# On Sparse Modern Hopfield Model

**Anonymous Author(s)**
**Affiliation**
**Address**
**email**

## Abstract

We introduce the sparse modern Hopfield model as a sparse extension of the modern Hopfield model. Like its dense counterpart, the sparse modern Hopfield model equips a memory-retrieval dynamics whose one-step approximation corresponds to the sparse attention mechanism. Theoretically, a key contribution of this work is the derivation of a closed-form sparse Hopfield energy using the convex conjugate of the sparse entropic regularizer. Building upon this, we derive the sparse memory retrieval dynamics from the sparse energy function and show its one-step approximation is equivalent to the sparse-structured attention. Importantly, we provide a sparsity-dependent memory retrieval error bound which is provably tighter than its dense analog. The conditions for the benefits of sparsity to arise are therefore identified and discussed. In addition, we show that the sparse modern Hopfield model maintains the robust theoretical properties of its dense counterpart, including rapid fixed point convergence and exponential memory capacity. Empirically, we exploit both synthetic and real-world datasets to demonstrate that the sparse Hopfield model outperforms its dense counterpart in many situations.

## 1 Introduction

We address the computational challenges of modern Hopfield models by introducing a sparse Hopfield model. Our sparse continuous Hopfield model equips a memory-retrieval dynamics that aligns with the sparse-structured attention mechanism. By establishing a connection to sparse attention, the proposed model not only offers a theoretically-grounded energy-based model for associative memory but also enables robust representation learning and seamless integration with deep learning architectures. This approach serves as an initial attempt of pushing the correspondence[1] between Hopfield models and attention mechanism [Ramsauer et al., 2020] toward sparse region, both theoretically and empirically, resulting in data-dependent sparsity for meaningful and robust pattern representations, and a focus on the most relevant information for each specific instance.

Hopfield networks are classic Ising-based associative memory models for both biological and artificial neural networks [Hopfield, 1982, 1984], which utilize statistical-mechanical retrieval dynamics to store and retrieve the memory pattern nearest to a specified query. Recently, their modern versions, the modern Hopfield models [Ramsauer et al., 2020], have been proposed and integrated into deep learning architectures via a strong connection with transformer attention, offering enhanced performance, theoretically guaranteed exponential memory capacity, and the ability to handle continuous patterns. In addition, the modern Hopfield models have found success in various applications, such as immunology [Widrich et al., 2020] and large language model [Fürst et al., 2022]. Apart from the elegant connection to attention, theoretical advantages and empirically successes, the modern Hopfield models have been shown to be computationally heavy and vulnerable against noisy queries [Millidge

[1]While this equivalence only holds when the retrieval dynamics is applied exactly once, as originally shown in [Ramsauer et al., 2020] and later emphasized in [Krotov and Hopfield, 2020], it allows us to view modern Hopfield models as generalized attentions with additional functionalities and hence opens new avenues for Hopfield-based architecture designs. See Appendix B for more discussions.

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

et al., 2022]. In particular, the dense output alignments of the retrieval dynamics in modern Hopfield models [Ramsauer et al., 2020] can be computationally inefficient, making models less interpretable and noise-sensitive by assigning probability mass to many implausible outputs (patterns/keys).

To combat above, incorporating sparsity is an essential and common strategy. While there is a vast body of work on sparsifying attention mechanisms [Tay et al., 2022, Beltagy et al., 2020, Qiu et al., 2019, Child et al., 2019, Peters et al., 2019, Martins and Astudillo, 2016], similar developments for the Hopfield models remain less explored. To bridge this gap, we present a sparse Hopfield model that corresponds to the sparsemax attention mechanism [Martins and Astudillo, 2016]. In this paper, we study the sparsification of the modern Hopfield model. The challenges are three-fold:

(C1) **Non-Trivial Sparsification — Sparse Hopfield $\leftrightarrow$ Sparse Attention:** To enable the use of sparse Hopfield models as computational devices (DNN learning models) akin to [Ramsauer et al., 2020], it is essential to achieve *non-trivial* sparsifications that exhibit equivalence to specific sparse attention models. In other words, any meaningful sparsification should extend the established equivalence [Ramsauer et al., 2020] between modern Hopfield models and attention to encompass the sparse domain. While generalizing such equivalence is potentially impactful as it may lay the groundwork for future Hopfield-based methodologies, architecture designs and bio-computing systems (as in [Kozachkov et al., 2022]), the *heuristic* design of the modern Hopfield model poses great difficulty to developing desired sparse models.

(C2) **Introducing Sparsity into Hopfield Models:** Unlike attention mechanisms where sparsification is typically achieved either on the attention matrix (e.g., structured-sparsity [Tay et al., 2020, Child et al., 2019]) or on the element-wise normalization map (e.g., sparsity-inducing maps [Correia et al., 2019, Peters et al., 2019, Martins and Astudillo, 2016]), the sparsification of Hopfield models is applied to *both* the energy function and the memory-retrieval dynamics, where the latter monotonically decreases the Hopfield energy over time. Since attention mechanisms (transformers) are typically not equipped with such a dynamical description, introducing sparsity into Hopfield models while retaining the connection to attention is a less straightforward process.

(C3) **Properties of the Sparse Hopfield Model:** Further, it is unclear how the introduced sparsity may affect different aspects of the model, such as memory capacity, fixed point convergence, retrieval accuracy, and so on. Ideally, we are looking for sparsities that offer provable computational benefits, such as enhanced robustness and increased memory capacity, among others.

Challenges (C1) and (C2) are inherent in Hopfield model, and certain requirements on the design of energy function and retrieval dynamics are inevitable to obtain non-trivial sparse models. Hence, we suppose the sparsified models should satisfy some conditions and verify them accordingly. Concretely, motivated by the anti-Hebbian learning rule for Hopfield models [Földiak, 1990], a formulation for deriving desired sparse Hopfield energy via convex conjugation of entropic regularizers is proposed. Furthermore, by applying Danskin's theorem and convex-concave procedure [Yuille and Rangarajan, 2003, 2001] on the sparse Hopfield energy function, we obtain sparse retrieval dynamics linked to sparse attention. For (C3), the convergence of energy stationary points and retrieval dynamics fixed points are connected via Zangwill's method [Zangwill, 1969]. The sparse retrieval error bound is derived and used to determined the well-separation condition for successful memory storage and retrieval. Lastly, the fundamental limit of memory capacity is derived using the expected separation of random points on spheres [Cai and Jiang, 2012, Brauchart et al., 2018, Ramsauer et al., 2020].

In summary, this work handles sparsification of modern Hopfield models while linking them to sparse attention by addressing the following question:

> Is it possible to develop a theoretically-grounded (non-trivial) sparse Hopfield model capable of storing information or learned prototypes throughout various layers of DNN models?

**Contributions.** We propose the Sparse Modern Hopfield Model. Our contributions are as follows:

- We propose a novel sparse Hopfield model whose retrieval dynamics corresponds to sparsemax attention mechanism. It leads to sparse patterns by design, inheriting both noise robustness and potential computational efficiency[2] from [Martins and Astudillo, 2016], compared to its dense

[2]Note that, the proposed model's sparsity falls under the category of *sparsity-inducing normalization maps* and hence only improves the computational efficiency if the sparsity can be exploited to avoid unnecessary computation, while the forward pass still requires $\mathcal{O}(n^2)$ space complexity, see Appendix B and [Martins and Astudillo, 2016, Sec. 2] for more discussion.

counterparts. This work extends the theoretical understanding of the correspondence between artificial and biological neural networks to sparse region. In addition, the sparse Hopfield layer, a new deep learning component, is introduced with data-dependent sparsity.

- Theoretically, we establish provably blessings from sparsity and identify the conditions under which these benefits arise. We begin by deriving the closed-form sparse Hopfield energy from the convex conjugation of sparse entropic regularizer. Next, we demonstrate the correspondence between sparse Hopfield retrieval dynamics and sparsemax attention. In addition, we prove the fast convergence of the fixed points (also known as memory patterns, attactor states in literature) for the retrieval dynamics and establish the exponential (in pattern size) memory capacity lower bound with *tighter* retrieval error bound, *compared* with modern Hopfield models.
- Empirically, we conduct synthetic and realistic experiments to verify our theoretical results and proposed methodology. Specifically, the sparse Hopfield model outperforms the dense Hopfield model and machine learning baselines in *sparse* Multiple Instance Learning (MIL) problems. This is observed with both *sparse* synthetic and real-world datasets, where the baselines tend to fall short. Moreover, even in cases without data sparsity, our proposed model delivers performance on par with its dense counterpart.

To the best of our knowledge, we are the first to propose a sparse Hopfield model whose retrieval dynamics is equivalent to sparse attention mechanism with provably computational advantages. Methodologically, the proposed model complements existing Hopfield-based DNN architectures [Hoover et al., 2023, Paischer et al., 2022, Seidl et al., 2022, Fürst et al., 2022, Ramsauer et al., 2020] by introducing a sparse Hopfield layer into deep learning models.

**Organization.** In Section 2, the sparse Hopfield model is introduced. In Section 3, the memory capacity is discussed. In Section 4, experimental studies are conducted. In Section 5, concluding discussions are provided. Finally, related works and limitations are discussed in Appendix B.

**Notations.** We write $\langle \mathbf{a}, \mathbf{b} \rangle := \mathbf{a}^\mathsf{T}\mathbf{b}$ as the inner product for vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$. The index set $\{1, \cdots, I\}$ is denoted by $[I]$, where $I \in \mathbb{N}^+$. The spectral norm is denoted by $\|\cdot\|_2$, which is equivalent to the $l_2$-norm when applied to a vector. Throughout this paper, we denote the memory patterns (keys) by $\boldsymbol{\xi} \in \mathbb{R}^d$ and the state/configuration/query pattern by $\mathbf{x} \in \mathbb{R}^d$ with $n := \|\mathbf{x}\|$, and $\boldsymbol{\Xi} := (\boldsymbol{\xi}_1, \cdots, \boldsymbol{\xi}_M) \in \mathbb{R}^{d \times M}$ as shorthand for stored memoery (key) patterns $\{\boldsymbol{\xi}_\mu\}_{\mu \in [M]}$. Moreover, we set $m := \mathrm{Max}_{\mu \in [M]} \|\boldsymbol{\xi}_\mu\|$ be the largest norm of memory patterns.

## 2 Sparse Hopfield Model

In this section, we first introduce the sparse Hopfield energy from convex conjugate of entropic regularizer, and then the sparse retrieval dynamics. In this paper we only consider the Gini entropic regularizer corresponding to the sparsemax distribution [Martins and Astudillo, 2016].

### 2.1 Sparse Hopfield Energy

Let $\mathbf{x} \in \mathbb{R}^d$ be the query pattern, and $\boldsymbol{\Xi} := (\boldsymbol{\xi}_1, \cdots, \boldsymbol{\xi}_M) \in \mathbb{R}^{d \times M}$ be the memory patterns. We introduce the sparse Hopfield energy as

$$\mathcal{H}(\mathbf{x}) = -\Psi^\star\left(\beta \boldsymbol{\Xi}^\mathsf{T}\mathbf{x}\right) + \frac{1}{2}\langle \mathbf{x}, \mathbf{x} \rangle, \tag{2.1}$$

with $\Psi^\star(\mathbf{z}) := \frac{1}{2}\|\mathbf{z}\|^2 - \frac{1}{2}\|\mathrm{Sparsemax}(\mathbf{z}) - \mathbf{z}\|^2 + \frac{1}{2}$, where $\mathrm{Sparsemax}(\cdot)$ is defined as follows. Let $\mathbf{z}, \mathbf{p} \in \mathbb{R}^M$, and $\Delta^M := \{\mathbf{p} \in \mathbb{R}^M_+ \mid \sum_\mu^M p_\mu = 1\}$ be the $(M-1)$-dimensional unit simplex.

**Definition 2.1** (Sparsemax in Variational Form [Martins and Astudillo, 2016], also see Remark E.1)**.**

$$\mathrm{Sparsemax}(\mathbf{z}) := \underset{\mathbf{p} \in \Delta^M}{\mathrm{ArgMin}} \|\mathbf{p} - \mathbf{z}\|^2, = \underset{\mathbf{p} \in \Delta^M}{\mathrm{ArgMax}} \left[\mathbf{p}^\mathsf{T}\mathbf{z} - \Psi(\mathbf{p})\right], \tag{2.2}$$

where $\Psi(\mathbf{p}) := -\frac{1}{2}\sum_\nu^M p_\nu(1 - p_\nu)$ is the negative Gini entropy or Gini entropic regularizer.

At first glance, the energy function (2.1) may seem peculiar. However, it indeed represents a non-trivial sparse model that we desire with appealing properties, including: (i) In response to challenge (C1) & (C2), as we shall see in Section 2.2, it leads to a sparse retrieval dynamics that not only retrieves memory by monotonically decreasing the energy function (2.1) to its stationary points, but also associates with sparsemax attention through its single-step approximation; (ii) In response to challenge (C3), as we shall see in Section 3, it indulges fast convergence (of retrieval), exponential-in-$d$ memory capacity akin to modern Hopfield models. Notably, it accomplishes these with a tighter retrieval error bound. We will reveal each of these properties in the following sections.

## 2.2 Sparse Retrieval Dynamics and Connection to Sparse Attention

The optimization problem $\mathrm{ArgMax}_{\mathbf{p}\in\Delta^M}\left[\mathbf{p}^\mathsf{T}\mathbf{z}-\Psi(\mathbf{p})\right]$ does not necessarily have a closed-form solution for arbitrary $\Psi$. However, a family of $\Psi$ has been investigated in literature [Correia et al., 2019, Martins and Astudillo, 2016] with closed-form solutions derived, including the $\mathrm{Sparsemax}(\cdot)$.

**Sparsemax in Closed-Form** (Proposition 1 of [Martins and Astudillo, 2016]). Let $\mathbf{z}\in\mathbb{R}^M$. Denote $[a]_+ \coloneqq \mathrm{Max}\{0,a\}$, $z_{(\nu)}$ the $\nu$'th element in a sorted descending $z$-sequence $\mathbf{z}_{\text{sorted}} \coloneqq z_{(1)}\geq z_{(2)}\geq \ldots \geq z_{(M)}$, and $\kappa(\mathbf{z}) \coloneqq \mathrm{Max}\left\{k\in[M] \;\middle|\; 1+kz_{(k)} > \sum_{\nu\leq k} z_{(\nu)}\right\}$. The optimization problem(s) (2.2) has closed-form solution

$$\mathrm{Sparsemax}(\mathbf{z}) = \left[\mathbf{z}-\tau(\mathbf{z})\right]_+, \tag{2.3}$$

where $\tau:\mathbb{R}^M\to\mathbb{R}$ is the threshold function $\tau(\mathbf{z}) = \left[\left(\sum_{\nu\leq\kappa(\mathbf{z})} z_{(\nu)}\right)-1\right]/\kappa(\mathbf{z})$, satisfying $\sum_{\mu=1}^M\left[z_\mu-\tau(\mathbf{z})\right]_+ = 1$ for all $\mathbf{z}$. Notably, $\kappa(\mathbf{z}) = |S(\mathbf{z})|$ where $S(\mathbf{z}) = \{\mu\in[M]\mid \mathrm{Sparsemax}_\mu(\mathbf{z})>0\}$ is the support set of $\mathrm{Sparsemax}(\mathbf{z})$.

In this case, we present the following theorem to derive the convex conjugate of $\Psi$ in closed-form:

**Theorem 2.1** (Convex Conjugate of Negative Gini Entropy). Let $F(\mathbf{p}) \coloneqq \langle\mathbf{p},\mathbf{z}\rangle - \Psi(\mathbf{p})$ with $\Psi$ being the negative Gini entropy, $\Psi(\mathbf{p}) = \frac{1}{2}\|\mathbf{p}\|^2-\frac{1}{2}$. The convex conjugate of $\Psi(\mathbf{p})$ is

$$\Psi^\star(\mathbf{z}) \coloneqq \operatorname*{Max}_{\mathbf{p}\in\Delta^M} F(\mathbf{p},\mathbf{z}) = \frac{1}{2}\|\mathbf{z}\|^2 - \frac{1}{2}\|\mathbf{p}^\star-\mathbf{z}\|^2+\frac{1}{2}, \tag{2.4}$$

where $\mathbf{p}^\star = \mathrm{Sparsemax}(\mathbf{z})$ is given by Section 2.2.

**Corollary 2.1.1.** By Danskin's Theorem, $\boldsymbol{\nabla}\Psi^\star(\mathbf{z}) = \mathrm{ArgMax}_{\mathbf{p}\in\Delta^M} F(\mathbf{p},\mathbf{z}) = \mathrm{Sparsemax}(\mathbf{z})$.

*Proof.* A detailed proof is shown in Appendix D.1. □

Theorem 2.1 and Corollary 2.1.1 not only provide the intuition behind the sparse Hopfield energy (2.1) — the memory patterns are stored in local minima aligned with the overlap-function constructions (i.e. $\left\|\boldsymbol{\Xi}^\mathsf{T}\mathbf{x}\right\|^2 = \sum_{\mu=1}^M\langle\boldsymbol{\xi}_\mu,\mathbf{x}\rangle^2$) in [Ramsauer et al., 2020, Demircigil et al., 2017, Krotov and Hopfield, 2016] — but also prepare us for the following corresponding sparse retrieval dynamics.

**Lemma 2.1** (Sparse Retrieval Dynamics). Let $t$ be the iteration number. The energy (2.1) can be monotonically decreased by the following sparse retrieval dynamics over $t$:

$$\mathcal{T}(\mathbf{x}_t) \coloneqq \boldsymbol{\nabla}_\mathbf{x}\Psi\left(\beta\boldsymbol{\Xi}^\mathsf{T}\mathbf{x}\right)\Big|_{\mathbf{x}_t} = \boldsymbol{\Xi}\,\mathrm{Sparsemax}\left(\beta\boldsymbol{\Xi}^\mathsf{T}\mathbf{x}_t\right) = \mathbf{x}_{t+1}. \tag{2.5}$$

*Proof Sketch.* To show monotonic decreasing property, we first derive the sparse retrieval dynamics by utilizing Theorem 2.1, Corollary 2.1.1, along with the convex-concave procedure [Yuille and Rangarajan, 2003, 2001]. Then, we show the monotonicity of $\mathcal{H}$ by constructing a iterative upper bound of $\mathcal{H}$ which is convex in $\mathbf{x}_{t+1}$ and thus, can be lowered iteratively by the convex-concave procedure. A detailed proof is shown in the Appendix D.2. □

**Remark 2.1.** Similar to [Ramsauer et al., 2020], (2.5) is equivalent to sparsemax attention [Martins and Astudillo, 2016] when the $\mathcal{T}$ is applied only once, see Appendix C for more details.

Notably, since $\left\|\boldsymbol{\Xi}^\mathsf{T}\mathbf{x}\right\|^2 = \sum_{\mu=1}^M\langle\boldsymbol{\xi}_\mu,\mathbf{x}\rangle^2$, (2.5) implies that the local optimum of $\mathcal{H}$ are located near the patterns $\boldsymbol{\xi}_\mu$. Different from previous studies on binary Hopfield models [Demircigil et al., 2017, Krotov and Hopfield, 2016], for continuous patterns, we adopt the relaxed definition from [Ramsauer et al., 2020][3] to rigorously analyze the memory retrieval, and the subsequent lemma.

**Definition 2.2** (Stored and Retrieved). Assuming that every pattern $\boldsymbol{\xi}_\mu$ surrounded by a sphere $S_\mu$ with finite radius $R \coloneqq \frac{1}{2}\,\mathrm{Min}_{\mu,\nu\in[M]}\|\boldsymbol{\xi}_\mu-\boldsymbol{\xi}_\nu\|$, we say $\boldsymbol{\xi}_\mu$ is *stored* if there exists a generalized fixed point of $\mathcal{T}$, $\mathbf{x}_\mu^\star\in S_\mu$, to which all limit points $\mathbf{x}\in S_\mu$ converge to, and $S_\mu\cap S_\nu=\emptyset$ for $\mu\neq\nu$. We say $\boldsymbol{\xi}_\mu$ is $\epsilon$-*retrieved* by $\mathcal{T}$ with $\mathbf{x}$ for an error[a] $\epsilon$, if $\|\mathcal{T}(\mathbf{x})-\boldsymbol{\xi}_\mu\|\leq\epsilon$.

[a] The retrieval error has a naive bound $\epsilon \coloneqq \mathrm{Max}\left\{\|\mathbf{x}-\boldsymbol{\xi}_\mu\|, \left\|\boldsymbol{\xi}_\mu-\mathbf{x}_\mu^\star\right\|\right\}$ by interpolating from $\mathbf{x}$ to $\boldsymbol{\xi}_\mu$.

[3] Recall that a fixed point of $\mathcal{T}$ with respect to $\mathcal{H}$ is a point where $\mathbf{x}=\mathcal{T}(\mathbf{x})$, and a generalized fixed point is a point where $\mathbf{x}\in\mathcal{T}(\mathbf{x})$. For more details, refer to [Sriperumbudur and Lanckriet, 2009].

The next lemma states the convergence results of the proposed sparse retrieval dynamics (2.5).

**Lemma 2.2** (Convergence of Retrieval Dynamics $\mathcal{T}$)**.** Suppose $\mathcal{H}$ is given by (2.1) and $\mathcal{T}(\mathbf{x})$ is given by (2.5). For any sequence $\{\mathbf{x}_t\}_{t=0}^{\infty}$ defined by $\mathbf{x}_{t'+1} = \mathcal{T}(\mathbf{x}_{t'})$, all limit points of this sequence are stationary points if they are obtained by iteratively applying $\mathcal{T}$ to $\mathcal{H}$.

*Proof Sketch.* We verify and utilize Zangwill's global convergence theory [Zangwill, 1969] for iterative algorithms $\mathcal{T}$, to first show that all the limit points of $\{\mathbf{x}_t\}_{t=0}^{\infty}$ are generalized fixed points and $\lim_{t\to\infty} \mathcal{H}(\mathbf{x}_t) = \mathcal{H}(\mathbf{x}^\star)$, where $\mathbf{x}^\star$ are some generalized fixed points of $\mathcal{T}$. Subsequently, by [Sriperumbudur and Lanckriet, 2009, Lemma 5], we show that $\{\mathbf{x}^\star\}$ are also stationary points of $\mathrm{Min}_{\mathbf{x}}[\mathcal{H}]$, and hence $\mathcal{H}$ converges to local optimum. A detailed proof is shown in Appendix D.4. □

Intuitively, Lemma 2.2 indicates that the energy function converges to local optimum, i.e. $\lim_{t\to\infty} \mathcal{H}(\mathbf{x}_t) \to \mathcal{H}(\mathbf{x}^\star)$, where $\mathbf{x}^\star$ are stationary points of $\mathcal{H}$. Consequently, it offers formal justifications for the retrieval dynamics (2.5) to retrieve stored memory patterns $\{\boldsymbol{\xi}_\mu\}_{\mu\in[M]}$: for any query (initial point) $\mathbf{x}$, $\mathcal{T}$ monotonically and iteratively approaches stationary points of $\mathcal{H}$, where the memory patterns $\{\boldsymbol{\xi}_\mu\}_{\mu\in[M]}$ are stored. As for the retrieval error, we provide the following theorem stating that $\mathcal{T}$ achieves a lower retrieval error compared to its dense counterpart.

**Theorem 2.2** (Retrieval Error)**.** Let $\mathcal{T}_{\text{Dense}}$ be the retrieval dynamics of the dense modern Hopfield model [Ramsauer et al., 2020]. It holds $\|\mathcal{T}(\mathbf{x}) - \boldsymbol{\xi}_\mu\| \le \|\mathcal{T}_{\text{Dense}}(\mathbf{x}) - \boldsymbol{\xi}_\mu\|$ for all $\mathbf{x} \in S_\mu$. Moreover,

$$\|\mathcal{T}(\mathbf{x}) - \boldsymbol{\xi}_\mu\| \le m + d^{1/2} m\beta \left[\kappa\left(\max_{\nu\in[M]} \langle \boldsymbol{\xi}_\nu, \mathbf{x}\rangle - \left[\boldsymbol{\Xi}^\mathsf{T}\mathbf{x}\right]_{(\kappa)}\right) + \frac{1}{\beta}\right], \tag{2.6}$$

where $\left[\boldsymbol{\Xi}^\mathsf{T}\mathbf{x}\right]_{(\kappa)}$ is the $\kappa$th element of $\boldsymbol{\Xi}^\mathsf{T}\mathbf{x} \in \mathbb{R}^M$ following closed-form sparsemax function (2.3).

*Proof.* A detailed proof is shown in Appendix D.3. □

Interestingly, (2.6) is a sparsity dependent bound[4]. By denoting $n := \|\mathbf{x}\|$, the second term on the RHS of (2.6) is dominated by the sparsity dimension $\kappa$ as it can be expressed as $\kappa\left(1 - \left[\boldsymbol{\Xi}^\mathsf{T}\mathbf{x}\right]_{(\kappa)}/(nm)\right) \propto \alpha\kappa$ with a constant $0 \le \alpha \le 2$. When $\boldsymbol{\Xi}^\mathsf{T}\mathbf{x}$ is sparse (i.e. $\kappa$ is small), the bound is tighter, vice versa. Moreover, in cases of contaminated patterns, i.e. $\widetilde{\mathbf{x}} = \mathbf{x} + \boldsymbol{\eta}$ or $\widetilde{\boldsymbol{\xi}}_\mu = \boldsymbol{\xi}_\mu + \boldsymbol{\eta}$, the impact of noise $\boldsymbol{\eta}$ on the sparse retrieval error (2.6) is linear, while its effect on the dense retrieval error (2.7) is exponential. This suggests the robustness advantage of the sparse Hopfield model, as evidenced in Figure 1.

### 2.3 Sparse Hopfield Layers for Deep Learning

The sparse Hopfield model can serve as a versatile component for deep learning frameworks, given its continuity and differentiability with respect to parameters. Corresponding to three types of Hopfield Layers proposed in [Ramsauer et al., 2020], we introduce their sparse analogs: **(1)** `SparseHopfield`, **(2)** `SparseHopfieldPooling`, **(3)** `SparseHopfieldLayer`. Layer `SparseHopfield` has memory (stored or key) patterns $\boldsymbol{\Xi}$ and query (state) pattern $\mathbf{x}$ as inputs, and associates these two sets of patterns via the sparse retrieval dynamics (2.5). This layer regards the transformer attention layer as its one-step approximation, while utilizing the sparsemax [Martins and Astudillo, 2016] on attention matrix. Layer `SparseHopfieldPooling` and Layer `SparseHopfieldLayer` are two variants of `SparseHopfield`, whose input patterns are memory patterns and query patterns from previous layers or external plugin, respectively. `SparseHopfieldPooling`, whose query patterns are learnable parameters, can be interpreted as performing a pooling operation over input memory patterns. `SparseHopfieldLayer`, by contrast, has learnable memory patterns and can be interpreted as a two-layer fully connected networks with sparsemax activation function. See (C.12) in Appendix C and [Ramsauer et al., 2020, Section 3] for more details of these associations. In Section 4, we apply these layers and compare them with their dense counterparts in [Ramsauer et al., 2020] and other baseline machine learning methods.

## 3 Fundamental Limits of Memory Capacity of Sparse Hopfield Models

How many patterns can be stored and reliably retrievable in the proposed model? We address this by decomposing it into to two sub-questions and answering them separately:

(A) What is the condition for a pattern $\boldsymbol{\xi}_\mu$ considered well stored in $\mathcal{H}$, and correctly retrieved?

[4] Notably, $\|\mathcal{T}(\mathbf{x}) - \boldsymbol{\xi}_\mu\|$ is also upper-bounded by a sparsity-independent but $M, \beta$-dependent bound

$$\|\mathcal{T}(\mathbf{x}) - \boldsymbol{\xi}_\mu\| \le \|\mathcal{T}_{\text{Dense}}(\mathbf{x}) - \boldsymbol{\xi}_\mu\| \le 2m(M-1)\exp\left\{-\beta\left(\langle \boldsymbol{\xi}_\mu, \mathbf{x}\rangle - \max_{\nu\in[M]} \langle \boldsymbol{\xi}_\mu, \boldsymbol{\xi}_\nu\rangle\right)\right\}. \tag{2.7}$$

(B) What is the number, in expectation, of the the patterns satisfying such condition?

For (A), we first introduce the notion of separation of patterns following [Ramsauer et al., 2020],

**Definition 3.1** (Separation of Patterns)**.** The separation of a memory pattern $\boldsymbol{\xi}_\mu$ from all other memory patterns $\boldsymbol{\Xi}$ is defined as its minimal inner product difference to any other patterns:

$$\Delta_\mu \coloneqq \operatorname*{Min}_{\nu,\nu\neq\mu} \left[\langle \boldsymbol{\xi}_\mu, \boldsymbol{\xi}_\mu\rangle - \langle \boldsymbol{\xi}_\mu, \boldsymbol{\xi}_\nu\rangle\right] = \langle \boldsymbol{\xi}_\mu, \boldsymbol{\xi}_\mu\rangle - \operatorname*{Max}_{\nu,\nu\neq\mu} \left[\langle \boldsymbol{\xi}_\mu, \boldsymbol{\xi}_\nu\rangle\right]. \tag{3.1}$$

Similarly, the separation of $\boldsymbol{\xi}_\mu$ at a given $\mathbf{x}$ from all memory patterns $\boldsymbol{\Xi}$ is given by

$$\widetilde{\Delta}_\mu \coloneqq \operatorname*{Min}_{\nu,\nu\neq\mu} \left[\langle \mathbf{x}, \boldsymbol{\xi}_\mu\rangle - \langle \mathbf{x}, \boldsymbol{\xi}_\nu\rangle\right]. \tag{3.2}$$

and then the well-separation condition for a pattern being well-stored and retrieved.

**Theorem 3.1** (Well-Separation Condition)**.** Following Definition 2.2, the memory patterns $\{\boldsymbol{\xi}_\mu\}_{\mu\in[M]}$ locate inside the sphere $S_\mu \coloneqq \left\{\mathbf{x} \;\middle|\; \|\mathbf{x}-\boldsymbol{\xi}_\mu\| \leq R\right\}$, with finite radius $R \coloneqq \frac{1}{2}\operatorname{Min}_{\mu,\nu\in[M]} \|\boldsymbol{\xi}_\mu - \boldsymbol{\xi}_\nu\|$ for all $\mu$. Then, the retrieved dynamics $\mathcal{T}$ maps $S_\mu$ to itself if

1. The starting point $\mathbf{x}$ is inside $S_\mu$: $\mathbf{x} \in S_\mu$.
2. The *well-separation* condition: $\Delta_\mu \geq mn + 2mR - \left[\boldsymbol{\Xi}^\mathsf{T}\mathbf{x}\right]_{(\kappa)} - \frac{1}{\kappa}\left(\frac{R-m-md^{1/2}}{m\beta d^{1/2}}\right)$.

**Corollary 3.1.1.** Let $\delta \coloneqq \|\mathcal{T}_{\text{Dense}} - \boldsymbol{\xi}_\mu\| - \|\mathcal{T} - \boldsymbol{\xi}_\mu\|$. The well-separation condition can be expressed as $\Delta_\mu \geq \frac{1}{\beta}\ln\left(\frac{2(M-1)m}{R+\delta}\right) + 2mR$, which reduces to that of the dense Hopfield model when $\delta = 0$.

*Proof Sketch.* Intuitively, the proofs proceed by connecting $\Delta_\mu$ with $\|\mathcal{T}(\mathbf{x}) - \boldsymbol{\xi}_\mu\|$. To do so, we utilize Theorem 2.2 to incorporate the $\Delta_\mu$-dependent bound on the retrieval error of both sparse and dense Hopfield models [Ramsauer et al., 2020]. A detailed proof is shown in Appendix D.5. □

Together with Lemma 2.2, the well-separated condition serves as the necessary condition for pattern $\boldsymbol{\xi}_\mu$ to be well-stored at the stationary points of $\mathcal{H}$, and can be retrieved with at most $\epsilon = R$ by $\mathcal{T}$, as per Definition 2.2. We make the following three observations about the blessings from sparsity.

1. In general, to appreciate the blessings of sparsity, we observe the two competing terms, $\alpha nm$ and $(R-m-md^{1/2})/(\kappa m\beta d^{1/2})$. Sparsity proves advantageous when the latter term surpasses the former, i.e. the sparse well-separation condition is consistently lower than its dense counterpart. The conditions under which sparsity benefits are more likely to emerge (i.e., when the well-separation condition is more readily satisfied) can be expressed as:
$$\frac{1}{2}\operatorname*{Min}_{\mu,\nu\in[M]} \|\boldsymbol{\xi}_\mu - \boldsymbol{\xi}_\nu\| \geq md^{1/2}\left(1 + \alpha\beta nm\kappa\right) + m, \quad \text{with } 0 \leq \alpha \leq 2. \tag{3.3}$$
Intuitively, the sparser $\boldsymbol{\Xi}^\mathsf{T}\mathbf{x}$ is, the easier it is for the above condition to be fulfilled.
2. **Large $M$ limit:** For large $M$, the dense well-separation condition (Corollary 3.1.1) explodes while the sparse one (Theorem 3.1) saturates to the first three $M$-independent terms. This suggests that the hardness of distinguishing patterns can be tamed by the sparsity, preventing an increase of $\Delta_\mu$ with $M$ as observed in the dense Hopfield model. We numerically confirm this in Figure 1.
3. $\beta \to \infty$ **Limit:** In the region of low temperature, where $\beta \to \infty$ and hence all patterns can be *error-free* retrieved as per (2.7), we have $\Delta_\mu \geq 2mR + \alpha nm$ with $0 \leq \alpha \leq 2$. Here, the second term on the RHS represents the sparsity level of $\boldsymbol{\Xi}^\mathsf{T}\mathbf{x}$, i.e. a smaller $\alpha$ indicates a higher degree of sparsity in $\boldsymbol{\Xi}^\mathsf{T}\mathbf{x}$. Hence, the higher the sparsity, the easier it is to separate patterns.

For (B), equipped with Theorem 3.1 and Corollary 3.1.1, we provide a lower bound for the number of patterns being well-stored and can be *at least* $R$-retrieved in the next lemma[5]:

**Lemma 3.1** (Memory Capacity Lower Bound)**.** Let $1-p$ be the probability of successfully storing and retrieving a pattern. The number of patterns randomly sampled from a sphere of radius $m$ that the sparse Hopfield model can store and retrieve is lower-bounded by

$$M \geq \sqrt{p}C^{\frac{d-1}{4}}, \tag{3.4}$$

where $C$ is the solution to $C = b/W_0(\exp\{a+\ln b\})$ with $W_0(\cdot)$ being the principal branch of Lambert $W$ function, $a \coloneqq 4/d-1\left\{\ln\left[2m(\sqrt{p}-1)/R+\delta\right] + 1\right\}$ and $b \coloneqq 4m^2\beta/5(d-1)$. For sufficiently large $\beta$, the sparse Hopfield model exhibits a larger lower bound on the exponential memory capacity compared to its dense counterpart [Ramsauer et al., 2020]: $M \geq M_{\text{Dense}}$.

[5]Following the convention in memory capacity literature [Ramsauer et al., 2020, Demircigil et al., 2017, Krotov and Hopfield, 2016], we assume that all memory patterns $\{\boldsymbol{\xi}_\mu\}$ are sampled from a $d$-sphere of radius $m$.

*Proof Sketch.* Our proof is built on [Ramsauer et al., 2020]. The high-level idea is to utilizes the separation of random patterns sampled from spheres [Cai and Jiang, 2012, Brauchart et al., 2018] and the asymptotic expansion of the Lambert $W$ function [Corless et al., 1996]. Firstly, we link the well-separation condition to cosine similarity distance, creating an inequality for the probability of a pattern being well-stored and retrieved. Next, we identify and prove conditions for the exponential memory capacity $M = \sqrt{p}C^{(d-1)/4}$ to hold. Finally, we analyze the scaling behaviors of $C$ using its asymptotic expansion and show that $M \geq M_{\text{Dense}}$. A detailed proof is shown in Appendix D.6. □

Intuitively, the benefits of sparsity arises from the increased energy landscape separation provided by the sparse Hopfield energy function, which enables the separation of closely correlated patterns, resulting in a tighter well-separation condition for distinguishing such patterns and hence a larger lower bound on the memory capacity. Moreover, the sparse Hopfield model also enjoys the properties of fast convergence and exponentially suppressed retrieval error provided by the following corollary.

**Corollary 3.1.2** (Fast Convergence and Exponentially Suppressed Retrieval Error)**.** For any query $\mathbf{x}$, $\mathcal{T}$ approximately retrieves a memory patterns $\boldsymbol{\xi}_\mu$ with retrieval error $\epsilon$ exponentially suppressed by $\Delta_\mu$: $\|\mathcal{T}(\mathbf{x}) - \boldsymbol{\xi}_\mu\| \leq 2m(M-1)\exp\{-\beta\left(\Delta_\mu - 2m\,\text{Max}\left[\|\mathbf{x} - \boldsymbol{\xi}_\mu\|, \left\|\mathbf{x} - \mathbf{x}_\mu^\star\right\|\right]\right)\}$.

*Proof.* This results from Theorem 2.2, Lemma 2.2, and [Ramsauer et al., 2020, Theorem 4]. □

Corollary 3.1.2 suggests that, with a sufficient $\Delta_\mu$, $\mathcal{T}$ can approximately retrieve patterns after a single *activation*, allowing the integration of sparse Hopfield models into deep learning architectures similarly to [Hoover et al., 2023, Seidl et al., 2022, Fürst et al., 2022, Ramsauer et al., 2020].

# 4 Proof of Concept Experimental Studies

We demonstrate the validity of our theoretical results and method by testing them on various experimental settings with both synthetic and real-world datasets.

## 4.1 Experimental Validation of Theoretical Results

We conduct experiments to verify our theoretical findings, and report the results in Figure 1. For the memory capacity (the top row of Figure 1), we test the proposed sparse model on retrieving half-masked patterns comparing with the Dense (Softmax) and 10th order polynomial Hopfield models [Millidge et al., 2022, Krotov and Hopfield, 2016] on MNIST (high sparsity), Cifar10 (low sparsity) and ImageNet (low sparsity) datasets. For all Hopfield models, we set $\beta = 1$.[6] A query is regarded as correctly retrieved if its cosine similarity error is below a set threshold. In addition, for the robustness against noisy queries (the bottom row of Figure 1), we inject Gaussian noises with varying variances ($\sigma$) into the images. Plotted are the means and standard deviations of 10 runs. The results show that the proposed sparse Hopfield model excels when memory patterns exhibit a high degree of sparsity and the signal-to-noise ratio in patterns is low, aligning with our theoretical results.

## 4.2 Multiple Instance Learning Tasks

Ramsauer et al. [2020] point out that the memory-enhanced Hopfield layers present a promising approach for Multiple Instance Learning (MIL) tasks. Therefore, we implement our sparse Hopfield layers and applied them to MIL tasks on one synthetic and four real-world settings.

### 4.2.1 Synthetic Experiments

We use a synthetic MIL dataset, the bit pattern dataset, to demonstrate the effectiveness of the sparse Hopfield model. Each bag in this synthetic dataset contains a set of binary bit strings. The positive bag includes at least one of the positive bit patterns. We compare the performance of the `SparseHopfield` and `SparseHopfieldPooling` to their dense counterparts and vanilla attention [Vaswani et al., 2017]. We report the mean test accuracy of 10 runs. To demonstrate the effectiveness of sparse Hopfield model, we vary two hyperparameters of the bit pattern dataset corresponding to two perspectives: bag sparsity (sparsity in data) and bag size (number of memory patterns, $M$). For **bag sparsity**, we fix the bag size as 200, and inject from 2 to 80 positive patterns in a positive bag, results in 1 to 40 percent of positive patterns in each positive bag. For **bag size**, we fix the number of positive pattern in a bag to be 1, and vary bag size from 20 to 300. We report results of `SparseHopfieldPooling` in Table 1, and implementation details in Appendix G.1.1. A more

[6] However, as pointed out in [Millidge et al., 2022], this is in fact *not* fair to compare modern Hopfield with $\beta = 1$ with higher order polynomial Hopfield models.

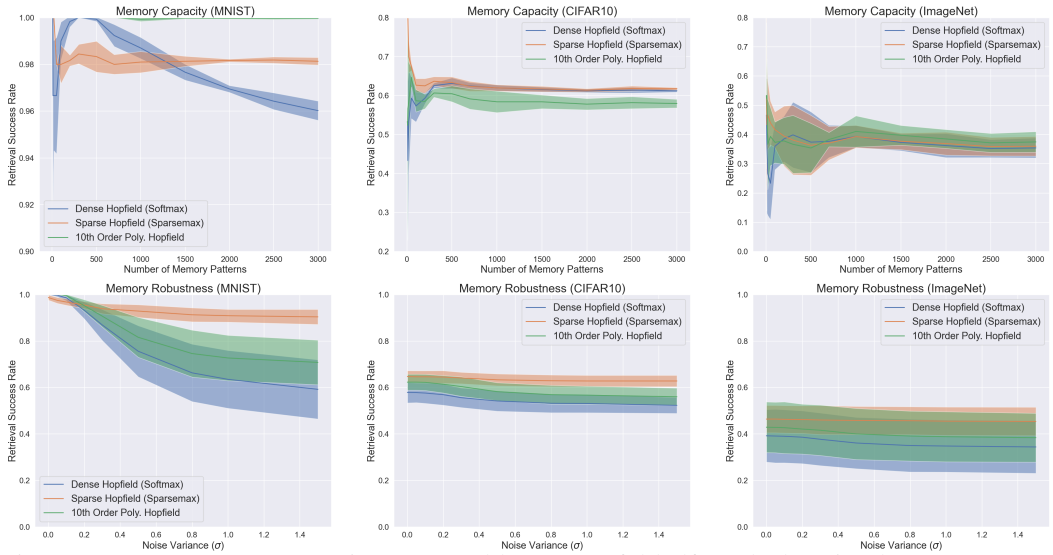


Figure 1: **Top:** Memory Capacity measured by successful half-masked retrieval rates. **Bottom:** Memory Robustness measured by retrieving patterns with varying levels of Gaussian noise. For all Hopfield models, we set $\beta = 1$. A query pattern is deemed correctly retrieved if its cosine similarity error is below a set threshold. For MNIST/CIFAR10/ImageNet datasets, we set the error thresholds to be 50/10/10% to cope with different sparse levels in data. Plotted are the means and standard deviations of 10 runs. The results suggest that the sparse Hopfield model excels when memory patterns exhibit a high degree of sparsity and the signal-to-noise ratio in patterns is low.

complete version of Table 1, including the results of `Hopfield` and attention, is in Appendix F. The sparse Hopfield model demonstrates a better performance across all sparsity and all bag sizes.

Table 1: **Top (Bag Size):** Accuracy comparison on bit pattern dataset for sparse and dense Hopfield model. We report the average accuracy over 10 runs. The results suggest that the sparse Hopfield model demonstrates a better performance when facing a bag size increase. **Bottom (Bag Sparsity):** Performance comparison on bit pattern dataset for sparse and dense Hopfield model with varying bag sparsity. We report the average accuracy over 10 runs. The results suggest that the sparse Hopfield model demonstrates a better performance across all sparsity.

| Bag Size | 20 | 50 | 100 | 150 | 200 | 300 |
|---|---|---|---|---|---|---|
| Dense Hopfield Pooling | $100.0 \pm 0.00$ | $100.0 \pm 0.00$ | $100.0 \pm 0.00$ | $76.44 \pm 0.23$ | $49.13 \pm 0.01$ | $52.88 \pm 0.01$ |
| Sparse Hopfield Pooling | $100.0 \pm 0.00$ | $100.0 \pm 0.00$ | $100.0 \pm 0.00$ | $\mathbf{99.76 \pm 0.00}$ | $\mathbf{99.76 \pm 0.00}$ | $\mathbf{99.76 \pm 0.00}$ |

| Bag Sparsity | 1% | 5% | 10% | 20% | 40% |
|---|---|---|---|---|---|
| Dense Hopfield Pooling | $49.20 \pm 0.00$ | $85.58 \pm 0.10$ | $100.0 \pm 0.00$ | $100.0 \pm 0.00$ | $99.68 \pm 0.00$ |
| Sparse Hopfield Pooling | $\mathbf{73.40 \pm 0.06}$ | $\mathbf{99.68 \pm 0.00}$ | $100.0 \pm 0.00$ | $100.0 \pm 0.00$ | $\mathbf{100.0 \pm 0.00}$ |

**Convergence Analysis.** In Figure 3, we numerically examine the convergence of the sparse and dense Hopfield models, plotting their loss and accuracy for the **bag size** tasks in above on the bit pattern dataset. We include multiple bag sizes to assess the effect of increasing memory patterns (i.e. $M$) on the loss curve. The plotted are the loss and accuracy curves of `SparseHopfieldPooling`. We refer results of `Hopfield` and more details to Appendix F.2. The results (Figure 3) show that, sparse Hopfield model surpasses its dense counterpart in all bag sizes. Moreover, for the same bag size, the sparse Hopfield model always reaches the minimum validation loss faster than dense Hopfield model, validating our Theorem 2.2.

**Sparsity Generalization.** We also evaluate the models' generalization performance with shifting information sparsity, by training dense and sparse Hopfield models with a specific bag sparsity and testing them on the other. We report the results in Table 3 and refer more details to Appendix F.2.

### 4.2.2 Real-World MIL Tasks

Next, we demonstrate that the proposed method achieves near-optimal performance on four realistic (*non-sparse*) MIL benchmark datasets: Elephant, Fox and Tiger for image annotation [Ilse

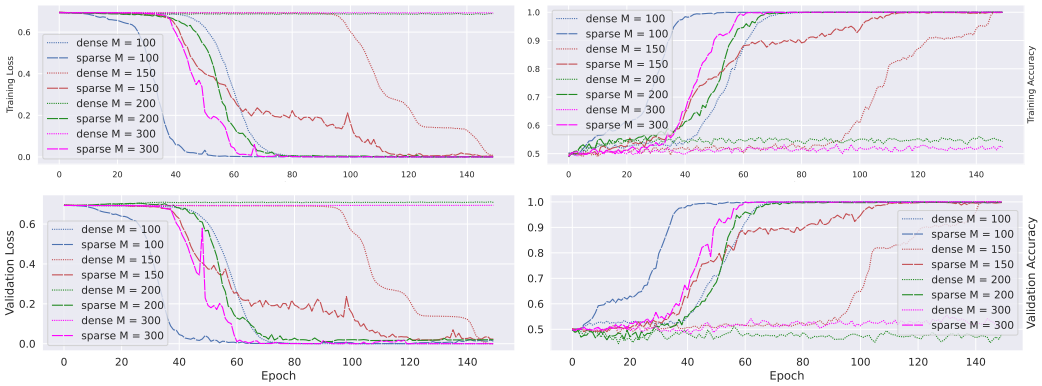


Figure 2: **Top:** The training loss and accuracy curve of dense and sparse Hopfield models with different bag sizes. **Bottom:** The validation loss and accuracy curve of dense and sparse Hopfield models with different bag sizes. The plotted are the mean of 10 runs. The results indicate that the sparse Hopfield model converges faster than the dense model and also yields superior accuracy.

et al., 2018], UCSB breast cancer classification [Kandemir et al., 2014]. We use `Hopfield` and `SparseHopfield` to construct a similar model architecture proposed in [Ramsauer et al., 2020] and a detailed description of this model as well as its training and evaluating process can be found in Appendix G.1.2. As shown in Table 2, both Sparse and Dense Hopfield achieve near-best results on Tiger, Elephant and UCSB datasets, despite the low sparsity in data. The sparse Hopfield model outperforms the dense Hopfield model by a small margin on three out of four datasets.

## 5 Conclusion

We present a sparse Hopfield model with a memory-retrieval dynamics that corresponds to the sparse-structured attention mechanism. This model is capable of merging into deep learning architectures with data-dependent sparsity. Theoretically, we explore the advantages of the sparse Hopfield model, delineating conditions that favor its use. Empirically, we demonstrate our theoretical results and methodology to be effective on various synthetic and realistic settings. This work extends the correspondence between artificial and biological neural networks to sparse domain, potentially paving the way for future Hopfield-based methodologies and bio-inspired computing systems.

Table 2: Results for MIL benchmark datasets in terms of AUC score. The baselines are Path encoding [Küçükaşcı and Baydoğan, 2018], MInD [Cheplygina et al., 2015], MILES [Chen et al., 2006], APR [Dietterich et al., 1997], Citation-KNN [Wang and Zucker, 2000] and DD [Maron and Lozano-Pérez, 1997]. Results for baselines are taken from [Ramsauer et al., 2020]. The results suggest the proposed model achieves near-optimal performance even when the data is not sparse.

| Method | Tiger | Fox | Elephant | UCSB |
|---|---|---|---|---|
| Dense Hopfield | $0.878 \pm 0.028$ | $0.600 \pm 0.011$ | $0.907 \pm 0.022$ | $0.880 \pm 0.013$ |
| Sparse Hopfield | $0.892 \pm 0.021$ | $0.611 \pm 0.010$ | $0.912 \pm 0.016$ | $0.877 \pm 0.009$ |
| Path encoding | $0.910 \pm 0.010$ | $0.712 \pm 0.014$ | $0.944 \pm 0.007$ | $0.880 \pm 0.022$ |
| MInD | $0.853 \pm 0.011$ | $0.704 \pm 0.016$ | $0.936 \pm 0.009$ | $0.831 \pm 0.027$ |
| MILES | $0.872 \pm 0.017$ | $0.738 \pm 0.016$ | $0.927 \pm 0.007$ | $0.833 \pm 0.026$ |
| APR | $0.778 \pm 0.007$ | $0.541 \pm 0.009$ | $0.550 \pm 0.010$ | – |
| Citation-kNN | $0.855 \pm 0.009$ | $0.635 \pm 0.015$ | $0.896 \pm 0.009$ | $0.706 \pm 0.032$ |
| DD | 0.841 | 0.631 | 0.907 | – |

**Boarder Impacts and Future Directions: Brain Science and Foundation Models.** The primary theme of our research is to perceive any data representation (set of patterns) as analogous to the neural responses of a global brain reacting to a vast range of external stimuli (queries). This perspective presents exciting opportunities to study large generative foundational models, such as large language models, within a rigorous scientific framework inspired by contemporary brain science research.

We believe this work could be impactful in several respects, even though it is foundational research and not tied to specific applications: **(Cognition.)** This research could contribute to our understanding of a memory-enhanced model's predictive capacity when given either in-context input (like historical data) or external stimuli (such as real-time events). **(Memory.)** It may also shed light on the inherent limits of artificial neural networks' memorization capabilities and how to augment them with external memory modules for rapid responses to potential external stimuli. **(Network.)** This research could enable models to better assess the intricate network of cross-sectional brain activity among different variables and infer its dynamic structural alterations to identify possible systematic properties.

# References


Shaojie Bai, J Zico Kolter, and Vladlen Koltun. Deep equilibrium models. *Advances in Neural Information Processing Systems*, 32, 2019.

Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*, 2015.

Johannes Brandstetter. Blog post: Hopfield networks is all you need, 2021. URL https://ml-jku.github.io/hopfield-layers/. Accessed: April 4, 2023.

Johann S Brauchart, Alexander B Reznikov, Edward B Saff, Ian H Sloan, Yu Guang Wang, and Robert S Womersley. Random point sets on the sphere—hole radii, covering, and separation. *Experimental Mathematics*, 27(1):62–81, 2018.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

T Tony Cai and Tiefeng Jiang. Phase transition in limiting distributions of coherence of high-dimensional random matrices. *Journal of Multivariate Analysis*, 107:24–39, 2012.

Yixin Chen, Jinbo Bi, and James Ze Wang. Miles: Multiple-instance learning via embedded instance selection. *IEEE transactions on pattern analysis and machine intelligence*, 28(12):1931–1947, 2006.

Veronika Cheplygina, David MJ Tax, and Marco Loog. Dissimilarity-based ensembles for multiple instance learning. *IEEE transactions on neural networks and learning systems*, 27(6):1379–1391, 2015.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

Robert M Corless, Gaston H Gonnet, David EG Hare, David J Jeffrey, and Donald E Knuth. On the lambert w function. *Advances in Computational mathematics*, 5:329–359, 1996.

Gonçalo M Correia, Vlad Niculae, and André FT Martins. Adaptively sparse transformers. *arXiv preprint arXiv:1909.00015*, 2019.

Herbert A David and Haikady N Nagaraja. *Order statistics*. John Wiley & Sons, 2004.

Mete Demircigil, Judith Heusel, Matthias Löwe, Sven Upgang, and Franck Vermet. On a model of associative memory with huge storage capacity. *Journal of Statistical Physics*, 168:288–299, 2017.

Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89(1-2):31–71, 1997.

Michael Elad. *Sparse and redundant representations: from theory to applications in signal and image processing*, volume 2. Springer, 2010.

Peter Földiak. Forming sparse representations by local anti-hebbian learning. *Biological cybernetics*, 64(2):165–170, 1990.

Andreas Fürst, Elisabeth Rumetshofer, Johannes Lehner, Viet T Tran, Fei Tang, Hubert Ramsauer, David Kreil, Michael Kopp, Günter Klambauer, Angela Bitto, et al. Cloob: Modern hopfield networks with infoloob outperform clip. *Advances in neural information processing systems*, 35: 20450–20468, 2022.

Bolin Gao and Lacra Pavel. On the properties of the softmax function with application in game theory and reinforcement learning. *arXiv preprint arXiv:1704.00805*, 2017.

Asela Gunawardana, William Byrne, and Michael I Jordan. Convergence theorems for generalized alternating minimization procedures. *Journal of machine learning research*, 6(12), 2005.

Benjamin Hoover, Yuchen Liang, Bao Pham, Rameswar Panda, Hendrik Strobelt, Duen Horng Chau, Mohammed J Zaki, and Dmitry Krotov. Energy transformer. *arXiv preprint arXiv:2302.07253*, 2023.

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.

John J Hopfield. 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, 1984.

Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018.

Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. *Bioinformatics*, 37 (15):2112–2120, 2021.

Melih Kandemir, Chong Zhang, and Fred A Hamprecht. Empowering multiple instance histopathology cancer diagnosis by cell graphs. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014: 17th International Conference, Boston, MA, USA, September 14-18, 2014, Proceedings, Part II 17*, pages 228–235. Springer, 2014.

Leo Kozachkov, Ksenia V Kastanenka, and Dmitry Krotov. Building transformers from neurons and astrocytes. *bioRxiv*, pages 2022–10, 2022.

Dmitry Krotov and John Hopfield. Large associative memory problem in neurobiology and machine learning. *arXiv preprint arXiv:2008.06996*, 2020.

Dmitry Krotov and John J Hopfield. Dense associative memory for pattern recognition. *Advances in neural information processing systems*, 29, 2016.

Emel Şeyma Küçükaşcı and Mustafa Gökçe Baydoğan. Bag encoding strategies in multiple instance learning problems. *Information Sciences*, 467:559–578, 2018.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 11(1), 2010.

Alireza Makhzani and Brendan J Frey. Winner-take-all autoencoders. *Advances in neural information processing systems*, 28, 2015.

Oded Maron and Tomás Lozano-Pérez. A framework for multiple-instance learning. *Advances in neural information processing systems*, 10, 1997.

Andre Martins and Ramon Astudillo. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *International conference on machine learning*, pages 1614–1623. PMLR, 2016.

Beren Millidge, Tommaso Salvatori, Yuhang Song, Thomas Lukasiewicz, and Rafal Bogacz. Universal hopfield networks: A general framework for single-shot associative memory models. In *International Conference on Machine Learning*, pages 15561–15583. PMLR, 2022.

Bruno A Olshausen and David J Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision research*, 37(23):3311–3325, 1997.

Frank WJ Olver, Daniel W Lozier, Ronald F Boisvert, and Charles W Clark. *NIST handbook of mathematical functions hardback and CD-ROM*. Cambridge university press, 2010.

Fabian Paischer, Thomas Adler, Vihang Patil, Angela Bitto-Nemling, Markus Holzleitner, Sebastian Lehner, Hamid Eghbal-Zadeh, and Sepp Hochreiter. History compression via language models in reinforcement learning. In *International Conference on Machine Learning*, pages 17156–17185. PMLR, 2022.

Günther Palm. Neural associative memories and sparse coding. *Neural Networks*, 37:165–171, 2013.

Ben Peters, Vlad Niculae, and André FT Martins. Sparse sequence-to-sequence models. *arXiv preprint arXiv:1905.05702*, 2019.

Jiezhong Qiu, Hao Ma, Omer Levy, Scott Wen-tau Yih, Sinong Wang, and Jie Tang. Blockwise self-attention for long document understanding. *arXiv preprint arXiv:1911.02972*, 2019.

Hubert Ramsauer, Bernhard Schafl, Johannes Lehner, Philipp Seidl, Michael Widrich, Thomas Adler, Lukas Gruber, Markus Holzleitner, Milena Pavlovic, Geir Kjetil Sandve, et al. Hopfield networks is all you need. *arXiv preprint arXiv:2008.02217*, 2020.

Ron Rubinstein, Alfred M Bruckstein, and Michael Elad. Dictionaries for sparse representation modeling. *Proceedings of the IEEE*, 98(6):1045–1057, 2010.

Philipp Seidl, Philipp Renz, Natalia Dyubankova, Paulo Neves, Jonas Verhoeven, Jorg K Wegner, Marwin Segler, Sepp Hochreiter, and Gunter Klambauer. Improving few-and zero-shot reaction template prediction using modern hopfield networks. *Journal of chemical information and modeling*, 62(9):2111–2120, 2022.

Bharath K Sriperumbudur and Gert RG Lanckriet. On the convergence of the concave-convex procedure. In *Nips*, volume 9, pages 1759–1767, 2009.

Yi Tay, Dara Bahri, Liu Yang, Donald Metzler, and Da-Cheng Juan. Sparse sinkhorn attention. In *International Conference on Machine Learning*, pages 9438–9447. PMLR, 2020.

Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. Efficient transformers: A survey. *ACM Computing Surveys*, 55(6):1–28, 2022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Jun Wang and Jean-Daniel Zucker. Solving multiple-instance problem: A lazy learning approach. 2000.

Michael Widrich, Bernhard Schäfl, Milena Pavlović, Hubert Ramsauer, Lukas Gruber, Markus Holzleitner, Johannes Brandstetter, Geir Kjetil Sandve, Victor Greiff, Sepp Hochreiter, et al. Modern hopfield networks and attention for immune repertoire classification. *Advances in Neural Information Processing Systems*, 33:18832–18845, 2020.

Yongyi Yang, Zengfeng Huang, and David Wipf. Transformers from an optimization perspective. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=VT0Y4PlV2m0.

Alan L Yuille and Anand Rangarajan. The concave-convex procedure (cccp). *Advances in neural information processing systems*, 14, 2001.

Alan L Yuille and Anand Rangarajan. The concave-convex procedure. *Neural computation*, 15(4): 915–936, 2003.

Willard I Zangwill. *Nonlinear programming: a unified approach*, volume 52. Prentice-Hall Englewood Cliffs, NJ, 1969.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 11106–11115, 2021.

Tian Zhou, Ziqing Ma, Qingsong Wen, Liang Sun, Tao Yao, Wotao Yin, Rong Jin, et al. Film: Frequency improved legendre memory model for long-term time series forecasting. *Advances in Neural Information Processing Systems*, 35:12677–12690, 2022.