Bayesian Telephone: Memory Consolidation and Recall as Generative Processes

Dongyu Gong
Yale University
dongyu.gong@yale.edu

Micah Gold
Yale University
micah.gold@yale.edu

Abstract

The hippocampus serves a key role in memory acquisition and consolidation, yet it is unknown whether the hippocampus stores raw sensory inputs or merely generative reconstructions of those inputs. In this paper we examined these competing hypotheses of memory representation in the hippocampus. To do so we modeled the hippocampus as a modern Hopfield network and the entorhinal cortex as a variational autoencoder (VAE). We used Mitsuba 3 to generate a Cornell box dataset. In our first model, we passed these scenes directly into our Hopfield network and trained our VAE on the Hopfield network's output when prompted by stimulus. In the second, the model first probabilistically inferred latent parameters for the observations and generated reconstructed observations which were then passed into the Hopfield network to aid in the training of our VAE. We tested the capacity of our models for generative recall of these scenes and found reliable minimization of reconstruction error during recall in both models. We concluded that either representation scheme or a combination of the two might be at work in the human brain. Future studies should explore implementing features such as forgetting and recall vulnerability in our base model and comparing model performance to human performance on recall tasks.

1 Introduction

Memory is at the core of human cognition. Understanding the algorithms of how we encode, store, retrieve, and use memory representations is essential for understanding human intelligence.

Simply put, human memory is characterized by the lossy compression and imperfect reconstruction of an initial set of neural signals associated with sensory experiences. In order to construct and maintain a memory, the mind must first learn "schemas" (latent structures) from sensory inputs [10]. The hippocampus is responsible for directing the initial parsing of stimuli into memories. Over time, the hippocampus mediates the transfer of relevant patterns of neural activity through the entorhinal cortex (EC) to the medial prefrontal cortex and anterolateral temporal cortex via "neural replay" [4, 21, 22]. This process of consolidation entails the learning of schemas and enables the creation of longer-term memory. Long-term memories can persist in the neocortex for a lifetime, but the process of "accessing" these memories is anything but straightforward [3]. Recall entails utilizing these learned schemas stored in the neocortex to realize a noisy estimation of aspects of the initial brain state at the time of first memory acquisition [18].

This whole process can be framed as the training of a generative model and the subsequent generative processes mediated by the latent variables learned by the model. However, a complete understanding of the inner-workings of the memory formation process remains remote [9, 15]. For one, it is unclear just exactly how short-term memories are represented in the hippocampus and when along the complex process of memory consolidation the initial armada of neural signals are stripped down to that more fine set responsible for encoding latents. Perhaps there are even multiple layers of Bayesian inference and image regeneration inherent in the pathways of memory formation.

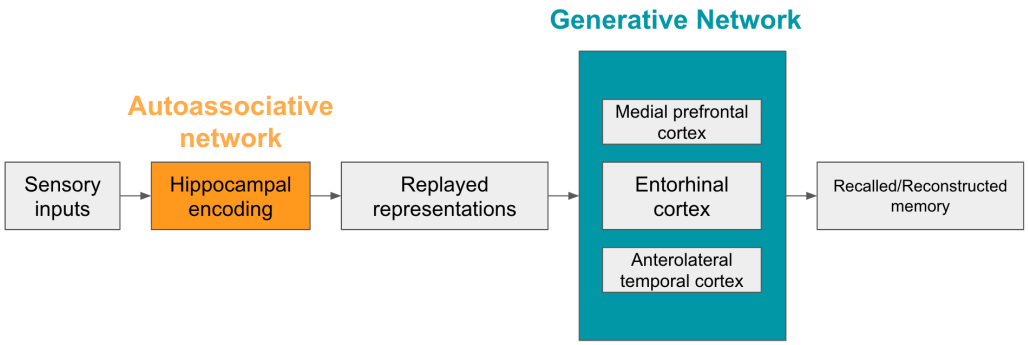

Figure 1: Model of the consolidation-recall pathway that allows for long-term memory formation. The hippocampus acts as an autoassociative network which stores encodings of sensory inputs. Over time, replay of these representations facilitates the learning of latent structures by the neocortex. The neocortex then acts as a generative network, producing noisy recollections of memories from these learned parameters on demand.

To grapple with these questions, recent studies have reverse engineered computationally plausible models of the consolidation-recall pathway. Many of these studies model the hippocampus as an autoassociatve network and the neocortex as a generative network [17]. One such study of interest, devised a system by which a modern Hopfield network, representing the hippocampus, employed teacher-student learning [11] to train a variational autoencoder (VAE) representing the EC [19]. This model proved capable of robust generative recall when the Hopfield network encoded exact representations of images.

Here, we examine whether and how altering the encoding mechanism in the Hopfield network might impact the capacity of the model for generative recall. Specifically we compare the performance of an exact reimplementation of the original model with one in which the Hopfield network instead stores reconstructions of the scenes after an initial layer of Bayesian inference and image regeneration. In doing so, we attempt to understand whether the hippocampus aids the neocortex in learning latent structures from raw sensory inputs or stochastic reconstructions of sensory inputs (i.e. perception). In other words, we seek to shed light on whether consolidation is merely the learning of latents from observations or rather the learning of latents from reconstructed observations which themselves result from a long game of Bayesian telephone.

2 Methods

Code for the methods described below can be accessed in this repository: `https://github.com/Daniel-Gong/final_project`.

2.1 Custom Cornell Box Dataset

Using Mitsuba 3 we constructed a dataset (see Figure 3) of 270 Cornell box images each with a distinct combination of wall color ({`red, white, green`}), floor color ({`red, white, green`}), feature shape ({`sphere, cube`}), feature color ({`red, white, green`}), and feature scale ($\{.3, .4, .5, .6, .7\}$).

2.2 Inverse Graphics Engine

Using the Gen package for Julia [6], we implemented an inverse graphic engine capable of inferring latent parameters from each observation in our dataset and reconstructing noisy images from these latents. Inference was conducted by constraining the generative function with bitmaps of the original observations selectively compressed to ensure representative description of the key-features of the

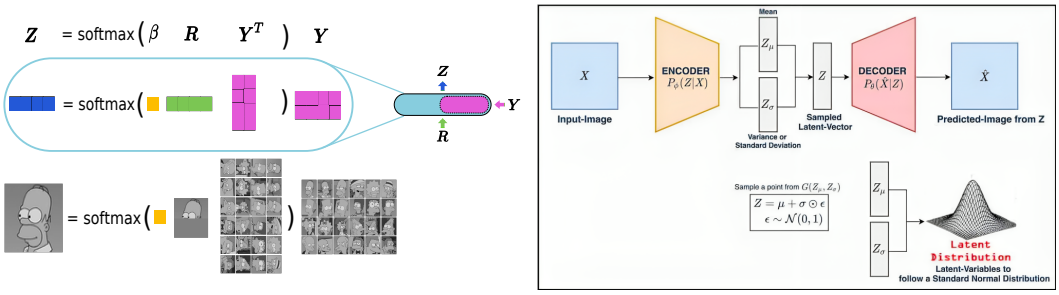

Figure 2: Model illustrations. **Left:** pattern retrieval process in the modern Hopfield network. Z is the pattern to be retrieved, Y are all stored patterns, R is the input pattern, and β is the inverse temperature parameter. **Right:** architecture of the variational autoencoder. The process begins with the input image fed into the encoder, which encodes the image into a latent representation characterized by a mean vector and a variance or standard deviation vector. From this, a sampled latent vector is generated, where the latent variable Z is sampled as $Z = \mu + \sigma \odot \epsilon$, with ϵ drawn from a standard normal distribution $\mathcal{N}(0, 1)$. The decoder then takes the sampled latent vector to reconstruct the predicted image from Z, which is the VAE's approximation of the original image.

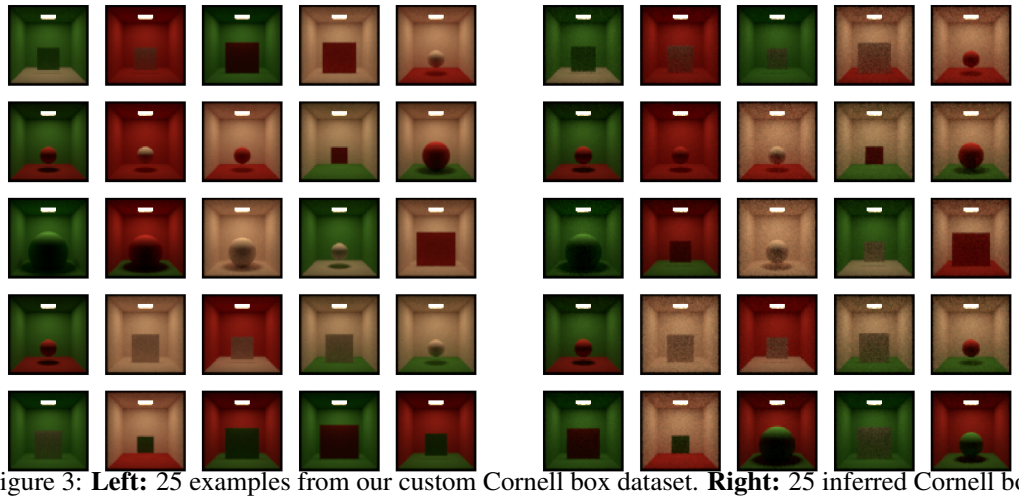

Figure 3: **Left:** 25 examples from our custom Cornell box dataset. **Right:** 25 inferred Cornell box images, corresponding to the ones on the left.

observations. Using Bayes rule a posterior for the latents was generated over these bitmaps:

$$P(\theta|b(I)) \propto P(b(I)|\theta)P(\theta) \tag{1}$$

where θ are latents, and $b(I)$ are bitmap observations. Random-walk Metropolis-Hastings algorithm was employed to sample latent values for each image from their respective posteriors.

2.3 Modern Hopfield Network

We implemented a modern Hopfield network to model the hippocampus using the TensorFlow package for Python (accesed via PyCall in Julia) [1]. Modern Hopfield netoworks are a variety of recursive neural network designed to model dense associative memory. In a modern Hopfield network, stable state corresponds to a particular memories [16, 13]. Compared to their classical alternatives, modern Hopfield networks are better suited for modeling biological memory because they maintain exponential storage capacity per neuron and differentiability across use cases [13]. A core equation for the modern Hopfield network we are using (see Figure 2 left panel) is the update rule for the state of the stored patterns, which is based on the concept of continuous states and uses an attention mechanism [16]. This can be represented as:

$$\mathbf{Z} = \text{softmax}(\beta \mathbf{R}\mathbf{Y}^T)\mathbf{Y} \tag{2}$$

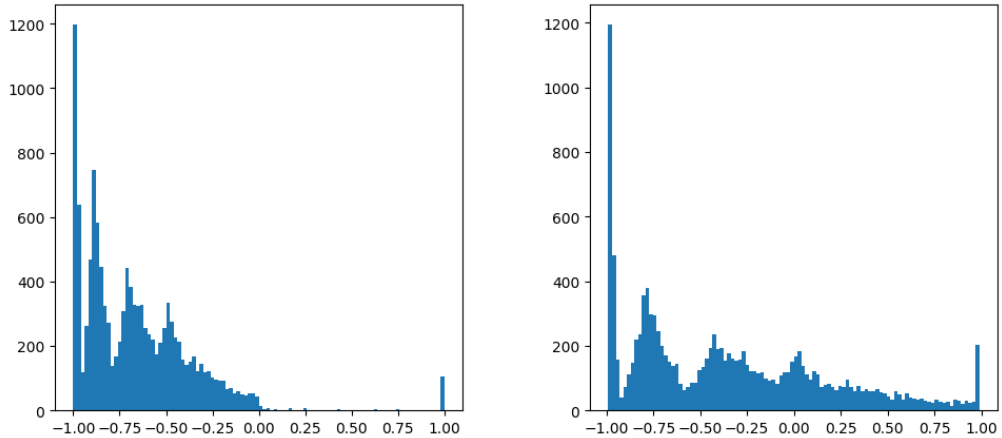

Figure 4: **Left:** Histogram of pixel values of one example image from our custom Cornell box dataset before normalization. **Right:** histogram of pixel values of the same image after normalization.

where $\mathbf{Z}$ is the state vector of the Hopfield network, $\mathbf{Y}$ is the matrix of stored patterns (weights), $\mathbf{R}$ is the input pattern, β is the inverse temperature parameter, controlling the sharpness of the softmax function. The softmax function is applied to ensure that the output is a probability distribution. This formula demonstrates how the attention mechanism is used in modern Hopfield networks, in which the similarity between the input pattern and the stored patterns determines the weights used to update the network's state.

We used uniform random noise inputs (between -1 and 1) as the input pattern to initiate pattern retrieval ("neural replay") from the Hopfield network. To make sure all stored patterns (no matter they are spheres or cubes) have equal probabilities to be replayed, we normalized the distribution of pixel values in the stored patterns to be between -1 and 1 (see Figure 4).

2.4 Variational Autoencoder

We implemented a VAE to model the EC using the TensorFlow package for Python (accesed via PyCall in Julia) [1]. Variational autoencoders are generative networks that learn to deconstruct observations into latent parameters from which those initial observations can be nosily reconstructed [12]. Our VAE (see Figure 2 right panel) was trained to minimize the reconstructed error between input images (i.e., replayed memories from the modern Hopfield network) and predicted images from the decoder, as well as the Kullback-Leibler (KL) divergence between the latent vector sampled in VAE and a standard normal distribution. The loss function of our VAE can be represented as:

$$L(\phi, \theta, x) = \frac{1}{N} \sum_{i=1}^{N} (X_i - \hat{X}_i)^2 + KL\left(G(Z_\mu, Z_\sigma) \,||\, \mathcal{N}(0,1)\right) \tag{3}$$

where ϕ are the parameters of the encoder, θ are the parameters of the decoder, X_i is the ith input image replayed from the Hopfield network, $\hat{X}_i$ is the ith reconstructed image produced by the decoder, and N is the total number of images in the dataset ($N = 270$ in our case). $G(Z_\mu, Z_\sigma)$ is the Gaussian distribution defined by the encoder's output, which includes a mean vector Z_μ and a standard deviation vector Z_σ. The first term on the right side of this equation is a Mean-Squared-Error (MSE) loss, and the second term (the KL divergence) acts as a regularizer, which keeps the encodings sufficiently diverse.

We used the AMSGrad variant of the Adam optimiser as the stochastic gradient descent method for the training of the VAE. A latent variable vector length of 20, learning rate of 0.001 and KL weighting of 1 were used in the main results.

To test the performance of our VAE after training, we created perturbed partial Cornell box images by randomly selecting 10% of pixels in each image and then replacing them with 0s. These partial images are used as the input to the trained VAE to check their generative recall.

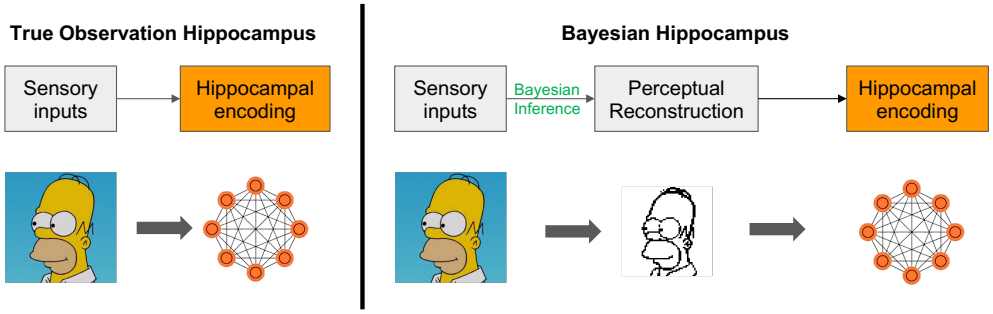

Figure 5: Two hypotheses. **Left:** The memory storage mechanism for the true observation hippocampus hypothesis simply entails storing the original sensory inputs to the Hopfield network. **Right:** The memory storage mechanism for the Bayesian hippocampus hypothesis entails passing the sensory inputs into the inverse graphics engine and capturing the perceptual reconstruction in the Hopfield network.

2.5 Two Hypotheses

2.5.1 True Observation Hippocampus

In this hypothesis (Figure 5 left panel), sensory inputs are stored in the hippocampus exactly as they are observed. In order to model this, the original 270 images from our Cornell Box dataset were stored in the modern Hopfield network. Next, we exposed the Hopfield network to random noises and recorded which observations were replayed. We passed in the replayed representations and trained the VAE over 50 epochs (with early stopping applied if there was no loss improvement for three epochs) to learn the encoding of those stimuli as latents which could then be decoded to generatively "recall" observations similar to those original observations stored in the Hopfield network. The reconstruction error of the VAE with respect to the replayed representations from the Hopfield network was analyzed to shed light on the of utility of the model.

2.5.2 Bayesian Hippocampus

In this hypothesis (Figure 5 right panel), sensory inputs are first passed into our inverse graphics engine for preprocessing. The inverse graphics engine infers latent parameters from those images and constructs noisy perceptions of the images. These reconstructions are passed into the modern Hopfield network and stored as memories. We exposed the Hopfield network to random noises and recorded replayed observations. As before, we passed in the replayed representations and trained the VAE over 50 epochs (with early stopping applied if there was no loss improvement for three epochs) to learn the latents underlying these representations so as to reconstruct them. The reconstruction error of the VAE was analyzed to shed light on the appropriateness of the Bayesian hippocampus hypothesis as compared with the true observation hippocampus hypothesis.

3 Results

3.1 True Observation Hypothesis

After 270 images from our custom Cornell box dataset were encoded in the modern Hopfield network, random noise inputs to the network successfully reactivated its memories (Figure 6). Notably, our normalization procedure on the distribution of pixel values ensured that all stored patterns had a chance to be replayed, no matter what shape they are, what color they have, etc. It is possible that this normalization procedure is also implemented in the brain to ensure better encoding and retrieval of memories [8, 5], considering that the neural signals of sensory experiences can vary a lot in terms of their salience merely because of their physical properties.

The training of our VAE on the replayed true observations converged after around 15 epochs and triggered early stopping after around 30 epochs. The performance test for the trained VAE showed

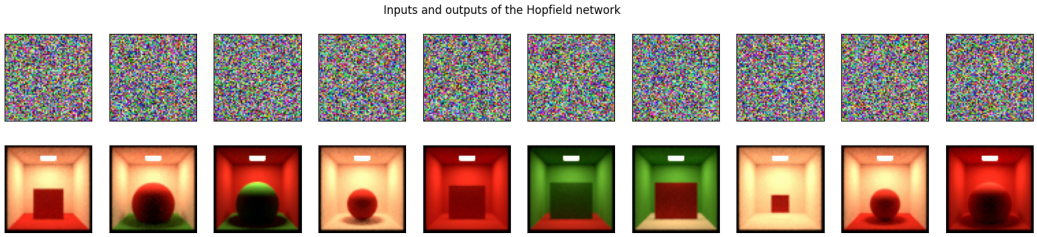

Figure 6: True observation hypothesis: random inputs to the Hopfield network and the corresponding replayed true observations.

Figure 7: True observation hypothesis. **Left:** reconstruction errors of the VAE across training epochs. **Right:** VAE reconstructions based on noisy Cornell box images.

that it was able to noisily reconstruct complete observations from perturbed partial images in the original custom Cornell box dataset (Figure 7). The recalled memories are to some degree sketchy and blurry given that there are only 270 training images. However, we believe scaling up the number training samples will lead to significantly better performance.

3.2 Bayesian Hypothesis

First, our inverse inverse graphics engine for the most part generated images that differed only slightly from the original images (Figure 3 right panel) , even though we only used a simple random walk Metropolis Hastings algorithm. As before, we applied the same normalization procedure to make sure every inferred image has a chance to be replayed after they are encoded in the Hopfield network. Random noise inputs to the network successfully reactivated diverse memories (Figure 8).

Furthermore, the training of our VAE on the replayed inferred images converged also after around 15 epochs and triggered early stopping after around 40 epochs. The performance test for the trained VAE showed that it was also able to noisily reconstruct complete observations from perturbed partial images in the original custom Cornell box dataset (Figure 9).

4 Discussion

Although the quality of our reconstructed images is somewhat impressive, we must first acknowledge a few constraints on our experiment that reduced the performance of both models and influenced our results. Firstly, given issues with accessing large compute resources on short notice we opted to train our models on smaller dataset of 270 images whereas other similar studies using VAE's recommend upwards of 10,000 images [12, 19]. Secondly, we built our inverse graphics engine in our Bayesian hypothesis to infer a measly six parameters, five of which were drawn from discrete distributions. As a result, the reconstructed images were in some ways too similar to our original images and in some ways unrealistically different. For example, in our model either the exact hue of red, white, or green was inferred or the incorrect color was inferred altogether. In reality however it is much more plausible that a human would infer a lightish-red ball to be a darkish-pink ball than a white ball.

Still, despite these issues both the true observation and Bayesian hypotheses demonstrated equal proficiency in recalling memories from noisy inputs. Considering that the brain's inverse graphic engine likely possesses comparable or greater strength than our implemented graphics engine for this task, this suggests that the hippocampus could represent memories in either fashion while yielding similar success in recall. Unfortunately, this leads to the conclusion that both models would

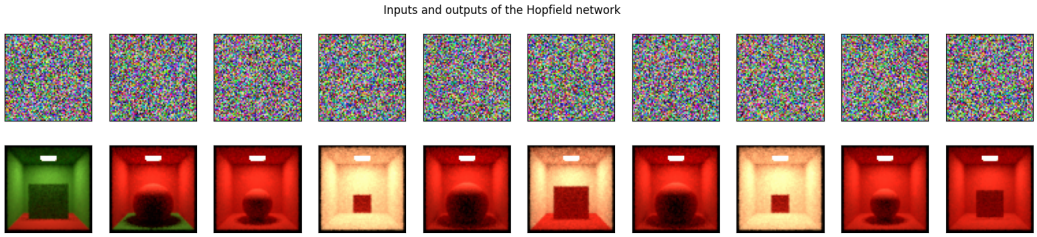

Figure 8: Bayesian hypothesis: random inputs to the Hopfield network and the corresponding replayed Bayesian inferences.

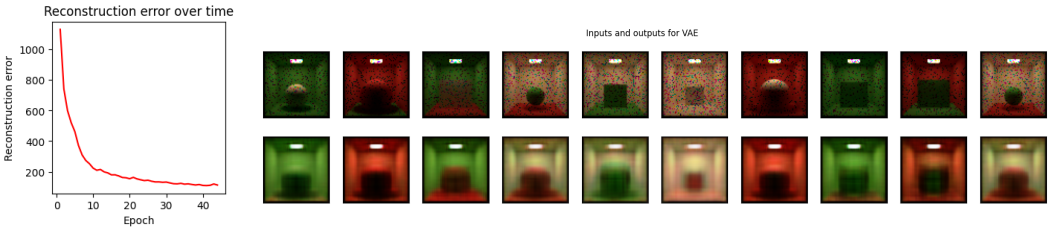

Figure 9: Bayesian hypothesis. **Left:** reconstruction errors of the VAE across training epochs. **Right:** VAE reconstructions based on noisy Cornell box images.

perform similarly if we compared their recall performance against human subjects. Thus, employing behavioral comparison metrics between our model and human subject would be insufficient to fully pick open the neural black-box and understand the fine mechanisms of the hippocampus.

On the flip side, the success of both hypotheses highlights an intriguing principle regarding how and when the brain may opt to deal with true observations as opposed to reconstructed observations. Since the computational outcome remains consistent, the brain might decide whether to utilize Bayesian reconstructions or original observations based on minimizing the temporal and spatial complexity of the system. In other words, algorithmically speaking, the brain might choose the representation that is more practically convenient. If this holds true, then considering our two models of hippocampal memory representation as mutually exclusive would be inaccurate. It is plausible that the hippocampus could store images, reconstructions of images from inferred latents, and perhaps even the latents themselves [19]. To test this hypothesis, future work should attempt to integrate all three forms of memory representation into the same Hopfield network. Surmising which hippocampal representation is in use any given moment in consolidation could be a matter of recording neural population dynamics during memory acquisition tasks in humans and comparing these dynamics to simulated traces from this integrated model [14].

Collecting behavioral data on similar recall tasks could also be useful to help direct the refinement of our overall model of generative recall and build in even more complex functionalities. For example, the current model only implicitly handles the possibility of forgetting images. True, consolidating additional memories in our model reduces the probability of recalling a prior memory. However, this is only because given a stimuli our model recalls a memory and the more similar memories there are stored in the VAE, the smaller the chance of a particular memory being recalled. A more robust system of forgetting, must consider the larger biological picture of the issue: there is a limited quantity of data that can physiologically be stored in both the hippocampus and the neocortex and a limited amount of time memories are stored in the hippocampus during which they can be replayed and consolidated and transferred to the neocortex. As a result, forgetting could be the result of either temporal constraints (the hippocampus selectively participating in the consolidation of one decaying memory at the expense of another) or spatial constrains (rewiring neural pathways in the neocortex resulting in the loss of particular activation dynamics) [7, 20]. Accessing behavioral data on the rate and nature of "forgetting" and the influence of recall on reconsolidating images (and thus increasing their likelihood of being recalled on future occasions) could guide the creation of a more robust probabilistic forgetting system.

In a similar vein, the current model does not implement any sort of system emulating recall vulnerability [2], the phenomenon by which generatively recalling memory results in increased malleability of the associated learned latent structures in the neocortex. Although we experimented with an

implementation of a probabilistic system capable of occasionally adding the products of generative recall to the Hopfield network and retraining the VAE with these noisy "hallucinated" memories, we faced difficulties in determining how to (1) tune the parameters of such a system to accurately reflect the realities of recall vulnerability (which has not been extensively quantitatively studied) and (2) rebuild our memory acquisition model in a sequential manner to avoid having to retrain the entire VAE every time recall affected our Hopfield network.

Hence, our base model represents a strong jumping-off point for future exploration of the generative processes that underlie episodic and semantic memory, imagination and recall. Refining our model piece-by-piece to bring it in line with behavioral results could yield key insights into the neural fabric of memory and the computational and algorithmic systems that govern our lives.

References

[1] Martín Abadi, Paul Barham, Jianmin Chen, and et al. Tensorflow: A system for large-scale machine learning. 2016. `arXiv:1605.08695`.

[2] Cristina M Alberini. The role of reconsolidation and the dynamic process of long-term memory formation and storage. *Frontiers in Behavioral Neuroscience*, 5:12, 2011. `doi:10.3389/fnbeh.2011.00012`.

[3] Craig H. Bailey, Eric R. Kandel, and Kausik Si. The persistence of long-term memory: A molecular approach to self-sustaining changes in learning-induced synaptic growth. *Neuron*, 44:49–57, 2004. `doi:10.1016/j.neuron.2004.09.017`.

[4] Michael F Bonner and Alfonso R Price. Where is the anterior temporal lobe and what does it do? *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 33(10):4213–4215, 2013. `doi:10.1523/JNEUROSCI.0041-13.2013`.

[5] Matteo Carandini and David J. Heeger. Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, 13(1):51–62, January 2012. `doi:10.1038/nrn3136`.

[6] Marco Cusumano-Towner, Feras Saad, Alexander Lew, and Vikash Mansinghka. Gen: A general-purpose probabilistic programming system with programmable inference. 11 2018.

[7] Ronald L Davis and Yi Zhong. The biology of forgetting-a perspective. *Neuron*, 95(3):490–503, 2017. `doi:10.1016/j.neuron.2017.05.039`.

[8] Narges Doostani, Gholam-Ali Hossein-Zadeh, and Maryam Vaziri-Pashkam. The normalization model predicts responses in the human visual cortex during object-based attention. *eLife*, 12:e75726, April 2023. `doi:10.7554/eLife.75726`.

[9] Tom Hartley, Chris M. Bird, Dennis Chan, Lisa Cipolotti, Masud Husain, Faraneh Vargha-Khadem, and Neil Burgess. The hippocampus is required for short-term topographical memory in humans. *Hippocampus*, 17(1):34–48, 2007. `doi:10.1002/hipo.20240`.

[10] P. Hemmer and M. Steyvers. A bayesian account of reconstructive memory. *Topics in Cognitive Science*, 1(1):189–202, 2009.

[11] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, and et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015.

[12] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. 2022. `arXiv:1312.6114`.

[13] Dmitry Krotov and John J Hopfield. Dense associative memory for pattern recognition, 2016. `arXiv:1606.01164`.

[14] Hsin-Hung Li and Clayton E. Curtis. Neural population dynamics of human working memory. *Current Biology*, 33(17):3775–3784.e4, 2023. URL: `https://www.sciencedirect.com/science/article/pii/S0960982223010394`, `doi:10.1016/j.cub.2023.07.067`.

[15] Alison R. Preston and Howard Eichenbaum. Interplay of hippocampus and prefrontal cortex in memory. *Current Biology*, 23(17):R764–R773, 2013. `doi:10.1016/j.cub.2013.05.041`.

[16] Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Thomas Adler, Lukas Gruber, Markus Holzleitner, Milena Pavlović, Geir Kjetil Sandve, Victor Greiff, David Kreil, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. Hopfield networks is all you need. 2021. arXiv:2008.02217.

[17] Edmund T Rolls. A quantitative theory of the functions of the hippocampal ca3 network in memory. *Frontiers in cellular neuroscience*, 7:98, 2013. doi:10.3389/fncel.2013.00098.

[18] D. L. Schacter. Adaptive constructive processes and the future of memory. *American Psychologist*, 67(8):603–613, 2012. doi:10.1037/a0029869.

[19] Eleanor Spens and Neil Burgess. A generative model of memory construction and consolidation. *bioRxiv*, page 2023.01.19.524711, 01 2023. doi:10.1101/2023.01.19.524711.

[20] Larry R Squire, Lisa Genzel, John T Wixted, and Richard G Morris. Memory consolidation. *Cold Spring Harbor perspectives in biology*, 7(8):a021766, 2015. doi:10.1101/cshperspect.a021766.

[21] Joel L. Voss, Donna J. Bridge, Neal J. Cohen, and Jeffrey A. Walker. A closer look at the hippocampus and memory. *Trends in Cognitive Sciences*, 21(8):577–588, 2017. doi:10.1016/j.tics.2017.05.008.

[22] H Freyja Ólafsdóttir, Daniel Bush, and Caswell Barry. The role of hippocampal replay in memory and planning. *Current Biology*, 28(1):R37–R50, 2018. doi:10.1016/j.cub.2017.10.073.