task, the output of Conv5 after ReLU of each DCNN was used to identify 33,250 face images into 133 identities in the discrimination dataset (see section Materials and Methods). As expected, the AlexNet was capable of face discrimination (65.9%), well above the chance level (0.75%), consistent with previous studies (AbdAlmageed et al., 2016; Grundstrom et al., 2016). Critically, the d-AlexNet also showed the capability of discriminating faces, with an accuracy of 69.30% that was even significantly higher than that of the AlexNet, $t_{(132)} = 3.16$, $p = 0.002$, Cohen's $d = 0.20$, (**Figure 1D**). Taken together, visual experiences on faces seemed not necessary for developing basic functions of processing faces.

Was a face module formed in the d-AlexNet to support these functions? To answer this question, we searched all the channels in Conv5 of the d-AlexNet, where face-selective channels have been previously identified in the AlexNet (Baek et al., 2019). To do this, we calculated the activation of each channel in Conv5 after ReLU in response to each category of the classification dataset, and then identified channels that showed significantly higher response to faces than non-face images with Mann-Whitney U test (ps < 0.05, Bonferroni corrected). Two face-selective channels (Ch29 and Ch50) met this criterion in the d-AlexNet (for an example channel, see **Figure 2A**, right), whereas four face-selective channels (Ch185, Ch125, Ch60, and Ch187) were identified in the AlexNet (for an example channel, see **Figure 2A**, left). The face-selective channels in two DCNNs differed in selectivity. The averaged selective ratio, the ratio of the activation magnitude to faces by that to the most activated non-face object category, was 1.29 (range: 1.22–1.36) in the d-AlexNet, much lower than that in the AlexNet (average ratio: 3.63, range: 1.43–6.66). The lifetime sparseness, which measures the breadth of tuning of a channel in response to a set of categories, also showed a similar result. The average lifetime sparseness index of the face channels in the AlexNet (mean = 0.25, range: 0.11–0.51) was smaller than that in the d-AlexNet (mean = 0.71, range: 0.70–0.71), indicating higher face selectivity in the AlexNet than that in the d-AlexNet. To confirm that the emergence of the face-selective channels in the d-AlexNet was not because of chance factors in network training, another five instances of face-deprived networks were independently initiated and trained respectively. One or two face-selective channels emerged in each of these face-deprived network instances, though the level of face selectivity was lower as compared to the AlexNet. In addition, we tested the face selectivity of the face channels in all face-deprived networks with the stimuli used to localize face-selective regions in fMRI studies, and found that the responses in these face-selective channels were significantly higher to the faces than those to the objects (Mann-Whitney U test, ps < 0.05, Bonferroni corrected). Taken together, this finding suggested that the face-selective channels indeed emerged in the d-AlexNet, though the face selectivity was weaker than the AlexNet.

To test the generalizability of these findings, we applied the same deprivation manipulation to another representative DCNN architecture, the ResNet-18, and the resultant d-ResNet reached top-1 accuracy (69.57%) and the top-5 accuracy (89.47%), comparable with those of the ResNet. Further, the face categorization accuracy of the d-ResNet (92.90%) was comparable to that of the ResNet (96.02%), and the discrimination accuracy of d-ResNet (65.34%) comparable to that of the pre-trained ResNet (59.80%). These findings were similar to those achieved with the d-AlexNet and the AlexNet.

How did the face-selective channels correspond to face-selective cortical regions in humans, such as the FFA and OFA? To answer this question, we calculated the coefficient of determination (R^2) of the multiple regression with the output of the face-selective channels as regressors and the fMRI signals from human visual cortex in response to movies on natural vision as the regressand (see section Materials and Methods). As shown in **Figure 2B** (right), the face-selective channels identified in the d-AlexNet corresponded to the bilateral FFA, OFA, and the posterior superior temporal sulcus face area (pSTS-FA). Similar correspondence was also found with the top two face-selective channels in the AlexNet (**Figure 2B**, left). Direct visual inspection revealed that the deprivation weakened the correspondence between the face-selective channels and face-selective regions in human brain. The increased channel-brain correspondence in the face-selective regions in the AlexNet compared with the d-AlexNet was confirmed by a two-way ANOVA of visual experience (d-AlexNet vs. AlexNet) by categorical selectivity (fMRI defined object-selective vs. face-selective regions, see section Methods). In addition to a main effect of categorical selectivity [$F_{(1, 398)} = 53.04$, $p < 0.001$, partial $\eta^2 = 0.12$], we also observed a two-way interaction [$F_{(1, 398)} = 79.99$, $p < 0.001$, partial $\eta^2 = 0.17$]. Follow-up simple effect analyses revealed that the correspondence to the face-selective regions decreased in the d-AlexNet as compared with the AlexNet in the face-selective regions (MD = −0.01, $p < 0.001$), but increased in the object-selective regions (MD = 0.013, $p < 0.001$), further indicating that the changes between the face-selective channels and human face-selective regions cannot be attributed to a global decrease in the channel-brain correspondence in the d-AlexNet comparing with the AlexNet.

We then examined whether this decrease in channel-brain correspondence affected different face-selective regions equally. A two-way ANOVA of visual experience (d-AlexNet vs. AlexNet) by regional correspondence (the OFA vs. the FFA) confirmed the decrease of channel-brain correspondence in the d-AlexNet compared with the AlexNet with a significant main effect of visual experiences [$F_{(1, 798)} = 161.97$, $p < 0.001$, partial $\eta^2 = 0.17$]. In addition, the main effect of the regional correspondence showed that the response profile of the face-selective channels in the DCNNs fitted better with the activation of the FFA than that of the OFA [$F_{(1, 798)} = 98.69$, $p = 0.001$, partial $\eta^2 = 0.11$], suggesting that the face-selective channels in DCNNs may in general tend to process faces as a whole than face parts. Critically, the two-way interaction was significant [$F_{(1, 798)} = 84.9$, $p < 0.001$, partial $\eta^2 = 0.10$], indicating that the experience affected the correspondence to the FFA and OFA disproportionally. A

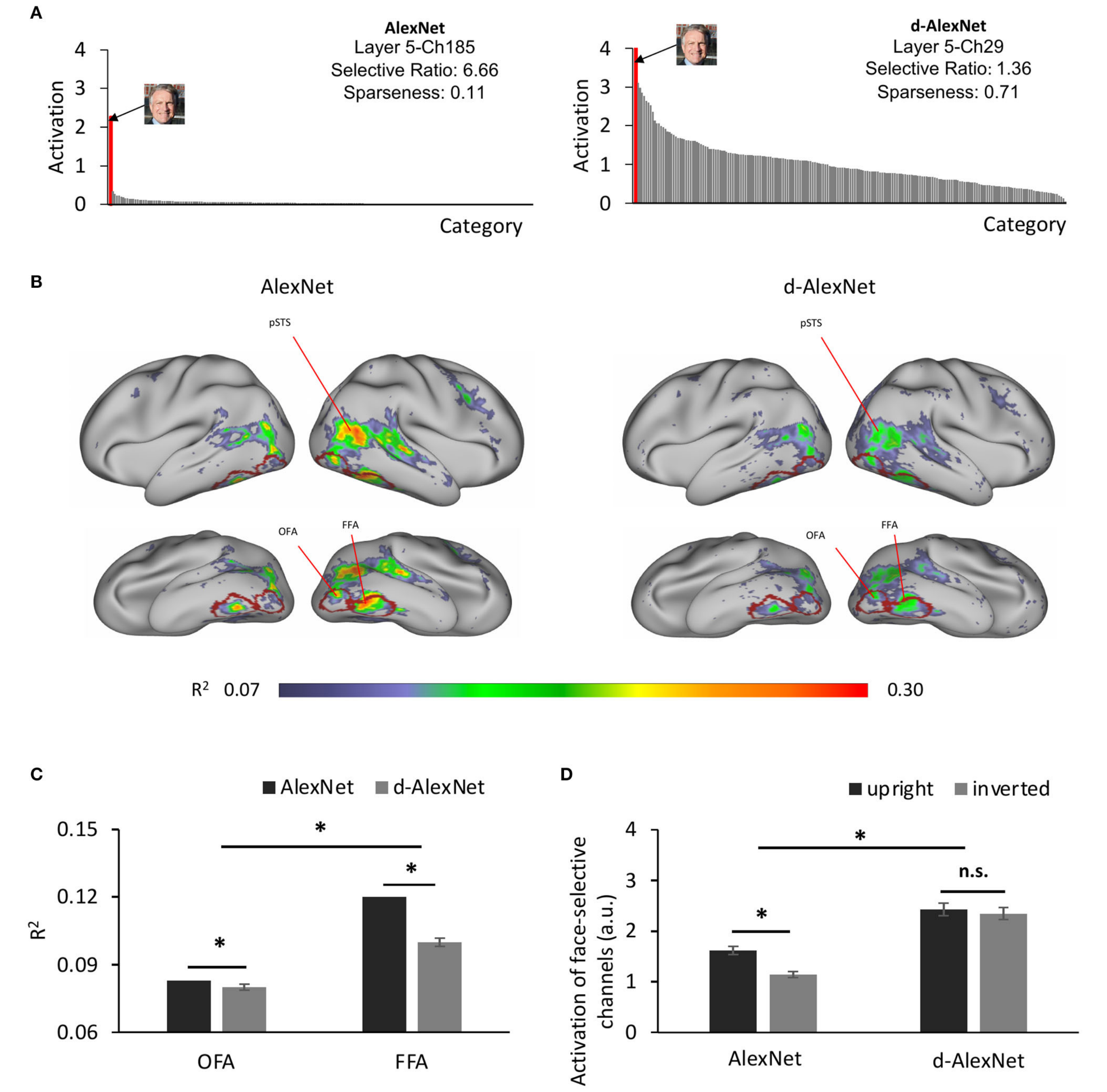


FIGURE 2 | (A) The category-wise activation profiles of example face-selective channels of the AlexNet (left) and the d-AlexNet (right). The "faces" shown here were AI-generated for illustration purpose only. **(B)** The R^2 maps of the regression with the activation of the d-AlexNet's (right) or the AlexNet's face-selective channels (left) as the independent variables. The higher R^2 in multiple regression, the better correspondence between the face channels in the DCNNs and the face-selective regions in the human brain. The crimson lines delineate the ROIs of the OFA and the FFA. **(C)** The face-channels of both DCNNs corresponded better with the FFA than the OFA, and the difference between the AlexNet and the d-AlexNet was larger in the FFA. **(D)** Face inversion effect. The average activation amplitude of the top two face-selective channels differed in response to upright and inverted faces in the AlexNet but not the d-AlexNet. The error bar denotes standard error. The asterisk denotes statistical significance ($\alpha = 0.05$). n.s. denotes no significance.

simple effect analysis revealed that the correspondence to the FFA (MD = 0.023, $p < 0.001$) was increased by face-specific experiences to a significantly larger extent than that to the OFA (MD = 0.004, $p = 0.013$, **Figure 2C**). Since the FFA is more involved in holistic processing of faces and the OFA is more dedicated to the part-based analysis, the disproportional decrease in correspondence between the face-selective channels in the d-AlexNet and the FFA implied that the role of the experience on faces was to facilitate the processing of faces as a whole.

To test this conjecture, we examined whether the d-AlexNet responded stronger to upright than inverted faces, since human studies suggested that the upright faces were processed in a more holistic manner than inverted faces. As expected, there was a face inversion effect in the AlexNet's face-selective channels, with the magnitude of the activation to upright faces significantly larger than that to inverted faces [$t_{(19)} = 6.45$, $p < 0.001$, Cohen's $d = 1.44$] (**Figure 2D**). However, no inversion effect was observed in the d-AlexNet, as the magnitude of the activation to upright faces was not significantly larger than that to inverted faces [$t_{(19)} = 0.86$, $p = 0.40$]. The lack of the inversion effect in the d-AlexNet was further supported by a two-way interaction of visual experience by orientation of faces, [$F_{(1,\ 19)} = 7.79$, $p = 0.012$, partial $\eta^2 = 0.29$]. That is, unlike the AlexNet, the d-AlexNet processed upright faces in the same fashion as inverted faces.

Previous studies on human suggested that inverted faces are processed in an object-like fashion. That is, it relies more on the parts-based analysis than the holistic processing. Therefore, we speculated that in the d-AlexNet faces were also represented more like non-face objects. To test this speculation, we first compared the representational similarity among responses in Conv5 to faces and bowling-pins, which were not present as a category in the training dataset of either DCNNs, and therefore alien to both DCNNs. As expected, the two-way interaction of experience (AlexNet vs. d-AlexNet) by category (faces vs. bowling-pins) was significant [$F_{(1,\ 6{,}318)} = 4{,}110.88$, $p < 0.001$, partial $\eta^2 = 0.39$], and the simple effect analysis suggested that the representation for faces in the AlexNet was more similar between each other than in the d-AlexNet (MD = 0.16, $p < 0.001$), whereas the within-category representation similarity for bowling-pins showed the same but numerically smaller between-DCNN difference (MD = 0.005, $p = 0.002$) (**Figure 3A**).

A more critical test was to examine how face-specific experiences made faces being processed differently from objects. Here we calculated between-category similarities between faces and bowling-pins. We found that the between-category similarity between faces and bowling-pins was significantly higher in the d-AlexNet than that in the AlexNet [$t_{(3{,}159)} = 42.42$, MD = 0.07, $p < 0.001$, Cohen's $d = 0.76$] (**Figure 3B**), suggesting that faces in the d-AlexNet were indeed represented more like objects. In short, although d-AlexNet was able to perform face tasks similar to the one with face-specific experiences, it represented faces in an object-like fashion.

Finally, we asked how faceness was achieved in DCNNs with face-specific experiences. Neurophysiological studies on monkeys demonstrate experience-associated sharpening of neural response, with fewer neurons activated after learning. Here we performed a similar analysis by measuring the number of non-zero units (i.e., units with above-zero activation) of the face-selective channels activated by natural images containing faces. As shown in the activation map (**Figure 3C**), a smaller number of units were activated by faces in the AlexNet than that in the d-AlexNet [$t_{(19)} = 3.317$, MD = 15.78, Cohen's $d = 0.74$] (**Figure 3D**), suggesting that the experience on faces made the representation to faces sparser, and thus allowing for more efficient coding. Another effect of visual experiences observed in neurophysiological studies is that experiences reduce the size of neurons' receptive field. Here we also mapped the empirical receptive field of the face-selective channels (see section Materials and Methods). Similarly, we found that the empirical receptive field of the AlexNet was smaller than that of the d-AlexNet. That is, within the theoretical receptive field, the empirical receptive field of the face-selective channels in the AlexNet was tuned to focus on a smaller region by face-specific experiences (**Figure 3E**).

DISCUSSION

This study presented a DCNN model of selective visual deprivation of faces. Specifically, we chose the AlexNet as a test platform because of the functional correspondence along the hierarchy between the AlexNet and primates' ventral visual pathway (e.g., Krizhevsky et al., 2012; Cadieu et al., 2014; Wen et al., 2017; Pospisil et al., 2018; Baek et al., 2019). We found that without genetic predisposition and face-specific visual experiences, DCNNs were still capable of face perception. In addition, face-selective channels were also present in the d-AlexNet, which corresponded to human face-selective regions. That is, the visual experience of faces was not necessary for an intelligent system to develop a face-selective module. On the other hand, besides the slightly compromised selectivity of the module, the deprivation led the d-AlexNet to process faces in a fashion more similar to that of processing objects. Indeed, unlike the AlexNet, face inversion did not affect the response magnitude of the face-selective channels in the d-AlexNet, and the representation of faces was more similar to objects as compared to the AlexNet. Finally, face-specific experiences might affect face processing by fine-tuning the sparse coding and the size of the receptive field of the face-selective channels. In sum, our study addressed a long-standing debate on nature vs. nurture in developing the face-specific module, and illuminated the role of visual experiences in shaping the module.

Given the main-stream viewpoint that faces are special and therefore cannot be compensated by the presence of non-face objects, it may seem surprising that without domain-specific visual experience, the face-selective processing and modules still emerged in the d-AlexNet. These observations were further replicated with another well-known DCNN architecture, the ResNet-18, suggesting the generalizability of our findings. However, our finding is consistent with previous studies on non-human primates and new-born human infants (Bushneil et al., 1989; Valenza et al., 1996; Sugita, 2008), where the face-specific experience is found not necessary for face detection and recognition. Therefore, our study argues against the experience-independent hypothesis that face specificity is largely attributed to either innate face-specific mechanisms (Morton and Johnson, 1991; McKone et al., 2012) or domain-general processing with predisposed biases (Simion et al., 2001; Simion and Di Giorgio, 2015). Our study argues against this conjecture, because unlike

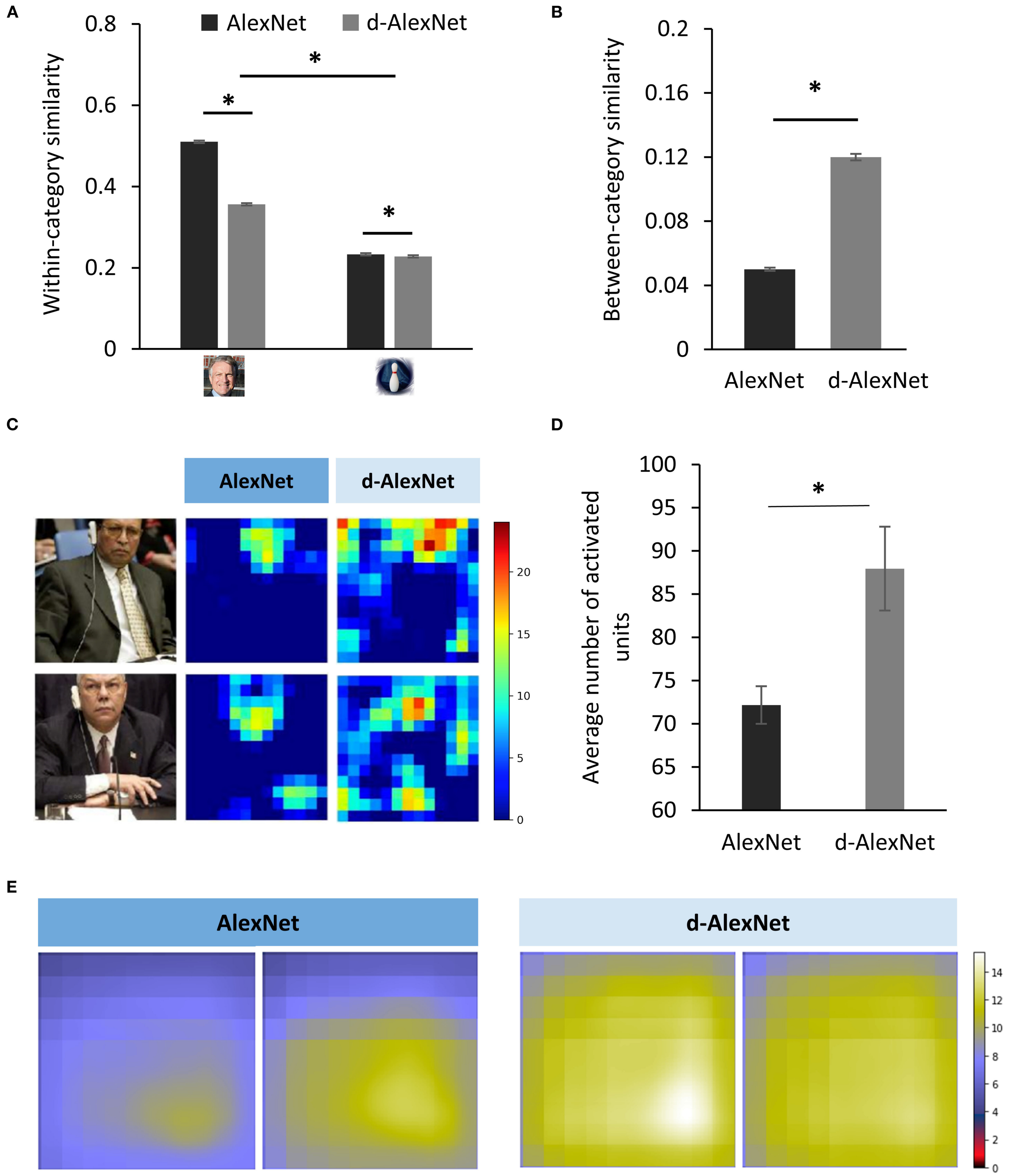


FIGURE 3 | (A) The within-category similarity in the face category and an unseen non-face category (bowling pins) in the DCNNs. **(B)** The between-category similarity between faces and bowling pins. **(C)** The activation maps of a typical face-selective channel of each DCNN in responses to natural images containing faces. Each pixel denotes activation in one unit. The color denotes the activation amplitude (a.u.). **(D)** The extent of activation of the face-selective channels of each DCNN in responses to natural images containing faces. **(E)** The empirical receptive fields of the top two face-selective channels of each DCNN. The color denotes the average activation amplitude (a.u, see section Sparse Coding and Empirical Receptive Field). The error bar denotes standard error. The asterisk denotes statistical significance ($\alpha = 0.05$). The real faces used in this figure are adapted from the FITW dataset.

any biological system, DCNNs have no domain-specific genetic inheritance or processing biases. Therefore, the face-specific processing observed in DCNNs had to derive from domain-general factors. From this sense, the present study provides one of the first direct evidence against the main-stream viewpoint and suggests that face specificity may emerge from domain-general visual experience.

We speculated that the face-selective processing and module in the d-AlexNet may result from the rich features represented in the multiple layers of the network; face-like features might be utilized when the neural network was forced to categorize faces even though these features were not learned for this purpose. In fact, previous studies on DCNNs have shown that DCNN's lower layers showed sensitivity to myriad visual features similar to primates' primary visual cortex (Krizhevsky et al., 2012), while the higher layers are tuned to complex features resembling those represented in the ventral visual pathway (Yamins et al., 2014; Güçlü and van Gerven, 2015). With such a repertoire of rich features, a representational space for faces, or any natural object, may be constructed by selecting features that are potentially useful in face tasks. With such repertoire of rich features, a representational space for faces, or for any natural object, may be constructed by selecting features that are potentially useful in face tasks.

Supporting evidence for this conjecture came from the observation that the d-AlexNet processed faces in an object-like fashion. For example, the face inversion effect, a signature of face-specific processing in human (Yin, 1969; Kanwisher et al., 1998) was absent in the d-AlexNet. Distinct from other non-face stimuli, faces are recognized better when they are upright than inverted (Yin, 1969), and the neural response to upright faces is stronger than that to inverted ones (e.g., Kanwisher et al., 1998; Rossion and Gauthier, 2002). This face inversion effect is attributed to that face processing relies particularly heavily on configural processing—processing of the relations among features instead of individual features. Since the configural information is difficult to perceive in inverted faces in a system with face specificity, inverted faces cannot engage face-specific processing as upright faces. Therefore, the finding of the lack of the face inversion effect in the DNN without face experience strengthened our argument that the lack of face experience leads to the compromise of face specificity. That is, similar to inverted faces, upright faces may also be processed like objects in the d-AlexNet. A more direct illustration of the object-like representation of faces came from the analysis of the representational similarity between faces and objects. As compared to the AlexNet, faces in the representational space of the d-AlexNet were less congregated among each other; instead they were more intermingled with non-face object categories. The finding that face representation was no longer qualitatively different from object representation may help to explain the performance of the d-AlexNet. Because faces were less segregated from objects in the representational space, the d-AlexNet's accuracy of face categorization was worse than that of the AlexNet. In contrast, within the face category, individual faces were less congregated in the representational space; therefore, the discrimination of individual faces became easier instead, suggested by the slightly higher face discrimination accuracy in the d-AlexNet than the AlexNet. In short, when the representational space of the d-AlexNet was formed exclusively based on features from non-face stimuli, faces were represented no longer qualitatively different from non-face objects, which inevitably led to "object-like" face processing.

The face-specific processing is likely achieved through prior exposure to faces. At first glance, the effect of face-specific experiences seemed quantitative, as in the AlexNet, both the selectivity to faces and the number of the face-selective channels were increased, and the correspondence between the face-selective channels and the face-selective regions in human brain was tighter. However, careful scrutiny of the difference between the two DCNNs revealed that the changes led by the experience may be qualitative. For example, the deprivation of visual experiences disproportionally weakened the DCNN-brain correspondence in the FFA as comparing to the OFA, and the FFA is engaged more in the configural processing and the OFA in parts-based analysis (Liu et al., 2010; Nichols et al., 2010; Zhao et al., 2014). Therefore, the "face-like" face processing may come from the fact that face-specific experiences led the representation of faces more congregated within face category and more separable from the representation of non-face objects stimuli (see also Gomez et al., 2019). In this way, a relative encapsulated representation may help developing a unique way of processing faces, qualitatively different from non-face objects.

The computational transparency of DCNNs may shed light on the development of domain specificity of the face module. First, we found that face-specific experiences increased the sparseness of face representation, as fewer units of the face channels were activated by faces in the AlexNet. The experience-dependent sparse coding has been widely discovered in the visual cortex (for reviews, see Desimone, 1996; Grill-Spector et al., 2006). The experience-induced increase of sparseness is thought to reflect a preference-narrowing process that tunes neurons to a smaller range of stimuli (Kohn and Movshon, 2004); therefore, with sparse coding faces are less likely to be intermingled with non-face objects, which may lead to more congregated representations in the representational space in the AlexNet, as compared to the d-AlexNet. Second, we found that the empirical receptive field of the face channels in the AlexNet was smaller than that in the d-AlexNet, suggesting that the visual experience on faces decreased the size of the receptive field of the face channels. This finding fits perfectly with neurophysiological studies that the size of receptive fields of visual neurons is reduced after eye-opening (Braastad and Heggelund, 1985; Tavazoie and Reid, 2000; Cantrell et al., 2010). Importantly, along with the refined receptive fields, the selectivity of neurons increases (Spilmann, 2014), possibly because neurons can avoid distracting information by focusing on a more restricted part of stimuli, which may further allowed finer representation of the selected regions. This is especially important for processing faces because faces are highly homogeneous, and some information is identical across faces, such as parts composition (eyes, noses, and mouth) and their configural arrangements. Therefore, the reduced receptive field of the face channels may facilitate selective analyses of discriminative face features while avoiding irrelevant information. Further, the sharpening of the receptive field and the fine-tuned selectivity may result in superior discrimination ability on faces, and allow faces to be processed at the subordinate level (i.e., identification), whereas the rest of objects are largely processed at the basic level (i.e., categorization).

It has long been assumed that domain-specific visual experiences and inheritance are the pre-requisites in the development of the face module in the brain. In our study with DCNN as a model, we completely decoupled the genetic predisposition and face-specific visual experiences, and found that the representation for faces can be constructed with features from non-face objects to realize basic functions for face recognition. Therefore, in many situations, the difference between faces and objects is "quantitative" rather than "qualitative," as they are represented in a continuum of the representational space. In addition, we also found that face-specific experiences likely fine-tuned the face representation, and thus transformed the "object-like" face processing into "face-specific" processing. However, we shall be cautious that our finding may not be applicable for the development of face module in human, as in the biological brain experience-induced changes are partly attributed to the inhibition from lateral connections (Norman and O'Reilly, 2003; Grill-Spector et al., 2006), whereas there is no lateral or feedback connection in DCNNs. However, despite structural differences, recent studies have shown similar representation for faces between DCNNs and humans (Song et al., 2021), suggesting that a common mechanism may be shared by both artificial and biological intelligent systems. Future studies are needed to examine the applicability of our finding to humans. In addition, higher cognitive functions such as attractiveness judgement and social-traits inference are also important components of face processing, but the present study followed the literature on face deprivation in humans and non-human primates and therefore focused on the sensory and perceptual stages of face processing. Future study may consider investigating the experiential effects on the social and affective aspects of face processing to comprehensively understand the effect of experience.

On the other hand, our study illustrated the advantages of using DCNN as a model to understand human mind because of its computational transparency and its dissociation of factors in nature and nurture. Thus, our study invites future studies with DCNNs to understand the development of domain specificity in particular and a broad range of cognitive modules in general.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found at: https://github.com/sdgds/Deprivation_dataset; http://www.vision.caltech.edu/Image_Datasets/Caltech256/; http://vis-www.cs.umass.edu/lfw/; https://openneuro.org/datasets/ ds001761.

ETHICS STATEMENT

This study used human fMRI data, the acquisition of which was reviewed and approved by Institutional Review Board at Purdue University Institutional Review Board.

AUTHOR CONTRIBUTIONS

JL conceived and designed the study. YZ analyzed the data with input from all authors. SX wrote the manuscript with input from JL, YZ, and ZZ.

FUNDING

This study was funded by the National Natural Science Foundation of China (31861143039, 31771251, and 31600925), the Fundamental Research Funds for the Central Universities, and the National Basic Research Program of China (2018YFC0810602). Data were provided [in part] by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University.

REFERENCES

AbdAlmageed, W., Wu, Y., Rawls, S., Harel, S., Hassner, T., Masi, I., et al. (2016). "Face recognition using deep multi-pose representations." in *2016 IEEE Winter Conference on Applications of Computer Vision* (Lake Placid, NY).

Arcaro, M. J., Schade, P. F., Vincent, J. L., Ponce, C. R., and Livingstone, M. S. (2017). Seeing faces is necessary for face-domain formation. *Nat. Neurosci.* 20:1404. doi: 10.1038/nn.4635

Baek, S., Song, M., Jang, J., Kim, G., and Paik, S.-B. (2019). *Spontaneous generation of face recognition in untrained deep neural networks. bioRxiv, 857466*. Available online at: https://www.biorxiv.org/content/10.1101/857466v1 (accessed November 29, 2019).

Berg, T. L., Berg, A. C., Edwards, J., and Forsyth, D. A. (2005). Who's in the picture. *Adv. Neural Inf Process. Syst.* 17, 137–144. Retrieved from: https://papers.nips.cc/paper/2004/file/03fa2f7502f5f6b9169e67d17cbf51bb-Paper.pdf

Braastad, B. O., and Heggelund, P. (1985). Development of spatial receptive-field organization and orientation selectivity in kitten striate cortex. *J. Neurophysiol.* 53, 1158–1178. doi: 10.1152/jn.1985.53.5.1158

Bushneil, I., Sai, F., and Mullin, J. (1989). Neonatal recognition of the mother's face. *Br. J. Dev. Psychol.* 7, 3–15. doi: 10.1111/j.2044-835X.1989.tb00784.x

Cadieu, C. F., Hong, H., Yamins, D. L. K., Pinto, N., Ardila, D., Solomon, E. A., et al. (2014). Deep neural networks rival the representation of primate it cortex for core visual object recognition. *PLoS Comput. Biol.* 10:e1003963. doi: 10.1371/journal.pcbi.1003963

Cantrell, D. R., Cang, J., Troy, J. B., and Liu, X. (2010). Non-centered spike-triggered covariance analysis reveals neurotrophin-3 as a developmental regulator of receptive field properties of ON-OFF retinal ganglion cells. *PLoS Comput. Biol.* 6:e1000967. doi: 10.1371/journal.pcbi.1000967

Chen, X., Zhou, M., Gong, Z., Xu, W., Liu, X., Huang, T., et al. (2020). DNNBrain: a unifying toolbox for mapping deep neural networks and brains. *Front. Comput. Neurosci.* 14:580632. doi: 10.3389/fncom.2020.580632

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). "Imagenet: a large-scale hierarchical image database," in *Paper Presented at the 2009 IEEE Conference on Computer Vision and Pattern Recognition* (Miami, FL).

Desimone, R. (1996). Neural mechanisms for visual memory and their role in attention. *Proc. Natl. Acad. Sci. U.S.A.* 93, 13494–13499. doi: 10.1073/pnas.93.24.13494

Freiwald, W., Duchaine, B., and Yovel, G. (2016). Face processing systems: from neurons to real-world social perception. *Annu. Rev. Neurosci.* 39, 325–346. doi: 10.1146/annurev-neuro-070815-013934

Gomez, J., Barnett, M., and Grill-Spector, K. (2019). Extensive childhood experience with Pokemon suggests eccentricity drives organization of visual cortex. *Nat. Hum. Behav.* 3, 611–624. doi: 10.1038/s41562-019-0592-8

Griffin, G., Holub, A., and Perona, P. (2007). *Caltech-256 Object Category Dataset*. Pasadena, CA: California University of Technology.

Grill-Spector, K., Henson, R., and Martin, A. (2006). Repetition and the brain: neural models of stimulus-specific effects. *Trends Cogn. Sci.* 10, 14–23. doi: 10.1016/j.tics.2005.11.006

Grundstrom, J., Chen, J., Ljungqvist, M. G., and Astrom, K. (2016). "Transferring and compressing convolutional neural networks for face representations," in *Image Analysis and Recognition, Vol*. 9730, eds A. Campilho and F. Karray (Cham: Springer), 20–29.

Güçlü, U., and van Gerven, M. A. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *J. Neurosci.* 35, 10005–10014. doi: 10.1523/JNEUROSCI.5023-14.2015

He, K., Zhang, X., Ren, S., Sun, J., and Ieee. (2016). "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV), 770–778.

Kanwisher, N., Tong, F., and Nakayama, K. (1998). The effect of face inversion on the human fusiform face area. *Cognition* 68, B1–B11. doi: 10.1016/S0010-0277(98)00035-3

Kanwisher, N., and Yovel, G. (2006). The fusiform face area: a cortical region specialized for the perception of faces. *Philos. Trans. R. Soc. B Biol. Sci.* 361, 2109–2128. doi: 10.1098/rstb.2006.1934

Kohn, A., and Movshon, J. A. (2004). Adaptation changes the direction tuning of macaque MT neurons. *Nat. Neurosci.* 7, 764–772. doi: 10.1038/nn1267

Kriegeskorte, N. (2015). Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annu. Rev. Vis. Sci.* 1, 417–446. doi: 10.1146/annurev-vision-082114-035447

Krizhevsky, A. (2014). One *weird trick for parallelizing convolutional neural networks. arXiv preprint arXiv:1404.5997*. Available online at: https://arxiv.org/abs/1404.5997 (accessed April 29, 2014).

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf Process. Syst.* 25, 1097–1105. doi: 10.1145/3065386

Liu, J., Harris, A., and Kanwisher, N. (2010). Perception of face parts and face configurations: an fMRI study. *J. Cogn. Neurosci.* 22, 203–211. doi: 10.1162/jocn.2009.21203

McKone, E., Crookes, K., Jeffery, L., and Dilks, D. D. (2012). A critical review of the development of face recognition: experience is less important than previously believed. *Cogn. Neuropsychol.* 29, 174–212. doi: 10.1080/02643294.2012.660138

Morton, J., and Johnson, M. H. (1991). CONSPEC and CONLERN: a two-process theory of infant face recognition. *Psychol. Rev.* 98:164. doi: 10.1037/0033-295X.98.2.164

Nichols, D. F., Betts, L. R., and Wilson, H. R. (2010). Decoding of faces and face components in face-sensitive human visual cortex. *Front. Psychol.* 1:28. doi: 10.3389/fpsyg.2010.00028

Norman, K. A., and O'Reilly, R. C. (2003). Modeling hippocampal and neocortical contributions to recognition memory: a complementary-learning-systems approach. *Psychol. Rev.* 110, 611–646. doi: 10.1037/0033-295X.110.4.611

Parkhi, O. M., Vedaldi, A., and Zisserman, A. (2015). Deep face recognition. *Proc. Br. Mach. Vis.* 1, 1–12. doi: 10.5244/C.29.41

Pospisil, D. A., Pasupathy, A., and Bair, W. (2018). "Artiphysiology" reveals V4-like shape tuning in a deep network trained for image classification. *Elife* 7:e38242. doi: 10.7554/eLife.38242

Rossion, B., and Gauthier, I. (2002). How does the brain process upright and inverted faces? *Behav. Cogn. Neurosci. Rev.* 1, 63–75. doi: 10.1177/1534582302001001004

Simion, F., and Di Giorgio, E. (2015). Face perception and processing in early infancy: inborn predispositions and developmental changes. *Front. Psychol.* 6:969. doi: 10.3389/fpsyg.2015.00969

Simion, F., Macchi Cassia, V., Turati, C., and Valenza, E. (2001). The origins of face perception: specific versus non-specific mechanisms. *Infant Child Dev. Int. J. Res. Pract.* 10, 59–65. doi: 10.1002/icd.247

Song, Y., Qu, Y., Xu, S., and Liu, J. (2021). Implementation-independent representation for deep convolutional neural networks and humans in processing faces. *Front. Comput. Neurosci.* 14:601314. doi: 10.3389/fncom.2020.601314

Spilmann, L. (2014). Receptive fields of visual neurons: the early years. *Perception* 43, 1145–1176. doi: 10.1068/p7721

Sugita, Y. (2008). Face perception in monkeys reared with no exposure to faces. *Proc. Natl. Acad. Sci. U.S.A.* 105, 394–398. doi: 10.1073/pnas.0706079105

Tavazoie, S. F., and Reid, R. C. (2000). Diverse receptive fields in the lateral geniculate nucleus during thalamocortical development. *Nat. Neurosci.* 3, 608–616. doi: 10.1038/75786

Valenza, E., Simion, F., Cassia, V. M., and Umilt,à, C. (1996). Face preference at birth. *J. Exp. Psychol. Hum. Percept. Perform.* 22:892. doi: 10.1037/0096-1523.22.4.892

Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E. J., Yacoub, E., Ugurbil, K., et al. (2013). The WU-Minn human connectome project: an overview. *Neuroimage* 80, 62–79. doi: 10.1016/j.neuroimage.2013.05.041

VanRullen, R., and Reddy, L. (2019). Reconstructing faces from fMRI patterns using deep generative neuralnetworks. *Commun. Biol.* 2, 193–193. doi: 10.1038/s42003-019-0438-y

Wen, H., Shi, J., Zhang, Y., Lu, K.-H., Cao, J., and Liu, Z. (2017). Neural encoding and decoding with deep learning for dynamic natural vision. *Cereb. Cortex* 28, 4136–4160. doi: 10.1093/cercor/bhx268

Wilmer, J. B., Germine, L., Chabris, C. F., Chatterjee, G., Williams, M., Loken, E., et al. (2010). Human face recognition ability is specific and highly heritable. *Proc. Natl. Acad. Sci. U.S.A.* 107, 5238–5241. doi: 10.1073/pnas.0913053107

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., and DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci.* 111, 8619–8624. doi: 10.1073/pnas.1403112111

Yi, D., Lei, Z., Liao, S., and Li, S. Z. (2014). *Learning face representation from scratch. arXiv preprint arXiv:1411.7923*. Available online at: https://arxiv.org/abs/1411.7923 (accessed November 28, 2014).

Yin, R. K. (1969). Looking at upside-down faces. *J. Exp. Psychol.* 81, 141–145. doi: 10.1037/h0027474

Zhao, M., Cheung, S.-H., Wong, A. C. N., Rhodes, G., Chan, E. K. S., Chan, W. W. L., et al. (2014). Processing of configural and componential information in face-selective cortical areas. *Cogn. Neurosci.* 5, 160–167. doi: 10.1080/17588928.2014.912207

Zhen, Z., Yang, Z., Huang, L., Kong, X.-,z., Wang, X., Dang, X., et al. (2015). Quantifying interindividual variability and asymmetry of face-selective regions: a probabilistic functional atlas. *Neuroimage* 113, 13–25. doi: 10.1016/j.neuroimage.2015.03.010

Zhu, Q., Song, Y., Hu, S., Li, X., Tian, M., Zhen, Z., et al. (2010). Heritability of the specific cognitive ability of face perception. *Curr. Biol.* 20, 137–142. doi: 10.1016/j.cub.2009.11.067

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read
for greatest visibility
and readership

FAST PUBLICATION
Around 90 days
from submission
to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative,
and constructive
peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers
acknowledged by name
on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data
and methods to enhance
research reproducibility

DIGITAL PUBLISHING
Articles designed
for optimal readership
across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics
track visibility across
digital media

EXTENSIVE PROMOTION
Marketing
and promotion
of impactful research

LOOP RESEARCH NETWORK
Our network
increases your
article's readership