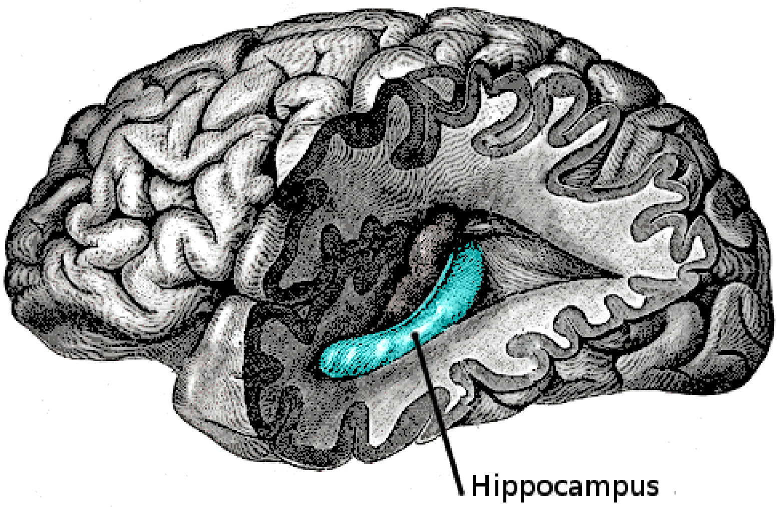
SSCP

SUMMER SCHOOL IN COMPUTATIONAL PHYSIOLOGY

Project 2025

Exploring Neural Models of Hippocampal Memory

Supervisors:
Nicolai Haug (nicolaih@simula.no)
Mikkel Lepperød (mikkel@simula.no)



Project Goal

The goal of this project is to explore computational models of memory, specifically focusing on mechanisms thought to be implemented in the hippocampus and related brain regions. You will investigate how artificial neural network architectures can capture key aspects of biological memory function, such as forming cognitive maps or recalling information based on learned associations.

1 Introduction

Memory is the faculty by which the brain encodes, stores, and retrieves information. This complex process serves as a record of experience that is essential for learning, guiding present actions, and planning for the future. Located deep within the brain's medial temporal lobes, the hippocampal formation is one of the most studied neuronal systems and has long been a central focus in memory research (Knierim, 2015). Its critical role was famously and tragically illustrated by the case of patient H.M., who, after surgical removal of the hippocampus, was left with an inability to form new long-term memories (Scoville and Milner, 1957). This case provided the foundational evidence that the hippocampus is essential for transforming fleeting short-term experiences into lasting memories, a process known as *memory consolidation*. The hippocampus is now understood to be critical for both episodic memory—the rich, contextual record of personal experiences—and spatial memory, which involves creating cognitive maps of the environment (Burgess et al., 2002). To accomplish these functions, the hippocampus performs complementary computations linked to different hippocampal subfields. It uses *pattern separation* to store similar experiences as distinct memories, preventing interference, and *pattern completion* to retrieve a full memory from a partial or noisy cue (Rolls, 2013).

A central challenge in neuroscience is understanding precisely how the hippocampus supports memory through its computations. To explore this question, researchers increasingly use computational modeling, in particular deep neural networks, to simulate and dissect the underlying biological mechanisms. This project serves as a practical introduction to this modern approach, in which you will implement and experiment with influential deep neural network architectures developed to simulate key hippocampal memory functions.

2 Models and Project Tasks

At the project's onset, your group and the supervisors will decide on the project's focus together. This collaborative approach ensures the project is both exciting and appropriately tailored to the team's collective experience and interests.

To help kick off the discussion, the following sections introduces some influential models of memory that your group may choose to explore.

2.1 Hopfield Networks

2.1.1 The Classical Hopfield Network

One of the earliest and most influential computational models of memory is the *Hopfield network*, introduced by John Hopfield in 1982. A Hopfield network is a type of *recurrent* *neural network* (RNN) where all neurons are connected to each other. It functions as a form of content-addressable memory, meaning it can retrieve a full memory based on the content of a partial or noisy cue. The network operates through two distinct phases. In the storage phase, memories are encoded into the synaptic weights according to a Hebbian-like principle ("neurons that fire together, wire together"), establishing each pattern as a stable attractor state in the network's dynamics. Subsequently, in the retrieval phase, the network dynamically converges to the nearest attractor state when presented with a partial or noisy cue, thereby completing the pattern and recalling the full memory.

The Hopfield network's retrieval mechanism provides a powerful and elegant implementation of pattern completion, one of the key computations attributed to the hippocampus. For this reason, the Hopfield network has long been considered a foundational model for understanding how the extensive recurrent connections in the hippocampal CA3 subfield might support memory recall (Rolls, 2013). While the classical Hopfield network has known limitations, such as a finite storage capacity and difficulty distinguishing between similar patterns, its principles remain fundamental for illustrating how associative memory can emerge from recurrent circuits.

2.1.2 Modern Hopfield Networks

For decades, the classical Hopfield network was viewed primarily as a foundational model with practical limitations, most notably its linear storage capacity—it could not store many more patterns than it had neurons without significant recall errors. However, a recent resurgence of interest, spearheaded by Krotov and Hopfield himself starting in 2016, has given rise to *modern Hopfield networks*, sometimes called *Dense Associative Memories*. The key innovation in these modern networks is a more complex energy function that allows for higher-order interactions between neurons. This architectural change creates sharper and more numerous attractor states, breaking the old linear scaling limit and enabling the network to store reliably retrieve a super-linear, or even exponential, number of memories relative to its size.

The massive increase in capacity has repositioned the Hopfield network from a historical model to a powerful tool in modern deep learning. Interestingly, researchers have shown that the *attention* mechanism, which is the cornerstone of the highly successful *Transformer* architecture, can be understood as a form of modern Hopfield network (Ramsauer et al., 2021). This insight has revitalized the architecture, establishing it as a cutting-edge method for building highly efficient and high-capacity associative memories.

2.2 Kanerva Machines

Sparse distributed memory was proposed by Pentti Kanerva in 1988 as a model for human long-term memory. It is an associative memory model that operates in a very high-dimensional binary space. Kanerva's original model is characterized by its distributed reading and writing operations, utilizing a fixed table of addresses and a modifiable memory matrix. The model can be used for storing and retrieving large amounts of information without focusing on the accuracy but rather on similarity of information. Despite its robustness and ability to retrieve patterns even from corrupted queries through iterative reading, its practical application is limited by the assumption of uniform and binary data distributions, which is rarely true for real-world data.

Building upon Kanerva's foundational ideas, the *Kanerva machine* (Wu et al., 2018) was developed to overcome the limitations of the original model by creating a fully differentiable, generative memory system that can learn and adapt directly to complex data. Unlike traditional slot-based memories, a Kanerva machine interpret memory operations, specifically writes and reads, as *Bayesian inference* within a generative model, where the memory itself is represented as a probabilistic distribution. Because the memory is represented as a probabilistic distribution, it can be updated online using an optimal Bayesian rule. A key feature of the Kanerva machine is its use of iterative retrieval. When presented with a cue, the model's underlying attractor dynamics cause its internal state to converge toward a clean, stable representation, thereby denoising the input and completing the pattern. Conceptually, the Kanerva machine can be viewed as an extension of a *Variational Autoencoder* (VAE) (Kingma and Welling, 2014), but one that uses its memory to learn a rich, data-dependent prior distribution. This design allows the model to effectively combine top-down knowledge from its learned prior with bottom-up sensory information when encoding an observation.

References

Knierim, J. J. (2015). “The Hippocampus”. In: *Current Biology* 25.23, R1116–R1121. DOI: 10.1016/j.cub.2015.10.049 (cit. on p. 1).

Scoville, W. B. and B. Milner (1957). “Loss of recent memory after bilateral hippocampal lesions”. In: *Journal of Neurology, Neurosurgery, and Psychiatry* 20.1, pp. 11–21. DOI: 10.1136/jnnp.20.1.11 (cit. on p. 1).

Burgess, N., E. A. Maguire, and J. O’Keefe (2002). “The human hippocampus and spatial and episodic memory”. In: *Neuron* 35.4, pp. 625–641. DOI: 10.1016/S0896-6273(02)00830-9 (cit. on p. 1).

Rolls, E. T. (2013). “The mechanisms for pattern completion and pattern separation in the hippocampus”. In: *Frontiers in Systems Neuroscience* 7, p. 74. DOI: 10.3389/fnsys.2013.00074 (cit. on pp. 1, 2).

Hopfield, J. J. (1982). “Neural networks and physical systems with emergent collective computational abilities”. In: *Proceedings of the National Academy of Sciences* 79.8, pp. 2554–2558. DOI: 10.1073/pnas.79.8.2554 (cit. on p. 1).

Krotov, D. and J. J. Hopfield (2016). “Dense Associative Memory for Pattern Recognition”. In: *Advances in Neural Information Processing Systems*. Ed. by D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett. Vol. 29. URL: https://proceedings.neurips.cc/paper_files/paper/2016/file/eaae339c4d89fc102edd9dbdb6a28915-Paper.pdf (cit. on p. 2).

Ramsauer, H. et al. (2021). “Hopfield Networks is All You Need”. In: *ICLR*. URL: https://arxiv.org/abs/2008.02217 (cit. on p. 2).

Kanerva, P. (1988). *Sparse Distributed Memory*. 1st ed. MIT Press. ISBN: 978-0-262-11132-4 (cit. on p. 2).

Wu, Y., G. Wayne, A. Graves, and T. and Lillicrap (2018). “The Kanerva Machine: A Generative Distributed Memory”. In: *ICLR*. URL: https://openreview.net/pdf?id=S1HlA-ZAZ (cit. on p. 3).

Kingma, D. P. and M. Welling (2014). *Auto-Encoding Variational Bayes*. arXiv: 1312.6114 (cit. on p. 3).