frontiers in Computational Neuroscience

ORIGINAL RESEARCH
published: 17 December 2015
doi: 10.3389/fncom.2015.00149

Neural Network Model of Memory Retrieval

Stefano Recanatesi[1], *Mikhail Katkov*[1], *Sandro Romani*[2] *and Misha Tsodyks*[1, 3]*

[1] Department of Neurobiology, Weizmann Institute of Science, Rehovot, Israel, [2] Janelia Farm Research Campus, Howard Hughes Medical Institute, Ashburn, VA, USA, [3] Department of Neurotechnologies, Lobachevsky State University of Nizhny Novgorod, Nizhny Novgorod, Russia

Human memory can store large amount of information. Nevertheless, recalling is often a challenging task. In a classical free recall paradigm, where participants are asked to repeat a briefly presented list of words, people make mistakes for lists as short as 5 words. We present a model for memory retrieval based on a Hopfield neural network where transition between items are determined by similarities in their long-term memory representations. Meanfield analysis of the model reveals stable states of the network corresponding (1) to single memory representations and (2) intersection between memory representations. We show that oscillating feedback inhibition in the presence of noise induces transitions between these states triggering the retrieval of different memories. The network dynamics qualitatively predicts the distribution of time intervals required to recall new memory items observed in experiments. It shows that items having larger number of neurons in their representation are statistically easier to recall and reveals possible bottlenecks in our ability of retrieving memories. Overall, we propose a neural network model of information retrieval broadly compatible with experimental observations and is consistent with our recent graphical model (Romani et al., 2013).

Keywords: attractor neural networks, recall, oscillations, memory, neural representations

OPEN ACCESS

Edited by:
Hava T. Siegelmann,
Rutgers University, USA

Reviewed by:
Paolo Del Giudice,
Italian National Institute of Health, Italy
Gianluigi Mongillo,
Paris Descartes University, France
Alberto Bernacchia,
Jacobs University Bremen, Germany

****Correspondence:***
Misha Tsodyks
misha@weizmann.ac.il

Received: *21 June 2015*
Accepted: *26 November 2015*
Published: *17 December 2015*

Citation:
Recanatesi S, Katkov M, Romani S and Tsodyks M (2015) Neural Network Model of Memory Retrieval. Front. Comput. Neurosci. 9:149. doi: 10.3389/fncom.2015.00149

1. INTRODUCTION

Human long-term memory capacity for names, facts, episodes and other aspects of our lives is practically unlimited. Yet recalling this information is often challenging, especially when no precise cues are available. A striking example of this deficiency is provided by classical studies of free recall, where participants are asked to recall lists of unrelated words after a quick exposure (Murdock, 1962; Kahana, 1996). Even for short lists of 5–10 words most of participants are unable to reproduce them without omissions (Murdock, 1960; Tulving, 1966).

Several influential models of recall were developed. Some of them were driven by the description of behavioral aspects (Glenberg and Swanson, 1986; Howard and Kahana, 1999; Davelaar et al., 2005; Brown et al., 2007); while others were rooted in biological principles (Grossberg and Stone, 1986; Ruppin and Yeshurun, 1991; Wong et al., 1991; Hasselmo and Wyble, 1997; Verduzco-Flores et al., 2012; Lansner et al., 2013).

According to the influential "search of associative memory" (SAM) model, items presented for recall acquire a set of mutual associations when stored temporarily in working memory buffers (Raaijmakers and Shiffrin, 1980). These acquired associations are then used to retrieve words from memory. SAM can be fit to reproduce recall data with great precision (Raaijmakers and Shiffrin, 1981), but since it has many parameters it cannot provide the first-principle explanation for very limited recall capacity observed in experiments. A recent model of memory retrieval (Romani et al., 2013; Katkov et al., 2015) introduced the notion that long-term associations between items determined by overlaps between their neuronal representations in memory networks, rather than short-term associations acquired during the experiment, are primarily responsible for recall process. With a simple phenomenological implementation of recall, this assumption results in a generic limit for the recall capacity compatible with the data (Romani et al., 2013). Moreover, the neuronal representations determine the recall probability of different items ("easy" vs. "difficult" words) and the order of their recall. In the current contribution, we develop a more realistic neural network model where recall is mediated by the sequential reactivation of neuronal ensembles encoding different items in memory. We show existence of stable states of the network corresponding to the activation of neuronal ensembles encoding single memory state and to the activation of intersection of neuronal ensembles encoding two or more memory states. We identify these different phases of the model with mean-field analysis of the network dynamics. We, further, show that the transitions between the memories may be driven by periodic modulation of the feedback inhibition that pushes the network to oscillate between the attractor memory states and intersections between these states, as was suggested in Romani et al. (2013). We identify these different phases of the model with mean-field analysis of the network dynamics. Based on this analysis we perform extensive numerical simulations to characterize the recall behavior of the model. In addition, we modeled short-term associations between memory items formed during the acquisition and characterize their effects. Finally, we systematically characterize the effects of neuronal noise on recall. The main goal of this paper is to present the network model of recall compatible with well-known features of free recall observed over many years of research. Some of the predictions of the model are also tested against a large recent dataset of free recall that was collected and made available by the lab of Prof. Kahana from University of Pennsylvania (see details in Section 2).

2. MATERIALS AND METHODS

2.1. The Dynamics

We consider a Hopfield neural network of N rate-neurons (Hopfield, 1984, see also Grossberg, 1988). The dynamics of neuron i is represented by the equation:

$$\tau \dot{c}_i(t) = -c_i(t) + \sum_{j=1}^{N} J_{ij} \cdot r_j(t) + \xi_i(t), \tag{1}$$

$$r_i = g(c_i)\,. \tag{2}$$

where c, r are respectively the synaptic currents and the firing rates, J the connectivity matrix, each $\boldsymbol{\xi}_i$ is an independent random variable having a gaussian distribution with mean zero and variance ξ_0 and τ is a constant[1].

The gain function is:

$$\begin{cases} g = (x+\theta)^{\gamma} & x+\theta > 0\,, \\ g = 0 & x+\theta \leq 0\,. \end{cases} \tag{3}$$

where $\theta > 0$ is a threshold for the activation of a neuron while γ defines the gain and is constrained to $\gamma < 1$ for the gain function to be sublinear.

Each of the P memory items is represented by bynary vectors of N bits:

$$\eta^{\mu\in\{1..P\}} = \underbrace{100011101001..1001}_{\text{N neurons}}\,. \tag{4}$$

where each bit has an indipendent random binary value, being 1 with probability f and 0 with probability $1-f$ (Kanerva, 1988; Treves and Rolls, 1991). We use these vectors to define the connectivity matrix J according to the Hebbian rule (Tsodyks, 1989):

$$J_{ij} = \frac{\kappa}{N}\left(\sum_{\mu=1}^{P}(\eta_i^{\mu}-f)(\eta_j^{\mu}-f) - \varphi\right)\,. \tag{5}$$

where κ and φ are two parameters that respectively define the strength of excitation and the relative strength of inhibition in the network. When simulating the network, all parameters are held constant except for the relative strength of the inhibition φ. We say that a particular memory is "recalled" when the corresponding memory pattern is active.

Memory μ is recalled if the average firing rate of neurons corresponding to memory μ (i such that $\eta_i^{\mu} = 1$) is above the threshold value r_{thresh}. This threshold is chosen so that two memories are never recalled simultaneously. If in a given time interval, e.g., from time 0 to T, the state of the network was in memories μ_1, μ_2, μ_3.. at different times, we say that the network has "retrieved" these memories in a time T.

A slight modification of the model allows to account for short-term associations as in the SAM model. For example, temporal contiguity is the tendency to recall neighboring presented items in temporal proximity. To account for this effect we add a new term to the connectivity matrix J_{ij}:

$$J_{ij}^{+-} = J_{ij} + \delta J_{ij} = J_{ij} + J_{+}\sum_{\mu=1}^{P-1}\eta_i^{\mu}\eta_j^{\mu+1} + J_{-}\sum_{\mu=2}^{P}\eta_i^{\mu}\eta_j^{\mu-1}. \tag{6}$$

The new part δJ_{ij} consists of two terms which respectively connect a given memory μ with the memories presented immediately before and after it ($\mu-1$ and $\mu+1$) (Sompolinsky and Kanter, 1986; Griniasty et al., 1993). In doing so the memories are chained one to the other in the 'forward' and 'backward' direction with an asymmetry which depends on the values of J_+ and J_-.

2.2. Meanfield Theory

We analyze the network in the absence of noise ($\xi_0 = 0$) and temporal contiguity ($J_+ = J_- = 0$). To quantify the degree

[1] Here and in the following a bold character, e.g., $\boldsymbol{c}$, identifies the entire vector

of memory activations we introduce the "overlaps" defined as in Amit and Tsodyks (1991):

$$\begin{cases} m^{\mu}(t) = \frac{1}{N}\sum\limits_{i=1}^{N}(\eta_i^{\mu} - f)r_i(t)\,, \qquad \mu \in \{1..P\} \\ m^0(t) = \frac{1}{N}\sum\limits_{i=1}^{N} r_i(t)\,. \end{cases} \tag{7}$$

While $m^0(t)$ measures the average firing rate in the network at time t, each $m^{\mu}(t)$ measures the difference between the average firing rate of neurons encoding memory μ and all other neurons:

$$\begin{aligned} m^{\mu}(t) &= \frac{1}{N}\sum_{i=1}^{N}(\eta_i^{\mu} - f)r_i(t) = \\ &= \frac{1}{N}\sum_{i=1}^{N}\left((1-f)\eta_i^{\mu}r_i(t) - f(1-\eta_i^{\mu})r_i(t)\right) = \\ &= (1-f)f\sum_{i=1}^{N}\frac{\eta_i^{\mu}r_i(t)}{fN} - \frac{(1-\eta_i^{\mu})r_i(t)}{(1-f)N}\,. \end{aligned} \tag{8}$$

At a fix point of the network dynamics (Equation 2) the synaptic currents can be expressed via the values of the overlaps:

$$c_i = \sum_{j=1}^{N} J_{ij}r_j = \sum_{\mu=1}^{P}\kappa((\eta_i^{\mu} - f)m^{\mu} - \varphi m^0)\,, \tag{9}$$

$$r_i = g(c_i)\,; \tag{10}$$

given by Equation (2) in Equation (7). This shows that one can calculate r for each neuron i given the set of $m's$. Pluggin Equation (10) into Equation (7) we obtain a a system of $P + 1$ equations for the overlaps $m's$. The solutions to such a system are the possible fixed points of the network. Consider a vector $\boldsymbol{\eta}_i \in \{0, 1\}^P$ representing the encoding of each memory item by neuron i. There are 2^P possible realizations of vector $\boldsymbol{\eta}_i$ that are denoted by a random vector $\nu \in \{0, 1\}^P$ where each component is indipendent from any other being 1 with probability f and zero otherwise. Each realization of ν identifies a population of neurons. We say that neuron i belongs to a population ν if $\boldsymbol{\eta}_i = \nu$ that is $\eta_i^{\mu} = \nu^{\mu}\ \ \forall\mu$. Furthermore, we say that a population ν belongs to a memory μ if $\nu^{\mu} = 1$ (Curti et al., 2004).

The cardinality of a vector is defined as

$$|\nu| = \sum_{\mu}\nu^{\mu}\,. \tag{11}$$

The probability for each vector ν is:

$$S_{\nu} = (1-f)^{P-|\nu|}\cdot f^{|\nu|}\,, \tag{12}$$

while the synaptic current for each neuron in population ν is:

$$c_{\nu} = \sum_{\nu=1}^{P}\kappa((\nu^{\nu} - f)m^{\nu} - \varphi m^0)\,. \tag{13}$$

The fixed point solutions can then be characterized in the limit $N \to \infty$ in terms of these population vectors. Plugging Equation (10) into Equation (7) and summing up we obtain in the limit $N \to \infty$:

$$\begin{cases} m^{\mu} = \left\langle(\nu^{\mu} - f)\cdot g\left(c_{\nu}\right)\right\rangle_{\nu} \\ m^0 = \left\langle g\left(c_{\nu}\right)\right\rangle_{\nu}\,. \end{cases} \tag{14}$$

where the average can be expressed in terms of the probability S_{ν} as:

$$\begin{cases} m^{\mu} = \sum\limits_{\nu}(\nu^{\mu} - f)S_{\nu}\cdot g\left(c_{\nu}\right) \\ m^0 = \sum\limits_{\nu} S_{\nu}\cdot g\left(c_{\nu}\right) \end{cases} \tag{15}$$

This system determines the fixed points of the network in the meanfield limit. It cannot be solved in general but for a given *ansatz* of the solution it is possible to determine the region, in the parameter space, for its existence and stability. The type of solutions that we analyze are those that represent either a single memory or the intersection between memories. The correct *ansatz* for these solutions are easily expressed in terms of the synaptic currents. A single memory solution is then defined by the following conditions:

- the currents to each population ν that belongs to the active memory μ are uniformly above threshold $c_{\nu} + \theta > 0$ if $\nu^{\mu} = 1$;
- the currents to each population that doesn't belong to the active memory μ are below threshold $c_{\nu} + \theta < 0$ if $\nu^{\mu} = 0$;

This two conditions define our ansatz for a single memory state. From this definition it follows that in the state of single memory the the only overlap m different from zero is the one of the active memory m^{μ}. Similarly we define the ansatz for the intersection between two or more memories. In this state only two overlaps m are different from zero. For each of these ansatz one can find its region of existence and stability in parameter space. In such a region the solution is steady state of the system. A detailed theoretical analysis of these regions goes beyond the scope of this paper and will be presented in a future publications.

2.3. Simulation Technique

To study the influence of finite size effects and noise on the dynamics of the network we simulate the dynamic of a network of $N = 10^5$ neurons. To achieve this goal we simplify the system in Equation (2). This is a dimensionality reduction of the network that reduces the number of simulated units. All the neurons that have the same vector $\boldsymbol{\eta}_i$ (i.e., are in the same population ν such that $\boldsymbol{\eta}_i = \nu$) can be described by a single unit. For these neurons the afferent connections given by the matrix J are identical. Each neuron receives the same input and projects equally on other neurons. It is not possible to differentiate their activity except for the effect of the noise term $\boldsymbol{\xi}$. But in Equation (2) we can average terms which share the same connections averaging also their noise. For a given realization of the network we can write the fraction of neurons in a given population ν as:

$$S_{\nu} = \frac{1}{N}\times\{\textit{number of i such that}\ \boldsymbol{\eta}_i = \nu\}\,, \tag{16}$$

which converges to the definition of Equation (12) in the limit of $N \rightarrow \infty$. Defining $c_v(t)$, the averaging synaptic current $c(t)$ for a neuron in population v at time t, it is then possible to write an equation for the dynamics of $c_v(t)$. By summing Equation (2) over all neurons which belong to the same population v we obtain:

$$\dot{c}_v(t) = -c_v(t) + \sum_w \tilde{J}_{vw} \cdot S_w \cdot g(c_w(t)) + \tilde{\xi}_v(t)\,, \qquad (17)$$

where $\tilde{\xi}_v$ is a gaussian white noise with mean zero and amplitude $\tilde{\xi}_v = \xi_0 \cdot S_v \cdot N$, while $\tilde{J}_{vw}$ is given by:

$$\begin{aligned}\tilde{J}_{vw} &= \frac{\kappa}{N}\sum_{\mu=1}^{P}((v^{\mu}-f)(w^{\mu}-f)-\varphi) + J_{+}\sum_{\mu=1}^{P-1} v^{\mu} w^{\mu+1} \\ &+ J_{-}\sum_{\mu=2}^{P} v^{\mu} w^{\mu-1}\,. \end{aligned} \qquad (18)$$

The vectors v and w are binary vectors of length P identifying different populations. The system of Equation (17) is a reduction of the original system of Equation (2), it has 2^P equations instead of the N. In this reduction the only piece of information which is not accessible is the precise value of the firing rate of each single neuron. Only the average firing rate of the population it belongs to is now accessible. The actual number of equations to simulate depends on the particular realization of the network given by the choice of $\eta^{\mu\in\{1..P\}}$. Although in principle the system has 2^P equations, in practice, due to the finite size of the network and its sparse connectivity, there are much less populations since $S_v = 0$ for most v (Curti et al., 2004). The total number of equations in the system will depend on N and f but will always be less than N, tending to N only for very large P. In this framework, for $P = 16$, we are able to simulate easily a large network of $N = 10^5$ neurons. Indeed taking $f = 0.1$, the number of equations to simulate drops from the 10^5 of the original system in Equation (2) to the $\approx$1000 of the reduced one of Equation (17).

Simulations are run according to Equation (17) employing the parameters in **Table 1**. The number of simulated networks is N_{trials} . For each simulation the network is initialized in the state of a single, randomly chosen memory μ. In this state all the populations v which belong to memory μ are initialized to a rate r_{ini} while the others are initialized to a zero rate. In the model the transitions between memories are triggered by oscillations of the variable φ. This oscillates sinusoidally between the values φ^{max} and φ^{min}. The oscillations have a period τ_o which is much larger than τ so that the network is undergoing an adiabatic process. Integrations of Equation (17) are performed with the Euler method with a time step of dt and the simulated interval is $[0..T]$. The total number of cycles of oscillations is T/τ_o.

TABLE 1 | Reference values for the parameters in the simulation.

	Parameters and hyperparameters	
Name	**Description**	**Value**
N	Number of neurons	100,000
P	Number of memories	16
f	Sparsity	0.1
τ	Decay time	0.01
κ	Excitation parameter	13, 000
φ^{max}	Max inhibition parameter	1.06
φ^{min}	Min inhibition parameter	0.7
γ	Gain function exponent	2/5
θ	Gain function threshold	0
τ_O	Oscillation time	1
T_{tot}	Total time	450
dt	Integration time step	0.001
J_+	Forward contiguity	1500
J_-	Backward contiguity	400
ξ_0	Noise variance	65
r_{thresh}	Recall threshold	15
N_{trials}	Number of trials	10, 000
r_{ini}	Initial rate	1

2.4. Experimental Methods and Data Analysis

The data analyzed in this manuscript were collected in the lab of M. Kahana as part the Penn Electrophysiology of Encoding and Retrieval Study. Here we analyzed the results from the 141 participants (age 17–30) who completed the first phase of the experiment, consisting of 7 experimental sessions. Participants were consented according the University of Pennsylvanias IRB protocol and were compensated for their participation. Each session consisted of 16 lists of 16 words presented one at a time on a computer screen and lasted approximately 1.5 h. Each study list was followed by an immediate free recall test. Words were drawn from a pool of 1638 words. For each list, there was a 1500 ms delay before the first word appeared on the screen. Each item was on the screen for 3000 ms, followed by jittered 800–1200 ms inter-stimulus interval (uniform distribution). After the last item in the list, there was a 1200–1400 ms jittered delay, after which the participant was given 75 s to attempt to recall any of the just-presented items. Only trials without errors (no intrusions and no repeated recalls of the same words) were used in the analysis.

We analyze this dataset to validate our model. We investigated several aspects of the dataset as described in Katkov et al. (2014, 2015). Here we show the plots concerning semantic similarity in **Figures 5B,D**. Of all the trials we exclude those where items not belonging to the presented list were reported (intrusions) and those where at least one word was retrieved twice (repetitions). For each list we then associate to each pair of words their LSA score as obtained from online datasets. We then consider the pairs formed by orderly associating two consecutively reported items. For each of these pairs we obtain the transition rank by ranking the LSA score the pair among all the scores of the first item with any other word in the list. As there are 16 words the maximum rank is 15 and the minimum is 1. This is the quantity shown on the x-axis of **Figure 5B**.

For each pair of consecutive reported items we compute the IRT by the difference of their times of retrieval. This is the quantity shown on the y-axis of **Figure 5B** vs. the LSA score of the same pair.

3. RESULTS

3.1. Meanfield Theory Vs. Network Simulations

The main principle of recall that was suggested in Romani et al. (2013) is that externally generated control signal, expressed in periodic modulation of the strength of feedback inhibition, drives the network to oscillate between two states; one state is characterized by activation of single attractors, which correspond to a recall of the corresponding item (Hasselmo and Wyble, 1997; Gelbard-Sagiv et al., 2008; Romani et al., 2013); the second state is the intersection between pairs of attractors, which is a step toward transitions between different items. In this way each retrieved item acts as an internal cue for the next one (Raaijmakers and Shiffrin, 1981). Here we use the meanfield analysis of the network (see Section 2) to confirm that these two state types are indeed present. We identify the parameter regimes for their existence and stability. The meanfield theory greatly simplifies the analysis of the network by reducing the dynamics from that of single neurons (Equation 2) to overlaps, which are variables that describe the degree to which the network state corresponds to one of the memory attractors (see Equations 7 and 15 in Section 2). In the state of single attractors, only one overlap is positive while other ones are zeros. In the intersection states, pairs of overlaps are positive. We therefore use the meanfield equations that determine the possible values of overlaps (Equation 15) to find solutions corresponding to the intersection of Q memories. These solutions are characterized by Q positive overlaps: $m^1 = ... = m^Q = m^{active}$. The overlaps have all the same values as all the active neurons in the intersection of Q memories fire at the same firing rate. The precise solution depends on the choice of the gain function in Equation (3). For concreteness, we chose a saturating gain function with threshold, with the exponent of $\gamma = 1/2$ that allows analytical solution. The solution to Equation (15) is

$$\begin{cases} m^0 = \frac{1}{2}\left(k^2 f^{2Q}\left((f-1)^2 Q - \varphi\right) + \right. \\ \qquad\qquad \left. \sqrt{k^4 f^{4Q}\left(\varphi - (f-1)^2 Q\right)^2 + 4\theta k^2 f^{2Q}}\right) , \\ m^{active} = (1-f)\cdot m^0 , \\ m^{inactive} = 0 \end{cases} \qquad (19)$$

where m^{active} and $m^{inactive}$ are respectively the value of the overlap for an active and inactive memory and m^0 denotes the average activity of the network. f denotes the sparseness of memory representations, k scales the strength of the recurrent associative synapses and φ defines the relative strength of inhibition, **Figure 1A** (see Section 2 for more details). The existence of these solutions requires the term in the square root to be positive, which results in the phase diagram shown in **Figure 1B**. Increasing the relative strength of feedback inhibition, the network state goes from the regime with only single attractor states to the one where single attractor and intersection of pairs of attractors coexist. More elaborated analysis of stability, which will be presented elsewhere, shows that these solutions are stable in the whole region of their existence, but the relative stability of single attractor states relative to the intersection states is decreasing with the increase in φ.

Based on this analysis, we simulate the network while modulating the inhibition to cause the transitions between these two states (see Section 2 for details of simulations). We also add noise in order to trigger the transitions to the intersections between two attractors when inhibition rises. To mimic the experimental protocol (see Section 2), we simulate multiple recall trials where random samples of 16 items are selected for each trial. One sample epoch of simulations is shown in **Figures 2A,B**.

Each of the colored line in **Figure 2B** shows the average firing rate of neurons representing a specific memory. When one of these is above the threshold value of r_{thresh} we regard the corresponding memory as retrieved. We note that the precise sequence of retrieved items is not predictable for a given list of presented words, as it strongly depends on the first item being recalled (here assumed to be chosen randomly) and is sensitive to noise.

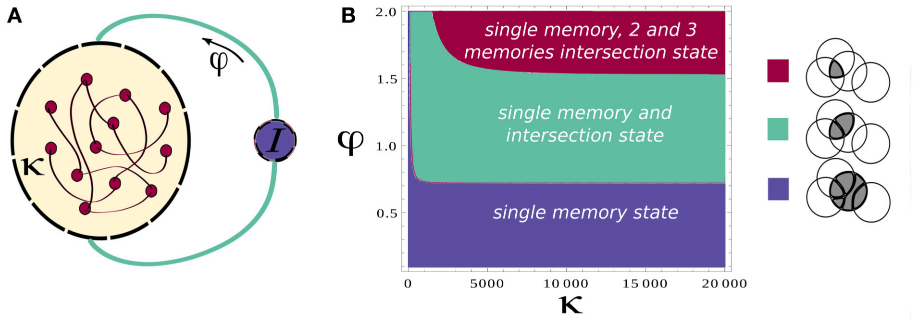
A
φ
I
κ
B
2.0
1.5
1.0
0.5
φ
0
5000
10 000
15 000
20 000
κ
single memory, 2 and 3 memories intersection state
single memory and intersection state
single memory state

FIGURE 1 | Network architecture and Mean-field phase diagram. (A) Neurons in the network are connected through simmetric connections induced by hebbian learning. Homeostatic control is induced by the inhibition strength determined by φ. **(B)** Mean-field phase diagram for the parameters κ and φ. The legend illustrates different phases. Circles denote a pool of neurons encoding a particular memory. For low values of φ the single attractor solution is found, as φ is increased other solutions appear. Parameters values are according to **Table 1**.

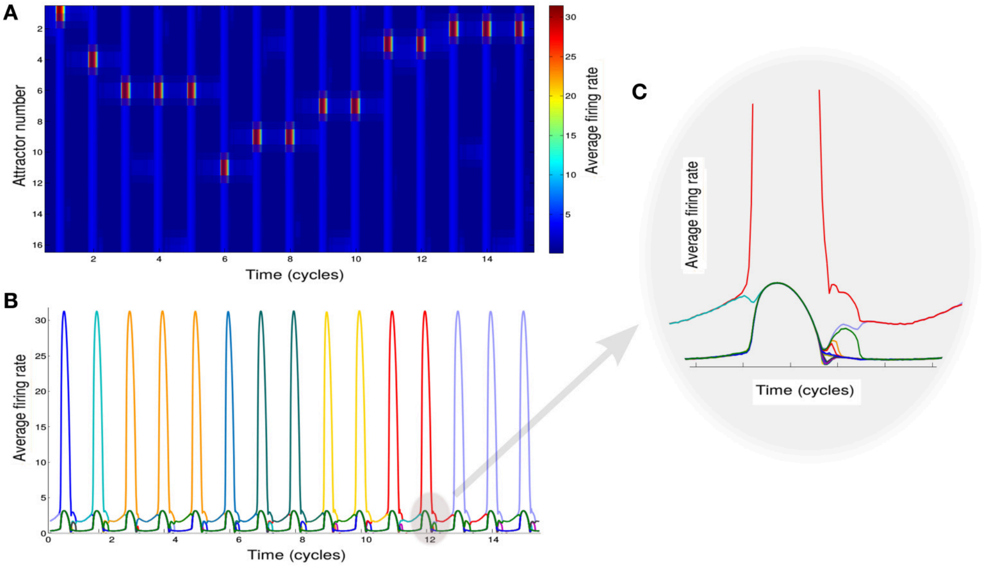

FIGURE 2 | Neural network activity. (A) Activity of the attractors in the network. Different rows correspond to the average firing rate of different memories for 15 cycles of oscillation of φ. **(B)** Activity of the attractors in the network. Each colored line correspond to the average firing rate of a different memory. **(C)** Details of the neuronal dynamics.

The effect of the oscillations is to modulate the overall activity in such a way that at each cycle the state of the network can potentially move from one attractor to another. The details of the underlying dynamics are shown in the plot of **Figure 2C** which zooms on the shadowed region in **Figure 2B** to show the transition from a single attractor to an intersection. This will lead to the retrieval of a new memory.

Although a switch between different states of the network is induced at every oscillation cycle, not always the state of the network shifts toward a new memory (**Figures 2A,B**). Rather it can remain in the same state or shifts toward an already explored memory so that only stochastically new memories are retrieved.

3.2. Time Course of Retrieval

Since the recall of subsequent memories is a stochastic process triggered by noise in the input, we perform multiple simulations to characterize the average accumulation of recalled memories with time (**Figure 3A**). We observe that after a quick initial accumulation, the retrieval process slows down sharply, however the number of memories recalled continues to increase. This behavior is compatible with experimental observations (Rohrer and Wixted, 1994; Wixted and Rohrer, 1994) and with results obtained by stochastic implementation of the free recall model presented in Katkov et al. (2015). The time between the recall of subsequent items (inter-retrieval time, IRT) is highly variable as shown in **Figure 3B**. Even after very long time-intervals it is possible to retrieve new items, in line with the experimental findings. We note that while the average accumulation curve is monotonic and smooth, each trial is characterized by a highly irregular set of IRTs, with short IRT interspersed between long ones due to cyclic transitions between items with relatively large overlaps. This is broadly consistent with experimental data (results not shown). Following the experimental study of Murdock and Okada (1970), Rohrer and Wixted (1994), we analyzed the average time progression of recall for trials with a certain number of words recalled (in a time window of 500 oscillation cycles). An interesting observation is that the corresponding curves separate already at the beginning of the recall, i.e., in the trials where more items are recalled eventually, the recall begins faster than in less successful trials, **Figure 3C**. This observation is also in line with the experimental results and with the stochastic model of Katkov et al. (2015).

3.3. Effects of Long-Term Memory Representations

Here we study the dependence of the recall process on the statistics of memory representations as defined by the memory patterns introduced in Section 2 (see Equation 4). In particular we consider the effects of representation size (number of neurons encoding a given item) and the size of intersections between the representations of two memories (number of neurons encoding both of the items). The representation size higly influences the probability of recall for a given memory. Our simulations show that simulating the network many times with items having a randomly drawn size, the probability to recall an item is monotonically increasing with the size of the corresponding representation (**Figure 4**). This is predominantly due to the fact that items represented by more neurons have on average a larger intersections with other items, since we assumed random encoding. Indeed as we show below, the intersection sizes play a major role in determining the subsequent items to be recalled. Therefore, our model is in agreement with the graph model of Romani et al. (2013), Katkov et al. (2015) where items with larger

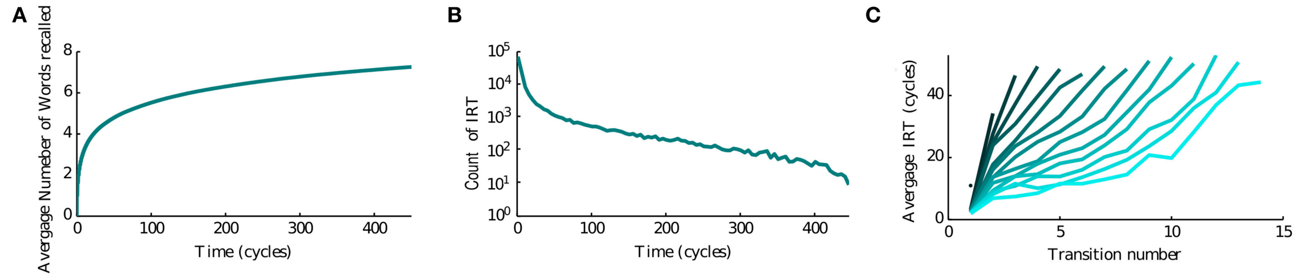

FIGURE 3 | Temporal properties of recall. (A) Average number of words recalled vs. time. **(B)** Distribution of the IRTs. **(C)** IRT average (y axis) for ordered transitions between words (x-axis). Each line represents the average over the set of trials in which a different number of words were recalled: going from left to right (or dark to light blue) less to more words up to the maximum of 16.

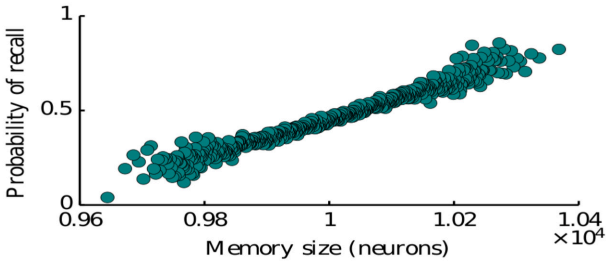

FIGURE 4 | Probability of recalling an item of a given size. The size is the number of neurons encoding for that particular memory.

representations have higher probability to be recalled (easy vs. difficult items).

Intersections between memory representations play a crucial role in our model of recall. In Romani et al. (2013) intersection sizes (the number of neurons encoding a pair of items) were assumed to govern the transitions between the recall of successive items. To evaluate the role of intersection sizes in the transitions between items we ranked intersection sizes for each presented list of 16 words, from low to high (1–15), and computed the probability of transition for each intersection rank (**Figure 5A**). Thirty percent of transitions occurred for largest intersection with the currently recalled item, the probability of other transitions monotonically decreases with the rank of intersections. Moreover, we found that the inter-recall time between the successive items also exhibited monotonic relation to the intersection size, with larger intersections leading to faster transitions (**Figure 5C**). These results indicate that the sizes of inter-item neuronal intersections to a large extent determine the temporal evolution of recall. It is therefore tempting to speculate that they are neuronal correlates of semantic similarity between the items (Baddeley, 1966; Mandler et al., 1969; Howard and Kahana, 2002b). To further elaborate on this hypothesis, we analyzed the dataset of free recall of lists of unrelated words collected and made available by Prof. Kahana from the University of Pennsylvania. We considered a measure of semantic similarity called (*Latent Semantic Analysis*, or LSA), which represent the number of times two words appear together in a representative corpora of natural text (Landauer and Dumais, 1997). We then used this measure to evaluate the effect of semantic similarity on the probability and speed of inter-item transitions in experimental observations, and obtained a remarkable agreement with the corresponding model predictions (compare **Figures 5A,C** with **Figures 5B,D**).

3.4. Performance

We now focus on factors which influence the recall performance, namely the number of items that can be retrieved in a given time window, between time 0 and time T. This window is chosen to be long enough such that the recall slowed down considerably (see **Figure 3A**). In particular we will consider the effects of temporal contiguity and noise.

The performance of the network is limited as item representations that control the retrieval dynamics are random and hence same items are recalled numerous times before the network can retrieve a new memory. It is known however that the order of recall is not completely random, e.g., words that have neighboring positions in the list have a tendency to be recalled in close proximity (Sederberg et al., 2010). This phenomenon is known as temporal contiguity and we model it by adding a special term in the connectivity pattern that links neighboring items to each other favoring the transitions between them (see Section 2, Equation 6), thereby overcoming the effects of randomness. Hence when the forward contiguity term is stronger, the network retrieves more items (**Figure 6A**). Although if it is too strong it becomes the only mechanism for triggering a transition and the average number of items retrieved will be half of the total number (8 items in **Figure 6A**). Indeed in this regime the network retrieves all items that come after the random initial one. Once it retrieves the last presented item it keeps retrieving it. The loop of connectivities via the second last item, which strongly projects on it, prevents the activation of any other memory.

Another crucial element of the model is the noise that causes the recall dynamics to escape the short loops and retrieve new items. We thus computed the network performance for increasing noise levels (**Figure 6B**). As expected, the performance is very poor for low noise amplitudes and increases for higher amplitudes. This growth is terminated at some optimal level of noise, after which the number of recalled items is slowly decreasing. The reason for this behavior is that at high noise

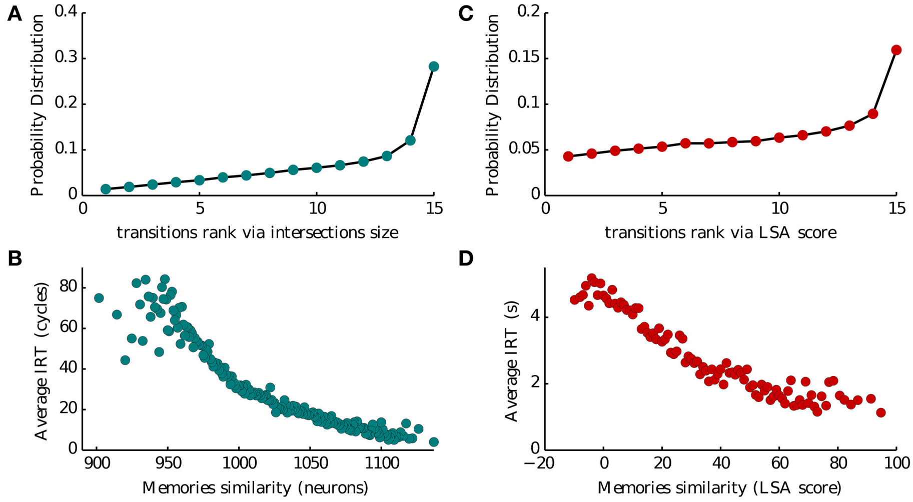

FIGURE 5 | Memory transitions. (A,C) Probability density of transitions between two subsequent recalled memories as a function of the ranked size of their intersection (1–15 going from the less to the most similar) and of their Latent Semantic Analysis score (LSA). **(B,D)** Average IRT between two subsequent recalled memories as a function of the size of their intersection (in number of neurons) and of their Latent Semantic Analysis score (LSA).

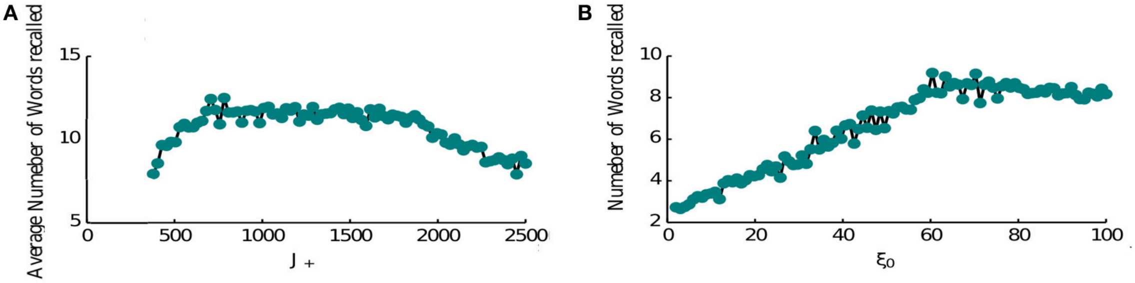

FIGURE 6 | Recall performance. (A) Temporal contiguity and performance: average number of words recalled as a function of J_+. J_+ ranges between the fixed value of $J_- = 400$ and 2500. The number of memories is $P = 16$. The number of memories is $P = 16$. **(B)** Noise and performance: the average number of words recalled is plotted at the vary of noise variance ξ. A small amount of noise helps the retrieval process triggering transitions from memory to memory. For high noise levels the retrieval mechanism is hindered.

levels, the network does not converge to inter-item intersections at high levels of inhibition, rather to noisy mixtures of different memories, which results in less robust transitions to other items when inhibition is reduced (results not shown). Based on these observations, we propose that noise amplitude could be regulated during the retrieval phase (e.g., with neuromodulators that control cortical synchrony) in order to facilitate the recall of items from long-term memory.

4. DISCUSSION

We presented a neural network model of information retrieval from long-term memory that is based on stochastic attractor dynamics controlled by periodically modulated strength of feedback inhibition. The model provides a more realistic implementation of the mechanisms behind associative recall based on neuronal representations of memory items, as proposed in Romani et al. (2013); Katkov et al. (2015). The network behavior is broadly compatible with some of the long-standing observations on free recall, in particular the slow-down of recall speed, highly variable inter-recall times and strong effects of semantic similarity between words.

In classical models of recall, such as SAM (Raaijmakers and Shiffrin, 1980) or TCM (Howard and Kahana, 2002a; Polyn et al., 2009), performance is mainly influenced by the temporal associations acquired during stimulus presentation. These effects were also considered in a possible network implementation (Bradski et al., 1994). In contrast, our model is based on long-term memory representations. Simple modification of the model (see Equation 6) allows to account for the effect of temporal contiguity (Sederberg et al., 2010). Therefore, we show that effects due to long-term memory representations and to presentation order can be implemented in a single neural network. It is important to note that effects due to long-term representations are masked by temporal association effects, being visible only in large data sets having many trials over lists composed of randomly selected words from a large preselected pool of words. In such datasets the same word is roughly uniformly distributed across temporal positions and their neighborhood words. Consequently, temporal association effects on the level of individual words are averaged out, and effects due to long-term representations become clearly visible. There are two major effects that historically were not considered neither experimentally nor in models: (1) intrinsic difficulty of words to be recalled—existence of "easy" and "difficult" words for recall; (2) masking of "difficult" words by "easy" words—"easy" words are statistically recalled earlier in the trial and suppress the recall of "difficult" words (Katkov et al., 2015). This work is a first attempt to implement a neural network that is taking into account long-term representation of memorized items.

Our network model is based on the basic assumption that when a word is recalled, a corresponding neuronal ensemble that represents this word in long-term memory is temporarily activated. The issue that we dont explicitly address is how the words that are presented for recall are selected, or primed and why other word representations are not reactivated (excluding rare instances of erroneous recall of words from previous lists). In the spirit of Kahanas TCM model (Howard and Kahana, 2002a), such a priming could be mediated by the excitation arriving from a separate "context" network where representation of the experimental setting is active throughout the recall trial. We therefore ignored the neuronal representations of words that are not in the list and considered a network with effectively very low "loading" level ($P \ll N$). More realistic implementation of the model with high loading levels should be considered in future.

Another simplifying unrealistic assumption of the model concerns the statistics of long-term representations that are taken as random uncorrelated binary vectors of fixed average sparsity. Real statistics of word representations is not clear but can be safely assumed to be much more complicated, possibly reflecting the rich semantic associations between words and the frequency of their usage. With our assumptions, overlaps between different representations exhibit Gaussian distribution with variance to mean ratio decaying in the limit of infinitely large networks. Considering the effects of overlap distribution in this limit requires an extended mean-field analysis that will be presented elsewhere.

Very often the same attractor is repeatedly activated before noise causes the transition to a new one, and it can still be activated again at a later time. Since participants are instructed to only recall each word once, we assume that they suppress the report of a word after it is already recalled. In some experiments, subjects are explicitly instructed to report a word as many times as it comes to mind during a recall. Comparing the model to the results of such experiments could be of interest for a future work.

We considered modulated inhibition as a driving force for transitions between network attractors. Other mechanisms could potentially play this role, e.g., neuronal adaptation or synaptic depression. We believe that oscillatory mechanism is more plausible as it allows the system to regulate the transitions by controlling the amplitude and frequency of oscillations. The oscillations of network activity could correspond to increased amplitude of theta rhythm observed in human subjects during recall (Kahana, 2006; Osipova et al., 2006) and other types of working memory experiments (Tesche and Karhu, 2000; Raghavachari et al., 2001; Jensen and Tesche, 2002). The way we implemented feedback inhibition is not fully biologically plausible. Feedback inhibition in the cortex is mediated by several major types of interneurons (Markram et al., 2004). In particular, one type of interneurons (VIP), was proposed as a gateway for regulating the local inhibition since it receives inputs from remote cortical and subcortical regions and preferentially targets other types of interneurons (Pi et al., 2013). More realistic neural network models of recall should include this kind of inhibition.

At the current level of realism, we propose to view our model as a platform for further development of realistic neural network models of information retrieval and other related types of cognitive tasks. Future modifications should include effects of positional order on recall, or positional chunking, i.e., the tendency to divide the presented lists on groups of contiguous words (Miller, 1956; Gobet et al., 2001), as well as primacy (tendency to recall earlier words with higher probability, see e.g., Grossberg and Pearson, 2008), or effects obtained in serial recall, such as e.g., encoding gradient or similar tasks (Averbeck et al., 2002, 2003; Farrell and Lewandowsky, 2004), where participants are forced to recall items in presented order, implying stricter tests on temporal associations.

AUTHOR CONTRIBUTIONS

MT and SR designed the study; SR developed and simulated the model; MT, MK, and SR performed a mathematical analysis, SR and MK performed data analysis; all the authors wrote the paper.

ACKNOWLEDGMENTS

We are grateful to M. Kahana for generously sharing the data obtained in his laboratory with us. The lab of Kahana is supported by NIH grant MH55687. MT is supported by EU FP7 (Grant agreement 604102) and the Foundation Adelis. A part of this work related to network recall performance (Section 3.4) was supported by The Russian Science Foundation No.14-11-00693.

REFERENCES

Amit, D. J., and Tsodyks, M. V. (1991). Quantitative study of attractor neural networks retrieving at low spike rates: II. Low-rate retrieval in symmetric networks. *Netw. Comput. Neural Syst.* 2, 275–294. doi: 10.1088/0954-898X_2_3_004

Averbeck, B. B., Chafee, M. V., Crowe, D. A., and Georgopoulos, A. P. (2002). Parallel processing of serial movements in prefrontal cortex. *Proc. Natl. Acad. Sci. U.S.A.* 99, 13172–13177. doi: 10.1073/pnas.162485599

Averbeck, B. B., Chafee, M. V., Crowe, D. A., and Georgopoulos, A. P. (2003). Neural activity in prefrontal cortex during copying geometrical shapes. i. single cells encode shape, sequence, and metric parameters. *Exp. Brain Res.* 150, 127–141. doi: 10.1007/s00221-003-1417-5

Baddeley, A. D. (1966). Short-term memory for word sequences as a function of acoustic, semantic and formal similarity. *Q. J. Exp. Psychol.* 18, 362–365. doi: 10.1080/14640746608400055

Bradski, G., Carpenter, G. A., and Grossberg, S. (1994). Store working memory networks for storage and recall of arbitrary temporal sequences. *Biol. Cybern.* 71, 469–480. doi: 10.1007/BF00198465

Brown, G. D. A., Neath, I., and Chater, N. (2007). A temporal ratio model of memory. *Psychol. Rev.* 114, 539–576. doi: 10.1037/0033-295X.114.3.539

Curti, E., Mongillo, G., La Camera, G., and Amit, D. J. (2004). Mean field and capacity in realistic networks of spiking neurons storing sparsely coded random memories. *Neural Comput.* 16, 2597–2637. doi: 10.1162/0899766042321805

Davelaar, E. J., Goshen-Gottstein, Y., Ashkenazi, A., Haarmann, H. J., and Usher, M. (2005). The demise of short-term memory revisited: empirical and computational investigations of recency effects. *Psychol. Rev.* 112, 3–42. doi: 10.1037/0033-295X.112.1.3

Farrell, S., and Lewandowsky, S. (2004). Modelling transposition latencies: constraints for theories of serial order memory. *J. Mem. Lang.* 51, 115–135. doi: 10.1016/j.jml.2004.03.007

Gelbard-Sagiv, H., Mukamel, R., Harel, M., Malach, R., and Fried, I. (2008). Internally generated reactivation of single neurons in human hippocampus during free recall. *Science* 322, 96–101. doi: 10.1126/science.1164685

Glenberg, A. M., and Swanson, N. G. (1986). A temporal distinctiveness theory of recency and modality effects. *J. Exp. Psychol. Learn. Mem. Cogn.* 12, 3–15. doi: 10.1037/0278-7393.12.1.3

Gobet, F., Lane, P. C. R., Croker, S., Cheng, P. C.-H., 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

Griniasty, M., Tsodyks, M. V., and Amit, D. J. (1993). Conversion of temporal correlations between stimuli to spatial correlations between attractors. *Neural Comput.* 5, 1–17. doi: 10.1162/neco.1993.5.1.1

Grossberg, S. (1988). Nonlinear neural networks: principles, mechanisms, and architectures. *Neural Netw.* 1, 17–61. doi: 10.1016/0893-6080(88)90021-4

Grossberg, S., and Pearson, L. R. (2008). Laminar cortical dynamics of cognitive and motor working memory, sequence learning and performance: toward a unified theory of how the cerebral cortex works. *Psychol. Rev.* 115, 677–732. doi: 10.1037/a0012618

Grossberg, S., and Stone, G. (1986). Neural dynamics of word recognition and recall: attentional priming, learning, and resonance. *Psychol. Rev.* 93, 46–74. doi: 10.1037/0033-295X.93.1.46

Hasselmo, M. E., and Wyble, B. P. (1997). Free recall and recognition in a network model of the hippocampus: simulating effects of scopolamine on human memory function. *Behav. Brain Res.* 89, 1–34. doi: 10.1016/S0166-4328(97)00048-X

Hopfield, J. J. (1984). Neurons with graded response have collective computational properties like those of two-state neurons. *Proc. Natl. Acad. Sci. U.S.A.* 81, 3088–3092. doi: 10.1073/pnas.81.10.3088

Howard, M. W., and Kahana, M. J. (1999). Contextual variability and serial position effects in free recall. *J. Exp. Psychol. Learn. Memory Cogn.* 25, 923–941. doi: 10.1037/0278-7393.25.4.923

Howard, M. W., and Kahana, M. J. (2002a). A distributed representation of temporal context. *J. Math. Psychol.* 46, 269–299. doi: 10.1006/jmps.2001.1388

Howard, M. W., and Kahana, M. J. (2002b). When does semantic similarity help episodic retrieval? *J. Mem. Lang.* 46, 85–98. doi: 10.1006/jmla.2001.2798

Jensen, O., and Tesche, C. D. (2002). Frontal theta activity in humans increases with memory load in a working memory task. *Eur. J. Neurosci.* 15, 1395–1399. doi: 10.1046/j.1460-9568.2002.01975.x

Kahana, M. J. (1996). Associative retrieval processes in free recall. *Mem. Cogn.* 24, 103–109. doi: 10.3758/BF03197276

Kahana, M. J. (2006). The cognitive correlates of human brain oscillations. *J. Neurosci.* 26, 1669–1672. doi: 10.1523/JNEUROSCI.3737-05c.2006

Kanerva, P. (1988). *Sparse Distributed Memory*. Bradford: MIT Press.

Katkov, M., Romani, S., and Tsodyks, M. (2014). Word length effect in free recall of randomly assembled word lists. *Front. Comput. Neurosci.* 8:129. doi: 10.3389/fncom.2014.00129

Katkov, M., Romani, S., and Tsodyks, M. (2015). Effects of long-term representations on free recall of unrelated words. *Learn. Mem.* 22, 101–108. doi: 10.1101/lm.035238.114

Landauer, T. K., and Dumais, S. T. (1997). A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychol. Rev.* 104, 211–240. doi: 10.1037/0033-295X.104.2.211

Lansner, A., Marklund, P., Sikström, S., and Nilsson, L.-G. (2013). Reactivation in working memory: an attractor network model of free recall. *PLoS ONE* 8:e73776. doi: 10.1371/journal.pone.0073776

Mandler, G., Pearlstone, Z., and Koopmans, H. S. (1969). Effects of organization and semantic similarity on recall and recognition. *J. Verb. Learn. Verb. Behav.* 8, 410–423. doi: 10.1016/S0022-5371(69)80134-9

Markram, H., Toledo-Rodriguez, M., Wang, Y., Gupta, A., Silberberg, G., and Wu, C. (2004). Interneurons of the neocortical inhibitory system. *Nat. Rev. Neurosci.* 5, 793–807. doi: 10.1038/nrn1519

Miller, G. A. (1956). The magical number seven plus or minus two: some limits on our capacity for processing information. *Psychol. Rev.* 63, 81–97. doi: 10.1037/h0043158

Murdock, Jr., B. B. (1960). The immediate retention of unrelated words. *J. Exp. Psychol.* 60, 222–234. doi: 10.1037/h0045145

Murdock, Jr., B. B. (1962). The serial position effect of free recall. *J. Exp. Psychol.* 64, 482–488. doi: 10.1037/h0045106

Murdock, B. B., and Okada, R. (1970). Interresponse times in single-trial free recall. *J. Verb. Learn. Verb. Behav.* 86, 263–267. doi: 10.1037/h0029993

Osipova, D., Takashima, A., Oostenveld, R., Fernández, G., Maris, E., and Jensen, O. (2006). Theta and gamma oscillations predict encoding and retrieval of declarative memory. *J. Neurosci.* 26, 7523–7531. doi: 10.1523/JNEUROSCI.1948-06.2006

Pi, H.-J., Hangya, B., Kvitsiani, D., Sanders, J. I., Huang, Z. J., and Kepecs, A. (2013). Cortical interneurons that specialize in disinhibitory control. *Nature* 503, 521–524. doi: 10.1038/nature12676

Polyn, S. M., Norman, K. A., and Kahana, M. J. (2009). A context maintenance and retrieval model of organizational processes in free recall. *Psychol. Rev.* 116, 129–156. doi: 10.1037/a0014420

Raaijmakers, J. G., and Shiffrin, R. M. (1981). Search of associative memory. *Psychol. Rev.* 88, 93–134. doi: 10.1037/0033-295X.88.2.93

Raaijmakers, J. G. W., and Shiffrin, R. M. (1980). SAM: a theory of probabilistic search of associative memory. *Psychol. Learn. Motiv.* 14, 207–262. doi: 10.1016/S0079-7421(08)60162-0

Raghavachari, S., Kahana, M. J., Rizzuto, D. S., Caplan, J. B., Kirschen, M. P., Bourgeois, B., et al. (2001). Gating of human theta oscillations by a working memory task. *J. Neurosci.* 21, 3175–3183.

Rohrer, D., and Wixted, J. T. (1994). An analysis of latency and interresponse time in free recall. *Mem. Cogn.* 22, 511–524. doi: 10.3758/BF03198390

Romani, S., Pinkoviezky, I., Rubin, A., and Tsodyks, M. (2013). Scaling laws of associative memory retrieval. *Neural Comput.* 25, 2523–2544. doi: 10.1162/NECO_a_00499

Ruppin, E., and Yeshurun, Y. (1991). Recall and recognition in an attractor neural network model of memory retrieval. *Connect. Sci.* 3, 381–400. doi: 10.1080/09540099108946594

Sederberg, P. B., Miller, J. F., Howard, M. W., and Kahana, M. J. (2010). The temporal contiguity effect predicts episodic memory performance. *Mem. Cogn.* 38, 689–699. doi: 10.3758/MC.38.6.689

Sompolinsky, H., and Kanter, I. (1986). Temporal association in asymmetric neural networks. *Phys. Rev. Lett.* 57, 2861–2864. doi: 10.1103/PhysRevLett.57.2861

Tesche, C. D., and Karhu, J. (2000). Theta oscillations index human hippocampal activation during a working memory task. *Proc. Natl. Acad. Sci. U.S.A.* 97, 919–924. doi: 10.1073/pnas.97.2.919

Treves, A., and Rolls, E. T. (1991). What determines the capacity of autoassociative memories in the brain? *Netw. Comput. Neural Syst.* 2, 371–397. doi: 10.1088/0954-898X_2_4_004

Tsodyks, M. (1989). Associative Memory in neural networks with the Hebbian learning rule. *Modern Phys. Lett. B* 03, 555–560. doi: 10.1142/S021798498900087X

Tulving, E. (1966). Subjective organization and effects of repetition in multi-trial free-recall learning. *J. Verb. Learn. Verb. Behav.* 5, 193–197. doi: 10.1016/S0022-5371(66)80016-6

Verduzco-Flores, S. O., Bodner, M., and Ermentrout, B. (2012). A model for complex sequence learning and reproduction in neural populations. *J. Comput. Neurosci.* 32, 403–423. doi: 10.1007/s10827-011-0360-x

Wixted, J. T., and Rohrer, D. (1994). Analyzing the dynamics of free recall: an integrative review of the empirical literature. *Psychon. Bull. Rev.* 1, 89–106. doi: 10.3758/BF03200763

Wong, K. Y. M., Kahn, P. E., and Sherrington, D. (1991). A neural network model of working memory exhibiting primacy and recency. *J. Phys. A Math. Gen.* 24, 1119. doi: 10.1088/0305-4470/24/5/025

Conflict of Interest Statement: 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.