Investigating Hebbian Alternatives to Dense Associative Memory

Connor Hanley

October 10, 2025

Abstract

Dense Associative Memories generalize traditional Hopfield networks while providing a substantially increased capacity. To do this, they show that by increasing the separation of similarity scores between query patterns and stored patterns. The cost of increased capacity is simplicity and biological plausibility. We propose to investigate the capacity characteristics of all Hebbian associative memories, to see if there exists any alternative to Dense Associative Memories which maintains the simplicity and appeal of Hopfield networks.

Humans are able to recognize and retrieve patterns of data using distorted, noisy, and partial patterns (Rumelhart et al., 1986). This capacity of human memory is known as *content-addressability*: patterns which are stored in memory are able to be "looked up" by themselves or their parts. Modeling this property is a classical task in computational cognitive and neuroscience (see Amari (1972), Little (1974), Marr (1971), Nakano (1972), and Stanley (1976)). The family of models which implement content-addressability are known as *associative memory models* (AMs). A recent revival of interest AMs in machine learning research, driven by their equivalence with "attention" layers in the transformer architecture (Ramsauer et al., 2021; Vaswani et al., 2023), has led to drastic advances in the storage capacity of AMs (Demircigil et al., 2017; Hu et al., 2024; Krotov & Hopfield, 2016).

The foundational model for modern associative memory research is the *Hopfield network* (J. J. Hopfield, 1984; J. Hopfield, 1982). Hopfield networks are single-layer neural networks with full, lateral connections. This means that each artificial neuron in the network is connected with every other neuron, except itself. Patterns that we wish to recall from the network are stored in the connection weights between every other unit. Hopfield networks learn new patterns to recall through a simple, biologically plausible learning rule called the *Activity Product Rule* (Haykin, 2009), which is a kind of Hebbian update rule (Hebb, 1949). The appeal of Hebbian update rules is that they are simple, in that they are totally defined by local interactions between layers in neural networks, and they are biologically plausible, having been observed in interactions between biological neurons (Bi & Poo, 1998; Markram et al., 1997; Rolls, 2013).

In spite of their simplicity and biological plausibility, Hopfield networks are inherently flawed. The number of patterns that Hopfield networks can store is pitifully small: with estimates in the range of 14% to 15% of the number of neurons in the network (Amit et al., 1987; J. Hopfield, 1982). In order to remedy the gap between the simplicity and plausibility of Hopfield networks and their child Dense Associative Memory models, we will theoretically derive and empirically test the storage capacities of all Hebbian update rules when used in single-layer neural networks with full lateral connections.

Discovering the limits of Hebbian alternatives to Dense Associative Memories is not only desirable because of simplicity and biological plausibility (while those are both laudable goals). For example, the study of Vector-Symbolic Architectures, used for representing high-level cognitive-tasks (Gayler, 1998; Kelly et al., 2013; Plate, 1995; Smolensky, 1990), requires an auto-associative memory "clean-up memory". This means that the expressive capabilities of Vector-Symbolic Architectures are affected by the capacity of the associative memory used.

In the following we will discuss the problem in more depth. To do this, we will first lay out the theoretical motivation: beginning with an introduction to the literature in section 1, covering associative memories in subsection 1.1, Hopfield networks in subsection 1.2, their generalization with Dense Associative Memories in subsection 1.3, and Hebbian learning and learning rules in subsection 1.4. We will restate the problem of Hebbian Dense Associative Memories in section 2, and provide a work plan in subsection 2.1. Finally, we will discuss the timeline for completing this project in section 3.

1 Background

In order to understand the need for a Hebbian alternative to Dense Associative Memories, we must first familiarize ourselves with the literature up to this point. Associative Memory research has a wealth of literature and alternatives. In order to limit the scope of this research, we limit ourselves only to the family of models based on Hopfield networks (J. Hopfield, 1982).

1.1 Associative Memories

An associative memory, in particular an *auto*-associative memory, is a general structure that can be implemented by many mathematical and computational objects. Broadly speaking, an associative memory is a tuple of a set of objects $\mathcal{X}$, called the set of *traces* or *stored patterns*, and a (typically learned) identity map over the set of stored patterns, $T : \mathcal{X} \to \mathcal{X}$. More formally,

Definition 1.1.1 (Associative memories)**.** An (auto) *associative memory* is a tuple $\langle \mathcal{X}, T \rangle$, where

(a) $\mathcal{X}$ is a set of objects, called the *pattern* set, or set of *traces*, typically high-dimensional vectors; and,

(b) The *recall function*, T, maps from the set of traces back to the set of traces, such that $T(x) \approx x$, for all $x \in \mathcal{X}$, and $T(\overline{x}) \approx x$, where $\overline{x}$ is a perturbed, masked, or degraded form of $x \in \mathcal{X}$.

We will be assuming throughout that the pattern set is the set of rows of the *pattern matrix* $\Xi = [\xi^1, \xi^2, \ldots, \xi^N]$, where each *stored pattern* $\xi^i \in \{-1, 1\}^D$. It is sufficient to specify the pattern matrix, or the sequence of patterns, for an associative memory, and it should be assumed that the pattern set is just composed of the rows of the pattern matrix. Associative memories are *content-addressable*, as their recall function is able to recover desired associated patterns from partial information (Haykin, 2009; McClelland et al., 1986). This is in contrast to memory models which recall information based on arbitrary associations, like indices into memory locations.

While it is not essential to the most general definition of associative memories, we usually desire that the recall function be learned. Furthermore, associative memories have a maximum capacity of traces that they can store.

Definition 1.1.2 (Retrievability; Capacity)**.** Following Bao et al. (2022), let us have an associative memory with patterns $\xi^1, \xi^2, \ldots, \xi^N$, where each element ξ^i_j is -1 or 1 with equal likelihood. For any $\delta < 0.5$, we define the δ-perturbation of ξ^i, $\bar{\xi}^i$, as the D-dimensional vector ξ^i with each element flipped with a likelihood of δ. Then, we say that the set $\xi^1, \xi^2, \ldots, \xi^N$ is (δ, ε)-retrievable if for every ξ^i, $i = 1, 2, \ldots, N$ it is such that:

$$\mathbb{P}(T(\bar{\xi}^i) \neq \xi^i) < \varepsilon. \tag{1}$$

The *capacity* C of the associative memory is the maximum cardinality of the pattern set such that all patterns are (δ, ϵ)-retrievable.

In Krotov and Hopfield (2016), capacity is denoted by $N^{\max}$, where N is the first dimension of the pattern matrix. In Correlation Matrix associative memories, capacity is affected by the (1) kind of data being stored, i.e. if the data has a high pairwise correlation, which leads to "cross-talk" (Kohonen, 1988), (2) the dimensionality of the data, and (3) the implementation of the recall function T. For example, if the associative memory is a neural network with tensor weights (Kelly et al., 2017), then the capacity scales with the number of the weights (Little & Shaw, 1978) (expanded further in subsection 1.3).

Before Hopfield networks, there were Correlation Matrix associative memories. So-called, because they relied on dot-product correlations for recall.

Example 1.1.1 (Correlation Matrix associative memory)**.** *Let us have a pattern matrix $\Xi = [\xi^1, \xi^2, \ldots, \xi^N]$ of vectors sampled from $\{1, -1\}^D$. The Correlation Matrix associative memory recall function is of the form:*

$$\begin{aligned} T(x) &= g\left(\sum_{\mu=1}^{N} \xi^\mu \left(\sum_{i=1}^{D} \xi_i^\mu x_i\right)\right) \\ &= g\left(\Xi^\top \Xi x\right), \end{aligned} \tag{2}$$

where g is an "activation function", typically signum for binary vectors.

Example 1.1.2. *Like above, let us have a pattern matrix* $\Xi = [\xi^1, \xi^2, \ldots, \xi^N]$ *of vectors sampled from* $\{1, -1\}^D$*. Let our recall function* T *be:*

$$T(x) = \xi^i, \text{ where } i = \arg\max_{i \in [1,N]} [sim(\xi^i, x)], \tag{3}$$

where sim is some similarity function (Kelly et al., 2017), e.g. the cosine similarity or Hamming distance.

Example 1.1.3 (Minerva2)**.** *Minerva2 (Hintzman, 1984) is an associative memory used in cognitive science which is about as old as Hopfield networks. As per usual, we have binary patterns* $\Xi = [\xi^1, \xi^2, \ldots, \xi^N]$*. The recall function is:*

$$T(x) = sgn\left[\sum_{\mu=1}^{N} \xi^\mu \left(\frac{\sum_{i=1}^{D} \xi_i^\mu x_i}{\|\xi^\mu\| \|x\|}\right)^3\right]. \tag{4}$$

1.2 Hopfield Networks

Hopfield networks (J. J. Hopfield, 1984; J. Hopfield, 1982) are an associative memory with deep connections to 1.1.1, except that they understand their recall function in terms of an *energy* function:

Definition 1.2.1 (Hopfield networks)**.** Hopfield networks are single-layer neural networks of D computational units with full, lateral connections.

1.3 Dense Associative Memories

Dense Associative Memories generalize Hopfield networks, and improve capacity as a function of both the dimension of the network, and the exponent used in their polynomial function. Talk about related work with single-pass networks (UHNs), as well as neuron-astrocyte memory.

1.4 Hebbian Learning Rules

Hopfield networks learn with the Activity Product Rule (Haykin, 2009). Talk about biological evidence for Hebbian learning rules. Stress simplicity.

2 Hebbian Dense Associative Memory

Outline of the project. Hopfield networks which, instead of manipulating a polynomial function, rather manipulate how patterns are stored to begin with.

2.1 Work Plan

Two main components of the project: theoretical and experimental. Theoretical involves proving some theorems about the minimum and maximum capacity of

Hopfield networks with different Hebbian learning rules. Experimental work means implementing the different networks and testing them on a task. Suggest testing them on classification of degraded inputs.

3 Timeline

Theoretical work should be done by Christmas. Experimental work likewise requires only a couple of months. Data science certification requires emphasis here on large-scale projects. So, I would like to rigorously examine each model on a variety of tasks. One to two papers arises from this project: one talking about the theory of Hebbian DAM, and the other part including the experimental work. It would be better for a single paper, but if that runs too long, then experimental work can be a part 2.

References

Amari, S.-I. (1972). Learning patterns and pattern sequences by self-organizing nets of threshold elements. *IEEE Transactions on Computers*, *C-21*(11), 1197–1206. https://doi.org/10.1109/T-C.1972.223477

Amit, D. J., Gutfreund, H., & Sompolinsky, H. (1987). Statistical mechanics of neural networks near saturation. *Annals of Physics*, *173*(1), 30–67. https://doi.org/10.1016/0003-4916(87)90092-3

Bao, H., Zhang, R., & Mao, Y. (2022). The capacity of the dense associative memory networks. *Neurocomputing*, *469*, 198–208. https://doi.org/10.1016/j.neucom.2021.10.058

Bi, G.-q., & Poo, M.-m. (1998). Synaptic modifications in cultured hippocampal neurons: Dependence on spike timing, synaptic strength, and postsynaptic cell type. *The Journal of Neuroscience*, *18*(24), 10464–10472. https://doi.org/10.1523/JNEUROSCI.18-24-10464.1998

Demircigil, M., Heusel, J., Löwe, M., Upgang, S., & Vermet, F. (2017). On a model of associative memory with huge storage capacity. *Journal of Statistical Physics*, *168*(2), 288–299. https://doi.org/10.1007/s10955-017-1806-y

Gayler, R. W. (1998). *Multiplicative binding, representation operators & analogy (workshop poster)*. Retrieved October 10, 2025, from https://web-archive.southampton.ac.uk/cogprints.org/502/

Haykin, S. S. (2009). *Neural networks and learning machines* (3. ed). Prentice-Hall.

Hebb, D. O. (1949). *The organization of behavior: A neuropsychological theory* (1. issued in paperback). Routledge.

Hintzman, D. L. (1984). MINERVA 2: A simulation model of human memory. *Behavior Research Methods, Instruments, &amp Computers*, *16*(2), 96–101. https://doi.org/10.3758/BF03202365

Hopfield, J. J. (1984). Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the National Academy of Sciences*, *81*(10), 3088–3092. https://doi.org/10.1073/pnas.81.10.3088

Hopfield, J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, *79*, 2554–2558.

Hu, J. Y.-C., Wu, D., & Liu, H. (2024, October 31). Provably optimal memory capacity for modern hopfield models: Transformer-compatible dense associative memories as spherical codes. https://doi.org/10.48550/arXiv.2410.23126

Kelly, M. A., Blostein, D., & Mewhort, D. J. K. (2013). Encoding structure in holographic reduced representations [Place: US Publisher: Educational Publishing Foundation]. *Canadian Journal of Experimental Psychology / Revue canadienne de psychologie expérimentale*, *67*(2), 79–93. https://doi.org/10.1037/a0030301

Kelly, M. A., Mewhort, D., & West, R. L. (2017). The memory tesseract: Mathematical equivalence between composite and separate storage memory models. *Journal of Mathematical Psychology*, *77*, 142–155. https://doi.org/10.1016/j.jmp.2016.10.006

Kohonen, T. (1988). Correlation matrix memories. In J. A. Anderson & E. Rosenfeld (Eds.), *Neurocomputing, volume 1* (pp. 174–180). The MIT Press. https://doi.org/10.7551/mitpress/4943.003.0075

Krotov, D., & Hopfield, J. J. (2016, September 27). Dense associative memory for pattern recognition. https://doi.org/10.48550/arXiv.1606.01164

Little, W. (1974). The existence of persistent states in the brain. *Mathematical Biosciences*, *19*(1), 101–120. https://doi.org/https://doi.org/10.1016/0025-5564(74)90031-5.

Little, W., & Shaw, G. L. (1978). Analytic study of the memory storage capacity of a neural network. *Mathematical Biosciences*, *39*(3), 281–290. https://doi.org/10.1016/0025-5564(78)90058-5

Markram, H., Lübke, J., Frotscher, M., & Sakmann, B. (1997). Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. *Science*, *275*(5297), 213–215. https://doi.org/10.1126/science.275.5297.213

Marr, D. (1971). Simple memory: A theory for archicortex. *Philosophical Transactions of the Royal Society of London*, *262*(841), 23–81.

McClelland, J. L., Rumelhart, D. E., & Hinton, G. E. (1986, July 17). The appeal of parallel distributed processing. In *Parallel distributed processing: Explorations in the microstructure of cognition: Foundations* (pp. 3–44, Vols. 2, Vol. 1). The MIT Press. Retrieved August 28, 2025, from https://doi.org/10.7551/mitpress/5236.001.0001

Nakano, K. (1972). Associatron-a model of associative memory. *IEEE Transactions on Systems, Man, and Cybernetics*, *SMC-2*(3), 380–388. https://doi.org/10.1109/TSMC.1972.4309133

Plate, T. (1995). Holographic reduced representations. *IEEE Transactions on Neural Networks*, *6*(3), 623–641. https://doi.org/10.1109/72.377968

Ramsauer, H., Schäfl, B., Lehner, J., Seidl, P., Widrich, M., Adler, T., Gruber, L., Holzleitner, M., Pavlović, M., Sandve, G. K., Greiff, V., Kreil, D., Kopp, M., Klambauer, G., Brandstetter, J., & Hochreiter, S. (2021, April 28). Hopfield networks is all you need. https://doi.org/10.48550/arXiv.2008.02217

Rolls, E. T. (2013). The mechanisms for pattern completion and pattern separation in the hippocampus. *Frontiers in Systems Neuroscience*, *7*. https://doi.org/10.3389/fnsys.2013.00074

Rumelhart, D. E., Hinton, G. E., & McClelland, J. L. (1986, July 17). A general framework for parallel distributed processing. In *Parallel distributed processing: Explorations in the microstructure of cognition: Foundations* (pp. 45–76, Vols. 2, Vol. 1). The MIT Press. Retrieved August 28, 2025, from https://doi.org/10.7551/mitpress/5236.001.0001

Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, *46*(1), 159–216. https://doi.org/10.1016/0004-3702(90)90007-M

Stanley, J. C. (1976). Simulation studies of a temporal sequence memory model. *Biological Cybernetics*, *24*(3), 121–137. https://doi.org/10.1007/BF00364115

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023, August 2). Attention is all you need. https://doi.org/10.48550/arXiv.1706.03762