Attention rules Episodic Memory

Zahra Fayyaz (zahra.fayyaz@ini.rub.de)
Institute for Neural Computation, Computer Science Department, Universitätstr. 150
44801 Bochum, Germany

Sen Cheng (sen.cheng@ini.rub.de)
Institute for Neural Computation, Computer Science Department, Universitätstr. 150
44801 Bochum, Germany

Laurenz Wiskott (laurenz.wiskott@ini.rub.de)
Institute for Neural Computation, Computer Science Department, Universitätstr. 150
44801 Bochum, Germany

Abstract

Attention plays a crucial role in memory and learning by prioritizing relevant information and filtering out redundant input. This study explores how attention, guided by semantic memory, enhances memory encoding and retrieval. We present a neural network model to simulate generative episodic memory, comprising a VQ-VAE encoder, an attention module, and a transformer-based semantic decoder. Three attention strategies (random, selective, and additive) were evaluated. Random attention, lacking prioritization, led to lowest memory accuracy. Selective attention, informed by semantic prediction, improved performance by focusing on novel, informative inputs. Additive attention, inspired by biological saccades, offered the highest performance through iterative, predictive refinement of input encoding, albeit at a higher computational cost. Furthermore, experiments on both MNIST and ImageNet datasets demonstrate that semantically-guided attention leads to more structured and less prototypical memory traces. These findings underscore the dynamic interplay between attention and memory, suggesting that attentional mechanisms shaped by prior knowledge significantly optimize learning and memory.

Keywords: Generative episodic memory; Semantic memory; Attention; Active learning; Computational modeling

Introduction

Attention is a cornerstone of cognitive processes, playing a pivotal role in how we perceive, process, and interact with the world. As a mechanism for selectively focusing mental resources, attention ensures that relevant information is prioritized while extraneous details are filtered out. This selective focus is essential for managing the overwhelming influx of sensory inputs and is critical for higher-order functions like decision-making, learning, and memory (Lindsay et al., 2020). Memory, as both a repository and a process, is deeply intertwined with attentional mechanisms (Aly & Turk-Browne, 2017; Cowan et al., 2024). Decades of research have shown that attention enhances the encoding of information, bolsters its retention, and facilitates its retrieval.

An often overlooked aspect of this dynamics is the influence of semantic memory — the repository of general knowledge and facts — on attention. Prior knowledge serves as a guide to direct attention toward the most relevant aspects of an environment and suppress predictable or redundant input. Attention not only optimizes memory storage by reducing the cognitive load of processing familiar information but also enables more efficient memory by emphasizing novel, informative data.

This paper argues that effective attention patterns, shaped by prior knowledge, enhance memory outcomes by reducing the need to store predictable inputs. By exploring the interplay between attention, memory, and prior knowledge, we aim to shed light on the cognitive and computational mechanisms that underlie efficient memory.

Methods and Results

We use a neural network to model the encoding and retrieval of episodic memory as a generative process. Our network consists of (i) an encoder module (encoder part of a vector-quantized variational autoencoder (VQ-VAE)) modeling the visual system, which compresses an input image into a more abstract representation, (ii) an attention module that masks out the irrelevant parts of this latent representation and stores the attended part as a memory trace, and (iii) a decoder module, which performs semantic completion using a bidirectional transformer network on the memory trace and then decodes it through the VQ-VAE decoder to reconstruct the memory.

VQ-VAE and transformer models are high-level abstractions that omit many biological details and provide a useful computational perspective for our modeling and discussion. The VQ-VAE encoder parallels the feedforward processing of the visual systemtransforming raw inputs into abstract representations, reminiscent of object encoding in the inferior temporal cortex (Yamins et al., 2014; Kuzovkin et al., 2018; Lindsay, 2021). The decoder mimics top-down feedback that reconstructs visual information during recall (Xia et al., 2015; Takeda, 2019; Al-Tahan & Mohsenzadeh, 2021). This is a process we use for visualization rather than as a literal mapping of neural reactivation. Meanwhile, the transformer learns the statistical relationships among the VQ-VAE's latent vectors, analogous to how the brain acquires semantic information from repeated experiences (Michaelian, 2011) and can fill in missing elements in a semantically consistent manner (Tang et al., 2018; Carrillo-Reid & Yuste, 2020). for more information about the biological significance of the base model see Fayyaz et al. (2022) and Reyhanian et al. (2024).

To guide attention in a principled and context-sensitive manner, we use the predictions of the last layer of the transformer just before they are converted into probabilities of different features using the softmax function.These scores are known as logits. We show that these logits are a good proxy for the model's internal confidence in its predictions. Specifically, lower logit values – ususally corresponding to more evenly distributed softmax probabilities – were interpreted as indicators of higher epistemic uncertainty. This uncertainty signal is used to select the parts of the input the model was least confident about, prioritizing information that was harder for the model to predict and, therefore, more informative.

As a stimulus set we constructed a dataset of two-digit numerals by concatenating images from the MNIST dataset (Fayyaz et al., 2025). The model was trained exclusively on numerals divisible by three, adhering to specific digit pairing rules. Training was conducted in a self-supervised manner without access to explicit digit labels. To evaluate the model's performance, we employed a supervised convolutional neural network classifier trained on the original MNIST dataset. This approach allowed us to measure the accuracy of digit recognition on the reconstructed numerals for the three different attention methods. We tested on both congruent (divisible by three) and incongruent (not divisible by three, i.e., out of distri-

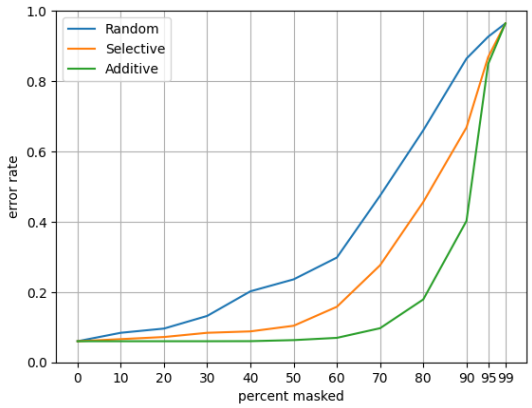
bution) to assess the effect of existing or violation of relevant semantic knowledge on memory performance.

In *random* attention, which is our baseline, the attention is distributed randomly across image representation, which led to low overall memory retrieval accuracy for all different attention levels (Fig. 1). The lack of prioritization meant that both predictable and unpredictable information is treated equally, overloading cognitive resources and diluting the quality of encoded representations. In contrast, *selective* attention feeds the entire encoded input into the semantic system and uses the transformer's logits to select the least predictable parts of the input. This input-dependent filtering significantly enhances memory accuracy by prioritizing informative parts and suppressing predictable ones, resulting in more efficient memory usage without incurring much extra computational cost. *Additive* attention, modeled after biological saccades, processes inputs one part at a time, allowing the semantic network to predict the next position that needs to be attended to. First, the fully masked latent representation was passed to the transformer, and then the part(s) with the lowest logit(s) were attended to. This process was repeated multiple times until the intended masking level was reached. While this strategy was more time-consuming, it achieved the highest accuracy for a given level of attention.

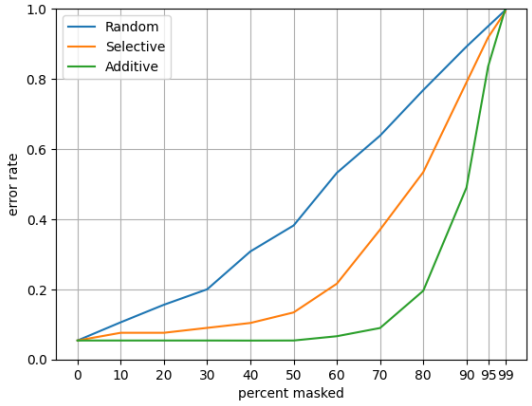
Attentional mechanisms guided by prior knowledge help optimize memory even with limited attention. This effect is especially pronounced for incongruent data (Fig. 1b), suggesting that optimized attention focuses on the most informative features, particularly when semantic predictions are unreliable. Generally, congruent cases show lower classification error, with the gap being largest for the random baseline. In contrast, selective and additive attention store key anchoring features that support accurate reconstruction, even when semantic cues are misleading. This reflects a competition between memory traces and semantic priors: when memory provides strong evidence for a specific digit, it can override the bias toward congruent interpretations. Consequently, we expect memories from selective and additive attention to show lower prototypicality -that is, they will be less similar to an average or generalized class representation-, indicating that the model retains distinct, task-relevant features of each stimulus.

Furthermore, tracking attention dynamics throughout training revealed a progressive refinement of attentional filters: initially diffuse, attention became increasingly precise as network accumulated semantic knowledge. This shift underscores the iterative interplay between learning semantics and attention, where learning continually reshapes attentional priorities to optimize future encoding.

We initially used the MNIST stimuli to study these different attention mechanisms and then repeated the same experiments with the ImageNet dataset in a much larger network. While the effects were more pronounced with simpler MNIST inputs, the attention-driven improvements persisted in the more complex ImageNet dataset, illustrating the robustness of the attention-memory interaction across varying input complexities.



(a) Congruent test data



(b) Incongruent test data

Figure 1: Classification error of the outputs for different masking levels using three attention methods for the double digit MNIST test data that was congruent (1a) with the training data and out of distribution data that is incongruent with the training data (1b)

Conclusion and Future Work

This study demonstrates that attention, guided by semantic information, significantly enhances the encoding and retrieval of episodic memory. Additive attention achieved the highest retrieval accuracy by iteratively refining memory representations, while selective attention offered substantial gains with lower computational demands. These findings highlight the role of attention as an active, knowledge-driven mechanism that optimizes learning efficiency and memory formation.

In the future, we will further investigate how attention can enhance the training process. Specifically, we will explore the use of learned attention to inform data sampling, curriculum learning, and uncertainty-aware training strategies. Although the current results are not yet finalized, we anticipate that integrating these dynamic attention mechanisms into the training process will improve generalization and learning efficiency, particularly in complex or data-scarce environments.

Acknowledgement

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), grant 397530566, FOR 2812, P5 (Wiskott,Cheng).

References

Al-Tahan, H., & Mohsenzadeh, Y. (2021). Reconstructing feedback representations in the ventral visual pathway with a generative adversarial autoencoder. *PLoS Computational Biology*, *17*(3), e1008775. doi: 10.1371/journal.pcbi.1008775

Aly, M., & Turk-Browne, N. B. (2017). How hippocampal memory shapes, and is shaped by, attention. *The hippocampus from cells to systems: Structure, connectivity, and functional contributions to memory and flexible cognition*, 369–403.

Carrillo-Reid, L., & Yuste, R. (2020). Playing the piano with the cortex: Role of neuronal ensembles and pattern completion in perception and behavior. *Current opinion in neurobiology*, *64*, 89–95. doi: 10.1016/j.conb.2020.03.014

Cowan, N., Bao, C., Bishop-Chrzanowski, B. M., Costa, A. N., Greene, N. R., Guitard, D., ... Ün, Z. E. (2024). The relation between attention and memory. *Annual Review of Psychology*, *75*, 183–214.

Fayyaz, Z., Altamimi, A., Zoellner, C., Klein, N., Wolf, O. T., Cheng, S., & Wiskott, L. (2022, August). A model of semantic completion in generative episodic memory. *Neural Computation*, *34*(9), 1841–1870. doi: 10.1162/neco_a_01520

Fayyaz, Z., Righetti, F., Wiskott, L., & Werning, M. (2025, Feb). *Remembering without (representational) memory: A neurocomputational study on regaining categoricity and compositionality from minimal traces.* OSF Preprints. Retrieved from osf.io/zjprq$_v$1 doi: 10.31219/osf.io/zjprq$_v$1

Kuzovkin, I., Vicente, R., Petton, M., Lachaux, J.-P., Baciu, M., Kahane, P., ... Aru, J. (2018). Activations of deep convolutional neural networks are aligned with gamma band activity of human visual cortex. *Communications biology*, *1*(1), 1–12. doi: 10.1038/s42003-018-0110-y

Lindsay, G. W. (2021). Convolutional neural networks as a model of the visual system: Past, present, and future. *Journal of cognitive neuroscience*, *33*(10), 2017–2031. doi: 10.1162/jocn$_{a0}$1544

Lindsay, G. W., et al. (2020). Attention in psychology, neuroscience, and machine learning. *Frontiers in Computational Neuroscience*, *14*, 29.

Michaelian, K. (2011). Generative memory. *Philosophical Psychology*, *24*(3), 323–342. doi: 10.1080/09515089.2011.559623

Reyhanian, S., Fayyaz, Z., & Wiskott, L. (2024, September). Analysis of a generative model of episodic memory based on hierarchical vq-vae and transformer. In *Proceedings of the 33rd international conference on artificial neural networks (icann 2024), lugano, switzerland.* Springer Nature Switzerland. doi: 10.1007/978-3-031-72341-4$_6$

Takeda, M. (2019). Brain mechanisms of visual long-term memory retrieval in primates. *Neuroscience research*, *142*, 7–15. doi: 10.1016/j.neures.2018.06.005

Tang, H., Schrimpf, M., Lotter, W., Moerman, C., Paredes, A., Caro, J. O., ... Kreiman, G. (2018). Recurrent computations for visual pattern completion. *Proceedings of the National Academy of Sciences*, *115*(35), 8835–8840. doi: 10.1073/pnas.1719397115

Xia, R., Guan, S., & Sheinberg, D. L. (2015). A multilayered story of memory retrieval. *Neuron*, *86*(3), 610–612. doi: 10.1016/j.neuron.2015.04.017

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, *111*(23), 8619–8624. doi: 10.1073/pnas.1403112111