



OPEN ACCESS

EDITED BY

Alex James,
Digital University Kerala, India

REVIEWED BY

Kamilya Smagulova,
King Abdullah University of Science
and Technology, Saudi Arabia
Tao Cai,
Jiangsu University, China

*CORRESPONDENCE

Modjtaba Rouhani
✉ rouhani@um.ac.ir

RECEIVED 09 January 2023

ACCEPTED 17 May 2023

PUBLISHED 07 June 2023

CITATION

Sanati S, Rouhani M and Hodtani GA (2023)
Information-theoretic analysis of Hierarchical
Temporal Memory-Spatial Pooler algorithm
with a new upper bound for the standard
information bottleneck method.
Front. Comput. Neurosci. 17:1140782.
doi: 10.3389/fncom.2023.1140782

COPYRIGHT

© 2023 Sanati, Rouhani and Hodtani. This is an
open-access article distributed under the terms
of the [Creative Commons Attribution License
\(CC BY\)](#). The use, distribution or reproduction
in other forums is permitted, provided the
original author(s) and the copyright owner(s)
are credited and that the original publication in
this journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted which
does not comply with these terms.

Information-theoretic analysis of Hierarchical Temporal Memory-Spatial Pooler algorithm with a new upper bound for the standard information bottleneck method

Shiva Sanati¹, Modjtaba Rouhani^{1*} and Ghosheh Abed Hodtani²

¹Department of Computer Engineering, Ferdowsi University of Mashhad, Mashhad, Iran, ²Department of Electrical Engineering, Ferdowsi University of Mashhad, Mashhad, Iran

Hierarchical Temporal Memory (HTM) is an unsupervised algorithm in machine learning. It models several fundamental neocortical computational principles. Spatial Pooler (SP) is one of the main components of the HTM, which continuously encodes streams of binary input from various layers and regions into sparse distributed representations. In this paper, the goal is to evaluate the sparsification in the SP algorithm from the perspective of information theory by the information bottleneck (IB), Cramer-Rao lower bound, and Fisher information matrix. This paper makes two main contributions. First, we introduce a new upper bound for the standard information bottleneck relation, which we refer to as modified-IB in this paper. This measure is used to evaluate the performance of the SP algorithm in different sparsity levels and various amounts of noise. The MNIST, Fashion-MNIST and NYC-Taxi datasets were fed to the SP algorithm separately. The SP algorithm with learning was found to be resistant to noise. Adding up to 40% noise to the input resulted in no discernible change in the output. Using the probabilistic mapping method and Hidden Markov Model, the sparse SP output representation was reconstructed in the input space. In the modified-IB relation, it is numerically calculated that a lower noise level and a higher sparsity level in the SP algorithm lead to a more effective reconstruction and SP with 2% sparsity produces the best results. Our second contribution is to prove mathematically that more sparsity leads to better performance of the SP algorithm. The data distribution was considered the Cauchy distribution, and the Cramer-Rao lower bound was analyzed to estimate SP's output at different sparsity levels.

KEYWORDS

Spatial Pooler (SP), Hierarchical Temporal Memory (HTM), sparsity, standard information bottleneck (IB), modified-information bottleneck (modified-IB), Fisher information matrix (FIM), Cramer-Rao lower bound (CRLB)

1. Introduction

Hierarchical Temporal Memory (HTM) is an unsupervised learning algorithm and a unique artificial intelligence method inspired by the neocortex (Hawkins et al., 2016). The neocortex plays an important role in the human cerebral cortex, accounting for about half of the brain's volume. It is responsible for behavioral and emotional responses and the greatest cognitive functions (Ghazanfar and Schroeder, 2006; Wang, 2022). The neocortex has a hierarchical and homogeneous structure in which higher parts learn general features and lower parts process stimuli (Clark et al., 2002; O'Reilly and Norman, 2002; Friston and Buzsáki, 2016). The neocortex consists of neurons, synapses, and segments (Menzel and Giurfa, 2001). Through synapses and segments, neurons can communicate with one another (Marois and Ivanoff, 2005). Essentially, two types of horizontal and vertical connections transmit information to the cell through the synapse. Horizontal connections represent context inputs, and vertical connections represent feedback and feedforward information (Barack and Krakauer, 2021). HTM is a theoretical model that resembles the neocortex in many respects; for example, it can memorize sequences and then recall them. With the help of a tree-shaped hierarchy neural network, The HTM algorithm extends and combines techniques used in Bayesian networks, spatial and temporal clustering algorithms, and sparse distributed memory. It is a new model of the deep learning process, which is a highly efficient technique in artificial intelligence algorithms. HTM is an online learning method that does not require multiple training epochs. It is a one-shot learning process because almost all the necessary synaptic connections are formed in the first learning round. This algorithm is able to predict and recognize sequences with such robustness without suffering from the usual limitations of conventional neural networks that hinder their training. HTM is a predictive framework, so upon the model receiving each new input, it tries to predict the next events of the world. The HTM algorithm is not only used to detect the next value in a sequence but also to detect anomalies in a sequence. There are four components in HTM: SDR Encoder, Spatial Pooler (SP), Temporal Memory (TM), and Classifier. SP is one of the main components of the HTM, which continuously encodes streams of binary sensory input from various layers and regions within the neocortex into sparse distributed representations (SDR) (Cui et al., 2017). So the information is processed sparsely and encoded inside the HTM neurons as in biological neuronal networks (Finelli et al., 2008).

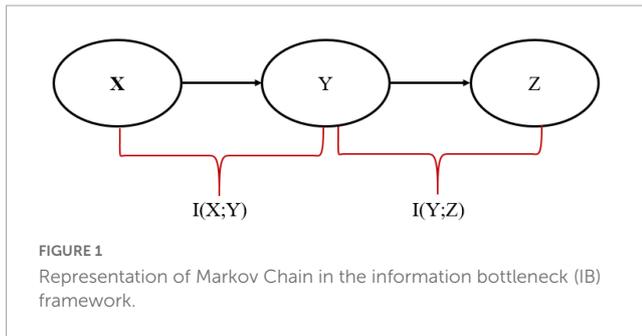
In Spatial Pooler, similar spatial patterns are grouped into highly sparse output representations of cortical mini-columns (Kaas, 2012; Ahmad and Scheinkman, 2019). In SP, there are two main tasks: the first task is to produce similar sparse outputs for similar inputs. The second critical task is to ensure that the output sparsity is fixed regardless of the number of bits in the binary input. Like normalization in other neural networks, these properties act as constraints on the behavior of neurons, which facilitates the training process (Hawkins et al., 2011). The SP algorithm is based on sparse coding techniques. According to the sparse coding theory, sparse activations in the brain's sensory cortex reduce brain energy consumption while maintaining most of the information (Foldiak, 2003; McClelland and Bayne, 2016).

In sparse coding, the cost function is optimized to combine a low reconstruction error with a high sparsity (Olshausen and Field, 1996). The receptive fields produced by sparse coding when applied to natural images are similar to those of brain V1 neurons (visual area); consequently, the sparse coding framework appears responsible for explaining early sensory neurons' functionality (Lee et al., 2006; Paiton et al., 2020). The sparse representation is noise-resistant, and it is suitable for face recognition (Wei et al., 2022), speech recognition (Kwek et al., 2022), and image reconstruction (Deeba et al., 2020).

The sparse coding aims to form associative memory with minimal crosstalk, reduce power consumption, and prevent information loss (Hu and Zeng, 2022). So, in addition to the usefulness of sparsity in the HTM-SP algorithm, the concept of sparsity is helpful in various fields, such as image processing (Peng et al., 2018), medical imaging (Li et al., 2021), machine learning algorithms (Li, 2013), dictionary learning (Wang et al., 2019), denoising (Zhou et al., 2022), sampling theory (Nagahara and Yamamoto, 2022), and signal recovery (Abiantun et al., 2019). In the sparse representation models, a small number of coefficients contain a significant amount of energy (Ravishankar and Bresler, 2012). The SP algorithm has a fixed-sparsity representation of the input (Ahmad and Hawkins, 2016). A fixed level of sparsity in presynaptic inputs results in reliable and robust recognition of presynaptic activation patterns (Olshausen and Field, 2004). In the case of highly variable sparsity, it is difficult to detect input patterns with low activation density. On the other hand, input patterns with high activation density cause action potentials in downstream neurons. Therefore, false negative errors will occur in the case of low-density patterns, and false positive errors will occur in the case of high-density patterns. In general, it is desirable to have a fixed sparsity since it ensures that all input patterns can be detected equally. The fixed sparsity is approximately 2%, it means that only 2% of columns in the SP algorithm are activated (Hawkins et al., 2016).

1.1. Our contributions

This paper analyzes the sparsification in the HTM-SP algorithm from an information theory perspective. So, we applied the IB relation to investigate the accuracy of SP output reconstruction at different sparsity levels and various amounts of noise for the first time. Moreover, we proposed a new MNIST relation, which was employed to resolve the reconstruction problem. Then the Fisher information matrix and Cramer-Rao lower bound are used to prove mathematically that more sparsity leads to better performance of the HTM-SP algorithm. The data distribution is considered the Cauchy distribution, although similar analyses could have been conducted with other distributions. Here, the modified-IB is introduced as an upper bound for IB. Furthermore, its applicability is tested on the sparsity-noise impact of SP algorithm by using both standard and modified-IB. However, the application of the modified-IB is by no means limited to the HTM or its SP algorithm. In fact, the modified-IB can be used in any other study, replacing the standard IB.



1.2. Paper organization

The structure of this paper is arranged as follows. Section “2 Preliminaries” reviews all the preliminary details to improve the paper’s readability. In section “3 Our work,” our work is explained. Numerical experiments and simulations of the proposed modified-IB method are in section “4 Numerical Results related to our new IB relation.” Finally, in section “5 Conclusion,” the paper is concluded.

2. Preliminaries

The purpose of this section is to explain some of the scientific terms used in the rest of the paper, including the information bottleneck¹ relation, Fisher information Matrix², and Cramer-Rao lower bounds³.

2.1. Information bottleneck

The Information bottleneck (IB) method is introduced by Tishby et al. (2000) and is used to find a maximally compressed representation in the $X \rightarrow Y \rightarrow Z$ Markov chain that transmits information from input random variable X to output random variable Z through a compressed representation Y to preserve as much relevant information as possible, as shown in Figure 1.

The objective of the IB is to find the optimal representation Y , which involves minimizing the following Lagrangian cost function:

$$L_{IB} = I(X; Y) - \beta I(Y; Z) \quad (1)$$

Where $I(_, _)$ is the mutual information. $I(X; Y)$ represents the compression or pruning term that discards irrelevant information by minimizing the mutual information between Y and source X and $I(Y; Z)$ preserving relevant information to ensure Y predicts Z by maximizing the mutual information between Y and Z . The multiplier $\beta \geq 0$ is a hyperparameter that controls the trade-off between these two terms (Alemi et al., 2016). A small β generally indicates more compression, whereas a large β indicates that more relevant information is being maintained.

2.2. Fisher information and Cramer–Rao lower bound definition

The Fisher information matrix (FIM) measures the amount of information the data can provide about the

unknown parameter in an estimation problem (Stein et al., 2014). In order to quantify the Fisher information (FI) based on log-likelihoods, the following definitions are presented:

$$I(\theta) = E_{\theta} \left\{ [l'(x; \theta)]^2 \right\} = \int [l'(x; \theta)]^2 f(x; \theta) dx \quad (2)$$

$$I(\theta) = -E_{\theta} \left\{ [l''(x; \theta)] \right\} = - \int \left[\frac{\partial^2}{\partial \theta^2} \log f(x; \theta) \right] f(x; \theta) dx \quad (3)$$

In these equations, $f(x; \theta)$ is the probability distribution function of a random variable X , where X depends on the parameter $\theta \in \Theta$ and $l(x; \theta)$ denotes the log-likelihood function.

The Fisher information of the probability family is a symmetric and positive semi-definite matrix valued function, where the ij th entry is as follows:

$$I_{i,j}(\theta) = E_{\theta} \left[\left(\frac{\partial}{\partial \theta_i} \log p_{\theta}(X) \right) \left(\frac{\partial}{\partial \theta_j} \log p_{\theta}(X) \right) \right] \quad (4)$$

The Cramer-Rao inequality, which is the right-hand side expression in (5), explains the relationship between FIM and error variance in the following manner. It is almost a direct result of a famous mathematics inequality known as the Cauchy-Schwartz inequality (Hoffmann and Kunze, 1971).

$$\text{var}[\hat{\theta}] \geq \frac{1}{nI(\theta)} \quad (5)$$

In this case, n represents the sample vector’s size, $I(\theta)$ represents the Fisher Information Matrix (FIM), and $\hat{\theta}$ represents the unbiased estimator of θ .

3. Our work

3.1. A new upper bound for IB relation on the SP algorithm

The learning aim is to find a function that minimizes the uncertainty of the output given the input while avoiding irrelevant information as much as possible. Tishby et al. (2000) introduced this viewpoint as the Information Bottleneck (IB), a fundamental concept in information theory.

This paper aims to apply the IB relation, for the first time, to analyze the effect of sparsity and noise on data reconstruction in the HTM-Spatial pooler algorithm and propose a modified-IB relation.

The IB method has an extraordinary application in various fields of machine learning and related domains (Riguzzi and Di Mauro, 2012; Goldfeld and Polyanskiy, 2020; Zuo et al., 2021; Musat and Andonie, 2022). It also applies to other areas, such as neuroscience (Schneidman et al., 2001; Buddha et al., 2013; Tucker et al., 2022). Tishby and Zaslavsky (2015) used the IB method to evaluate the deep neural networks’ performance and determine the reasons for their success. Despite the impressive successes of deep neural networks, they have been criticized for the lack of sufficient information from inside the network and for being unable to understand the internal structure and

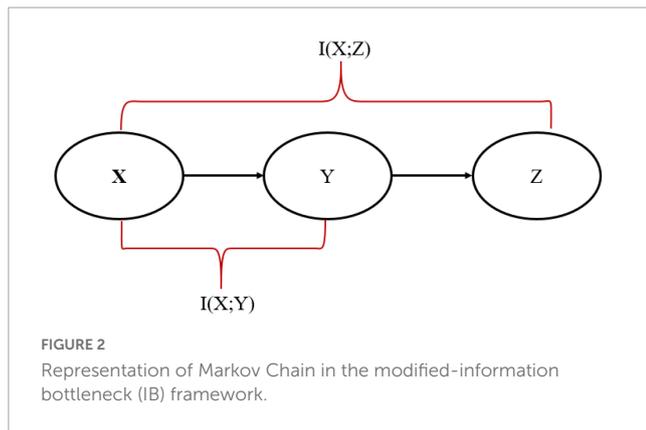


FIGURE 2 Representation of Markov Chain in the modified-information bottleneck (IB) framework.

optimization process in recent years. The IB method has almost solved this problem and opened the deep learning black box. The experimental evidence in [Shwartz-Ziv and Tishby \(2017\)](#) indicated that deep neural networks implicitly solve the information bottleneck optimization problem, i.e., compress the input while preserving the associated information related to the output. The IB principle has been applied to complex and high-dimensional data in numerous studies, including improving and analyzing the learning of deep neural networks ([Gu et al., 2020](#); [Raj et al., 2020](#); [Li and Liu, 2021](#); [Vera et al., 2022](#)), learning disentangled and invariant representations ([Achille and Soatto, 2018](#)), and enhancing robustness against adversarial attacks ([Fischer, 2020](#)). Scientists are still interested in this topic as it is still an open issue.

In this paper, the SP algorithm was fed with 60,000 MNIST data as a training set and produced the sparse representation. A probabilistic mapping method and Hidden Markov Model (HMM) were used to reconstruct the sparse SP output representation in the input space ([Mnatzaganian et al., 2017](#)). Although some information has been lost during the reconstruction process, it is unquestionable that the reconstructed data is similar to the original data. Using the standard IB relation (described in equation 6), we can accurately assess the similarity between input data and the data reconstructed by the SP algorithm at various sparsity levels and various amounts of noise. Therefore, we analyze how sparsity and noise affect reconstruction accuracy using the standard IB. It is the first time someone has calculated the standard IB relation for the SP algorithm and quantitatively compared the reconstructed data with the original data. Furthermore, we proposed a modified- IB relation ([Figure 2](#)), which was applied to measure the similarity in the SP’s reconstruction problem. Then we compare these two measures, as shown in [Figure 3](#).

$$L_{IB} = I(X; Y) - \beta_1 I(Y; Z) \tag{6}$$

$$L_{M-IB} = I(X; Y) - \beta_2 I(X; Z) \tag{7}$$

The mathematical comparison of L_{IB} and L_{M-IB} (assuming $\beta_1 = \beta_2$) easily demonstrates that L_{M-IB} is greater than L_{IB} . Because in the $X \rightarrow Y \rightarrow Z$ Markov chain, it is obvious

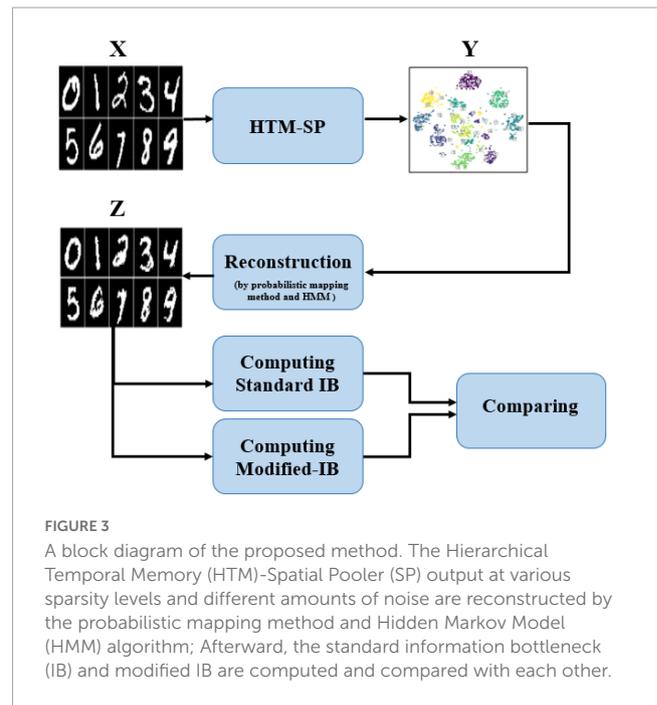


FIGURE 3 A block diagram of the proposed method. The Hierarchical Temporal Memory (HTM)-Spatial Pooler (SP) output at various sparsity levels and different amounts of noise are reconstructed by the probabilistic mapping method and Hidden Markov Model (HMM) algorithm; Afterward, the standard information bottleneck (IB) and modified IB are computed and compared with each other.

that as the distance between two nodes increases, more information is lost along the way:

$$I(X; Z) \leq I(Y; Z) \tag{8}$$

So,

$$L_{M-IB} \geq L_{IB} \tag{9}$$

It was stated in section “2.1 Information bottleneck” that L_{IB} represents the information bottleneck, which should be minimized. As a result, we have proved that $L_{M-IB} \geq L_{IB}$. If we decrease L_{M-IB} , L_{IB} must also decrease because it is less than L_{M-IB} .

So, $\min(L_{M-IB}) \geq \min(L_{IB})$. Therefore, we demonstrate L_{M-IB} is an upper bound of the information bottleneck method, and its minimization incorporates additional information. L_{M-IB} minimization also includes the information derived from the information bottleneck method.

3.2. The corresponding comparisons on the SP algorithm

A lower bound on variance is a handy feature of any unbiased estimator. Using this lower bound, selecting the most appropriate estimator based on possible minimal variance is possible ([Tune, 2012](#); [Huang et al., 2020](#); [Khorasani et al., 2020](#)). The Cramer-Rao Lower Bound (CRLB) is the most popular lower bounds in the literature due to its attractiveness and ease of evaluation ([Rao et al., 1973](#); [D’Amico et al., 2022](#)). As stated in the Cramer-Rao inequality ([Dogandzic and Nehorai, 2001](#)), the diagonal terms of the inverse of the Fisher information Matrix (FIM) (assuming it exists) represent asymptotic lower bounds on any unbiased estimator’s variance. The Fisher information matrix measures the amount of information the data can provide about the unknown parameter in an estimation problem ([Stein et al., 2014](#)). A parameter can be more accurately

estimated when the data contains more information about the parameter. Information amount is determined by considering how the likelihood of observing the acquired data changes with the parameter value. In the absence of a significant change in the likelihood of the data with respect to the parameter value, the data contains very little information about the parameter. The Cramer-Rao lower bound can be used as a benchmark to determine the accuracy of a method. When the estimator's variance is equal to the CRLB, it is considered the most efficient estimator. The purpose of this section is to demonstrate that CRLB will be decreased as a consequence of data sparsification. The claims are proved by using the Cauchy distribution. We first examine the sparsity variation of output caused by the filter in the SP algorithm and then explore three scenarios (as described below) in which the CRLB of the estimation error is calculated to examine the effect of SP's output sparsity on the estimation error.

- Compute CRLB when SP's output is not sparse.
- Compute CRLB when SP's output has the maximum Sparsity.
- Compute CRLB when adding noise to the SP's output.

Let's consider the first scenario where the input data (X) follows the Cauchy distribution, so $x_0 \sim C(x_0, \gamma)$ and $\theta = (x_0, \gamma)^T$; therefore, multiple parameters must be estimated. The first step in determining CRLB is defining a Fisher information matrix for the two main parameters of Cauchy distribution (x_0 and γ). A log-likelihood function $l(x; \theta)$ based on the Cauchy distribution is written as follows.

$$f(x; \theta) = f(x; x_0, \gamma) = \frac{1}{\pi\gamma[1 + (\frac{x-x_0}{\gamma})^2]} = \frac{\gamma}{\pi(\gamma^2 + (x - x_0)^2)} \quad (10)$$

Where x_0 indicates the location parameter of the peak and shifts the graph along the x-axis, and γ indicates the scale parameter of the graph, which can be either shorter or taller. And then

$$l(x; \theta) = \ln f(x; \theta) = \ln(\gamma) - \ln(\pi) - \ln(\gamma^2 + (x - x_0)^2) \quad (11)$$

The Fisher information matrix can be determined by computing the first and second derivatives of $\ln f(x; \theta)$:

$$\frac{\partial l(x; \theta)}{\partial \theta} = \left(\frac{\partial l(x; \theta)}{\partial x_0}, \frac{\partial l(x; \theta)}{\partial \gamma} \right)^T = \left(\frac{2(x - x_0)}{\gamma^2 + (x - x_0)^2}, \frac{1}{\gamma} - \frac{2\gamma}{\gamma^2 + (x - x_0)^2} \right) \quad (12)$$

And

$$\frac{\partial^2 l(x; \theta)}{\partial \theta^2} = \begin{bmatrix} \frac{\partial^2 l(x; \theta)}{\partial x_0^2} & \frac{\partial^2 l(x; \theta)}{\partial x_0 \partial \gamma} \\ \frac{\partial^2 l(x; \theta)}{\partial \gamma \partial x_0} & \frac{\partial^2 l(x; \theta)}{\partial \gamma^2} \end{bmatrix} = \begin{bmatrix} \frac{-4\gamma^2 + 2[\gamma^2 + (x - x_0)^2]}{[\gamma^2 + (x - x_0)^2]^2} & \frac{-4\gamma(x - x_0)}{[\gamma^2 + (x - x_0)^2]^2} \\ \frac{-4\gamma(x - x_0)}{[\gamma^2 + (x - x_0)^2]^2} & \frac{-1}{\gamma^2} - \frac{2}{\gamma^2 + (x - x_0)^2} + \frac{4\gamma^2}{[\gamma^2 + (x - x_0)^2]^2} \end{bmatrix} \quad (13)$$

Equation (14) provides Fisher's information matrix, and the calculations related to this part are contained in the **Supplementary Appendix 1**.

$$I(\theta) = -E_{\theta} \left[\frac{\partial^2 l(x; \theta)}{\partial \theta^2} \right] = \begin{bmatrix} \frac{1}{2\gamma^2} & 0 \\ 0 & \frac{1}{2\gamma^2} \end{bmatrix} \quad (14)$$

As a result, x_0 's Fisher information contained in the random variable X is equivalent to $I_{11}(\theta) = \frac{1}{2\gamma^2}$, and γ 's Fisher information is equal to $I_{22}(\theta) = \frac{1}{2\gamma^2}$.

So, according to their general definition, the Cramer-Rao lower bound for the estimation error for both parameters x_0 and γ is $CRLB = I(\theta)^{-1}$.

$$CBRLB_{x_0} = I_{11}^{-1}(\theta) = 2\gamma^2 \quad (15)$$

$$CBRLB_{\gamma} = I_{22}^{-1}(\theta) = 2\gamma^2 \quad (16)$$

This article examines the effect of Sparsity on the error bound of SP's output reconstruction. For this purpose, we explore an ideal state of Sparsity (the maximum Sparsity) in which all the information is contained in one SP's output element, and all the rest are zero after the sparsification process, $x = [x_{sparse}]$ so $x_0 = x_{sparse}$ and $\gamma = 0$.

TABLE 1 Some other distributions that can be used in our experiments.

Distribution	PDF/PMF	Parameter	Fisher information	Cramer-Rao lower bound
Gaussian	$\frac{1}{\delta\sqrt{2\pi}} \exp -\frac{(x - \mu)}{2\delta^2}$	μ, δ^2	$\begin{pmatrix} \frac{1}{\delta^2} & 0 \\ 0 & \frac{1}{2\delta^4} \end{pmatrix}$	$CRLB_{\mu} = \delta^2$ $CRLB_{\delta^2} = 2\delta^4$
Pareto	$\frac{\alpha x_m^\alpha}{x^{\alpha+1}}$	$x_m > 0$ $\alpha > 0$	$\begin{pmatrix} \frac{\alpha^2}{x_m^2} & 0 \\ 0 & \frac{1}{\alpha^2} \end{pmatrix}$	$CRLB_{x_m} = \frac{x_m^2}{\alpha^2}$ $CRLB_{\alpha} = \alpha^2$
Exponential	$\lambda \exp -\lambda x$	$\lambda > 0$	$\frac{1}{\lambda^2}$	λ^2
Poisson	$\frac{\lambda^k \exp -\lambda}{k!}$	$\lambda \in (0, \infty)$	$\frac{1}{\lambda}$	λ
Bernoulli	$\begin{cases} q = 1 - p & \text{if } k = 0 \\ p & \text{if } k = 1 \end{cases}$	$0 \leq p \leq 1$ $q = 1 - p$	$\frac{1}{pq}$	pq

By substituting $\gamma = 0$ in the Fisher information matrix in (14), the following result in (17) is obtained:

$$I_{11}(\theta) = I_{22}(\theta) = \infty \tag{17}$$

Therefore, the Cramer-Rao lower bound for these two parameters is equal to zero, which means that the SP's output reconstruction error is zero when we have the SP with maximum Sparsity.

$$CBRLB_{x_0} = CBRLB_{\gamma} = 0 \tag{18}$$

TABLE 2 The comparison of the modified-information bottleneck (IB) and standard information bottleneck (IB) methods with different β values in the reconstruction of the Hierarchical Temporal Memory (HTM)-Spatial Pooler (SP) algorithm using different sparsity (column-activation parameter) on MNIST dataset.

	Method	β			
		10^{-2}	10^{-1}	10^0	10^1
2% sparsity	Modified-IB	0.3247	0.3247	0.3242	0.3199
	Standard IB	0.3247	0.3244	0.3217	0.2947
10% sparsity	Modified-IB	0.3306	0.3306	0.3301	0.3256
	Standard IB	0.3306	0.3303	0.3277	0.3016
20% sparsity	Modified-IB	0.338	0.338	0.3378	0.3358
	Standard IB	0.338	0.3378	0.336	0.318
30% sparsity	Modified-IB	0.4013	0.4013	0.401	0.3988
	Standard IB	0.4013	0.4012	0.4001	0.3893
40% sparsity	Modified-IB	0.41	0.41	0.4097	0.407
	Standard IB	0.41	0.4099	0.409	0.4003

TABLE 3 The comparison of the modified-information bottleneck (IB) and standard information bottleneck (IB) methods with different β values in the reconstruction of the Hierarchical Temporal Memory (HTM)-Spatial Pooler (SP) algorithm using different sparsity (column-activation parameter) on Fashion-MNIST dataset.

	Method	β			
		10^{-2}	10^{-1}	10^0	10^1
2% sparsity	Modified-IB	0.3429	0.3429	0.3407	0.3209
	Standard IB	0.3429	0.3425	0.3389	0.3029
10% sparsity	Modified-IB	0.3608	0.3608	0.3586	0.3388
	Standard IB	0.3608	0.3604	0.3568	0.3208
20% sparsity	Modified-IB	0.3712	0.3712	0.3710	0.3692
	Standard IB	0.3712	0.3708	0.3672	0.3312
30% sparsity	Modified-IB	0.4355	0.4355	0.4351	0.4315
	Standard IB	0.4355	0.4353	0.4335	0.4155
40% sparsity	Modified-IB	0.4493	0.4493	0.4490	0.4463
	Standard IB	0.4493	0.4492	0.4482	0.4392

TABLE 4 The comparison of the modified-information bottleneck (IB) and standard information bottleneck (IB) methods with different β values in the reconstruction of the Hierarchical Temporal Memory (HTM)-Spatial Pooler (SP) algorithm using different sparsity (column-activation parameter) on NYC-Taxi dataset.

	Method	β			
		10^{-2}	10^{-1}	10^0	10^1
2% sparsity	Modified-IB	0.1121	0.1121	0.1099	0.0901
	Standard IB	0.1121	0.1116	0.1071	0.0621
10% sparsity	Modified-IB	0.1488	0.1488	0.1461	0.1218
	Standard IB	0.1488	0.1483	0.1438	0.0988
20% sparsity	Modified-IB	0.1560	0.1560	0.1558	0.1540
	Standard IB	0.1560	0.1557	0.1530	0.1260
30% sparsity	Modified-IB	0.2284	0.2284	0.2281	0.2254
	Standard IB	0.2284	0.2282	0.2264	0.2084
40% sparsity	Modified-IB	0.2512	0.2512	0.2509	0.2482
	Standard IB	0.2512	0.2511	0.2502	0.2412

Tests were conducted with other distributions, including Normal, Poisson, Exponential, and Bernoulli. Once again, the results indicated that the CRLB is zero when the Sparsity of the SP's output is maximum.

Next, we assume that the SP's output has decreased its Sparsity. For example, we decrease SP's output sparsity by adding a constant value λ to all of its elements. The aim is to investigate the relation between the Cramer-Rao lower bound and decreasing the SP's output sparsity.

Let's assume that the input data (X) follows the Cauchy distribution, so $x_0 \sim C(x_0, \gamma)$. In this case, the parameters of the new vector ($X + \lambda$) will change as follows:

$$x_{\lambda} = x_0(\lambda + X_0) = \lambda + x_0(X_0) = \lambda + x_0 \tag{19}$$

$$\gamma_{\lambda} = \gamma(\lambda + X_0) = \gamma(X_0) = \gamma \tag{20}$$

To incorporate these changes, the second derivative of the log-likelihood function in equation (13) is rewritten as follows:

$$\frac{\partial^2 I(x; \theta)}{\partial \theta^2} = \begin{bmatrix} \frac{-4\gamma^2 + 2[\gamma^2 + (x-x_0-\lambda)^2]}{[\gamma^2 + (x-x_0-\lambda)^2]^2} & \frac{-4\gamma(x-x_0-\lambda)}{[\gamma^2 + (x-x_0-\lambda)^2]^2} \\ \frac{-4\gamma(x-x_0-\lambda)}{[\gamma^2 + (x-x_0-\lambda)^2]^2} & \frac{-1}{\gamma^2} - \frac{2}{\gamma^2 + (x-x_0-\lambda)^2} + \frac{4\gamma^2}{[\gamma^2 + (x-x_0-\lambda)^2]^2} \end{bmatrix} \tag{21}$$

Therefore, The Fisher information matrix can be expressed as follows:

$$I(\theta) = -E_{\theta} \left[\frac{\partial^2 I(x; \theta)}{\partial \theta^2} \right] = \begin{bmatrix} \frac{1}{2\gamma^2} & 0 \\ 0 & \frac{1}{2\gamma^2} \end{bmatrix} \tag{22}$$

The CRLB, which is the inverse matrix of Fisher information, can be expressed as follows:

$$CBRLB_{x_0} = I_{11}^{-1}(\theta) = 2\gamma^2 \tag{23}$$

$$CBRLB_{\gamma} = I_{22}^{-1}(\theta) = 2\gamma^2 \tag{24}$$

This analysis allows us to compare the CRLB in equations (23) and (24) to the CRLBs calculated for the SP's output before adding constant λ value, equations (15) and (16). Explicitly, it is shown that a decrease in sparsity does not change the Cramer-Rao lower bound in the Cauchy distribution. It should be noted that, in this particular test, if the data followed a Gaussian distribution, the decrease in sparsity would increase the Cramer-Rao lower bound (Khorasani et al., 2020). Consequently, a Gaussian distribution is considered the worst-case scenario in estimating unknown parameters. As stated before, these experiments can be performed with other distributions, including Gaussian, Pareto, Poisson, Exponential, and Bernoulli. Table 1 summarizes the properties of these five distributions, as well as the Fisher information matrix and the Cramer-Rao lower bound.

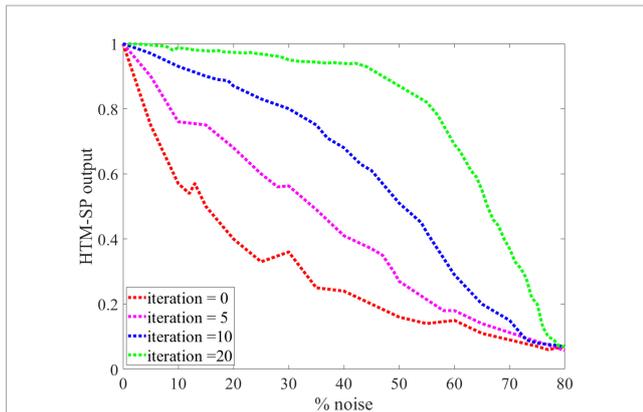


FIGURE 4
An analysis of the Spatial Pooler (SP) algorithm’s behavior against different levels of input noise during learning process by different iterations value is presented. The performance of the SP algorithm without learning (red line) is greatly affected by small amounts of noise. In contrast, even 40% noise does not significantly affect the output of the SP algorithm with learning (green line), after which the curve’s slope gradually changes.

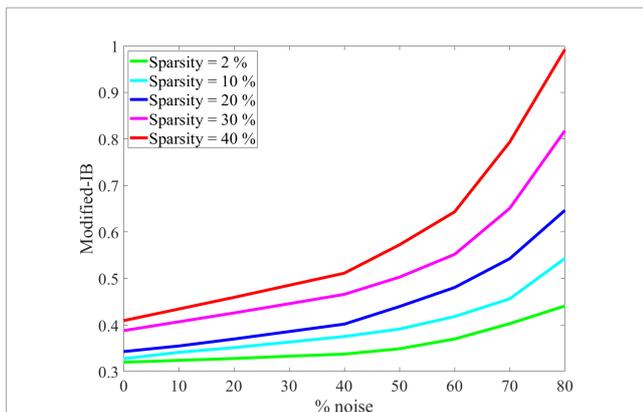


FIGURE 5
Noise effects on the modified-information bottleneck (IB) relation at $\beta = 10$ for different sparsity levels. With a higher sparsity level, the Spatial Pooler (SP) algorithm is more resistant to noise, and the modified-IB relation provides the best results.

4. Numerical results related to our new IB relation

This section examines the effect of different sparsity levels and various amounts of noise on the SP algorithm’s output reconstruction. The SP algorithm was fed with 60,000 MNIST data as a training set and produced the sparse representation. We add 0, 10, 20, 30, and 40% noise values to the input of the SP algorithm for each SP’s sparsity (Column-activation = 2, 10, 20, 30, and 40%), and the output is obtained. A probabilistic mapping method and Hidden Markov Model were used to reconstruct the sparse SP output representation in the input space (Mnatzaganian et al., 2017). The SP algorithm was simulated using the mHTM⁴ implementation. To examine the accuracy of the SP output reconstruction, we used the standard IB relation and proposed a new upper bound for it. The results of comparing the modified-IB

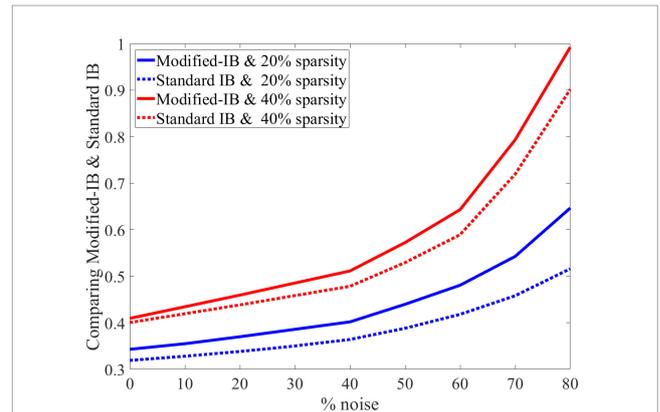


FIGURE 6
At different sparsity levels, standard information bottleneck (IB) is always lower than modified-IB.

and standard IB at different sparsity levels of the SP algorithm are reported in **Table 2**.

According to the above experiments, the most accurate reconstruction occurs when the sparsity of the SP algorithm is high (2% sparsity), whereas reducing the sparsity in the SP algorithm increases the reconstruction error. In this experiment, it can be seen that standard IB is always lower than or equal to modified-IB, which is in accordance with our previously stated fact that modified-IB is an upper bound for the standard IB.

Using the Fashion-MNIST and NYC-Taxi datasets, we repeated all of the experiments, and compared the Modified-IB and Standard IB methods with different β values in the reconstruction of the HTM-SP algorithm using different sparsity. As shown in **Tables 3, 4**, SP with 2% sparsity produces the best results and the Modified-IB is always greater than or equal to the standard IB in these two datasets, as expected.

Following this, we evaluate the noise robustness of the SP algorithm during the learning process by adding different amounts of noise to its input. The SP algorithm was trained using the noisy MNIST image as input for each noise level (between 0 and 75%) by randomly flipping active bits to inactive bits and inactive bits to active bits for the given percentage of the pixels. Twenty-one iterations are necessary for the learning process to ensure that the SP has a stable output representation. As shown in **Figure 4**, the SP algorithm with learning (green line) was more robust against noise than the SP algorithm without learning (red line). It was found that the learned outputs remained virtually unchanged (or with relatively small changes) after adding a significant amount of noise to the input, as it is almost smooth until 40% noise. Therefore, even adding up to 40% noise to the input resulted in no discernible change in the output of the SP algorithm with learning. The following experiments of this paper used the SP algorithm with learning due to its advantages over the SP without learning.

As shown in **Figure 5**, the effect of SP’s sparsity and input noise on SP’s output reconstruction was quantified using the modified-IB relation. This experiment demonstrated that the original data performs better than the noisy data, and by increasing the sparsity of the SP algorithm, the noise entropy decreases. So, the addition of noise causes a higher reconstruction error. The IB curve is almost

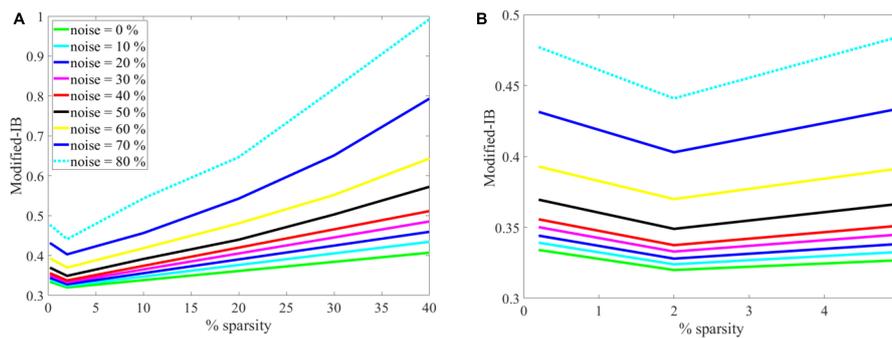


FIGURE 7 (A) The effects of sparsity on the modified-information bottleneck (IB) relation at $\beta = 10$ for different noise levels. With a lower noise level and sparsity = 2%, the modified-IB relation provides the best results. (B) The graph on the left has been zoomed in the range of 0–5 to gain more clarity and to show the sparsity = 2% is the most appropriate sparsity in the Spatial Pooler (SP) algorithm.

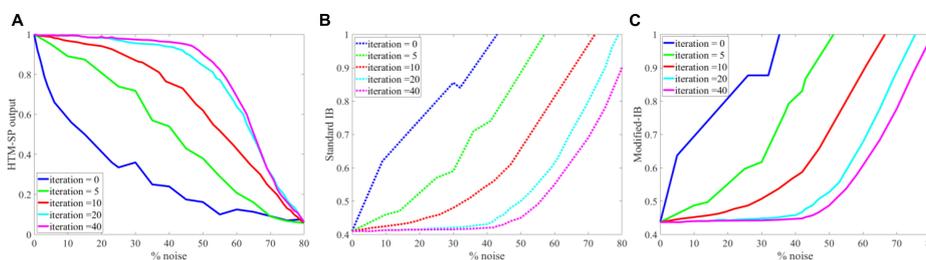


FIGURE 8 Analysis of the Spatial Pooler (SP) algorithm’s against different levels of input noise during learning process by different iterations value. (A) The change of the SP outputs is plotted as a function of the noise level (Cui et al., 2017). (B) Noise effects on the Standard information bottleneck (IB) relation at $\beta = 10$ for fixed sparsity 2% and in different iterations. (C) Noise effects on the Modified-IB relation at $\beta = 10$ for fixed sparsity 2% and in different iterations. In iteration 40, the SP algorithm is more resistant to noise, and the Standard IB relation provides the best results.

smooth, up to 40% noise, then changes dramatically. Therefore, adding up to 40% noise to the input results in no discernible change in the output of the SP algorithm. It is possible to measure the resistance to noise of the SP algorithm by the modified-IB relation (at different sparsity levels), and the result of Figure 5 is in accordance with the result of Figure 4.

As in the previous case, we concluded that the higher sparsity in the SP algorithm leads to a better reconstruction of the output, and standard IB is always lower than modified-IB (Figure 6).

Again, the SP algorithm is quantitatively evaluated using the modified-IB relation at different sparsity levels and noise, as shown in (A) in Figure 7. For more clarity, in Figure 7 the graph (A) has been zoomed in the range of 0 to 5, as shown in (B). Accordingly, the (A) and (B) graphs in Figure 7 show some interesting results:

- If there is no noise (green line), modified-IB is smaller, and the SP algorithm is more accurate.
- If the sparsity level is low and the noise level is high, modified IB expresses a higher number indicating that the SP algorithm is inaccurate.
- The optimal situation is sparsity = 2% since modified-IB has the lowest value for all noises at this point.
- In the case of sparsity = 2%, the noises of 0 to 40% are almost adjacent, whereas the noises of 50–80% are far apart. Thus, we can conclude that adding up to 40% noise to the input results in no significant change in the output of the SP algorithm.

To further analysis the applicability of our method, we perform the same experience as in Cui et al. (2017) on random sparse inputs and tested the effect of noise on the SP algorithm in different iterations. In Figure 8A, presents the results of Figure 4 in reference (Cui et al., 2017). Figures 8B, C, are the results of application of IB and modified-IB in the same experience. In this figure an analysis of the SP algorithm’s behavior at different levels of input noise during the learning process with different iteration values is shown. The performance of the SP algorithm without learning

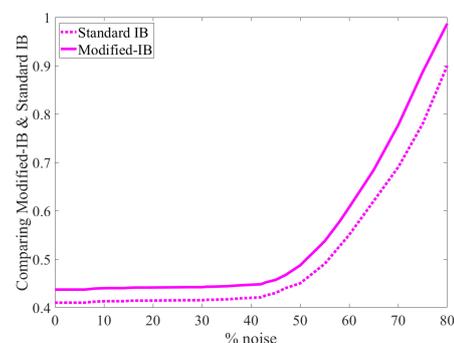


FIGURE 9 Modified-information bottleneck (IB) is an upper bound for the standard IB.

(blue line) is greatly affected by small amounts of noise. In contrast, even 40% noise does not significantly affect the output of the SP algorithm with learning (Purple line), after which the curve's slope gradually changes. As indicated in those figures, the results from reference (Cui et al., 2017) is in accordance with results from IB and Modified-IB. **Figures 8B, C** demonstrate that the SP algorithm performs better when algorithm iterations are higher and the SP is in the learning phase (Purple line). In this case, even 40% noise does not significantly affect the SP algorithm output. In addition, **Figure 9** shows that Modified-IB is an upper bound for Standard-IB, as we expected.

5. Conclusion

This paper aimed to evaluate the sparsification in the SP algorithm from the perspective of information theory as measured by the information bottleneck, Cramer-Rao lower bound, and Fisher information matrix. Two main contributions were made in this paper. First, we introduced a new upper bound for the standard information bottleneck relation. This measure has been used to evaluate the performance of the SP algorithm in different sparsity levels and various amounts of noise. The MNIST dataset was fed as input to the SP algorithm. The SP algorithm with learning was found to be resistant to noise. Adding up to 40% noise to the input resulted in no discernible change in the output. Using the probabilistic mapping method and Hidden Markov Model, the sparse SP output representation was reconstructed in the input space. The purpose was to assess the similarity between input data and the data reconstructed by the SP algorithm. SP with 2% sparsity produced the best results. The claims are numerically validated by the standard information bottleneck relation and its proposed new version (modified-IB). The results show that a lower amount of noise and a higher sparsity level in the SP algorithm improved reconstruction accuracy. This finding, which is based solely on information theory measures, is in coherence with empirical result. we have proved what was previously only an experimental observation. Our second contribution was to prove mathematically that more sparsity leads to better performance of the SP algorithm. So, we analyzed the relationship between the Cramer–Rao lower bound on the estimation of the SP's output and the Sparsity of the SP's output and also the relation of sparsity and adding noise to the SP's output. Accordingly, this paper investigated the effects of varying the sparsity of the SP, followed by comparing the error bounds before and after sparsification. The data distribution was considered the Cauchy distribution, although similar analyses could have been conducted with other distributions, including Gaussian, Pareto, Poisson, Exponential, and Bernoulli. As a result of this research, it will be possible to reduce recovery errors during compression and transmission procedures.

References

- Abiantun, R., Juefei-Xu, F., Prabhu, U., and Savvides, M. (2019). SSR2: Sparse signal recovery for single-image super-resolution on faces with extreme low resolutions. *Pattern Recognit.* 90, 308–324. doi: 10.1016/j.patcog.2019.01.032
- Achille, A., and Soatto, S. (2018). Emergence of invariance and disentanglement in deep representations. *J. Mach. Learn. Res.* 19, 1947–1980.

Data availability statement

Publicly available datasets were analyzed in this study, including the MNIST, Fashion-MNIST, and NYC-Taxi datasets. This data can be found here: <http://yann.lecun.com/exdb/mnist/>; <https://github.com/zalandoresearch/fashion-mnist>; and http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml.

Author contributions

SS: conceptualization, methodology, software, validation, and writing—original draft. MR and GH: conceptualization, methodology, validation, supervision, project administration, and review and editing. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the Cognitive Sciences and Technologies Council in Iran (Grant No. 10466).

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.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fncom.2023.1140782/full#supplementary-material>

- Ahmad, S., and Hawkins, J. (2016). How do neurons operate on sparse distributed representations? A mathematical theory of sparsity, neurons and active dendrites. *arXiv [Preprint]*

- Ahmad, S., and Scheinkman, L. (2019). How can we be so dense? The benefits of using highly sparse representations. *arXiv [Preprint]*

- Alemi, A. A., Fischer, I., Dillon, J. V., and Murphy, K. (2016). Deep variational information bottleneck. *arXiv [Preprint]*
- Barack, D. L., and Krakauer, J. W. (2021). Two views on the cognitive brain. *Nat. Rev. Neurosci.* 22, 359–371. doi: 10.1038/s41583-021-00448-6
- Buddha, S. K., So, K., Carmena, J. M., and Gastpar, M. C. (2013). Function identification in neuron populations via information bottleneck. *Entropy* 15, 1587–1608. doi: 10.3390/e15051587
- Clark, R. E., Manns, J. R., and Squire, L. R. (2002). Classical conditioning, awareness, and brain systems. *Trends Cogn. Sci.* 6, 524–531. doi: 10.1016/S1364-6613(02)02041-7
- Cui, Y., Ahmad, S., and Hawkins, J. (2017). The HTM spatial pooler—a neocortical algorithm for online sparse distributed coding. *Front. Comput. Neurosci.* 11:111. doi: 10.3389/fncom.2017.00111
- D'Amico, A. A., Morelli, M., and Moretti, M. (2022). Frequency estimation by interpolation of two fourier coefficients: Cramér-Rao bound and maximum likelihood solution. *IEEE Trans. Commun.* 70, 6819–6831. doi: 10.1109/TCOMM.2022.3200679
- Deeba, F., Kun, S., Ali Dharejo, F., and Zhou, Y. (2020). Sparse representation based computed tomography images reconstruction by coupled dictionary learning algorithm. *IET Image Process.* 14, 2365–2375. doi: 10.1049/iet-ipc.2019.1312
- Dogandzic, A., and Nehorai, A. (2001). Cramer-Rao bounds for estimating range, velocity, and direction with an active array. *IEEE Trans. Signal Process.* 49, 1122–1137. doi: 10.1109/78.923295
- Finelli, L. A., Haney, S., Bazhenov, M., Stopfer, M., and Sejnowski, T. J. (2008). Synaptic learning rules and sparse coding in a model sensory system. *PLoS Comput. Biol.* 4:e1000062. doi: 10.1371/journal.pcbi.1000062
- Fischer, I. (2020). The conditional entropy bottleneck. *Entropy* 22:999. doi: 10.3390/e22090999
- Foldiak, P. (2003). “Sparse coding in the primate cortex,” in *The handbook of brain theory and neural networks*, ed. M. Arbib (Cambridge, MA: MIT Press).
- Friston, K., and Buzsáki, G. (2016). The functional anatomy of time: What and when in the brain. *Trends Cogn. Sci.* 20, 500–511. doi: 10.1016/j.tics.2016.05.001
- Ghazanfar, A. A., and Schroeder, C. E. (2006). Is neocortex essentially multisensory? *Trends Cogn. Sci.* 10, 278–285. doi: 10.1016/j.tics.2006.04.008
- Goldfeld, Z., and Polyanskiy, Y. (2020). The information bottleneck problem and its applications in machine learning. *IEEE J. Sel. Areas Inf. Theory* 1, 19–38. doi: 10.1109/JSAIT.2020.2991561
- Gu, T., Xu, G., and Luo, J. (2020). Sentiment analysis via deep multichannel neural networks with variational information bottleneck. *IEEE Access* 8, 121014–121021. doi: 10.1109/ACCESS.2020.3006569
- Hawkins, J., Ahmad, S., and Dubinsky, D. (2011). *Cortical learning algorithm and hierarchical temporal memory*. Redwood City, CA: Numenta.
- Hawkins, J., Ahmad, S., Purdy, S., and Lavin, A. (2016). *Biological and machine intelligence (bami)*. Initial online release 0.4. Available online at: <https://numenta.com/resources/biological-and-machine-intelligence/> (accessed May 27, 2023).
- Hoffmann, K., and Kunze, R. A. (1971). *Linear algebra*. Hoboken, NJ: Prentice-Hall.
- Hu, X., and Zeng, Z. (2022). Bridging the functional and wiring properties of V1 neurons through sparse coding. *Neural Comput.* 34, 104–137. doi: 10.1162/neco_a_01453
- Huang, J., Gu, K., Wang, Y., Zhang, T., Liang, J., and Luo, S. (2020). Connectivity-based localization in ultra-dense networks: CRLB, theoretical variance, and MLE. *IEEE Access* 8, 35136–35149. doi: 10.1109/ACCESS.2020.2974320
- Kaas, J. H. (2012). Evolution of columns, modules, and domains in the neocortex of primates. *Proc. Natl. Acad. Sci.* 109(Suppl. 1), 10655–10660. doi: 10.1073/pnas.1201892109
- Khorasani, S. M., Hodtani, G. A., and Kakhki, M. M. (2020). Decreasing Cramer-Rao lower bound by preprocessing steps. *Signal Image Video Process.* 14, 781–789. doi: 10.1007/s11760-019-01605-2
- Kwek, L.-C., Tan, A. W.-C., Lim, H.-S., Tan, C.-H., and Alaghbari, K. A. (2022). Sparse representation and reproduction of speech signals in complex Fourier basis. *Int. J. Speech Technol.* 25, 211–217. doi: 10.1007/s10772-021-09941-w
- Lee, H., Battle, A., Raina, R., and Ng, A. (2006). Efficient sparse coding algorithms. *Adv. Neural Inf. Process. Syst.* 19:801.
- Li, J., and Liu, D. (2021). Information bottleneck theory on convolutional neural networks. *Neural Process. Lett.* 53, 1385–1400. doi: 10.1007/s11063-021-10445-6
- Li, Q., Wang, W., Chen, G., and Zhao, D. (2021). Medical image fusion using segment graph filter and sparse representation. *Comput. Biol. Med.* 131:104239. doi: 10.1016/j.combiomed.2021.104239
- Li, Y. (2013). “Sparse representation for machine learning,” in *Proceedings of the Canadian conference on artificial intelligence* (Regina: Springer), 352–357. doi: 10.1007/978-3-642-38457-8_38
- Marois, R., and Ivanoff, J. (2005). Capacity limits of information processing in the brain. *Trends Cogn. Sci.* 9, 296–305. doi: 10.1016/j.tics.2005.04.010
- McClelland, T., and Bayne, T. (2016). Ensemble coding and two conceptions of perceptual sparsity. *Trends Cogn. Sci.* 20, 641–642. doi: 10.1016/j.tics.2016.06.008
- Menzel, R., and Giurfa, M. (2001). Cognitive architecture of a mini-brain: The honeybee. *Trends Cogn. Sci.* 5, 62–71. doi: 10.1016/S1364-6613(00)01601-6
- Mnatzaganian, J., Fokoué, E., and Kudithipudi, D. (2017). A mathematical formalization of hierarchical temporal memory's spatial pooler. *Front. Robot. AI* 3:81. doi: 10.3389/frobt.2016.00081
- Musat, B., and Andonie, R. (2022). Information bottleneck in deep learning—a semiotic approach. *Int. J. Comput. Commun. Control* 17:4650. doi: 10.15837/ijccc.2022.1.4650
- Nagahara, M., and Yamamoto, Y. (2022). “Sparse representation for sampled-data filters,” in *Realization and model reduction of dynamical systems*, eds C. Beattie, P. Benner, M. Embree, S. Gugercin, and S. Lefteriu (Cham: Springer), 427–444. doi: 10.1007/978-3-030-95157-3_23
- Olshausen, B. A., and Field, D. J. (1996). Wavelet-like receptive fields emerge from a network that learns sparse codes for natural images. *Nature* 381, 607–609. doi: 10.1038/381607a0
- Olshausen, B. A., and Field, D. J. (2004). Sparse coding of sensory inputs. *Curr. Opin. Neurobiol.* 14, 481–487. doi: 10.1016/j.conb.2004.07.007
- O'Reilly, R. C., and Norman, K. A. (2002). Hippocampal and neocortical contributions to memory: Advances in the complementary learning systems framework. *Trends Cogn. Sci.* 6, 505–510. doi: 10.1016/S1364-6613(02)02005-3
- Paiton, D. M., Frye, C. G., Lundquist, S. Y., Bowen, J. D., Zarcone, R., and Olshausen, B. A. (2020). Selectivity and robustness of sparse coding networks. *J. Vis.* 20:10. doi: 10.1167/jov.20.12.10
- Peng, J., Li, L., and Tang, Y. Y. (2018). Maximum likelihood estimation-based joint sparse representation for the classification of hyperspectral remote sensing images. *IEEE Trans. Neural Netw. Learn. Syst.* 30, 1790–1802. doi: 10.1109/TNNLS.2018.2874432
- Raj, V., Nayak, N., and Kalyani, S. (2020). Understanding learning dynamics of binary neural networks via information bottleneck. *arXiv [Preprint]*
- Rao, C. R., Rao, C. R., Statistiker, M., Rao, C. R., and Rao, C. R. (1973). *Linear statistical inference and its applications*. New York, NY: Wiley. doi: 10.1002/9780470316436
- Ravishanker, S., and Bresler, Y. (2012). Learning sparsifying transforms. *IEEE Trans. Signal Process.* 61, 1072–1086. doi: 10.1109/TSP.2012.2226449
- Riguzzi, F., and Di Mauro, N. (2012). Applying the information bottleneck to statistical relational learning. *Mach. Learn.* 86, 89–114. doi: 10.1007/s10994-011-5247-6
- Schneidman, E., Slonim, N., Tishby, N., de Ruyter van Steveninck, R. R., and Bialek, W. (2001). “Analyzing neural codes using the information bottleneck method,” in *Advances in neural information processing systems*, eds M. I. Jordan, Y. LeCun, and S. A. Solla (Boston, MA: MIT Press).
- Shwartz-Ziv, R., and Tishby, N. (2017). Opening the black box of deep neural networks via information. *arXiv [Preprint]*
- Stein, M., Mezghani, A., and Nossek, J. A. (2014). A lower bound for the fisher information measure. *IEEE Signal Process. Lett.* 21, 796–799. doi: 10.1109/LSP.2014.2316008
- Tishby, N., Pereira, F. C., and Bialek, W. (2000). The information bottleneck method. *arXiv [Preprint]*
- Tishby, N., and Zaslavsky, N. (2015). “Deep learning and the information bottleneck principle,” in *Proceedings of the 2015 IEEE information theory workshop (ITW)* (Jerusalem: IEEE). doi: 10.1109/ITW.2015.7133169
- Tucker, M., Shah, J., Levy, R., and Zaslavsky, N. (2022). Towards human-agent communication via the information bottleneck principle. *arXiv [Preprint]*
- Tune, P. (2012). Computing constrained Cramér-Rao bounds. *IEEE Trans. Signal Process.* 60, 5543–5548. doi: 10.1109/TSP.2012.2204258
- Vera, M., Vega, L. R., and Piantanida, P. (2022). Information flow in deep restricted Boltzmann machines: An analysis of mutual information between inputs and outputs. *Neurocomputing* 507, 235–246. doi: 10.1016/j.neucom.2022.08.014
- Wang, H., Ren, B., Song, L., and Cui, L. (2019). A novel weighted sparse representation classification strategy based on dictionary learning for rotating machinery. *IEEE Trans. Instrum. Meas.* 69, 712–720. doi: 10.1109/TIM.2019.2906334
- Wang, X.-J. (2022). Theory of the multiregional neocortex: Large-scale neural dynamics and distributed cognition. *Annu. Rev. Neurosci.* 45, 533–560. doi: 10.1146/annurev-neuro-110920-035434
- Wei, X., Shi, Y., Gong, W., and Guan, Y. (2022). Improved image representation and sparse representation for face recognition. *Multimed. Tools Appl.* 81, 44247–44261. doi: 10.1007/s11042-022-13203-5
- Zhou, X., Zhou, H., Wen, G., Huang, X., Lei, Z., Zhang, Z., et al. (2022). A hybrid denoising model using deep learning and sparse representation with application in bearing weak fault diagnosis. *Measurement* 189:110633. doi: 10.1016/j.measurement.2021.110633
- Zuo, L., Dewey, B., Liu, Y., He, Y., Newsome, S., Mowry, E., et al. (2021). Unsupervised MR harmonization by learning disentangled representations using information bottleneck theory. *Neuroimage* 243:118569. doi: 10.1016/j.neuroimage.2021.118569