Face familiarity detection with complex synapses

Li Ji-An[1], Fabio Stefanini[1], Marcus K. Benna[1], and Stefano Fusi[1,*]

[1]Zuckerman Institute, Columbia University, New York, NY 10027, USA

[*]Correspondence: sf2237@columbia.edu

Abstract

Synaptic plasticity is a complex phenomenon involving multiple biochemical processes that operate on different timescales. We recently showed that this complexity can greatly increase the memory capacity of neural networks when the variables that characterize the synaptic dynamics have limited precision, as in biological systems. These types of complex synapses have been tested mostly on simple memory retrieval problems involving random and uncorrelated patterns. Here we turn to a real-world problem, face familiarity detection, and we show that also in this case it is possible to take advantage of synaptic complexity to store in memory a large number of faces that can be recognized at a later time. In particular, we show that the memory capacity of a system with complex synapses grows almost linearly with the number of the synapses and quadratically with the number of neurons. Complex synapses are superior to simple ones, which are characterized by a single variable, even when the total number of dynamical variables is matched. Our results indicate that a memory system with complex synapses can be used in real-world applications such as familiarity detection.

Keywords: memory capacity, face familiarity, familiarity detection, complex synapse, one-shot learning

Introduction

Synaptic memory is a complex phenomenon, which involves intricate networks of diverse biochemical processes that operate on different timescales. We recently showed that this complexity can be harnessed to increase the memory capacity greatly[1,2] in situations in

which the synaptic weights are stored with limited precision. More specifically, we proposed a complex synaptic model in which m variables that might correspond to different biochemical processes interact within each synapse such that the memory capacity of a population of synapses can increase almost linearly with its size (i.e., the number of synapses N_{syn}), even when both m and the number of states of each variable grow no faster than logarithmically with N_{syn}. This is the optimal scaling under some conditions (see[2]) and significantly better than what can be achieved by employing a simple synapse characterized by a single variable[3–5].

These previous studies on complex synapses focused on a class of problems that assumed that the memories are represented by random and uncorrelated patterns. Only recently, complex synapses started to be employed in more realistic problems (e.g., see[6]). Here we show that synaptic complexity can be important also in a real-world problem, face familiarity detection. The task is particularly difficult because we consider the version of the task in which each face is presented only once (one-shot learning) and has to be remembered for a long time. This is a typical situation in which complexity can play an important role. Indeed, the complex synapses of [2] that we incorporated into our model are characterized by dynamical variables that operate on multiple timescales. The fast ones can rapidly store information about a new visual stimulus such as a face, even when the stimulus is shown only once. This information is then progressively transferred to the slow variables, which can retain it for a long time. Because of these slow variables, which influence the synaptic efficacy, the older memories are protected from overwriting due to the storage of new faces. Synapses that are described by a single dynamical variable can either learn quickly if they are fast, but then they also forget quickly, or they can retain memories for a long time if they are slow, but then they cannot learn in one shot and require multiple stimulations. This plasticity-rigidity dilemma concerns a very broad class of realistic synaptic models whose dynamical variables have a limited precision[3,5,7].

Here we incorporated complex synapses into a neural network model that is able to perform face familiarity detection. Familiarity detection (sometimes called familiarity discrimination or novelty detection) is an important form of recognition memory. There are several biology-inspired computational models studying different aspects of recognition memory: some neural network models following the complementary learning systems approach were proposed to tease apart the hippocampal and neocortical contributions to recognition memory[8,9]; other models were concerned with the synaptic plasticity (learning) rules in the perirhinal cortex[10]. Finally, there are models that stress the distinct roles for familiarity and recollection in retrieving memories[11].

Analytical estimates of familiarity memory capacity showed that in the case of random uncorrelated patterns, the number of memories that can be correctly recognized as familiar

can scale quadratically with the number of neurons N in a recurrent network[12]. Not too surprisingly, this is a much better scaling than the linear scaling of the Hopfield model[13], in which random memories are actually reconstructed (see also the Discussion). The scaling is markedly worse and can be as low as $\sqrt{N}$ when the patterns representing the memories are correlated[10]. These computational models can replicate some interesting aspects of experiments on the capacity of human recognition memory[14].

We constructed a model for recognition memory that, for the first time, incorporates complex synapses characterized by variables that have limited dynamical range (number of distinguishable states). We show that a simple neural circuit designed to reconstruct the memorized face can take advantage of the complexity of synapses and can efficiently store a large number of faces. In particular, we show that the number of faces that can be successfully recognized as familiar scales approximately quadratically with the number of neurons, or linearly with the number of synapses. This is the same scaling achieved in[12], in which synaptic weights could be stored with unlimited precision. Moreover, this scaling is similar to the one predicted for random patterns in[2]. Interestingly, the network can recognize a face even when it is presented in a different pose, and the scaling is only slightly worse than in the case in which the exact same picture of the face is presented for familiarity testing. This ability to generalize is a distinctive feature of recognition memory and it is rarely modeled. We then compare the performance of the recognition system with complex synapses to one with the same architecture, but with a larger number of neurons and simple synapses characterized by a single dynamical variable. The number of neurons is chosen so that the total number of synaptic variables would be the same in the two systems. We show that the system with complex synapses outperforms the one with simple synapses, indicating that complexity provides a clear computational advantage.

Materials and methods

Face data set

We used a large-scale face data set called VGGFace2[15]. Compared to other public face data sets (such as Labelled Faces in the Wild data set[16], CelebFaces+ data set[17], VGGFace data set[18], MegaFace data set[19], and Ms-Celeb-1M data set[20]), it contains a relative large number of individuals (3.31 million images of 9131 individuals) and large intra-identity variations in pose, age, illumination and background (362.6 images per person on average), with available human-verified bounding boxes around faces. For each face image, the bounding box was then enlarged by 30% to include the whole head, resized such that the shorter side was 256

pixels long, and center-cropped to 224×224 pixels to serve as the input for our neural system described below.

Neural face familiarity detection system

Our face familiarity detection system consists of three modules: an input (embedding) module, a memory module, and a readout (detection) module.

Input (embedding) module

The embedding module consists of a deep convolutional neural network (SE-ResNet-50, SENet for short), which is a ResNet architecture integrated with Squeeze-and-Excitation (SE) blocks adaptively recalibrating channel-wise feature responses[21]. Such networks for face recognition with different architectures and different training protocols are publicly available online[15]. We used one specific version of SENet, which is pre-trained on the MS-Celeb-1M data set[20] and then fine-tuned on the VGGFace2 data set. This version was reported to have the best generalization power on face verification and identification among architectures (e.g., SENet and ResNet-50) and training protocols (e.g., training on different data sets with or without fine-tuning) tested[15].

The 2048 dimensional activity of the penultimate layer (adjacent to the classification layer) was extracted as the face feature vector for each face image input. Because the face feature vectors are sparse and non-negative, we took the following steps to transform them into a format that's suitable as the input to the memory module: (i) the dimensionality of the feature vector of each face was first reduced using principal component analysis (PCA); (ii) each dimension was then binarized with a threshold equal to the median (-1 for values less than the median and +1 for values larger than the median). The first N binarized principal components were taken as the binary face pattern $x = [x_1, \ldots, x_N]^T$, serving as the activity of the N input neurons of the memory module.

The data set only contains faces from 9131 different people. To have a larger number of independent face patterns, we also synthesized artificial face patterns. First, we extracted the mean values and the diagonal covariance matrix of the face feature vectors after PCA to get an estimate of the distribution of patterns generated by the faces of all the people in the data set. We then synthesized artificial face patterns for new people by passing new samples from the corresponding multivariate normal distribution through the binarization step.

Memory module

The memory module is the only part of our network containing plastic synapses. The synapses are continuously updated by the ongoing presentation of the face patterns, whereas the weights of the input module are frozen during the online learning phase. The memory module consists of N memory neurons, one for each unit of the embedding module providing inputs to the memory module. We will refer to these units as input neurons for short. The j-th input neuron connects to the i-th memory neuron (for $i \neq j$) with synaptic weight (efficacy) w_{ij} and bias term b_i. There is no connection between the i-th input neuron and the i-th memory neuron for any i (i.e., $w_{ii} = 0\ \forall i$). The activity of the i-th memory neuron is

$$y_i = \text{sign}\left(b_i + \sum_{j \neq i} w_{ij} x_j\right) \ , \tag{1}$$

and we denote the binary memory patterns retrieved in this manner as $y = [y_1, \ldots, y_N]^T$. This plastic layer of synapses implements a simple feedforward memory model that can perform an approximate one-step reconstruction of a stored input pattern from a noisy cue at test time.

To update the synaptic weights and biases we used:

$$\Delta w_{ij} = x_i x_j \ , \tag{2}$$

$$\Delta b_i = x_i \ . \tag{3}$$

These equations give the desirable plasticity steps to store each new pattern. However, simply applying these additive updates would eventually result in unbounded values of the w_{ij}. Therefore, we employed a mechanism to limit the weights to bounded dynamical ranges. For each synapse (i.e., for each weight w and bias term b), we implemented a complex synaptic model[2] with m dynamical variables $u_1, \ldots, u_m$ in discrete time. Here m denotes the total number of variables per synapse (a measure of synaptic complexity), each of which operates on a different timescale. Specifically, at each time step t the dynamical variables u_k ($2 \leq k \leq m$) are updated as follows (the indices i and j labeling the synapses are omitted for simplicity)

$$u_k(t+1) = u_k(t) + n^{-2k+2}\alpha(u_{k-1}(t) - u_k(t)) - n^{-2k+1}\alpha(u_k(t) - u_{k+1}(t)) \ . \tag{4}$$

For $k = m$, the last variable u_{k+1} is simply set to zero in this update equation, and for $k = 1$

we have

$$u_1(t+1) = u_1(t) + I(t) - n^{-1}\alpha(u_1(t) - u_2(t)) \ . \tag{5}$$

Here $I(t)$ is the desirable update (Δw or Δb) imposed by the pattern $x(t)$, which takes a value $+1$ or -1 and is computed from equations (2) or (3). The first variable u_1 is used as the actual value of the synaptic weight w or bias b at test time. The parameters α and n determine the overall timescale of the model dynamics and the ratio of timescales of successive synaptic variables (we set $\alpha = 0.25$ and $n = 2$ in our models; see[2] for additional details).

To study the situation in which variables can only be stored with limited precision, we discretized the m synaptic variables and truncated their dynamical range to a maximum and minimum value. Hence, each variable can take one of only a finite number of integer-spaced values arranged symmetrically around zero (namely $\{-V,\ -V+1,\ \ldots,\ V-1,\ V\}$, where in our simulations we chose $V = 33$). At every time step, if the $u_k(t+1)$ computed according to equations (4) and (5) falls between two adjacent levels, its new value is set to one of those two levels, based on the result of a biased coin flip with an odds ratio equal to the inverse ratio of the distances from $u_k(t+1)$ to the two levels.

Readout (detection) module

The readout (detection) module compares the output $x = [x_1, \ldots, x_N]^T$ of the embedding module and the output $y = [y_1, \ldots, y_N]^T$ of the memory module to assess the level of familiarity of a given pattern. This module computes the Hamming distance between x and y, and outputs "familiar" (or "unseen"/"unfamiliar"/"novel") if the distance is smaller (or larger) than some pre-set threshold. This approach is similar to the one proposed in[12].

Evaluating the memory signal and noise

One way to measure the strength of a memory is to take the perspective of an ideal observer, who can access directly all the synaptic weights[1,2,7]. Following this approach, we considered the expected ideal observer signal $\mathscr{S}_{\text{io}}$ and noise term $\mathscr{N}_{\text{io}}$, and computed the ideal observer signal-to-noise ratio (ioSNR) $\mathscr{S}_{\text{io}}/\mathscr{N}_{\text{io}}(\Delta t)$ as a measure of the expected fidelity of recall of the stored memories as a function of the time elapsed since storage. The $\mathscr{S}_{\text{io}}$ and $\mathscr{N}_{\text{io}}$ are computed as follows.

For a given face memory, the signal at time t of the input pattern $x(t')$ stored at an earlier time t' is defined as the overlap (inner product) between the synaptic modification

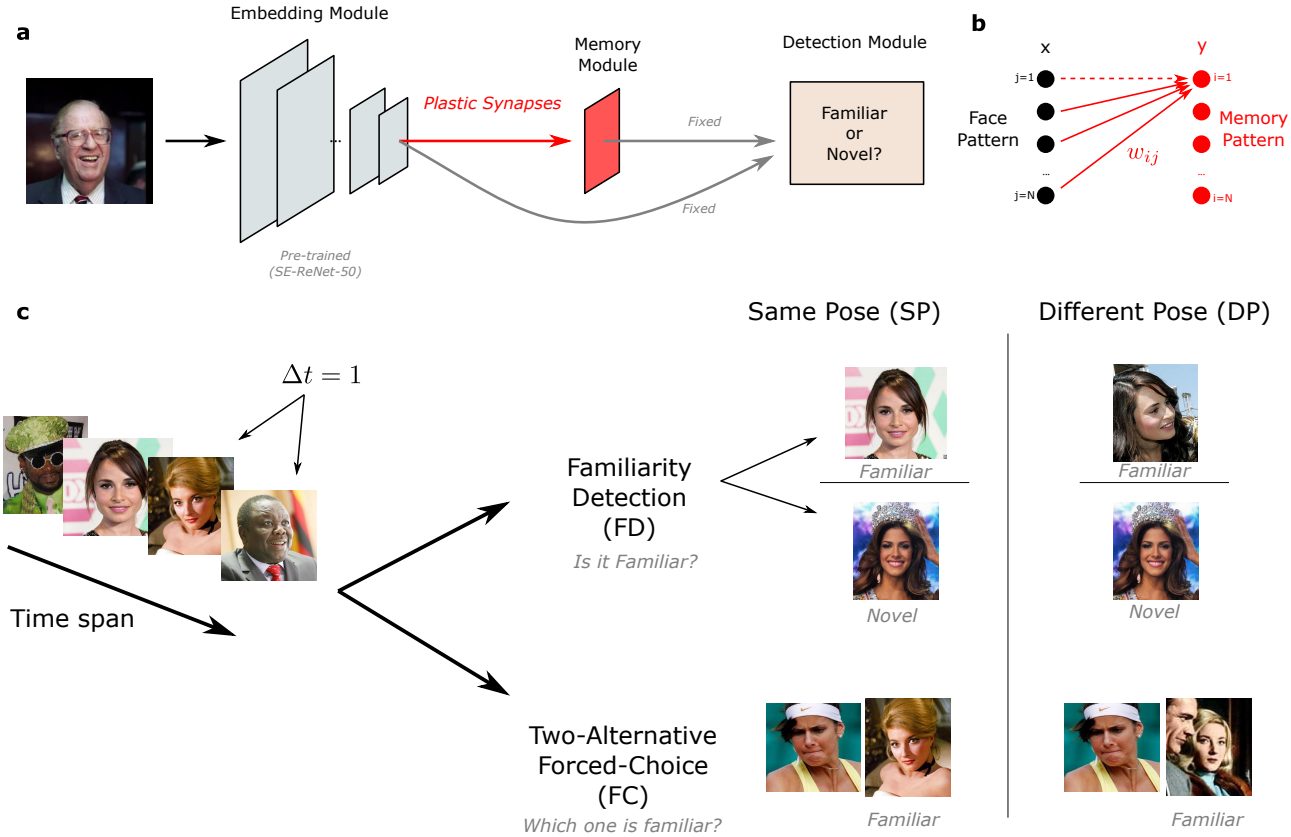

Figure 1: The architecture of our face familiarity detection system and the task diagram. (a) The neural system contains three modules: the input (embedding) module, the memory module, and the readout (detection) module. The synapses between the embedding module and the memory module (as well as the biases in the memory module) are plastic, while all other synapses are fixed (after being either set by hand or pre-trained) during the test phase, which requires online learning of face patterns. (b) The plastic connections between the input neurons in the embedding module and the memory neurons in the memory module. (c) A series of face images are presented to the neural system. In each familiarity detection (FD) test, the system is required to determine whether a presented face is familiar or unseen. A face is considered familiar if the test image is identical to a previously presented one (i.e., the same pose, SP) or a new pose of a previously presented face (i.e., a different pose, DP), and is considered novel if it is an image of an unseen person's face. In each two-alternative forced-choice (FC) test, the neural system is presented with a pair of face images (exactly one familiar and one unseen), and is required to choose which one of the two is familiar.

$\Delta w_{ij}(t')$ imposed at storage and the current ensemble of synaptic weights $w_{ij}(t)$:

$$S_{\mathrm{io}}(t-t') = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} \Delta w_{ij}(t') w_{ij}(t) \ . \tag{6}$$

We can then compute the average (denoted by $\langle\rangle$) over all memories with an age of $\Delta t = t - t'$ to obtain the expected signal

$$\mathscr{S}_{\mathrm{io}}(\Delta t) = \langle S_{\mathrm{io}}(\Delta t) \rangle \ , \tag{7}$$

and the corresponding noise term

$$\mathscr{N}^2_{\mathrm{io}}(\Delta t) = \langle (S_{\mathrm{io}}(\Delta t) - \langle S_{\mathrm{io}}(\Delta t) \rangle)^2 \rangle \ . \tag{8}$$

We also considered another measure of the memory signal that is more directly related to the ability of the system to read out the stored memory, the readout signal $\mathscr{S}_{\mathrm{r}}$. Similarly to the ioSNR, this signal is defined as the overlap between an input pattern $x(t')$ (stored at time t') and the retrieved memory pattern $y(t, t')$, the output of the memory module when the same pattern is presented again at time t without updating the synaptic weights. We have

$$S_{\mathrm{r}}(t-t') = \frac{1}{N} \sum_{i=1}^{N} x_i(t') y_i(t, t') \ . \tag{9}$$

As above, we can compute the expected signal $\mathscr{S}_{\mathrm{r}}$ and noise $\mathscr{N}_{\mathrm{r}}$ by averaging over memories of a given age, and obtain the readout signal-to-noise ratio (rSNR) $\mathscr{S}_{\mathrm{r}}/\mathscr{N}_{\mathrm{r}}(\Delta t)$.

Because the SNR decreases with time, we define the quantities t^*_{ioSNR} and t^*_{rSNR} as the memory ages t^* at which the ioSNR or rSNR, respectively, drop below a certain retrieval threshold as measures of the memory capacity of the system or of the expected memory lifetime.

Task protocol

To evaluate the performance of our system, we considered two tasks in which we presented a series of pre-processed face images to the neural system and tested its memory on randomly chosen faces (see Fig. 1c). We considered two types of tests. In the familiarity detection (FD) test, the neural system is required to determine whether the face image presented at test time is familiar or unseen by comparing the output of the detection module to a threshold. A familiar face image could be a previously stored one (same pose, SP) or a new pose of a

previously presented face (different pose, DP), while an unseen face image is an image of an unseen person. In the two-alternative forced-choice (FC) task, the neural system is presented with a pair of face images containing one familiar (either SP or DP) and one unseen face, and is required to choose which one of the two is familiar by comparing the output of the detection module for the two faces. These tasks are made particularly challenging by the fact that the familiar faces are presented only once.

Evaluating the task performance

In the FD task, the face images presented to the system are balanced, i.e., familiar faces previously presented within a certain age-range and unseen faces appear at test time with equal probability. We computed the average classification accuracy over this set of faces as a function of the threshold on the overlap computed by the detection module (see eqn. (9)) and the time elapsed since storage, which we refer to as the memory age. To maximize the overall performance, we chose the optimal overlap threshold for each age-range.

In the FC test, the task performance is defined as the probability of correctly choosing the familiar face (over the unseen one) for face memories of different ages. Similarly to our SNR analyses, we defined the memory age t^* as the age at which the task performance drops below some threshold (which defines the quantities $t^*_{\rm FD}$ and $t^*_{\rm FC}$).

In each simulation, all memory quantities, including the SNR and the task performance, are evaluated after the neural system reaches its steady state, i.e., when a large number of face patterns (with constant input statistics) have already been stored. In the steady state, the distribution of synaptic weights does not change any longer, although synapses continue to be updated as new face images are memorized. The system is then presented with two thousand real face images from different people, followed by the necessary number of artificial face image patterns. Our memory measures were evaluated only over these two thousand face images and further averaged over five independent simulations to reduce the noise floor.

Results

As we were interested in the scaling properties of the memory capacity in the case of familiarity detection, we systematically studied the performance of our neural system as functions of two key parameters: the number of memory neurons N and the number of dynamical variables m per synapse (i.e., the synaptic complexity). Increasing the synaptic complexity m for constant N or increasing N for constant m initially leads to a rapid improvement of the memory capacity but only up to a point where N is comparable to the longest timescale of

the synapse determined by m. Beyond this point, the growth of the memory capacity slows down (and it may even drop slightly in the case of increasing m further). To take advantage of a larger population of neurons, it is important to increase the longest timescale of the synapses, which is related to its complexity m. This can be achieved by choosing an m that grows logarithmically with N (such that $m = \log_2 N - 1$, as suggested in[2]).

SNR analysis and memory performance

We considered the SP and the DP tests separately (see Fig. 2). The DP performance cannot surpass the SP performance, because detecting familiarity for a different pose of the same person is clearly more difficult.

The ioSNR critically depends on the number of memories that are stored after the tracked face pattern, i.e., the memory age (see Materials and methods). Different curves correspond to synaptic models with different numbers of input features (and memory neurons) N and dynamical variables m. The curves are plotted on a log-log scale, for which a straight line represents a power-law dependence.

In the SP case, the ioSNR curves decay as a power-law over a time interval T corresponding to the longest timescale of the synapse before the decay becomes exponential. The ioSNR decays as slowly as the inverse square root of the memory age in the power-law regime. Changing N shifts the ioSNR curves in the log-log plot vertically, while increasing m primarily extends the power-law regime (i.e., increases T; see Fig. 2a). We determined the scaling of the familiarity memory lifetime with N (and m), where the lifetime t^*_{ioSNR} is represented by the memory age at which the ioSNR first drops below a given threshold. A value of 1 corresponds to a situation where the signal and the noise are of the same intensity. We chose a threshold of 0.1, though its precise value does not affect the scaling behavior much. We found that the familiarity memory lifetime scales approximately as N^2 (see Fig. 2c, in which the linear regression slope on a log-log scale is about 2.01 for the SP case, compared to 2.00 for random patterns (RD)). This scaling is very close to the theoretical result for optimal storage of random unstructured patterns[2]. Because m increases together with N (logarithmically), the familiarity memory lifetime scales exponentially with m (with the same linear regression slope on a $\log_2$-linear plot of t^*_{ioSNR} versus m).

For the DP case (see Fig. 2b), the ioSNR curves are lower than those in the SP case, due to the differences between the memorized and the tested face patterns. When there are more memory neurons, the initial ioSNR grows more slowly with N, and the shape of its initial decay with memory age becomes flatter. Nevertheless, the familiarity memory capacity still scales as a power of N (the regression slope is 1.71).

We also studied the properties of the rSNR. We found that the rSNR behaves similarly to the ioSNR at long time lags, but deviates from it for small memory ages, reflecting the effect of the neuronal nonlinearity. This nonlinear effect, which becomes more significant for larger N or smaller m (see also Fig. 5), leads to larger initial rSNR values, but does not substantially affect the memory lifetime compared to the ioSNR measure. The initial SNR enhancement quickly attenuates, leading to a similar scaling for the familiarity memory lifetime t^*_{rSNR}. These results further validate the ideal observer approach.

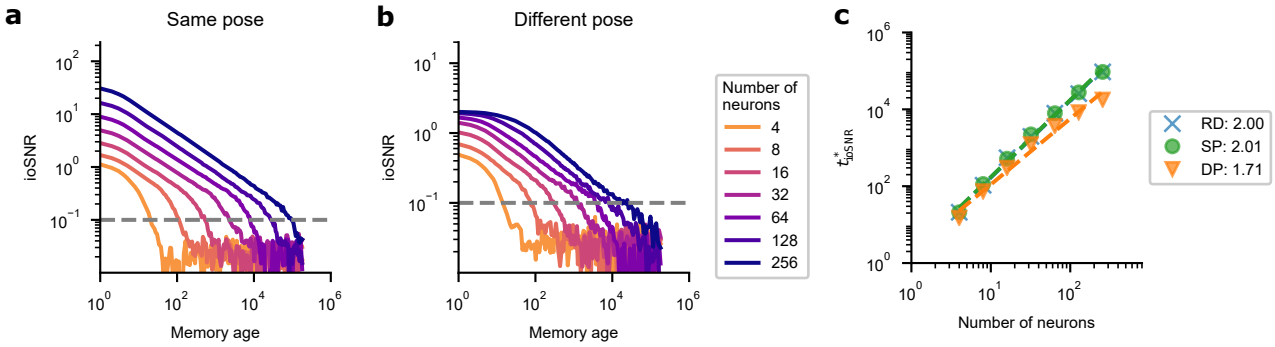

Figure 2: Ideal observer SNR (ioSNR) of the memory module as a function of face memory age and its scaling properties. (a) Doubly logarithmic plots of ioSNR versus the number of subsequently stored memories. Different curves correspond to models with a different number N of memory neurons and m of dynamical variables in the same pose (SP) case. The parameters N and m are varied by increasing N by factors of two and setting $m = \log_2 N - 1$. (b) The same as in the previous panel, but in the different pose (DP) case. (c) Log-Log plot of the ioSNR memory lifetime versus N in the SP, DP, and random-pattern (RD) cases. The legend indicates the best fit linear regression slopes (corresponding to the power of N in the scaling behavior).

Task performance

In Fig. 4, we plot the task performance in the FD test as a function of test age-range (along each curve, the points are a series of systems with thresholds optimized for different memory age-ranges). In the SP case, the initial task performance quickly improves as N and m increase. It saturates at 100% when the system has more than 64 neurons (see Fig. 4a). Moreover, increasing N and m leads to a substantial extension of the task-relevant familiarity memory lifetime t^*_{FD} (even beyond this value of $N = 64$). The memory lifetime was estimated assuming a performance threshold of 53% (this value was chosen to keep the initial task performance of all the simulations above the threshold). The power-law scaling behavior of the familiarity memory lifetime is revealed by plotting t^*_{FD} versus N on a log-log scale (linear

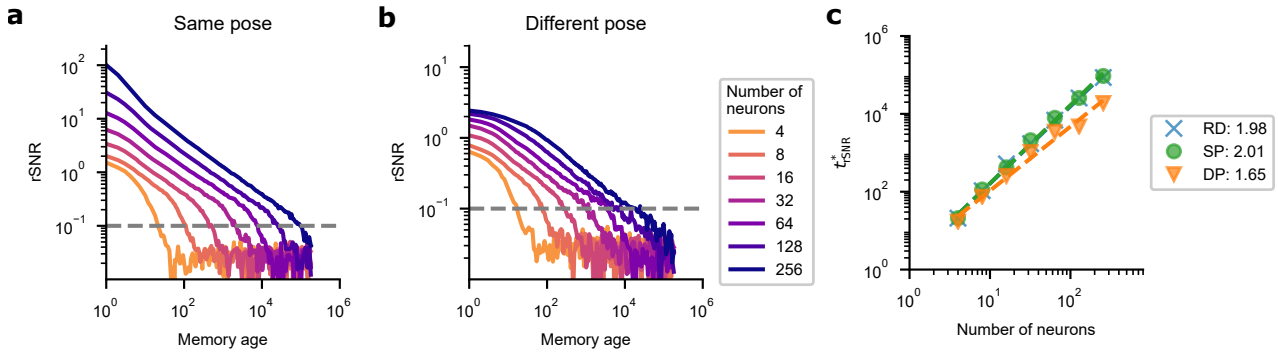

Figure 3: Readout SNR (rSNR) of the memory module as a function of face memory age and its scaling properties. (a) Doubly logarithmic plots of rSNR versus the number of subsequently stored memories. Different curves correspond to models with a different number N of memory neurons and m of dynamical variables in the same pose (SP) case. The parameters N and m are varied by increasing N by factors of two and setting $m = \log_2 N - 1$. (b) The same as in the previous panel, but in the different pose (DP) case. (c) Log-Log plot of the rSNR memory lifetime versus N in the SP, DP, and random-pattern (RD) cases. The legend indicates the best fit linear regression slopes (corresponding to the power of N in the scaling behavior).

regression slope 1.85; see Fig. 4e), which shows a very similar growth also in the RD case (linear regression slope 1.86).

As expected, in the DP case the task performance is worse than in the SP case (see Fig. 4c). However, we still found a reasonable power-law scaling with N (regression slope 1.44).

We also plotted the task performance in the FC test as a function of the memory age. The regression slope of the memory lifetime t^*_{FC} versus N on a log-log scale is 1.95 for the SP case and 1.49 for the DP case (see Fig. 4b, d and f).

Complex versus simple synapses

To obtain a fair comparison between complex synapses[2] and the well-studied, simple (multi-state) synapses[3–5], we evaluated the familiarity memory performance of a neural system with complex synapses and three models with simple synapses with an approximately equal number of dynamical variables. The complex model has 256 memory neurons, and every neuron has 255 incoming synapses and one bias with seven dynamical variables each (i.e., $256^2 * 7 = 458752$ variables in total). All of the three simple models have 677 memory neurons, and every neuron has 676 incoming synapses and one bias with one dynamical variable each (i.e., $677^2 * 1 = 458329$ variables in total). These simple synapses follow essentially the same model dynamics as the previously studied hard-bounded multi-state synapses[4]. They differ in their level of plasticity: the synapses in the first model are updated every time

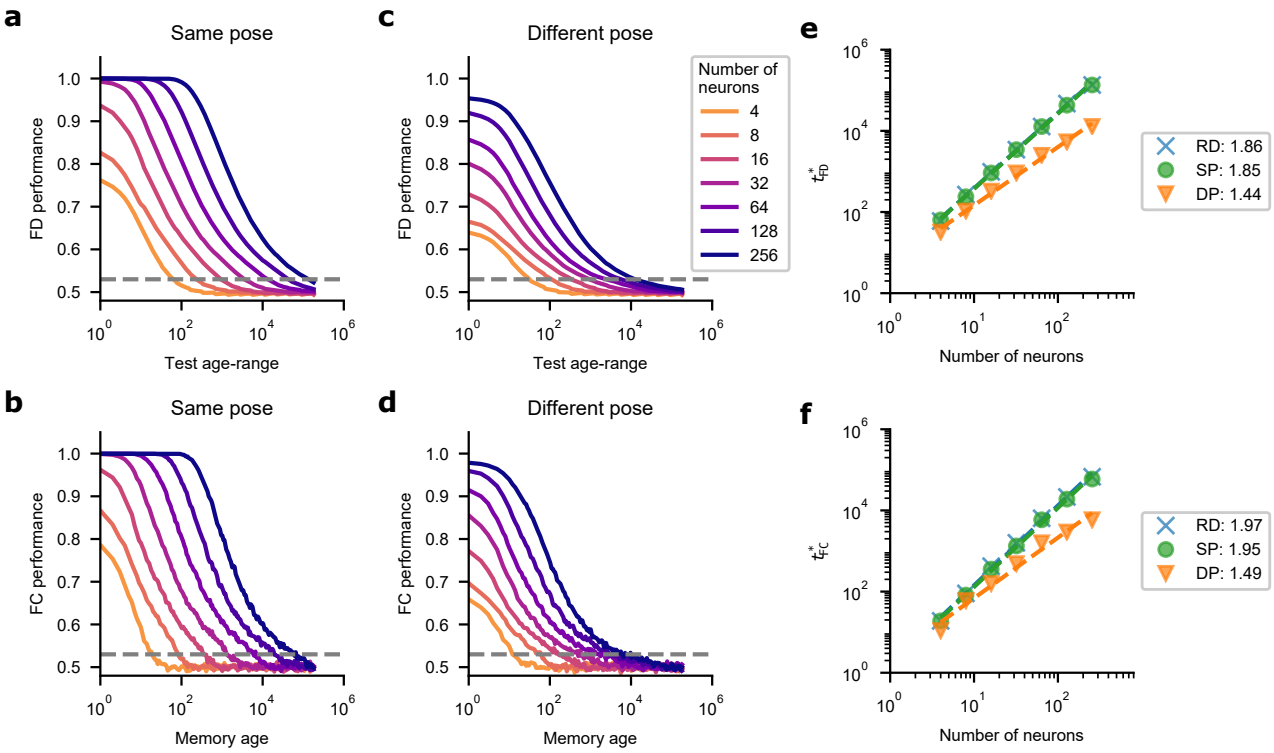

Figure 4: Familiarity detection (FD) and two-alternative forced-choice (FC) test performance of our system and their scaling properties. (a, b) Task performance as a function of the test age-range (FD) or the memory age (FC). Different curves correspond to models with a different number N of memory neurons (and number m of dynamical variables such that $m = \log_2 N - 1$) in the same pose (SP) case. (c, d) As in the previous panels, but for the different pose (DP) case. (e, f) FD and FC memory lifetimes versus N in the SP, DP, and random-pattern (RD) cases. The legends indicate the best fit linear regression slope, i.e., the power of N in the corresponding scaling.

an input pattern is stored, while the synapses in the second and third ones are changed stochastically according to a learning rate (encoding probability) less than one, and thus are more "rigid" [3,22]. Small learning rates lead to lower initial ioSNR values, but also to longer memory lifetimes. We choose $q = 0.128$ for the second model so that its initial ioSNR is comparable to the complex synapse system in the SP and DP cases. For the third model, we picked $q = 0.005$ to obtain the longest memory lifetimes possible for a system of simple synapses of this size, with an initial SNR just above the threshold (in the DP case).

Each variable in all of these models has the same number of discrete levels, and the total numbers of variables are approximately matched in the simple and the complex system. These simulations show that the complex system has a substantially better familiarity memory performance than the simpler systems (see Fig. 5), despite the smaller number of neurons. For the SP and RD cases, the memory lifetime of the system with complex synapses is ~ 1000 times longer; while for the DP case, the improvement factor is $\sim 100 - 400$. Slower simple synapses (with smaller q) can greatly extend the familiarity memory lifetime, but at the expense of the initial SNR and thus the generalization ability. Even so, they are far from matching the memory lifetime of the complex system. We can conclude that the memory model with complex synapses performs at least two orders of magnitude better in terms of memory capacity, and we expect the gap between simple and complex systems to grow even wider in networks with a larger number of neurons because of the different scaling behaviors.

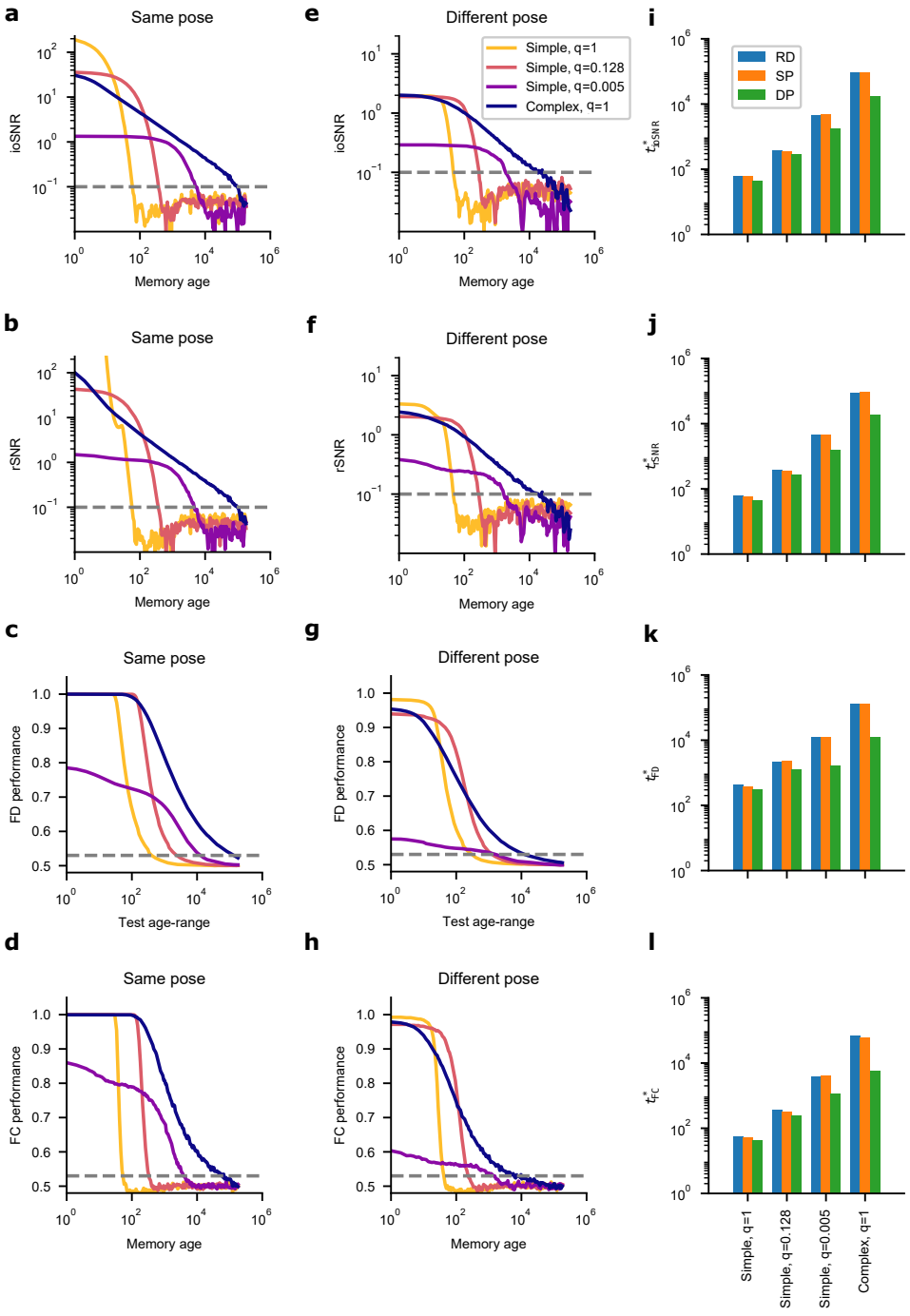

Figure 5: Comparison between models with simple synapses ($N = 677$, $m = 1$) and different learning rates ($q = 1$, 0.128, and 0.005, respectively) and a complex model ($N = 256$, $m = 7$, $q = 1$) with approximately the same total number of plastic variables. (a-d) Comparisons between models for the same pose (SP) case in terms of ioSNR, rSNR, familiarity detection (FD) performance, and two-alternative forced-choice (FC) performance. (e-h) Similar comparisons between models in the different pose (DP) case. (i-l) Comparisons between models in terms of different measures of familiarity memory lifetime (t^*_{ioSNR}, t^*_{rSNR}, t^*_{FD}, and t^*_{FC}, respectively) in the SP, DP, and random-pattern (RD) cases.

Discussion

We have presented a modular memory system that can solve a real-world problem such as face familiarity detection, which involves the ability to store in memory in one shot a large number of visual inputs. Thanks to the interactions between fast and slow variables of the complex synaptic model, the familiarity memory capacity grows almost linearly with the number of plastic synapses or quadratically with the number of neurons. The scaling of the system with simple synapses is only logarithmic with the number of synapses[3,22], though the memory performance can significantly increase when the learning rate q becomes small, or when the number of states per variable increases[3,4]. However, even when these parameters are properly tuned, the linear scaling can never be achieved, and the system with complex synapses outperforms the one with simple synapses in all cases, even when the total number of dynamical variables is the same for the two systems.

The advantage of complex synapses comes from two important properties: the first one is that they involve multiple timescales, enabling the system to learn quickly, using the fast components, and forget slowly, thanks to the slow components. The second one is that the dynamical components operating on different timescales can interact to transfer information from one component to another. In the case of our specific model the information diffuses from the fast components to the slow ones, and back (see[2] for more details). These two properties are important for any memory system that involves a process of consolidation, whether the process is synaptic or requires communication across multiple brain areas (memory consolidation at the systems level, see e.g.[23]).

Our previous work[2] systematically studied the scaling properties, the memory capacity, and the robustness of a broad class of complex synaptic models for random and uncorrelated synaptic modifications. One of the situations in which the synaptic modifications are random and uncorrelated is when the patterns of activity that represent the memories are also random and uncorrelated, which is what was assumed in all the early works on memory capacity (e.g.[13]). One of the reasons behind this assumption is that it allowed theorists to perform analytic calculations. However, it is a reasonable assumption even when more complex memories are considered. Indeed, storage of new memories is likely to exploit similarities with previously stored information. Hence, the information contained in a memory is likely to be pre-processed, so that only those components that are not correlated with previously stored memories are actually stored. In other words, it is more efficient to store only the information that is not already present in our memory. As a consequence, it is not unreasonable to consider memories that are unstructured (random) and do not have any correlations with previously stored information (uncorrelated). Unfortunately, these processes that lead to

uncorrelated representations are rarely modeled explicitly (but see[24]) and we currently do not have a general theory for dealing with more realistic, highly structured memories. In our model, the face stimuli, which are highly structured and correlated, are pre-processed by a simulated visual system, whose intermediate representations are then used as inputs to our memory module. This pre-processing seems to be sufficient to achieve the same scaling properties predicted for random patterns.

Another important difference between our previous and present work is related to the nature of the memory problem to be solved. In our previous work, we were dealing either with a classification task (a typical perceptron problem with only one output unit) or with a reconstruction memory problem in which a recurrent network would learn to reproduce a previously seen input at the time of memory retrieval. In this work, we considered familiarity detection, which is a recognition memory problem. To reconstruct each individual binary feature of a memorized pattern, we would employ $N-1$ synapses. Here we have designed a system in which N such output neurons are combined and read out to report a one-bit response, which is familiarity. We are using all $N(N-1)$ plastic synapses that are available to output only one bit of information. Hence it is not surprising that in the case of reconstruction memory, the number of memories that can be retrieved (reconstructed) scales linearly with the number of neurons N, while in the case of familiarity detection, the memory lifetime scales quadratically with N.

We also studied the generalization performance of the system by considering different poses of presented faces as retrieval cues (the DP case), using probe patterns that differ from the originally stored ones. Although the task performance for this DP case is worse than in the SP case, the power-law scaling properties are similar, and the drop in performance could be compensated by introducing more memory neurons and possibly increasing the synaptic complexity. The ability to generalize to different poses is presumably helped by the complexity of the synapses. Indeed, in the case of random patterns, generalization is related to the memory SNR[2]. In future studies, we will determine whether there is a similar relationship between the SNR and the ability to generalize to different poses.

Implications for neuromorphic engineering

Because the dynamical variables of the plastic synapses in our memory module are digital and relatively low-precision, they can be effectively implemented as a group of binary switches (bits) in neuromorphic devices, using b bits to obtain $M = 2^b$ discrete levels. Because the distribution of synaptic weights is approximately Gaussian (though discretized), M should be appropriately chosen such that the Gaussian distributions will not be substantially truncated.

It is known that for the hidden variables with longer time constants, the width of the distribution is smaller and fewer discrete levels are required[2]. For the hidden variable with the longest time constant, M can be as small as 2 without affecting the memory performance much. Another class of complex synaptic models that have been shown to exhibit essentially the same memory performance as the models studied here can be implemented with even fewer hardware bits[25]. We believe that the proposed system suggests a useful architecture for a new generation of neuromorphic devices suitable for on-chip online learning. One of the attractive features of the proposed system is that it can solve a real-world problem without requiring a large number of plastic synapses. Indeed, the vast majority of the synapses of the complete system, which would include the pre-processing part, are not plastic. Nevertheless, the system can perform an interesting form of continual learning.

Limitations of our system

One limitation of our work is the assumption that the memory neurons use exactly the same representations as the input neurons. In reality, the number of memory neurons is unlikely to be precisely the same as the number of input pattern dimensions, and they would in general use a different representation of a given face from the input neurons. The detection module has to essentially compare the reconstructed memory with the representation of the current cue. This is a computation that can be performed even when the representations in the detection module are completely different from those in the input. However, it will require a smarter readout system that is trained to perform this comparison. Generalizing our system to include a more biologically plausible mapping between the embedding module and the memory module, with a corresponding readout mechanism in the detection module, is an important direction for our future work.

In our hippocampus-like memory module, there is only one feedforward layer that uses dense neural representations. However, recurrent neural computations in the hippocampus can be beneficial in some memory tasks[26,27]. In addition, sparse representations of memory patterns have long been known to harbor computational benefits such as larger memory capacity and the capability to mitigate disruptive effects of correlations[2,3,28]. To what extent recurrent connections and sparse coding are beneficial in our neural system for familiarity detection are questions currently under investigation.

Funding

This work was supported by NSF NeuroNex Award DBI-1707398, the Gatsby Charitable Foundation and the Swartz Foundation.

Acknowledgments

We thank Xihan Li for his kindness in providing computational resources.

Data Availability

The VGGFace2 face data set[15] employed in our study can be found on the following website: `http://www.robots.ox.ac.uk/~vgg/data/vgg_face2/`.

References

[1] Stefano Fusi, Patrick J Drew, and Larry F Abbott. Cascade models of synaptically stored memories. *Neuron*, 45(4):599–611, 2005.

[2] Marcus K Benna and Stefano Fusi. Computational principles of synaptic memory consolidation. *Nature neuroscience*, 19(12):1697, 2016.

[3] Daniel J Amit and Stefano Fusi. Learning in neural networks with material synapses. *Neural Computation*, 6(5):957–982, 1994.

[4] Stefano Fusi and LF Abbott. Limits on the memory storage capacity of bounded synapses. *Nature neuroscience*, 10(4):485, 2007.

[5] Stefano Fusi. Computational models of long term plasticity and memory. *arXiv preprint arXiv:1706.04946*, 2017.

[6] Christos Kaplanis, Murray Shanahan, and Claudia Clopath. Continual reinforcement learning with complex synapses. *arXiv preprint arXiv:1802.07239*, 2018.

[7] Stefano Fusi. Hebbian spike-driven synaptic plasticity for learning patterns of mean firing rates. *Biological cybernetics*, 87(5-6):459–470, 2002.

[8] Kenneth A Norman. How hippocampus and cortex contribute to recognition memory: revisiting the complementary learning systems model. *Hippocampus*, 20(11):1217–1227, 2010.

[9] Kenneth A Norman and Randall C O'Reilly. Modeling hippocampal and neocortical contributions to recognition memory: a complementary-learning-systems approach. *Psychological review*, 110(4):611, 2003.

[10] Rafal Bogacz and Malcolm W Brown. Comparison of computational models of familiarity discrimination in the perirhinal cortex. *Hippocampus*, 13(4):494–524, 2003.

[11] Cristina Savin, Peter Dayan, and Máté Lengyel. Two is better than one: distinct roles for familiarity and recollection in retrieving palimpsest memories. In *Advances in Neural Information Processing Systems 24*, pages 1305–1313. 2011.

[12] Rafal Bogacz, Malcolm W Brown, and Christophe Giraud-Carrier. High capacity neural networks far familiarity discrimination. 1999.

[13] John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982.

[14] Zacharias Androulidakis, Andrew Lulham, Rafal Bogacz, and Malcolm W Brown. Computational models can replicate the capacity of human recognition memory. *Network: Computation in Neural Systems*, 19(3):161–182, 2008.

[15] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 67–74. IEEE, 2018.

[16] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. 2008.

[17] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation from predicting 10,000 classes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1891–1898, 2014.

[18] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, et al. Deep face recognition. In *bmvc*, volume 1, page 6, 2015.

[19] Ira Kemelmacher-Shlizerman, Steven M Seitz, Daniel Miller, and Evan Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4873–4882, 2016.

[20] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European Conference on Computer Vision*, pages 87–102. Springer, 2016.

[21] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[22] Srdjan Ostojic and Stefano Fusi. Synaptic encoding of temporal contiguity. *Frontiers in computational neuroscience*, 7:32, 2013.

[23] Alex Roxin and Stefano Fusi. Efficient partitioning of memory systems and its importance for memory consolidation. *PLoS computational biology*, 9(7):e1003146, 2013.

[24] Marcus K. Benna and Stefano Fusi. Are place cells just memory cells? memory compression leads to spatial tuning and history dependence. *bioRxiv*, 2019. doi: 10.1101/624239.

[25] Marcus K Benna and Stefano Fusi. Efficient online learning with low-precision synaptic variables. In *2017 51st Asilomar Conference on Signals, Systems, and Computers*, pages 1610–1614. IEEE, 2017.

[26] Dharshan Kumaran, Demis Hassabis, and James L McClelland. What learning systems do intelligent agents need? complementary learning systems theory updated. *Trends in cognitive sciences*, 20(7):512–534, 2016.

[27] Dharshan Kumaran and James L McClelland. Generalization through the recurrent interaction of episodic memories: a model of the hippocampal system. *Psychological review*, 119(3):573, 2012.

[28] M. V Tsodyks and M. V Feigel'man. The enhanced storage capacity in neural networks with low activity level. *Europhysics Letters (EPL)*, 6(2):101–105, May 1988. ISSN 0295-5075, 1286-4854. doi: 10.1209/0295-5075/6/2/002.