Outlier-Efficient Hopfield Layers for Large Transformer-Based Models

Jerry Yao-Chieh Hu [*1] **Pei-Hsuan Chang** [*2] **Haozheng Luo** [*1] **Hong-Yu Chen** [2] **Weijian Li** [1] **Wei-Po Wang** [2] **Han Liu** [1 3]

Abstract

We introduce an Outlier-Efficient Modern Hopfield Model (termed `OutEffHop`) and use it to address the outlier inefficiency problem of training gigantic transformer-based models. Our main contribution is a novel associative memory model facilitating *outlier-efficient* associative memory retrievals. Interestingly, this memory model manifests a model-based interpretation of an outlier-efficient attention mechanism (Softmax_1): it is an approximation of the memory retrieval process of `OutEffHop`. Methodologically, this allows us to introduce novel outlier-efficient Hopfield layers as powerful alternatives to traditional attention mechanisms, with superior post-quantization performance. Theoretically, the Outlier-Efficient Modern Hopfield Model retains and improves the desirable properties of standard modern Hopfield models, including fixed point convergence and exponential storage capacity. Empirically, we demonstrate the efficacy of the proposed model across large-scale transformer-based and Hopfield-based models (including BERT, OPT, ViT, and STanHop-Net), benchmarking against state-of-the-art methods like $\texttt{Clipped}_{\texttt{S}}\texttt{oftmax}$ and $\texttt{Gated}_{\texttt{A}}\texttt{ttention}$. Notably, `OutEffHop` achieves an average reduction of 22+% in average kurtosis and 26+% in the maximum infinity norm of model outputs across four models. Code is available at GitHub; future updates are on arXiv.

[*]Equal contribution [1]Department of Computer Science, Northwestern University, Evanston, USA [2]Department of Physics, National Taiwan University, Taipei, Taiwan [3]Department of Statistics and Data Science, Northwestern University, Evanston, USA. Correspondence to: Jerry Yao-Chieh Hu <jhu@u.northwestern.edu>, Pei-Hsuan Chang <b09202022@ntu.edu.tw>, Haozheng Luo <robinluo2022@u.northwestern.edu>, Hong-Yu Chen <b0976960890@gmail.com>, Weijian Li <weijianli@u.northwestern.edu>, Wei-Po Wang <b09202009@ntu.edu.tw>, Han Liu <hanliu@northwestern.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

1. Introduction

We address the outlier-inefficient problem in large Transformer-based models by debuting a novel outlier-efficient modern Hopfield model. This problem is of practical importance in the era of Large Foundation Models (Bommasani et al., 2021), i.e. huge transformer-based models, pretrained on massive datasets. They play a central role not only in machine learning but also in a wide range of scientific domains, such as ChatGPT (Brown et al., 2020; Floridi and Chiriatti, 2020) for natural language, BloombergGPT (Wu et al., 2023) for finance, DNABERT (Zhou et al., 2024; 2023; Ji et al., 2021) for genomics, and many others. Specifically, the problem of outlier inefficiency in these large models stems from their tendency to allocate attention to less informative tokens (the "no-op" outliers), including delimiters and punctuation marks. This tendency arises because these large models assign non-zero attention probabilities to low-information tokens, diluting the overall effectiveness of the attention mechanism (Bondarenko et al., 2023, Section 3). As training progresses, the influence of these "no-op" outliers magnifies due to the softmax function's inability to assign zero probability. Consequently, it leads to a scenario where even irrelevant tokens contribute to the model's outputs. Besides, it makes the model need unnecessarily large GPU memory space to host due to the extra bits that outliers take. This hampers the model's processing efficiency and potential accuracy.

To combat this, we take a route from the deep learning compatible modern Hopfield models (Wu et al., 2024a;b; Hu et al., 2024a;b; 2023; Ramsauer et al., 2020). Through the associative memory model interpretation of transformer attention, we introduce a novel outlier-efficient modern Hopfield model. This model's memory retrieval dynamics approximate an outlier-efficient attention mechanism (Softmax_1) (Miller, 2023). This allows us to debut novel outlier-efficient Hopfield layers as outlier-efficient alternatives for vanilla attention (Vaswani et al., 2017). The fundamental idea of our model is to add one extra "no-op classification" dimension into state/configuration space of the Hopfield energy function. This dimension classifies whether a stored memory pattern is a "no-op" outlier, see Figure 1 for a visualization. We regard the "no-op" outliers as distinct or rare patterns with no similarity to other memory patterns. Then, we present an

outlier-efficient Hopfield energy function with a refined log-sum-exponential function. Consequently, this energy-based associative memory model allocates this "no-op" pattern to the zero-energy point of the energy function, remaining unaffected by state updates (retrievals). Remarkably, by the standard CCCP derivation for modern Hopfield models, this new energy function leads to a memory-retrieval dynamics that not only retrieves stored memories in an outlier-efficient fashion but also subsumes the $\mathrm{Softmax}_1$ attention (Miller, 2023) as its special case (when limited to a single update).

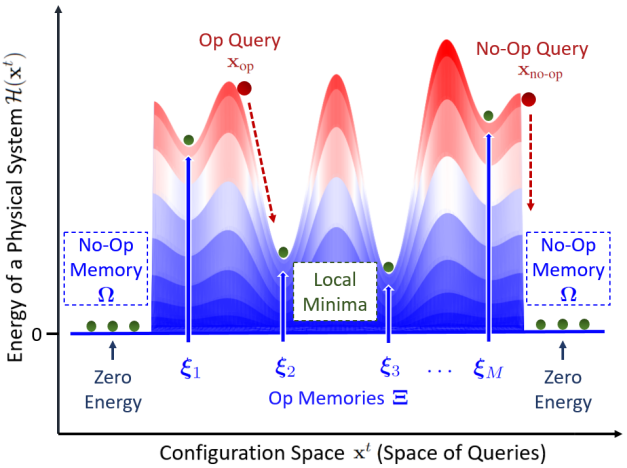


Figure 1. Visualization of Outlier-Efficient Hopfield Model.

Contributions. We propose the Outlier-Efficient Modern Hopfield Model. Our contributions are as follows:

- We propose an associative memory model capable of outlier-efficient memory retrievals with strong physics intuition. Theoretically, we analyze the proposed model equips the standard properties of modern Hopfield models: fixed point convergence (Theorem 3.1) and exponential memory capacity (Theorem 3.3). Importantly, we derive an outlier-efficient Hopfield layer `OutEffHop` as a promising attention alternative (Section 2.4). Moreover, we provide a model-based interpretation for the $\mathrm{Softmax}_1$ attention (Miller, 2023): it is an approximation of the memory retrieval dynamics of the outlier-efficient modern Hopfield model (Lemma 2.1).
- Methodologically, we introduce outlier-efficient Hopfield layers as new components in deep learning. These layers tackle the outlier problem of large models by reducing the probability assigned to low-information vectors. In addition to outlier reduction, we explore the generalization of `OutEffHop`. We establish a generalization bound (Theorem 3.4) that scales with $N^{-1/2}\log N$ in sample size and $\log(dM)$ in the pattern dimension d and the size of the stored memory set M. This positions `OutEffHop` as a promising alternative to transformer attention.
- Empirically, we validate the proposed method on 3 common large transformer-based and 1 Hopfield-based models (BERT (Devlin et al., 2019), Open Pre-trained Transformer (OPT) (Zhang et al., 2022), Vision Transformer (ViT) (Dosovitskiy et al., 2020) and STanHop-Net (Wu et al., 2024b)). Specifically, `OutEffHop` reduces average kurtosis and maximum infinity norm by ∼22+% and ∼26+%, respectively[1], and improves the same metrics by an average of 3% and 4% compared to 3 variants of STanHop-Net and ranks among the top two in outlier efficiency in 25 out of 30 settings.

2. Outlier-Efficient Hopfield Model

This section introduces the Outlier-Efficient Modern Hopfield Model. Section 2.2 presents an internal "no-op classification" mechanism for all memory patterns. Then, Section 2.3 utilizes this mechanism to construct a model facilitating outlier-efficient associative memory retrievals. Importantly, the retrieval dynamics of this model subsumes an outlier-efficient attention as its special case, and Section 2.4 debuts outlier-efficient Hopfield layers for deep learning.

2.1. Background

This section presents the ideas we build on.

"No-Op" Outliers in Attention Heads. Clark et al. (2019); Kovaleva et al. (2019) identify specific tokens in BERT, such as delimiters and punctuation mark, receive larger attention weights. Furthermore, Kobayashi et al. (2020) reveal that tokens with small value vectors tend to receive significantly large attention weights. As stated in (Bondarenko et al., 2023), low-information tokens within BERT and background patches in the Vision Transformer (ViT) attract large attention probability to achieve no-update.

To see this, we consider an input sequence $X = [x_1, \ldots, x_L] \in \mathbb{R}^{d\times L}$ and the attention mechanism

$$\mathrm{Attention}(X) = \mathrm{Softmax}\big(QK^\mathsf{T}\big)V = A.$$

We focus on the part of transformer right after attention

$$\mathrm{Output} = \mathrm{Residual}(X + A). \tag{2.1}$$

If the input X already has enough information and does not require further feature extraction, the attention mechanism tends to behave like an identity map, and output a zero A. This is known as the *no-update situation*: the output of (2.1) is the same as input X. A direct consequence of this is that — the attention mechanism forces tokens with large values (as in V) receive *close-to-zero* attention probability (as in $\mathrm{Softmax}\big(QK^\mathsf{T}\big)$), resulting small-value tokens to have large attention probability. By the normalization nature of softmax function, this operation forces its input QK^T to have a wide range. This is the fundamental source of outliers: there must be some tokens causing the "wide range"

[1] See Table 1 for details.

of QK^T, namely *outliers*. Since attention to these tokens behaves as a "no-op", as mentioned in (Clark et al., 2019), we term these outliers as "no-op" outliers. Furthermore, since the softmax function never reaches exact zero, it always sends back a gradient signal, leading to the magnification of outliers during training (Bondarenko et al., 2023).

Modern Hopfield Models. Let $x \in \mathbb{R}^d$ represent the query patterns and $\Xi = [\xi_1, \cdots, \xi_M] \in \mathbb{R}^{d\times M}$ the memory patterns. Krotov and Hopfield (2016) introduce the dense associative memory model encoding memory patterns Ξ into energy function $\mathcal{H}(x)$ using *overlap-construction*: $\mathcal{H}(x) = F(\Xi^\mathsf{T} x)$, where $F : \mathbb{R}^M \to \mathbb{R}$ is a smooth function. The choice of energy function and the corresponding retrieval dynamics results in different Hopfield models types (Krotov and Hopfield, 2016; 2021; Demircigil et al., 2017; Ramsauer et al., 2020; Hu et al., 2023; 2024a; Wu et al., 2024a;b). Inspired by the dense associative memory models, Ramsauer et al. (2020) introduce the modern Hopfield models with the energy function of the form

$$\mathcal{H}(x) = -\text{lse}\left(\beta, \Xi^\mathsf{T} x\right) + \frac{1}{2}\langle x, x\rangle + \text{Const.},$$

where $\text{lse}(\beta, z) \coloneqq \beta^{-1}\log\sum_{\mu=1}^M \exp\{\beta z_\mu\}$. In addition, they introduce the corresponding retrieval dynamics as

$$x_{\text{new}} \leftarrow \mathcal{T}(x) = \Xi\text{Softmax}(\beta\Xi^\mathsf{T} x), \tag{2.2}$$

for any input query $x \in \mathbb{R}^d$. The modern Hopfield model possesses several desirable properties, including:

1. **Exponential Memory Capacity:** Achieved by highly non-linear energy functions.
2. **One-step Retrieval Dynamics:** Achieved by guaranteeing monotonic energy function minimization.
3. **Compatibility with Deep Learning Architectures:** Achieved by the link between their retrieval dynamics and attention mechanisms.

2.2. One Dimension More

As models of associative memory, modern Hopfield models aim to retrieve a memory pattern x_{new} from the stored memories Ξ, closest to the input query x. By (2.2), they do this by computing the output x_{new} as the *expectation value* of Ξ over the distribution $\text{Softmax}(\Xi^\mathsf{T} x)$. Crucially, the weight of $\text{Softmax}(\Xi^\mathsf{T} x)$, i.e., $\Xi^\mathsf{T} x$, represents the inner-product similarity measure between the input query x and each stored memory ξ_μ. Namely, the greater $\langle \xi_\mu, x\rangle$ is, the stronger their correlation.

Under this interpretation, for a given query x, the memory patterns with low similarity inevitably deviate the expectation value from the ground truth. This occurs because the softmax function always assigns non-zero probability weights, even for near zero similarity $\langle \xi_\mu, x\rangle \simeq 0$. Consequently, this results to more iterative retrievals for the retrieval dynamics to converge to the ground truth memory (w.r.t x). We refer to these low-similarity memory patterns as "**no-op** patterns," as they are unrelated to the presented query and should **no**t **op**erate during the retrieval process.

Motivated by above, we introduce a new dimension into the pattern vectors to distinguish "no-op patterns" from the relevant ones, via the following "no-op classification."

No-Op Classification Mechanism. Given an input query pattern $x = (x_1, \ldots, x_d)$ and memory patterns $\xi^\mu = (\xi_1^\mu, \cdots, \xi_d^\mu)$ with $\mu \in [M]$. We extend their dimension such that

$$\bar{x} = (x_1, \ldots, x_d, 0), \quad \bar{\xi}^\mu = (\xi_1^\mu, \cdots, \xi_d^\mu, \omega),$$

with an extra $\omega \in \mathbb{R}$. In addition, for memory patterns, we set this extra dimension ω to be

- $\omega \neq 0$: non-zero for no-op outliers, and
- $\omega = 0$: zero for the rest memory patterns,

assuming we are aware of which patterns are outliers[2]. Then we introduce the following function:

$$\Lambda(\bar{\xi}_\mu) = \begin{cases} (\xi_1^\mu, \cdots, \xi_d^\mu, 0) = \bar{\xi}_{\text{op}}^\mu \in \mathbb{R}^{d+1}, & \text{if } \omega = 0, \\ (\underbrace{0, \cdots, 0}_{d}, C) = \Omega \in \mathbb{R}^{d+1}, & \text{if } \omega \neq 0\ , \end{cases} \tag{2.3}$$

with some $C \in \mathbb{R}$ and for all $\mu \in [M]$, to map all "no-op patterns" into an unique "*no-op memory class vector* Ω." By design, the inner product of the vector Ω with the query $\bar{x}$ is zero: $\langle \Omega, \bar{x}\rangle = 0$. We term the Λ function (2.3) the "no-op classification mechanism." It enforces all outlier memory patterns to have zero inner product with the input query.

Remark 2.1. It is also feasible to design Λ so that each outlier pattern maps to a distinct, non-repeated C. However, the form presented here offers better elegance and simplicity.

In sum, for any set of (d-dimensional patterns) x and $\Xi = [\xi_1, \ldots, \xi_M]$, we obtain a set of ($(d+1)$-dimensional patterns) $\bar{x}$ and $\bar{\Xi} = [\bar{\xi}_1, \ldots, \bar{\xi}_M]$. Suppose there are K outliers in $\bar{\Xi}$. Then, with Λ, we further categorize $\bar{\Xi}$ into $(M-K)$ $\{\bar{\xi}_{\text{op}}^\mu\}_{\mu\in[M-K]}$ and a single Ω.

2.3. Hopfield Energy and Retrieval Dynamics

Now, we utilize above to construct the Outlier-Efficient Modern Hopfield Model. For the ease of presentation, in the following, we set

$$x \leftarrow \bar{x}, \quad \xi_\mu \leftarrow \bar{\xi}_\mu \quad (\text{i.e., } \Xi \leftarrow \bar{\Xi}),$$

for query and memory patterns. Move rover, since we only need a single Ω for outliers, we set

$$d \leftarrow (d+1), \quad M \leftarrow (M-K),$$

[2]We can do this by either ad-hoc assignment or similarity measure thresholding (See B.1 for details).

for pattern dimension and the number of "op" memory patterns. We introduce the outlier-efficient Modern Hopfield energy as:

$$\mathcal{H}(x) = -\mathrm{lse}_1\left(\beta, \Xi^\mathsf{T} x\right) + \frac{1}{2}\langle x, x\rangle + \text{Const.}, \tag{2.4}$$

where lse_1 is a refined log-sum-exponential fucntion:

$$\begin{aligned}&\mathrm{lse}_1\left(\beta, \Xi^\mathsf{T} x\right)\\ &:= \beta^{-1}\log\left(\sum_{\mu=1}^{M}\exp\{\beta\langle\xi_\mu, x\rangle\} + \exp\{\beta\langle\Omega, x\rangle\}\right)\\ &= \beta^{-1}\log\left(\sum_{\mu=1}^{M}\exp\{\beta\langle\xi_\mu, x\rangle\} + 1\right).\end{aligned} \tag{2.5}$$

Remark 2.2. Since the log function is monotonic, (2.5) has a physical interpretation of an energy function with a *zero-energy point*[3] associated with $\exp\{\beta\langle\Omega, x\rangle\} = \exp\{0\}$. Naturally, this zero-energy point serves as one of the local minima of $\mathcal{H}$, and thus corresponds to a memory pattern. More precisely, we retrieve "no-op" memory from there with proper retrieval dynamics $\mathcal{T}$.

Remark 2.3. Echoing with Remark 2.1, if we twist (2.3) to have one Ω_C for each no-op patterns with non-repeated C's, then (2.5) simply becomes $\mathrm{lse}_K\left(\beta, \Xi^\mathsf{T} x\right) := \beta^{-1}\log\left(\sum_{\mu=1}^{M}\exp\{\beta\langle\xi_\mu, x\rangle\} + K\right)$.

Retrieval Dynamics. With (2.4), we derive the following memory retrieval dynamics:

Lemma 2.1 (Retrieval Dynamics)**.** Let $\mathrm{Softmax}_1(z) := \exp\{z\}/\left(\sum_{\mu=1}^{M}\exp\{z_\mu\} + 1\right)$ for any $z \in \mathbb{R}^M$ and t be the iteration number. The memory retrieval dynamics:

$$\mathcal{T}_{\text{OutEff}}(x_t) := \Xi\mathrm{Softmax}_1\left(\beta\Xi^\mathsf{T} x_t\right) = x_{t+1}, \tag{2.6}$$

monotonically minimizes the energy (2.4) over t.

Proof Sketch. Since (2.5) is concave by design, we prove this by standard CCCP derivation following (Hu et al., 2023). See Appendix C.1 for a detailed proof. □

Due to the monotonic decreasing property of Lemma 2.1, for any given input query x, (2.6) retrieves a memory closest to it by approaching to the nearest local minimum of $\mathcal{H}$. Interesting, when $\mathcal{T}_{\text{OutEff}}$ is applied only once, (2.6) is equivalent to an outlier-efficient attention (Miller, 2023).

Remark 2.4. Importantly, this set of $\mathcal{H}$ and $\mathcal{T}$ enables outlier-efficient associative memory retrievals. When we identify specific memory patterns as outliers relative to the input query and classify them into Ω, they no longer contribute to the retrieval output defined by (2.6).

[3]This zero-energy point does not mean $\mathcal{H} = 0$.

2.4. Connection to Deep Learning

Outlier-efficient Hopfield model is applicable to nowadays deep learning architectures, due to its connection to transformer attention mechanism when the retrieval dynamics $\mathcal{T}_{\text{OutEff}}$ undergoes a single iteration. Consider the raw query R and memory pattern Y. We define the *query* and *memory* associative (or embedded) spaces through transformations: $X^\mathsf{T} = RW_Q := Q$ and $\Xi^\mathsf{T} = YW_K := K$, with matrices W_Q and W_K. By transposing the retrieval dynamics (2.6) and multiplying with W_V (letting $V := KW_V$), we get: $Q^{\text{new}}W_V = \mathrm{Softmax}_1(\beta QK^\mathsf{T})V$.

This equation resembles the attention mechanism but with a $\mathrm{Softmax}_1$ activation. When substituting the original patterns R and Y, we present the Outlier-Efficient Hopfield (`OutEffHop`) layer:

$$\begin{aligned}Z &= \texttt{OutEffHop}(R, Y)\\ &= \mathrm{Softmax}_1\left(\beta RW_Q W_K^\mathsf{T} Y^\mathsf{T}\right) YW_K W_V.\end{aligned} \tag{2.7}$$

This layer is readily incorporated into deep learning models. To elaborate, the `OutEffHop` layer takes R and Y as input, paired with weight matrices W_Q, W_K, and W_V. Similar to (Hu et al., 2024a; Wu et al., 2024b; Hu et al., 2023; Ramsauer et al., 2020), its configuration determines its behavior:

- **Memory Retrieval:** This mode does not require learning. The matrices W_K, W_Q, and W_V are identity matrices. R acts as the query to retrieve memory patterns Y.
- **`OutEffHop`:** In this design, R and Y are inputs. The matrices W_K, W_Q, and W_V are adjustable, offering an alternative to the usual attention mechanism with outlier efficiency. R, Y, and Y function as the sources of query, key, and value, respectively. To mimic a self-attention mechanism, we set R equal to Y.
- **`OutEffHopPooling`:** Here, Y is the only input of the layer. Q acts as a learnable query that can search static prototype patterns in Y. We consider this layer as a pooling layer if only one static state pattern (query) exists.
- **`OutEffHopLayer`:** With just R as input (which denotes the query pattern), the adaptive matrices W_K and W_V act as repositories for stored patterns and pattern projections. This implies that keys and values are independent of input, suggesting an interpretation of Y as an identity matrix.

Remark 2.5. For outlier efficient Hopfield model with lse_K energy (Remark 2.3), the corresponding deep learning layer becomes $\mathrm{Softmax}_K(x_i) = \exp(x_i)/(K + \sum_j \exp(x_j))$.[4]

[4]https://github.com/softmax1/Flash-Attention-Softmax-N

3. Theoretical Analysis

In this section, we validate our model as a theoretically robust Hopfield model. Furthermore, by establishing a lower upper bound on retrieval error, we prove the proposed model's enhancements over its original counterpart, including expanded memory capacity.

3.1. Convergence Guarantee

We start our analysis with the notion of memory storage and retrieval[5] of modern Hopfield models following (Wu et al., 2024b; Hu et al., 2023; Ramsauer et al., 2020).

Definition 3.1 (Storage and Retrieval)**.** For all $\mu \in [M]$, let $R := \frac{1}{2}\operatorname{Min}_{\mu,\nu\in[M];\mu\neq\nu}\|\xi_\mu - \xi_\nu\|$ be the finite radius of each sphere $\mathcal{S}_\mu$ centered at memory pattern ξ_μ. We say ξ_μ is *stored* if all $x \in \mathcal{S}_\mu$ are generalized fixed points of $\mathcal{T}$, $x^\star_\mu \in \mathcal{S}_\mu$, and $\mathcal{S}_\mu \cap \mathcal{S}_\nu = \emptyset$ for $\mu \neq \nu$. We say ξ_μ is ϵ-*retrieved* by $\mathcal{T}$ with x for an error ϵ, if $\|\mathcal{T}(x) - \xi_\mu\| \leq \epsilon$.

Definition 3.1 does not guarantee alignment between $\mathcal{T}_{\text{OutEff}}$'s fixed points and $\mathcal{H}$'s stationary points. Additionally, the monotonicity of equation (2.6) does not ensure the existence of stationary points concerning energy $\mathcal{H}$ (Sriperumbudur and Lanckriet, 2009). In the following lemma, we establish our proposed model as a well-defined Hopfield model by demonstrating two types of convergence.

Theorem 3.1 (Convergence of $\mathcal{T}_{\text{OutEff}}$)**.** Suppose $\mathcal{H}$ is given by (2.4) and $\mathcal{T}_{\text{OutEff}}(x)$ is given by (2.6). For any sequence $\{\mathbf{x}_t\}_{t=0}^{\infty}$ defined by $\mathbf{x}_{t'+1} = \mathcal{T}_{\text{OutEff}}(\mathbf{x}_{t'})$, all limit points of this sequence are stationary points if they are obtained by iteratively applying $\mathcal{T}_{\text{OutEff}}$ to $\mathcal{H}$.

Proof Sketch. Following (Hu et al., 2023), we first show that $\mathcal{H}$ converges to its generalized fixed point $x^\star_\mu$ through $\mathcal{T}_{\text{OutEff}}$ (1st convergence guarantee). Then, we show that $x^\star_\mu$ corresponds to the stationary points of the energy minimization, and hence $\mathcal{H}$ converges to local optimum (2nd convergence guarantee). See Appendix C.2 for a proof. □

3.2. Retrieval Error Analysis

Calibrating against the standard results (Ramsauer et al., 2020), we prove the superiorities of the proposed model.

Theorem 3.2 (Retrieval Error)**.** Let $\mathcal{T}_{\text{original}}$ be the retrieval dynamics of the original modern Hopfield model (Ramsauer et al., 2020). $\|\mathcal{T}_{\text{OutEff}}(x) - \xi_\mu\|$ has lower upper bound than $\|\mathcal{T}_{\text{original}}(x) - \xi_\mu\|$ for all $x \in S_\mu$

Corollary 3.2.1 (Tighter Retrieval Error)**.** Assume all patterns x and $\{\xi_\mu\}_{\mu\in[M]}$ are normalized. Let $\gamma := \sum_{\mu=1}^{M}\left[\text{Softmax}_1\left(\beta\boldsymbol{\Xi}^\intercal x\right)\right]_\mu$ and α be the angle between $\mathcal{T}_{\text{original}}(x)$ and ξ_μ. It holds $\|\mathcal{T}_{\text{OutEff}}(x) - \xi_\mu\| \leq \|\mathcal{T}_{\text{original}}(x) - \xi_\mu\|$ when $(\gamma+1)/2 \geq \cos(\alpha)$.

Proof. See Appendix C.3 and Appendix C.4 for detailed proofs of Theorem 3.2 and Corollary 3.2.1. □

Remark 3.1. Corollary 3.2.1 is typically observed at the beginning of retrieval.

Theorem 3.3 (Memory Capacity Lower Bound, Informal)**.** Assume all memory patterns are randomly sampled from a sphere of radius m. For any $\beta > 0$, our proposed model's capacity to store and retrieve patterns scales exponentially with the pattern size d, and has a larger capacity lower bound than that of original modern Hopfield model (Ramsauer et al., 2020): $M \geq M_{\text{original}}$.

Proof. See Appendix C.5 for a detailed proof. □

Remark 3.2. Comparing previous asymptotic larger capacity results of sparse models (Hu et al., 2023; Wu et al., 2024b) with large β, Theorem 3.3 is exact for all $\beta > 0$.

3.3. Generalization Bound

Following notations from Section 2.4, we analyze the generalization of the proposed layers. Consider the input query $Q = [q_1, ..., q_T]^\top \in \mathbb{R}^{T\times d}$ and memory pattern $Y = [y_1, ..., y_M]^\top \in \mathbb{R}^{M\times a}$, where $y \in \mathbb{R}^a$ and $q \in \mathbb{R}^d$.

As standard supervised learning setting, we set the sample size (number of sequences) to be N, i.e. *input* queries $Q^{(1)}, Q^{(2)}, ..., Q^{(N)}$, and the corresponding *target* memory sets to be $Y^{(1)}, Y^{(2)}, ..., Y^{(N)}$. For vectors, $\|\cdot\|_p$ and $\|\cdot\|$ denote the ℓ_p-norm and ℓ_2-norm of vectors, respectively. For matrices, $\|\cdot\|_p$ denotes the ℓ_p-norm, and $\|\cdot\|_{p,q}$ the (p,q) matrix norm, which is q-norm of the p-norm of the columns of a matrix. Namely, $\|A\|_{p,q} = \|(\|a_1\|_p, \ldots, \|a_i\|_p, \ldots)\|_q$, where a_i is the i-th column vector of A.

Here, we consider the transpose of (2.7) and taking $\widetilde{W}_V := W_K W_V$, while taking query Q and raw memory pattern Y as inputs. We write the Outlier-Efficient Modern Hopfield mechanism as a function $f_{\text{hop}} : \mathbb{R}^{M\times a} \times \mathbb{R}^{T\times d} \to \mathbb{R}^{d\times T}$ (i.e. f_{hop} is the transpose of `OutEffHop` in (2.7)):

$$f_{\text{hop}}(Y, Q; W_K, \widetilde{W}_V) = \widetilde{W}_V^\intercal Y^\intercal \text{Softmax}_1\left(\beta Y W_K Q^\intercal\right),$$

where $W_K \in \mathbb{R}^{a\times d}$ and $\widetilde{W}_V \in \mathbb{R}^{a\times d}$. Also, we write the corresponding function class

$$\mathcal{F}_{\text{hop}} := \{(Y, Q) \mapsto f_{\text{hop}}(Y, Q; W_K, \widetilde{W}_V) \mid W_K \in \mathcal{W}_K, \widetilde{W}_V \in \widetilde{\mathcal{W}}_V\}.$$

and make the following mild assumptions.

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

Table 1. **Comparing OutEffHop with Vanilla Attention in BERT, OPT, ViT and STanHop-Net.** We showcase the outlier efficiency of `OutEffHop` in 3 large transformer-based and 1 Hopfield-based models, using Average Kurtosis and Maximum Infinity Norm $\|x\|_\infty$. Additionally, we showcase the quantization performance of `OutEffHop`, by comparing FP16 and W8A8 (Weight-8bit-Activation-8bit) performance. The best results are highlighted in bold, and the second-best results are underlined. In all settings, `OutEffHop` delivers significant outlier reduction, and further enhances its combinations with `Clipped_Softmax` and `Gated_Attention`. *For FP16 and W8A8, we report *Perplexity Score* for BERT and OPT, *Top-1 Accuracy* for ViT, and *Mean Square Error* (MSE) for STanHop-Net.

Model	Method	Avg. kurtosis	Max inf. norm	FP16*	W8A8*	Parameters
BERT	Vanilla	418.724 ± 0.814	255.859 ± 0.004	6.237 ± 0.001	7.154 ± 0.009	108.9m
	OutEffHop	26.564 ± 0.022	33.618 ± 0.000	6.209 ± 0.001	6.295 ± 0.001	
	Clipped Softmax	14.210 ± 0.003	33.619 ± 0.001	**6.118** ± 0.002	**6.189** ± 0.001	
	Clipped OutEffHop	**11.839** ± 0.001	**30.107** ± 0.001	6.133 ± 0.000	6.199 ± 0.001	
	Gated Attention	17.779 ± 0.014	34.082 ± 0.000	6.230 ± 0.001	6.299 ± 0.003	109m
	Gated OutEffHop	15.625 ± 0.012	32.777 ± 0.000	6.214 ± 0.001	6.279 ± 0.003	
OPT	Vanilla	23341.513 ± 27.363	92.786 ± 0.002	15.974 ± 0.001	42.012 ± 19.514	124.06m
	OutEffHop	21.542 ± 0.000	13.302 ± 0.001	15.916 ± 0.002	16.429 ± 0.013	
	Clipped Softmax	9731.110 ± 0.000	43.803 ± 0.000	16.042 ± 0.000	30.825 ± 0.330	
	Clipped OutEffHop	24127.332 ± 0.000	67.602 ± 0.000	16.118 ± 0.000	29.269 ± 0.184	
	Gated Attention	90.321 ± 0.000	13.704 ± 0.000	**15.677** ± 0.000	16.236 ± 0.074	124.07m
	Gated OutEffHop	**11.449** ± 0.000	**7.568** ± 0.000	15.751 ± 0.000	**16.148** ± 0.005	
ViT	Vanilla	37.104 ± 0.000	272.198 ± 0.000	76.810 ± 0.000	74.935 ± 0.046	22.03m
	OutEffHop	31.601 ± 0.001	249.163 ± 0.000	76.788 ± 0.000	**76.313** ± 0.012	
	Clipped Softmax	33.868 ± 0.00	257.613 ± 0.00	76.612 ± 0.000	75.179 ± 0.013	
	Clipped OutEffHop	24.642 ± 0.000	196.199 ± 0.001	**76.871** ± 0.001	76.083 ± 0.007	
	Gated Attention	45.145 ± 0.864	269.279 ± 1.426	69.922 ± 2.436	67.479 ± 1.447	22.04m
	Gated OutEffHop	**21.979** ± 0.254	**60.169** ± 1.153	74.089 ± 2.585	73.958 ± 3.126	
STanHop-Net	Vanilla	2.954 ± 0.063	5.048 ± 0.232	0.360 ± 0.008	0.362 ± 0.000	35.13m
	OutEffHop	2.897 ± 0.011	4.565 ± 0.209	**0.360** ± 0.004	0.355 ± 0.000	
	Clipped Softmax	2.995 ± 0.05	4.890 ± 0.17	0.553 ± 0.03	0.591 ± 0.000	
	Clipped OutEffHop	2.864 ± 0.06	**4.145** ± 0.23	0.506 ± 0.05	0.517 ± 0.000	
	Gated Attention	2.487 ± 0.017	4.277 ± 0.163	0.380 ± 0.006	0.375 ± 0.000	35.15m
	Gated OutEffHop	**2.459** ± 0.041	4.240 ± 0.155	0.376 ± 0.007	0.367 ± 0.000	

Assumption 3.1 (Norm Bounds)**.** We assume that

(A1). Query vectors r_τ are bounded by 1 in ℓ_2-norm

$$\|q_\tau\| \leq 1 \quad \forall \tau \in [T].$$

(A2). Memory vectors y_t are bounded in ℓ_2-norm

$$\|y_t\| \leq B_Y \quad \forall t \in [M].$$

(A3). W_K is bounded in $\ell_{2,1}$-norm

$$\mathcal{W}_K : \{W_K \in \mathbb{R}^{a\times d} \mid \left\|W_K^\top\right\|_2 \leq B_K, \|W_K\|_{2,1} \leq B_K^{2,1}\}.$$

(A4). $\widetilde{W}_V$ is bounded in ℓ_2-norm and $\ell_{2,1}$-norm

$$\widetilde{\mathcal{W}}_V : \{\widetilde{W}_V \in \mathbb{R}^{a\times d} \mid \left\|\widetilde{W}_V^\intercal\right\|_2 \leq B_V, \left\|\widetilde{W}_V\right\|_{2,1} \leq B_V^{2,1}\}.$$

Then, we state our generalization results.

Theorem 3.4 (Outlier Efficient Hopfield Layer Generalization Bound)**.** For any $\delta > 0$, with probability at least $1-\delta$,

$$\varepsilon_{\text{gen}}(f_{\text{hop}}) \leq \widetilde{\mathcal{O}}\left(\sqrt{N^{-1}}\left[\sqrt{(E_1+E_2)^3} + \sqrt{\log(1/\delta)}\right]\right),$$

where $E_1 = \left[4B_V^2 B_Y^2 (\beta B_K^{2,1})^2 \log(dNM)\right]^{1/3}$, $E_2 = \left[(B_V^{2,1})^2 \log(dNM)\right]^{1/3}$.

Proof Sketch. We first derive the covering number bound of the Outlier-Efficient Hopfield Layer by showing the Lipschitzness of f_{hop} (Lemma C.7). By Dudley's Theorem, we obtain the generalization bound via covering number (Lemma C.4). See Appendix C.6 for a detailed proof. □

Our results indicate that the generalization error remains controllable as long as the size of the data N *at least* scales logarithmically with the pattern dimension d and the size of stored memory set M. In addition, the length of the sequence, T, does not impact generalization, making Hopfield layers as a promising alternative to transformer attention.

4. Experimental Studies

We conduct a series of experiments to validate the Outlier-Efficient Modern Hopfield Model and layers. Specifically,

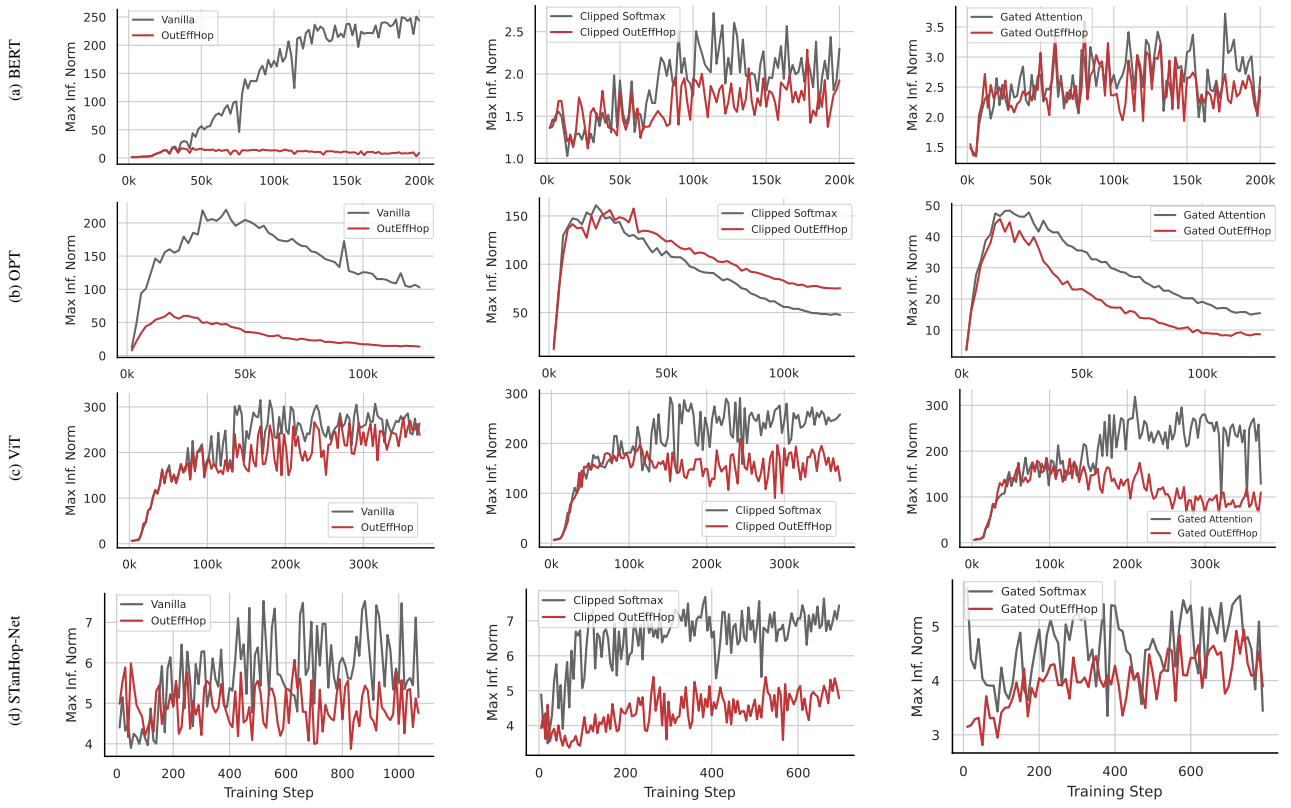


Figure 2. **The Impact of `OutEffHop` on Maximum Infinity Norm $\|\mathbf{x}\|_\infty$ Changes During Pretraining of (a) BERT, (b) OPT, (c) ViT, and (d) STanHop-Net.** The plots, from left to right, compare `OutEffHop` with the vanilla attention baseline and their combination with `Clipped_Softmax` and `Gated_Attention` as per (Bondarenko et al., 2023). Each figure's y-axis scale varies. For better visualization, we focus on the outlier reduction in layer 10 of the BERT, ViT and OPT model, and in layer 9 of the STanHop-Net. In all settings, `OutEffHop` delivers significant reduction of the $\|\mathbf{x}\|_\infty$ compared to the vanilla attention and improves `Clipped_Softmax` and `Gated_Attention`.

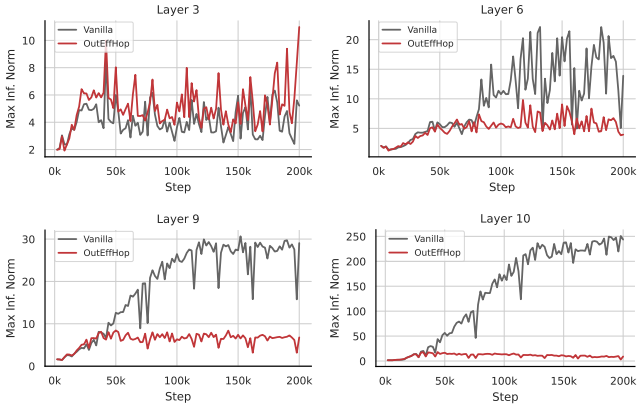


Figure 3. The trend of Feed-Forward Network (FFN) output maximum infinity norm values in layers 3, 6, 9, and 10 of a BERT encoder is analyzed using two softmax variations: `OutEffHop` (represented in red) and vanilla Softmax (in grey). The findings indicate that `OutEffHop` significantly reduces outliers in the model compared to the vanilla Softmax.

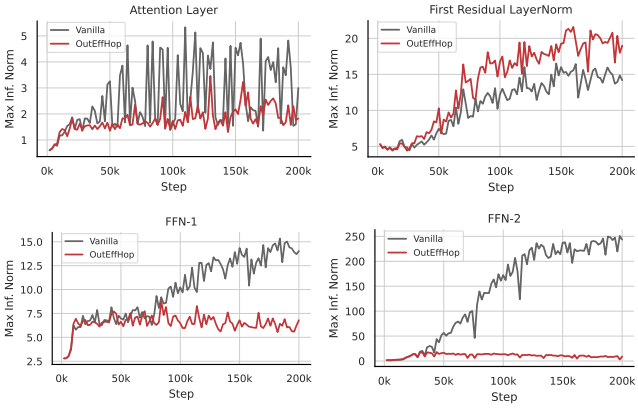


Figure 4. Maximum infinity norm $\|\mathbf{x}\|_\infty$ for different tensor components within layer 10 of BERT. Our work is analysed using two softmax variations: `OutEffHop` (represented in red) and vanilla Softmax (in grey). We find `OutEffHop` suppresses the outliers growing in both FFN layers.

we test our model in accordance with SOTA methods outlined in (Bondarenko et al., 2023), with 3 common large transformer-based models and 1 Hopfield-based model.

4.1. Outlier Efficiency of `OutEffHop`

To test the model's robustness against outliers, we use `OutEffHop` in BERT (Devlin et al., 2019), Open Pretrained Transformers (OPT) (Zhang et al., 2022), Vision Transformers (ViT) (Dosovitskiy et al., 2020) and STanHop-Net (Wu et al., 2024b) to replace the vanilla attention layer (Vaswani et al., 2017) and Hopfield layer (Hu et al., 2023; Ramsauer et al., 2020). We then train these models from scratch and evaluate them on the validation set. We conduct each evaluation three times with different random seeds and present the average and standard deviation for each metric.

Metrics. We report *maximum infinity norm* $\|\mathbf{x}\|_\infty$ of the activation tensors $\mathbf{x}$ across all the Transformer layers as the metrics for outliers' existence. Also, we report the *average kurtosis* of $\mathbf{x}$. For BERT, we only average on the outputs tensors from the Feed-Forward Network (FFN) layer and Layer Normalization. These two parts are known for outlier presence, as confirmed by our experiments and previous studies (Bondarenko et al., 2021; Wei et al., 2022; Bondarenko et al., 2023). In the case of OPT, ViT and STanHop, we average over every output component in the transformer layers. These two metrics have been shown to correlate well with the model quantizability (i.e., robustness against outliers) (Bondarenko et al., 2021; Shkolnik et al., 2020). Specifically, previous studies (Dettmers et al., 2022; Wei et al., 2022; Bondarenko et al., 2021) highlight a substantial decline in model performance attributed to quantization in the presence of outliers. As a result, we report the models' performance before and after quantization. For before quantization performance, we report (i) **FP16** (in 16-bit floating-point) *Perplexity Score* for BERT and OPT, (ii) **FP32** *Top-1 Accuracy* for ViT, and (iii) *Mean Square Error* (MSE) for STanHop-Net. For after quantization performance **W8A8** (in 8-bit floating-point), we report the same metrics.

Datasets. We use 4 real-world datasets: Bookcorpus (Zhu et al., 2015), wiki40b/en (Guo et al., 2020), ImageNet-1k (Russakovsky et al., 2015) and ETTh1 (Zhou et al., 2021). The first two are for language models, i.e. OPT and BERT, the third is for vision model, i.e. ViT, and the last is for time series model, i.e. STanHop-Net.

Models. Following Bondarenko et al. (2023), we validate our method (`OutEffHop` layers) with 4 popular models: 2 language models (BERT, OPT), 1 vision model (ViT) and 1 time series model (STanHop). For BERT, we adopt the BERT-base-uncased model of size 109 million parameters[6]. We pretrain this model with the masked language modeling (MLM) technique, following the original BERT paper (Devlin et al., 2019). As for OPT, we adopt a OPT model of size 125 million parameters[7]. For this model, we employ causal language modeling (CLM) as the pre-training objective. To optimize training efficiency, we set specific constraints on sequence length: sequence length of 128 for BERT and of 512 for OPT. As for ViT, we adopt the ViT-S_16 variant of size 22.03 million parameters[8]. We pretrain this model with standard image classification objective. As for STanHop-Net, we adopt a STanHop-Net of size 35.13 million parameters[9]. We pretrain this model on a multivariate time series prediction objective.

Results. In Table 1 and Figure 2, our results show that `OutEffHop` achieves performance in outlier reduction comparable to `Clipped_Softmax` and `Gated_Attention`. Moreover, combining `OutEffHop` with these two methods further improves the effect, achieving an average reduction of $\sim$22+% in average kurtosis and $\sim$26+% in maximum infinity norm across four test models. The only exception is Clipped `OutEffHop` in the OPT model. This anomaly aligns with the findings of Bondarenko et al. (2023), which suggest that the `Clipped_Softmax` approach does not perform well with OPT. In sum, the efficacy of `OutEffHop` is also apparent in the reduction of the maximum infinity norm value during the pre-training process, particularly noticeable in layer 10 of the BERT, ViT and OPT model, and in layer 9 of the STanHop models, as depicted in Figure 2. `OutEffHop` is more efficient at reducing outliers during the pretraining process compared to its baseline methods, with particularly notable improvements in the OPT model.

4.2. `OutEffHop` Improves Hopfield-Centric Deep Learning Model: A Case Study on STanHop-Net

We also test our method on STanHop-Net (Wu et al., 2024b), a Hopfield-based time series prediction model. We conduct a comparison between our method and common modern Hopfield layers (Hu et al., 2023; Ramsauer et al., 2020).

Data. Following Wu et al. (2024b), we use 3 realistic datasets for multivariate time series prediction tasks: ETTh1 (Electricity Transformer Temperature-hourly), ETTm1 (Electricity Transformer Temperature-minutely), WTH (Weather). We divide these datasets into training, validation, and test sets with a ratio of 14/5/5. For each dataset, we conduct evaluations across various prediction horizons.

Metrics. To evaluate the outlier efficiency, we use the same metrics as the above experiments: the maximum infin-

[6] https://huggingface.co/bert-base-uncased

[7] https://huggingface.co/facebook/opt-125m

[8] https://huggingface.co/WinKawaks/vit-small-patch16-224

[9] https://github.com/MAGICS-LAB/STanHop

Table 2. **STanHop-Net (Wu et al., 2024b): Outlier Reduction of Multivariate Time Series Predictions.** We implement 4 STanHop variants, **Hopfiled** with **D**ense `Hopfield` layer (Ramsauer et al., 2020), **SparseHopfiled** with **S**parse `SparseHopfield` layer (Hu et al., 2023), **STanHop-Net** with GSH layer (Wu et al., 2024b) and **OutEffHop** with our $\mathrm{Softmax}_1$ layer respectively. To evaluate outlier reduction performance, we report the maximum infinity norm and average kurtosis metrics. We also report the average Mean Square Error (MSE) and Mean Absolute Error (MAE) metrics with variance omitted as they are all $\leq 2\%$. We evaluate each dataset with different prediction horizons (shown in the second column). We have the best results **bolded** and the second best results underlined. In 25 out of 30 settings, `OutEffHop` ranks either first or second. Our results indicate that our proposed `OutEffHop` delivers consistent top-tier outlier-reduction performance compared to all the baselines.

Models		`Hopfield`				`SparseHopfield`				STanHop-Net (GSH)				`OutEffHop`			
Metric		MSE	MAE	Avg. kurtosis	Max inf. norm	MSE	MAE	Avg. kurtosis	Max inf. norm	MSE	MAE	Avg. kurtosis	Max inf. norm	MSE	MAE	Avg. kurtosis	Max inf. norm
ETTh1	24	0.360	0.401	2.954 ± 0.063	5.048 ± 0.232	0.388	0.411	3.311 ± 0.082	4.954 ± 1.064	0.395	0.415	3.269 ± 0.117	4.947 ± 0.173	0.361	0.397	**2.897** ± 0.011	**4.565** ± 0.209
	48	0.405	0.424	2.968 ± 0.039	4.969 ± 0.033	0.466	0.452	3.295 ± 0.136	4.749 ± 0.517	0.458	0.448	3.271 ± 0.200	4.644 ± 0.341	0.409	0.426	**2.965** ± 0.004	**4.570** ± 0.424
	168	0.881	0.710	2.545 ± 0.004	3.923 ± 0.115	1.422	0.921	3.149 ± 0.015	4.348 ± 0.085	1.422	0.926	3.093 ± 0.065	4.160 ± 0.285	0.872	0.704	**2.526** ± 0.011	**3.865** ± 0.035
	336	0.755	0.648	2.436 ± 0.003	3.536 ± 0.230	1.223	0.851	3.071 ± 0.009	4.156 ± 0.199	1.381	0.909	3.043 ± 0.021	4.248 ± 0.159	0.780	0.658	**2.433** ± 0.009	**3.416** ± 0.042
	720	0.852	0.709	**2.443** ± 0.006	3.266 ± 0.132	1.134	0.824	3.030 ± 0.015	4.179 ± 0.054	1.360	0.904	3.062 ± 0.089	4.238 ± 0.197	0.894	0.788	2.450 ± 0.035	**3.218** ± 0.142
ETTm1	24	0.272	0.339	3.617 ± 0.003	4.717 ± 0.353	0.265	0.331	**3.357** ± 0.045	4.334 ± 0.087	**0.261**	**0.328**	3.547 ± 0.096	4.696 ± 0.279	0.347	0.429	3.584 ± 0.136	**4.212** ± 0.262
	48	0.352	0.387	4.211 ± 0.113	5.603 ± 0.854	**0.304**	**0.355**	4.280 ± 0.102	6.296 ± 0.479	0.328	0.367	4.384 ± 0.415	**5.557** ± 4.188	0.375	0.409	**3.967** ± 0.253	5.816 ± 0.209
	96	0.396	0.412	3.102 ± 0.026	4.534 ± 0.328	0.345	0.383	3.568 ± 0.127	4.441 ± 0.650	**0.344**	0.375	3.609 ± 0.364	4.618 ± 0.319	0.529	0.487	**3.014** ± 0.042	**4.333** ± 0.394
	288	0.600	0.540	2.643 ± 0.005	3.179 ± 1.798	**0.500**	**0.471**	2.783 ± 0.075	3.172 ± 0.048	0.515	0.483	2.803 ± 0.101	3.228 ± 0.056	0.572	0.513	**2.498** ± 0.031	**3.151** ± 0.072
	672	0.784	0.627	2.674 ± 0.079	3.740 ± 0.318	**0.537**	**0.495**	3.429 ± 0.206	3.875 ± 0.380	0.571	0.519	3.427 ± 0.138	**3.439** ± 0.093	0.752	0.607	**2.553** ± 0.081	3.641 ± 0.091
WTH	24	0.357	0.404	**3.616** ±0.117	6.668 ± 1.102	0.378	0.429	3.656 ±0.082	5.609 ±0.154	0.370	0.394	3.726 ± 0.231	9.126 ± 0.322	0.378	0.423	3.711 ± 0.017	**5.428** ±0.093
	48	0.441	**0.464**	3.904 ± 0.090	**6.481** ±0.417	**0.441**	0.474	3.957 ± 0.184	7.409 ± 1.445	0.472	0.500	3.911 ± 0.282	6.730 ± 0.150	0.464	0.480	**3.663** ±0.144	6.649 ±0.586
	168	**0.549**	0.562	2.617 ±0.046	3.028 ± 0.097	0.575	0.575	2.835 ± 0.012	3.364 ± 0.045	0.561	0.565	2.712 ± 0.040	3.087 ± 0.089	0.562	**0.561**	**2.552** ± 0.031	**2.931** ± 0.068
	336	0.572	0.579	**2.565** ± 0.082	3.185 ± 0.055	0.598	0.593	2.849 ± 0.031	3.640 ± 0.078	**0.552**	**0.557**	2.710 ± 0.072	**3.087** ± 0.043	0.613	0.604	**2.516** ± 0.057	3.383 ± 0.063
	720	0.727	0.670	2.578 ± 0.027	3.617 ± 0.443	**0.591**	0.587	2.737 ± 0.009	3.228 ± 0.078	**0.571**	**0.573**	2.737 ± 0.009	**3.219** ± 0.073	0.794	0.710	**2.543** ± 0.006	3.524 ± 0.261

ity norm $\|x\|_\infty$ and *average kurtosis* over 12 decoder layers. To evaluate the prediction accuracy, we use Mean Squared Error (MSE) and Mean Absolute Error (MAE). We repeat each experiment 10 times and report the average results.

Results. In Table 2, our results demonstrate the effectiveness of `OutEffHop` in enhancing outlier efficiency of modern Hopfield network architectures. `OutEffHop` delivers significant improvements on outlier efficiency with marginal sacrifice of model performance. `OutEffHop` achieves top-tier outlier-efficiency in 25 out of 30 evaluated scenarios, ranking either first or second in these settings. In STanHop-Net, `OutEffHop` model demonstrates a notable enhancement in outlier efficiency compared to Vanilla and Sparse, Generalized Sparse Modern Hopfield Models. Specifically, there are 3% and 4% reductions in $\|x\|_\infty$ and average kurtosis, respectively.

4.3. Additional Experimental Results (Appendix D)

Figure 3 & Figure 4. To supplement Section 4.1, We conduct in-depth case studies on the BERT model. In Figure 3, we focus on the outlier performance in selected layers, and in Figure 4, we delve into the maximum infinity norm $\|\mathbf{x}\|_\infty$ within the 10th layer's various tensor components. `OutEffHop` offers evidence of its effectiveness in mitigating outliers within our approach. Additionally, we observe that this mitigation effect becomes particularly pronounced in the final several layers. See Appendix D.1 for more details.

Verifying Theoretical Results. Following (Hu et al., 2023; Wu et al., 2024b; Ramsauer et al., 2020), we validate the superiority of `OutEffHop`'s theoretical results on memory retrieval and MIL learning tasks on 3 datasets, benchmarking against (Krotov and Hopfield, 2016; Ramsauer et al., 2020; Hu et al., 2023; Wu et al., 2024b). See Appendix D.2 for more details.

5. Conclusion and Discussion

We present the Outlier-Efficient Modern Hopfield Model to manage the computational challenges posed by outliers in large transformer-based models. Our model not only inherits the appealing features of modern Hopfield models, but also introducing the `OutEffHop` layers as new deep learning components for large transformer-based models with strong outlier-reducing capabilities. Empirically, `OutEffHop` achieves an average reduction of $\sim$22+% in average kurtosis and $\sim$26+% in maximum infinity norm across four test models. Additionally, it improves the same metrics by an average of 3% and 4% compared to 3 variants of STanHop-Net and ranks among the top two in outlier efficiency in 25 out of 30 settings.

Limitation and Future Work. One limitation is that `OutEffHop` does not address outliers induced by LayerNorm (see First Residual LayerNorm in Figure 4). In fact, Wei et al. (2022) observe that LayerNorm outliers arise from mechanisms different from those of attention, as studied here. We plan to integrate these different types of outliers with `OutEffHop` in future research.

Impact Statement

We believe this methodology offers an opportunity to enhance the foundations of foundation models, including large language models, through insights from associative memory models. However, this approach could intensify biases in the training data, potentially resulting in unfair or discriminatory outcomes for underrepresented groups.

Acknowledgments

JH would like to thank Shang Wu, Yen-Ju Lu, Jing Liu, Jesus Villalba, Dino Feng and Andrew Chen for enlightening discussions, the Red Maple Family for support, and Jiayi Wang for facilitating experimental deployments. The authors would also like to thank the anonymous reviewers and program chairs for their constructive comments.

JH is partially supported by the Walter P. Murphy Fellowship. HL is partially supported by NIH R01LM1372201, NSF CAREER1841569, DOE DE-AC02-07CH11359, DOE LAB 20-2261 and a NSF TRIPODS1740735. This research was supported in part through the computational resources and staff contributions provided for the Quest high performance computing facility at Northwestern University which is jointly supported by the Office of the Provost, the Office for Research, and Northwestern University Information Technology. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

References

Josh Alman and Zhao Song. Fast attention requires bounded entries. In *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS)*, 2023. URL https://openreview.net/forum?id=KOVWXcrFIK.

Josh Alman and Zhao Song. The fine-grained complexity of gradient computation for training large language models. *arXiv preprint arXiv:2402.04497*, 2024a.

Josh Alman and Zhao Song. How to capture higher-order correlations? generalizing matrix softmax attention to kronecker computation. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024b. URL https://openreview.net/forum?id=v0zNCwwkaV.

Andreas Auer, Martin Gauch, Daniel Klotz, and Sepp Hochreiter. Conformal prediction for time series with modern hopfield networks. *Advances in Neural Information Processing Systems*, 36, 2023. URL https://arxiv.org/abs/2303.12783.

Aishwarya Bhandare, Vamsi Sripathi, Deepthi Karkada, Vivek Menon, Sun Choi, Kushal Datta, and Vikram Saletore. Efficient 8-bit quantization of transformer neural machine language translation model. *arXiv preprint arXiv:1906.00532*, 2019. URL https://arxiv.org/abs/1906.00532.

Alberto Bietti, Vivien Cabannes, Diane Bouchacourt, Herve Jegou, and Leon Bottou. Birth of a transformer: A memory viewpoint. *Advances in Neural Information Processing Systems (NeurIPS)*, 36, 2023. URL https://arxiv.org/abs/2306.00802.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. URL https://arxiv.org/abs/2108.07258.

Yelysei Bondarenko, Markus Nagel, and Tijmen Blankevoort. Understanding and overcoming the challenges of efficient transformer quantization, 2021. URL https://arxiv.org/abs/2109.12948.

Yelysei Bondarenko, Markus Nagel, and Tijmen Blankevoort. Quantizable transformers: Removing outliers by helping attention heads do nothing. *arXiv preprint arXiv:2306.12929*, 2023. URL https://arxiv.org/abs/2306.12929.

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

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. URL https://arxiv.org/abs/2005.14165.

Thomas F Burns. Semantically-correlated memories in a dense associative model. In *Forty-first International Conference on Machine Learning (ICML)*, 2024. URL https://arxiv.org/abs/2404.07123.

Thomas F Burns and Tomoki Fukai. Simplicial hopfield networks. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=_QLsH8gatwx.

Vivien Cabannes, Elvis Dohmatob, and Alberto Bietti. Scaling laws for associative memories. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=Tzh6xAJSll.

Hamza Chaudhry, Jacob Zavatone-Veth, Dmitry Krotov, and Cengiz Pehlevan. Long sequence hopfield memory. *Advances in Neural Information Processing Systems (NeurIPS)*, 36, 2023. URL https://arxiv.org/abs/2306.04532.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. revealt does BERT look at? an analysis of BERT's attention. In Tal Linzen, Grzegorz Chrupała, Yonatan Belinkov, and Dieuwke Hupkes, editors, *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages

276–286, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4828. URL https://aclanthology.org/W19-4828.

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. URL https://arxiv.org/abs/1702.01929.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. Gpt3.int8(): 8-bit matrix multiplication for transformers at scale. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 30318–30332. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/c3ba4962c05c49636d4c6206a97e9c8a-Paper-Conference.pdf.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. URL https://aclanthology.org/N19-1423.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. URL https://arxiv.org/abs/2010.11929.

Richard M Dudley. Central limit theorems for empirical measures. *The Annals of Probability*, pages 899–929, 1978. URL https://projecteuclid.org/journals/annals-of-probability/volume-6/issue-6/Central-Limit-Theorems-for-Empirical-Measures/10.1214/aop/1176995384.full.

Benjamin L Edelman, Surbhi Goel, Sham Kakade, and Cyril Zhang. Inductive biases and variable creation in self-attention mechanisms. In *International Conference on Machine Learning*, pages 5793–5831. PMLR, 2022. URL https://arxiv.org/abs/2110.10090.

Luciano Floridi and Massimo Chiriatti. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30:681–694, 2020. URL https://link.springer.com/article/10.1007/s11023-020-09548-1.

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. URL https://arxiv.org/abs/2110.11316.

Yeqi Gao, Zhao Song, Weixin Wang, and Junze Yin. A fast optimization view: Reformulating single layer attention in llm based on tensor and svm trick, and solving it in matrix multiplication time. *arXiv preprint arXiv:2309.07418*, 2023.

Mandy Guo, Zihang Dai, Denny Vrandečić, and Rami Al-Rfou. Wiki-40b: Multilingual language model dataset. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2440–2452, 2020. URL https://aclanthology.org/2020.lrec-1.297/.

Claus Hofmann, Simon Schmid, Bernhard Lehner, Daniel Klotz, and Sepp Hochreiter. Energy-based hopfield boosting for out-of-distribution detection. *arXiv preprint arXiv:2405.08766*, 2024.

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. URL https://arxiv.org/abs/2302.07253.

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. URL https://www.pnas.org/doi/10.1073/pnas.79.8.2554?trk=public_post_comment-text.

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. URL https://www.pnas.org/doi/10.1073/pnas.81.10.3088.

Mark Horowitz. 1.1 computing's energy problem (and what we can do about it). In *2014 IEEE international solid-state circuits conference digest of technical papers (ISSCC)*, pages 10–14. IEEE, 2014. URL https://ieeexplore.ieee.org/document/6757323.

Jerry Yao-Chieh Hu, Donglin Yang, Dennis Wu, Chenwei Xu, Bo-Yu Chen, and Han Liu. On sparse modern hopfield model. In *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS)*, 2023. URL https://arxiv.org/abs/2309.12673.

Jerry Yao-Chieh Hu, Bo-Yu Chen, Dennis Wu, Feng Ruan, and Han Liu. Nonparametric modern hopfield models.

arXiv preprint arXiv:2404.03900, 2024a. URL https://arxiv.org/abs/2404.03900.

Jerry Yao-Chieh Hu, Thomas Lin, Zhao Song, and Han Liu. On computational limits of modern hopfield models: A fine-grained complexity analysis. In *Forty-first International Conference on Machine Learning (ICML)*, 2024b. URL https://arxiv.org/abs/2402.04520.

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. URL https://academic.oup.com/bioinformatics/article/37/15/2112/6128680.

johnowhitaker. Blog post: Exploring softmax1, or "community research for the win!", 2023. URL https://datasciencecastnet.home.blog/2023/08/04/exploring-softmax1-or-community-research-for-the-win/. Accessed: August 4, 2023.

Marcin Junczys-Dowmunt, Kenneth Heafield, Hieu Hoang, Roman Grundkiewicz, and Anthony Aue. Marian: Cost-effective high-quality neural machine translation in c++. *arXiv preprint arXiv:1805.12096*, 2018. URL https://arxiv.org/abs/1805.12096.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. Attention is not only a weight: Analyzing transformers with vector norms, 2020. URL https://aclanthology.org/2020.emnlp-main.574/.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. Revealing the dark secrets of bert, 2019. URL https://arxiv.org/abs/1908.08593.

Leo Kozachkov, Ksenia V Kastanenka, and Dmitry Krotov. Building transformers from neurons and astrocytes. *bioRxiv*, pages 2022–10, 2022. URL https://www.pnas.org/doi/10.1073/pnas.2219150120.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. URL https://www.cs.utoronto.ca/~kriz/learning-features-2009-TR.pdf.

Dmitry Krotov and John J. Hopfield. Dense associative memory for pattern recognition. *CoRR*, 2016. URL https://arxiv.org/abs/1606.01164.

Dmitry Krotov and John J. Hopfield. Large associative memory problem in neurobiology and machine learning. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2008.06996.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. URL https://ieeexplore.ieee.org/document/726791.

Percy Liang. Cs229t/stat231: Statistical learning theory (winter 2016), 2016. URL https://web.stanford.edu/class/cs229t/notes.pdf.

M. Marchesi, G. Orlandi, F. Piazza, and A. Uncini. Fast neural networks without multipliers. *IEEE Transactions on Neural Networks*, 4(1):53–62, 1993. doi: 10.1109/72.182695. URL https://ieeexplore.ieee.org/document/182695.

Evan Miller. Blog post: Attention is off by one, 2023. URL https://www.evanmiller.org/attention-is-off-by-one.html. Accessed: August 4, 2023.

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. URL https://www.amazon.com/Handbook-Mathematical-Functions-Hardback-CD-ROM/dp/0521192250.

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. URL https://arxiv.org/abs/2205.12258.

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. URL https://arxiv.org/abs/2008.02217.

Alex Reneau, Jerry Yao-Chieh Hu, Chenwei Xu, Weijian Li, Ammar Gilani, and Han Liu. Feature programming for multivariate time series prediction. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pages 29009–29029. PMLR, 23–29 Jul 2023. URL https://arxiv.org/abs/2306.06252.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y. URL https://arxiv.org/abs/1409.0575.

Johannes Schimunek, Philipp Seidl, Lukas Friedrich, Daniel Kuhn, Friedrich Rippmann, Sepp Hochreiter, and Günter Klambauer. Context-enriched molecule representations improve few-shot drug discovery. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=XrMWUuEevr.

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. URL https://pubs.acs.org/doi/10.1021/acs.jcim.1c01065.

Moran Shkolnik, Brian Chmiel, Ron Banner, Gil Shomron, Yury Nahshan, Alex Bronstein, and Uri Weiser. Robust quantization: One model to rule them all, 2020. URL https://arxiv.org/abs/2002.07686.

Bharath K Sriperumbudur and Gert RG Lanckriet. On the convergence of the concave-convex procedure. In *Advances in neural information processing systems*, volume 9, pages 1759–1767, 2009. URL https://papers.nips.cc/paper_files/paper/2009/file/8b5040a8a5baf3e0e67386c2e3a9b903-Paper.pdf.

C.Z. Tang and H.K. Kwan. Multilayer feedforward neural networks with single powers-of-two weights. *IEEE Transactions on Signal Processing*, 41(8):2724–2727, 1993. doi: 10.1109/78.229903. URL https://ieeexplore.ieee.org/document/229903.

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. URL https://arxiv.org/abs/1706.03762.

Xiuying Wei, Yunchen Zhang, Xiangguo Zhang, Ruihao Gong, Shanghang Zhang, Qi Zhang, Fengwei Yu, and Xianglong Liu. Outlier suppression: Pushing the limit of low-bit transformer language models. *Advances in Neural Information Processing Systems*, 35:17402–17414, 2022. URL https://arxiv.org/abs/2209.13325.

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. URL https://arxiv.org/abs/2007.13505.

Dennis Wu, Jerry Yao-Chieh Hu, Teng-Yun Hsiao, and Han Liu. Uniform memory retrieval with larger capacity for modern hopfield models. In *Forty-first International Conference on Machine Learning (ICML)*, 2024a. URL https://arxiv.org/abs/2404.03827.

Dennis Wu, Jerry Yao-Chieh Hu, Weijian Li, Bo-Yu Chen, and Han Liu. STanhop: Sparse tandem hopfield model for memory-enhanced time series prediction. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024b. URL https://arxiv.org/abs/2312.17346.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*, 2023. URL https://arxiv.org/abs/2303.17564.

Chenwei Xu, Yu-Chao Huang, Jerry Yao-Chieh Hu, Weijian Li, Ammar Gilani, Hsi-Sheng Goan, and Han Liu. Bishop: Bi-directional cellular learning for tabular data with generalized sparse modern hopfield model. In *Forty-first International Conference on Machine Learning (ICML)*, 2024. URL https://arxiv.org/abs/2404.03830.

A. L. Yuille and Anand Rangarajan. The Concave-Convex Procedure. *Neural Computation*, 15(4):915–936, 04 2003. URL https://doi.org/10.1162/08997660360581958.

Ofir Zafrir, Guy Boudoukh, Peter Izsak, and Moshe Wasserblat. Q8bert: Quantized 8bit bert. In *2019 Fifth Workshop on Energy Efficient Machine Learning and Cognitive Computing-NeurIPS Edition (EMC2-NIPS)*, pages 36–39. IEEE, 2019. URL https://arxiv.org/abs/1910.06188.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022. URL https://arxiv.org/abs/2205.01068.

Tong Zhang. *Mathematical analysis of machine learning algorithms*. Cambridge University Press, 2023. URL https://tongzhang-ml.org/lt-book/lt-book.pdf.

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, 2021. URL https://arxiv.org/abs/2012.07436.

Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana Davuluri, and Han Liu. Dnabert-2: Efficient foundation model and benchmark for multi-species genome. *arXiv preprint arXiv:2306.15006*, 2023. URL https://arxiv.org/abs/2306.15006.

Zhihan Zhou, Weimin Wu, Harrison Ho, Jiayi Wang, Lizhen Shi, Ramana V Davuluri, Zhong Wang, and Han Liu. Dnabert-s: Learning species-aware dna embedding with genome foundation models. *ArXiv*, 2024. URL https://arxiv.org/abs/2402.08777.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015. URL https://arxiv.org/abs/1506.06724.

Supplementary Material

A. Related Works

Associative Memory Models for Deep Learning. The classical Hopfield models (Hopfield, 1984; 1982; Krotov and Hopfield, 2016) mirror the associative memory of the human brain, focusing on the storage and retrieval of specific memory patterns. Recently, a resurgence in associative memory model research is attributable to (i) advancements in memory storage capacities (Wu et al., 2024a; Chaudhry et al., 2023; Krotov and Hopfield, 2016; Demircigil et al., 2017), (ii) the innovative architectural designs (Wu et al., 2024b; Hoover et al., 2023; Seidl et al., 2022; Fürst et al., 2022; Ramsauer et al., 2020), and (iii) their biological plausibility (Burns, 2024; Kozachkov et al., 2022; Krotov and Hopfield, 2021). Notably, the associative memory networks a.k.a. the modern Hopfield models (Hu et al., 2024a;b; Wu et al., 2024b; Burns and Fukai, 2023; Hu et al., 2023; Brandstetter, 2021; Ramsauer et al., 2020) exhibit favorable properties, including fast convergence speed and exponential memory capacity. They form a bridge to Transformer architecture (Hu et al., 2024a; 2023; Wu et al., 2024b; Cabannes et al., 2024; Bietti et al., 2023; Ramsauer et al., 2020), positioning themselves as advanced extensions of attention mechanisms. Consequently, their applicability extends across various fields, including drug discovery (Schimunek et al., 2023), immunology (Widrich et al., 2020), time series forecasting (Wu et al., 2024b; Auer et al., 2023), tabular learning (Xu et al., 2024), out-of-distribution detection (Hofmann et al., 2024), reinforcement learning (Paischer et al., 2022), and vision (Fürst et al., 2022). Our study refines this research direction towards efficient models. We believe that this study is critical in guiding future research towards a Hopfield-driven design paradigm, especially for large-scale models.

Outlier-Efficient Methods. Quantization is a method to reduce the computational burden of large models via low-bit precision computing (Horowitz, 2014; Tang and Kwan, 1993; Marchesi et al., 1993). For instance, common quantization schemes, INT8 and INT4, compress the models' weights and activations by using 8-bit or 4-bit integers encoding (Zafrir et al., 2019; Bhandare et al., 2019; Junczys-Dowmunt et al., 2018). However, the presence of outliers challenges the quantization performance of transformer-based models due to outlier-induced exploding attention weights (Bondarenko et al., 2023; 2021). To combat this, Wei et al. (2022) modify LayerNorm to enable quantization on outlier-free activation tensors, and introduce Token-Wise Clipping to optimize clipping ranges for each token. Further, Dettmers et al. (2022) quantize outlier features and other features with different degrees of precision. Yet, since outliers stem from the softmax function (see Section 2.1 for details), neither of above methods address the outlier issue from its source. To this end, Bondarenko et al. (2023) introduce `Clipped_Softmax` and `Gated_Attention` to force attention mechanism to output exact zeros, thereby tackling the source of outliers. Specifically, `Clipped_Softmax` extends the output range of the softmax function from (0, 1) to larger span, and `Gated_Attention` determines to keep or nullify the update. However, these two methods need hyperparameters for optimal performance. Moreover, `Clipped_Softmax` underperforms with the OPT model and `Gated_Attention` introduces additional training parameters. In this paper, we present a novel modern Hopfield model such that it endows outlier-efficient computation. Surprisingly, its retrieval dynamics subsumes the $\mathrm{Softmax}_1$ outlier-efficient attention (Miller, 2023) as a special case[10]. We expect this work to shed light on research into (Hopfield-based) large foundation models, both theoretically and methodologically.

Outlier Related Transformer Theories. Recent works highlight the theoretical advantages of removing outliers in the attention heads of transformer-based large foundation models. Alman and Song (2023) show that efficient transformers (both vanilla and tensor (Alman and Song, 2024b)) require bounded attention weights using fine-grained reduction. Hu et al. (2024b) show that efficient modern Hopfield models and corresponding networks also need bounded query and key patterns for sub-quadratic time complexity via fine-grained reduction. Alman and Song (2024a); Gao et al. (2023) show that efficient training of transformer-based models necessitates bounded weight matrices.

[10] For any $x \in \mathbb{R}^d$, $\mathrm{Softmax}_1(x)_i = \frac{\exp\{x_i\}}{1+\sum_j \exp\{x_j\}}$. Preliminary experimental results (johnowhitaker, 2023) confirm its outlier efficiency.

B. Supplementary Backgrounds

B.1. How Does $\mathrm{Softmax}_1$ Solve the Outlier Problem?

The "outlier" challenge in (multi-head) attention arises from the inherent design of $\mathrm{Softmax}$. $\mathrm{Softmax}$ forces each attention head to attend to at least one position in the input sequence, even if there is no useful information throughout the entire input sequence. In the case where a **no-update** behavior is needed, since for $\mathrm{Softmax}$, producing close-to-zero probabilities for all positions is not an option, it has to produce a high probability to a spurious position (such as a comma sign in a sentence) and produces close-to-zero probabilities for all other positions. However, this is a workaround and it still introduces noises.

To solve this, Miller (2023) proposes $\mathrm{Softmax}_1$, which adds 1 to the denominator of $\mathrm{Softmax}$. This adjustment reduces the relative importance of each head. As a result, if a head provides less relevant or even misleading information, the model does not depend on it. This ensures the influence of "less-relevant heads" remains moderate. Therefore, $\mathrm{Softmax}_1$ allows a head to "abstain" or contribute minimally when its information is not beneficial for the current context.

In this paper, we introduce the Outlier-Efficient modern Hopfield model for two purposes:

I. Outlier Efficient Associative Memory Model

II. Outlier Efficient Attention-like Layer for deep learning

We only have to identify outlier when our model serves as Associative Memory Models . For similarity measure thersholding, it has following process:

1. Calculating similarity scores among patterns.
2. Setting a threshold, patterns with scores below this threshold are considered dissimilar.
3. Patterns with consistently low similarity scores across the board are identified as "no-op" outliers.

As for ad-hoc assignment, we create a provisional classification system to identify "no-op" outliers. This temporary framework allows us to segregate data that does not fit into predefined categories.

For Outlier Efficient model implement as attention-like layer for deep learning, the similarity measurement is automatically done by learning. Thus, it identifies outliers without extra effort. Patterns with small inner products with queries get almost zero attention probability, because of our retrieval dynamic design (2.6).

Explicitly, let $z := (z_1, ..., z_M) \in \mathbb{R}^M$. By (2.5) and (2.6), $\mathrm{Softmax}_1(z)$ automatically assigns ~ 0 output to $z_i \sim 0$ for all $i \in [M]$ without requiring other $z_{j \neq i}$ to be super huge, by associating them to zero-point energy state (no-op memories). Here z is learned according to (2.7) when OutEffHop is used as a learning layer. Hence, it's clear the outlier identification is done automatically through learning.

Consider an example involving a negligible input vector in the attention mechanism: n = [-10, -10, -10]. Upon passing n through the $\mathrm{Softmax}$ function, it yields relatively large weights:

$$\mathrm{Softmax}(n) \approx [0.33,\ 0.33,\ 0.33].$$

To achieve a **no-update**, the attention mechanism allocates increasing attention to low-information tokens, causing the probability of other tokens to approach zero (See Section 2.1 for details). For instance, if the first element in n represents a low-information token (e.g., [SEP]), the input vector might transform into

$$n' = [100, -10, -10].$$

This transformation causes the weights of all but the first token to converge to zero:

$$\mathrm{Softmax}(n') \approx [0.99,\ 2 \times 10^{-48},\ 2 \times 10^{-48}].$$

This procedure requires the wide range of input vector, leading the emergence of outliers. However, when n is processed by $\mathrm{Softmax}_1$, the result is as follows:

$$\mathrm{Softmax}_1(n) \approx [5 \times 10^{-5}, 5 \times 10^{-5}, 5 \times 10^{-5}]$$

In this case, all vector values diminish to a level close to zero. Consequently, the attention head does not need to assign a higher probability mass to specific tokens, resulting in a reduction in the memory space for the vector. Therefore, by construction, $\mathrm{Softmax}_1$ is outlier-robust.

C. Proofs of Main Text

C.1. Lemma 2.1

Proof of Lemma 2.1. To show monotonic decreasing property of the energy (2.4), we first derive the outlier-efficient retrieval dynamics by utilizing the convex-concave procedure (Yuille and Rangarajan, 2003) (CCCP). The total energy $\mathcal{H}(x)$ is split into convex term $\mathcal{H}_1 \coloneqq \frac{1}{2}\langle x, x\rangle$ and concave term $\mathcal{H}_2 \coloneqq -\mathrm{lse}_1(\beta, \Xi^\mathsf{T} x)$. In addition, $\mathcal{H}_1$ and $\mathcal{H}_2$ are both differentiable by definition. Every iteration of CCCP applied on $\mathcal{H}$ gives:

$$\underbrace{\nabla_x \mathcal{H}_1(x_{t+1})}_{=\frac{1}{2}\nabla_x\langle x_{t+1}, x_{t+1}\rangle} = -\nabla_x \underbrace{\mathcal{H}_2(x_t)}_{=-\mathrm{lse}_1(\beta, \Xi^\mathsf{T} x_t)},$$

such that

$$x_{t+1} = \nabla_x \mathrm{lse}_1(\beta, \Xi^\mathsf{T} x_t).$$

To derive the gradient of $\mathrm{lse}_1(\beta, \Xi^\mathsf{T} x_t)$, we set $\tau(\beta z_l) \coloneqq \sum_l^N \exp(\beta z_l)$.

Then,

$$\begin{aligned}
\nabla_x \mathrm{lse}_1\left(\beta, \Xi^\mathsf{T} x_t\right)|_{x_t} &= \nabla_{x_t}\left(\beta^{-1}\log\{\tau(\beta\Xi^\mathsf{T} x_t)+1\}\right) \\
&= \beta^{-1}\nabla_\tau \log(\tau+1)\cdot\nabla_{x_t}\tau(\beta\Xi^\mathsf{T} x_t) \\
&= \frac{1}{\tau(\beta\Xi^\mathsf{T} x_t)+1}\cdot\exp\left(\beta\Xi^\mathsf{T} x_t\right)\cdot\Xi^\mathsf{T} \\
&= \Xi\cdot\frac{\exp\left(\beta\Xi^\mathsf{T} x_t\right)}{1+\sum\exp\left(\beta\Xi^\mathsf{T} x_t\right)} \\
&= \Xi\cdot\mathrm{Softmax}_1\left(\beta\Xi^\mathsf{T} x_t\right).
\end{aligned}$$

Hence, we obtain

$$x_{t+1} = \nabla_x \mathrm{lse}_1(\beta\Xi^\mathsf{T} x_t) = \Xi\cdot\mathrm{Softmax}_1\left(\beta\Xi^\mathsf{T} x_t\right)$$

Due to the concave design of $\mathcal{H}_2$, we demonstrate that $\mathcal{H}$ can be monotonically decreased by $\mathcal{T}_{\text{OutEff}}(x)$ given by (2.6), following the proof in (Hu et al., 2023, Appendix E.2). □

C.2. Theorem 3.1

With the monotonic decreasing property from Lemma 2.1, we prove Theorem 3.1 following the same strategy as (Wu et al., 2024b, Lemma 3.3) and (Hu et al., 2023, Lemma 2.2).

C.3. Theorem 3.2

Proof of Theorem 3.2. Let $\mathcal{T}_{\text{original}}$ be the retrieval dynamics of the original modern Hopfield model (Ramsauer et al., 2020), and $\|\mathcal{T}_{\text{OutEff}}(x) - \xi_\mu\|$ and $\|\mathcal{T}_{\text{original}}(x) - \xi_\mu\|$ be the retrieval error of outlier-efficient and modern Hopfield model, respectively.

To prove $\|\mathcal{T}_{\text{OutEff}}(x) - \xi_\mu\|$ has tighter upper bound than $\|\mathcal{T}_{\text{original}}(x) - \xi_\mu\|$, we recall the upper bound on $[\mathrm{Softmax}(\beta\Xi^\mathsf{T} x)]_\nu$ from (Wu et al., 2024b, Eqaution C.37):

$$[\mathrm{Softmax}(\beta\Xi^\mathsf{T} x)]_\nu \leq \exp\left\{-\beta\widetilde{\Delta}_\mu\right\},$$

where $\widetilde{\Delta}_\mu \coloneqq \langle\xi_\mu, x\rangle - \mathrm{Max}_{\mu,\nu\in[M];\mu\neq\nu}\langle\xi_\mu, \xi_\nu\rangle$.

Since we observe the relation

$$\left[\text{Softmax}_1\left(\beta\Xi^\mathsf{T}x\right)\right]_\nu = \left(\sum_{\mu=1}^{M}\left[\text{Softmax}_1\left(\beta\Xi^\mathsf{T}x\right)\right]_\mu\right)\left[\text{Softmax}\left(\beta\Xi^\mathsf{T}x\right)\right]_\nu,$$

it holds

$$[\text{Softmax}_1(\beta\Xi^\mathsf{T}x)]_\nu \leq \exp\left\{-\beta\widetilde{\Delta}_\mu\right\}\cdot\gamma,$$

where $\gamma \coloneqq \sum_{\mu=1}^{M}\left[\text{Softmax}_1\left(\beta\Xi^\mathsf{T}x\right)\right]_\mu$. Note that $0<\gamma<1$.

For any $\beta>0$, there exist a $\delta>0$ such that $\gamma \coloneqq \exp\{-\beta\delta\}$.

Also, recall the bound of $\|\mathcal{T}_{\text{original}}-\xi_\mu\|$ from (Wu et al., 2024b, Equation C.41) :

$$\|\mathcal{T}_{\text{original}}-\xi_\mu\| \leq 2m(M-1)\exp\left\{-\beta\widetilde{\Delta}_\mu\right\} \leq 2m(M-1)\exp\{-\beta\left(\Delta_\mu-2mR\right)\}. \tag{C.1}$$

By (Wu et al., 2024b, Equation C.39), we know $\widetilde{\Delta}_\mu \geq \Delta_\mu - 2mR$ and R is the radius of the sphere S_μ. Then we have

$$\|\mathcal{T}_{\text{OutEff}}-\xi_\mu\| \leq 2m(M-1)\exp\{-\beta\left(\Delta_\mu-2mR+\delta\right)\}.$$

Comparing above with (C.1), this complete the proof. □

C.4. Corollary 3.2.1

Proof of Corollary 3.2.1. We aim to establish the validity of the following inequality:

$$\|\mathcal{T}_{\text{OutEff}}(x)-\xi_\mu\| \leq \|\mathcal{T}_{\text{original}}(x)-\xi_\mu\|.$$

It is equivalent to consider

$$\|\mathcal{T}_{\text{OutEff}}(x)-\xi_\mu\|^2 - \|\mathcal{T}_{\text{original}}(x)-\xi_\mu\|^2 \leq 0. \tag{C.2}$$

That is,

$$\left\|\sum_{\nu=1}^{M}\xi_\nu\left[\text{Softmax}_1\left(\beta\Xi^\mathsf{T}x\right)\right]_\nu-\xi_\mu\right\|^2 - \left\|\sum_{\nu=1}^{M}\xi_\nu\left[\text{Softmax}\left(\beta\Xi^\mathsf{T}x\right)\right]_\nu-\xi_\mu\right\|^2 \leq 0. \tag{C.3}$$

Let $\gamma \coloneqq \sum_{\mu=1}^{M}\left[\text{Softmax}_1\left(\beta\Xi^\mathsf{T}x\right)\right]_\mu$ $(0<\gamma<1)$.

For ease of presentation, we set $v_1 \coloneqq \sum_{\nu=1}^{M}\xi_\nu\left[\text{Softmax}\left(\beta\Xi^\mathsf{T}x\right)\right]_\nu$, $v_2 \coloneqq \sum_{\nu=1}^{M}\xi_\nu\left[\text{Softmax}_1\left(\beta\Xi^\mathsf{T}x\right)\right]_\nu = \gamma v_1$ and $w \coloneqq \xi_\mu$.

(C.3) becomes

$$\|v_2-w\|^2 - \|v_1-w\|^2 \leq 0,$$

expanding both terms

$$v_2^2 - 2v_2\cdot w + w^2 - v_1^2 + 2v_1\cdot w - w^2 \leq 0,$$

simplifying the expression

$$(\gamma^2-1)v_1^2 - 2(\gamma-1)v_1\cdot w \leq 0,$$

since vectors are normalized

$$(\gamma^2 - 1)\|v_1\| - 2(\gamma - 1)\|w\| \cos(\alpha) \leq 0,$$

and rearranging the terms

$$(\gamma + 1)\|v_1\| - 2\|w\| \cos(\alpha) \geq 0,$$

where $\cos(\alpha)$ quantifies the overlap between v_1 and w.

If all the memories and queries are normalized, (C.2) holds when

$$\frac{\gamma + 1}{2} \geq \cos(\alpha).$$

As memory patterns and queries exhibit smaller overlap at the beginning of the retrieval process, the proposed model experiences smaller retrieval errors than its original counterpart during the initial phase of memory retrieval. □

C.5. Theorem 3.3

Lemma C.1 (Memory Capacity Lower Bound, Formal)**.** Suppose the probability of successfully storing and retrieving memory pattern is given by $1 - p$. The number of memory patterns sampled from a sphere of radius m that the Outlier-Efficient Hopfield model can store and retrieve has a lower bound: $M \geq \sqrt{p} C^{\frac{d-1}{4}}$, where C is the solution for $C = b/W_0(\exp\{a + \ln b\})$ with $W_0(\cdot)$ being the principal branch of Lambert W function (Olver et al., 2010), $a := 4/(d-1)\left\{\ln[2m^2(\sqrt{p}-1)/R] + 1 - \delta/(2\beta m R)\right\}$ and $b := 4m^2\beta/5(d-1)$. For all β, we have larger memory capacity lower bound compared to the original modern Hopfield model (Ramsauer et al., 2020): $M \geq M_{\text{original}}$

To prove it, we first derive the well-separation condition for the outlier-efficient modern Hopfield model.

Lemma C.2 (Modified from Lemma C.3 of (Wu et al., 2024b))**.** Let $\gamma := \sum_{\mu=1}^{M} \left[\text{Softmax}_1\left(\beta \Xi^\mathsf{T} x\right)\right]_\mu$ and $1 > \gamma > 0$. For any $\beta > 0$, there exist a $\delta > 0$ such that $\gamma := \exp\{-\beta\delta\}$. Then, the well-seperation condition can be formulated as:

$$\Delta_\mu \geq \frac{1}{\beta} \ln\left(\frac{2(M-1)m}{R}\right) + 2mR - \delta.$$

Proof. From Appendix C.1 we obtain the result

$$\|\mathcal{T}_{\text{OutEff}} - \xi_\mu\| \leq 2m(M-1)\exp\{-\beta\left(\Delta_\mu - 2mR + \delta\right)\}$$

Therefore, for $\mathcal{T}_{\text{OutEff}}$ to be mapping $\mathcal{T}_{\text{OutEff}} : S_\mu \to S_\mu$, it is sufficient to obtain

$$2(M-1)\exp\{-\beta\left(\Delta_\mu - 2mR + \delta\right)\}m \leq R.$$

This leads to the separation condition for the proposed Outlier-Efficient Modern Hopfield Model

$$\Delta_\mu \geq \frac{1}{\beta} \ln\left(\frac{2(M-1)m}{R}\right) + 2mR - \delta. \tag{C.4}$$

Given that (C.4) possesses a stricter lower bound compared to its original counterpart (Ramsauer et al., 2020, Equation (300)),we complete the proof following the similar approach in (Wu et al., 2024b, Lemma 3.4). □

Lemma C.3. [(Ramsauer et al., 2020)] If the identity

$$ac + c\ln c - b = 0,$$

holds for all real numbers $a,\ b \in \mathbb{R}$, then c takes a solution:

$$c = \frac{b}{W_0(\exp(a + \ln b))}.$$

Proof. By looking at the proof in (Wu et al., 2024b). □

Then we start our formal proof of Theorem 3.3.

Proof. Since $\Delta_{\text{min}} = \min_{1 \le \mu \le M} \Delta_\mu$, we get

$$\Delta_{\text{min}} \ge \frac{1}{\beta} \ln \left(\frac{2(M-1)m}{R} \right) + 2mR - \delta.$$

Following the proof in ((Wu et al., 2024b), Appendix Theorem A5), we obtain

$$a \coloneqq \frac{4}{(d-1)} \left\{ \ln \left[\frac{2m^2(\sqrt{p}-1)}{R} \right] + 1 - \frac{\delta}{(2\beta m R)} \right\}, \quad b \coloneqq \frac{4m^2\beta}{5(d-1)}.$$

By Lemma C.3, C can be expressed as

$$C = \frac{b}{W(\exp\{a + \ln b\})}.$$

We expressed the original counterpart of a and b as

$$\widetilde{a} \coloneqq \frac{4}{(d-1)} \left\{ \ln \left[\frac{2m^2(\sqrt{p}-1)}{R} \right] + 1 \right\}, \ \widetilde{b} = b.$$

Since

$$\widetilde{a} > a$$

and

$$\widetilde{C} = \frac{b}{W(\exp\{\widetilde{a} + \ln b\})} < \frac{b}{W(\exp\{a + \ln b\})} = C,$$

we arrive at

$$M_{\text{original}} = \sqrt{p}\widetilde{C}^{\frac{d-1}{4}} < \sqrt{p}C^{\frac{d-1}{4}} = M.$$

This completes the proof. □

C.6. Theorem 3.4

To bound the generalization error of Outlier-Efficient Modern Hopfield, we utilize the generalization bound via covering number (Dudley, 1978; Edelman et al., 2022; Liang, 2016; Zhang, 2023).

Definition C.1 (Covering Number)**.** For a given class of vector-valued functions $\mathcal{F}$, the covering number $\mathcal{N}_\infty\left(\mathcal{F};\varepsilon;\left\{z^{(i)}\right\}_{i=1}^m;\|\cdot\|\right)$ is the smallest size of a collection (a cover) $\mathcal{C}\subset\mathcal{F}$ such that $\forall f\in\mathcal{F},\exists\widehat{f}\in\mathcal{C}$ satisfying

$$\max_i\left\|f\left(z^{(i)}\right)-\widehat{f}\left(z^{(i)}\right)\right\|\leq\varepsilon.$$

Also, define

$$\mathcal{N}_\infty(\mathcal{F},\varepsilon,m,\|\cdot\|)\quad:=\sup_{z^{(1)}\ldots z^{(m)}}\mathcal{N}_\infty\left(\mathcal{F};\varepsilon;z^{(1)},\ldots,z^{(m)},\|\cdot\|\right).$$

Lemma C.4 (Generalization Bound via Covering Number (Liang, 2016; Zhang, 2023))**.** Suppose $\mathcal{F}$ is a class of bounded functions, and $\log\mathcal{N}_\infty\left(\mathcal{F};\varepsilon;x^{(1)},\ldots,x^{(m)}\right)\leq C_\mathcal{F}/\varepsilon^2$ for all $x^{(1)},\ldots,x^{(m)}\in\mathcal{X}^m$. Then for any $\delta>0$, with probability at least $1-\delta$, simultaneously for all $f\in\mathcal{F}$, the generalization error ε_{gen} satisfies

$$\varepsilon_{\text{gen}}(f)\leq\widetilde{O}\left(\sqrt{\frac{C_\mathcal{F}}{m}}+\sqrt{\frac{\log(1/\delta)}{m}}\right).$$

We start by proving the Lipschitzness of Softmax_1. We first introduce Lemma C.5 from (Edelman et al., 2022).

Lemma C.5 (Lemma A.6. of (Edelman et al., 2022))**.** Consider a function $f:\mathbb{R}^d\to\mathbb{R}^d$ such that the Jacobian $\mathcal{J}_f:=\boldsymbol{\nabla}f$ of the function satisfies $\|\mathcal{J}_f(v)\|_{1,1}\leq c_f$ for all $v\in\mathbb{R}^d$, then for any vectors $v_1,v_2\in\mathbb{R}^d$,

$$\|f(v_1)-f(v_2)\|_1\leq c_f\|v_1-v_2\|_\infty.$$

With Lemma C.5, we obtain the Lipschitzness of Softmax_1.

Lemma C.6 (Lipschitzness of Softmax_1)**.** For vectors $x_1,x_2\in\mathbb{R}^d$,

$$\|\text{Softmax}_1(x_1)-\text{Softmax}_1(x_2)\|_1\leq 2\|x_1-x_2\|_\infty.$$

Proof of Lemma C.6. We prove that Softmax_1 satisfies $\|\mathcal{J}_f(\theta)\|_{1,1}\leq c_f$, and use Lemma C.5 to obtain the Lipschitzness.

Let $\mathcal{J}_{\text{Softmax}_1}(x):=\boldsymbol{\nabla}_x\text{Softmax}_1(x)$.

For any $x\in\mathbb{R}^d$, we first denote the elements of $\mathcal{J}_{\text{Softmax}_1}(x)$ as

$$\frac{\partial\text{Softmax}_1(x)_i}{\partial x_j},\text{ for }i,j\in[d].$$

Observe that for $i=j$:

$$\frac{\partial\text{Softmax}_1(x)_i}{\partial x_j}=\text{Softmax}_1(x)_i-\text{Softmax}_1(x)_i\text{Softmax}_1(x)_j,$$

and for $i\neq j$:

$$\frac{\partial\text{Softmax}_1(x)_i}{\partial x_j}=-\text{Softmax}_1(x)_i\text{Softmax}_1(x)_j.$$

Therefore, we have

$$\begin{aligned}\|\mathcal{J}_{\text{Softmax}_1}(x)\|_{1,1} &= \left|\sum_{i,j=1}^{d} \text{Softmax}_1(x)_i \mathbb{1}(i=j) - \text{Softmax}_1(x)_i \text{Softmax}_1(x)_j\right| \\ &= \left|\sum_{i,j=1}^{d} \text{Softmax}_1(x)_i \left(\mathbb{1}(i=j) - \text{Softmax}_1(x)_j\right)\right| \\ &< 2\sum_{i}^{d} \text{Softmax}_1(x)_i \left(1 - \text{Softmax}_1(x)_i\right) \\ &\leq 2.\end{aligned}$$

Finally, by Lemma C.5, we have

$$\|\text{Softmax}_1(x_1) - \text{Softmax}_1(x_2)\|_1 \leq 2\|x_1 - x_2\|_\infty .$$

This completes the proof. □

Next, we prove the Lipschitzness of f_{hop} in parameter.

Lemma C.7 (Lipschitzness of f_{hop})**.** For any $W_K, W_K' \in \mathcal{W}_K$, $\widetilde{W}_V, \widetilde{W}_V' \in \widetilde{\mathcal{W}}_V$, $\tau \in [T]$:

$$\begin{aligned}&\left\|f_{\text{hop}}(Y, q_\tau; W_K, \widetilde{W}_V) - f_{\text{hop}}(Y, q_\tau; W_K', \widetilde{W}_V')\right\| \\ &\leq 2B_V B_Y \|\beta Y W_K q_\tau - \beta Y W_K' q_\tau\|_\infty + \left\|\left(Y\widetilde{W}_V\right)^\mathsf{T} - \left(Y\widetilde{W}_V'\right)^\mathsf{T}\right\|_{2,\infty}.\end{aligned}$$

Proof of Lemma C.7.

$$\begin{aligned}&\left\|f_{\text{hop}}(Y, q_\tau; W_K, \widetilde{W}_V) - f_{\text{hop}}(Y, q_\tau; W_K', \widetilde{W}_V')\right\| \\ &= \left\|\widetilde{W}_V^\mathsf{T} Y^\mathsf{T} \text{Softmax}_1\left(\beta Y W_K q_\tau\right) - \widetilde{W}_V'^\mathsf{T} Y^\mathsf{T} \text{Softmax}_1\left(\beta Y W_K' q_\tau\right)\right\| \\ &= \left\|\widetilde{W}_V^\mathsf{T} Y^\mathsf{T}\left(\text{Softmax}_1\left(\beta Y W_K q_\tau\right) - \text{Softmax}_1\left(\beta Y W_K' q_\tau\right)\right) + \left(\widetilde{W}_V^\mathsf{T} Y^\mathsf{T} - \widetilde{W}_V'^\mathsf{T} Y^\mathsf{T}\right) \text{Softmax}_1\left(\beta Y W_K' q_\tau\right)\right\| \\ &\leq \left\|\widetilde{W}_V^\mathsf{T} Y^\mathsf{T}\left(\text{Softmax}_1\left(\beta Y W_K q_\tau\right) - \text{Softmax}_1\left(\beta Y W_K' q_\tau\right)\right)\right\| + \left\|\left(\widetilde{W}_V^\mathsf{T} Y^\mathsf{T} - \widetilde{W}_V'^\mathsf{T} Y^\mathsf{T}\right) \text{Softmax}_1\left(\beta Y W_K' q_\tau\right)\right\| \\ &\qquad \text{(By triangle inequality)} \\ &\leq \left\|\widetilde{W}_V^\mathsf{T} Y^\mathsf{T}\right\|_{2,\infty} \left\|\text{Softmax}_1\left(\beta Y W_K q_\tau\right) - \text{Softmax}_1\left(\beta Y W_K' q_\tau\right)\right\|_1 \\ &\qquad + \left\|\widetilde{W}_V^\mathsf{T} Y^\mathsf{T} - \widetilde{W}_V'^\mathsf{T} Y^\mathsf{T}\right\|_{2,\infty} \left\|\text{Softmax}_1\left(\beta Y W_K' q_\tau\right)\right\|_1 \qquad \text{(By } \|Ax\| \leq \|A\|_{2,\infty}\|x\|_1\text{)} \\ &\leq \left\|\widetilde{W}_V^\mathsf{T}\right\|_2 \left\|Y^\mathsf{T}\right\|_{2,\infty} \left\|\text{Softmax}_1\left(\beta Y W_K q_\tau\right) - \text{Softmax}_1\left(\beta Y W_K' q_\tau\right)\right\|_1 \\ &\qquad + \left\|\widetilde{W}_V^\mathsf{T} Y^\mathsf{T} - \widetilde{W}_V'^\mathsf{T} Y^\mathsf{T}\right\|_{2,\infty} \left\|\text{Softmax}_1\left(\beta Y W_K' q_\tau\right)\right\|_1 \quad \text{(By } \|PQ\|_{2,\infty} \leq \|P\|_2 \|Q\|_{2,\infty}\text{)} \\ &\leq 2B_V B_Y \|\beta Y W_K q_\tau - \beta Y W_K' q_\tau\|_\infty + \left\|\widetilde{W}_V^\mathsf{T} Y^\mathsf{T} - \widetilde{W}_V'^\mathsf{T} Y^\mathsf{T}\right\|_{2,\infty} \left\|\text{Softmax}_1\left(\beta Y W_K' q_\tau\right)\right\|_1 \\ &\qquad \text{(By Assumption 3.1-(A4) and Lemma C.6)} \\ &\leq 2B_V B_Y \|\beta Y W_K q_\tau - \beta Y W_K' q_\tau\|_\infty + \left\|\left(Y\widetilde{W}_V\right)^\mathsf{T} - \left(Y\widetilde{W}_V'\right)^\mathsf{T}\right\|_{2,\infty}.\end{aligned}$$

□

Next, with Lemma C.7, we construct a covering number bound for a Modern Hopfield model function class using the covering number of its composing functions. We write the composing functions as $f_K : \mathbb{R}^{M\times a} \times \mathbb{R}^d \to \mathbb{R}^M$ as:

$$f_K(Y, q; W_K) = \beta Y W_K q,$$

and $f_V : \mathbb{R}^{M\times a} \to \mathbb{R}^{d\times M}$ as:

$$f_V(Y; \widetilde{W}_V) = \left(Y\widetilde{W}_V\right)^\mathsf{T}.$$

With f_k and f_V, we prove that the covering number of $\mathcal{F}_{\text{hop}}$ is bounded as below.

Lemma C.8. Under the Assumption 3.1, for any $\alpha \in [0, 1]$ the covering number of $\mathcal{F}_{\text{hop}}$ satisfies

$$\begin{aligned}
&\log \mathcal{N}_\infty \left(\mathcal{F}_{\text{hop}}\,; \varepsilon; \left\{\left(Y^{(i)}, q_\tau^{(i)}\right)\right\}_{\tau\in[T]}^{i\in[N],}; \|\cdot\|_2\right) \\
&\le \log \mathcal{N}_\infty \left(\mathcal{F}_K; \varepsilon_K; \left\{\left(y_t^{(i)}, q_\tau^{(i)}\right)\right\}_{t\in[M],\tau\in[T]}^{i\in[N]}\right) + \log \mathcal{N}_\infty \left(\mathcal{F}_V; \varepsilon_V; \left\{y_t^{(i)}\right\}_{t\in[M]}^{i\in[N]}; \|\cdot\|_2\right),
\end{aligned}$$

where $\mathcal{F}_K = \left\{(y, q) \mapsto \beta y^\mathsf{T} W_K q : W_K \in \mathcal{W}_K\right\}$ and $\mathcal{F}_V = \left\{y^\mathsf{T} \mapsto \left(y^\top \widetilde{W}_V\right)^\mathsf{T} : \widetilde{W}_V \in \widetilde{\mathcal{W}}_V\right\}$.

Proof of Lemma C.8. We prove that for each $\varepsilon > 0$ and input sample $(Y^{(i)}, Q^{(i)})$ for all $i \in [N]$, there exists a cover $\mathcal{C}_{\text{hop}}$ for $\mathcal{F}_{\text{hop}}$. From Lemma C.7, we see that

$$\begin{aligned}
&\left\|f_{\text{hop}}(Y, q_\tau; W_K, \widetilde{W}_V) - f_{\text{hop}}(Y, q_\tau; W'_K, \widetilde{W}'_V)\right\| \\
&\le 2B_V B_Y \left\|f_K\left(Y^{(i)}, q_\tau^{(i)}; W_K\right) - f_K\left(Y^{(i)}, q_\tau^{(i)}; W'_K\right)\right\|_\infty + \left\|f_V\left(Y^{(i)}; \widetilde{W}_V\right) - f_V\left(Y^{(i)}; \widetilde{W}'_V\right)\right\|_{2,\infty}.
\end{aligned}$$

With the property of ℓ_∞-norm, we have

$$\max_{i\in[N]} \left\|f_K\left(Y^{(i)}, q_\tau^{(i)}; W_K\right) - f_K\left(Y^{(i)}, q_\tau^{(i)}; W'_K\right)\right\|_\infty = \max_{i\in[N],t\in[M]} \left|f_K\left(y_t^{(i)}, q_\tau^{(i)}; W_K\right) - f_K\left(y_t^{(i)}, q_\tau^{(i)}; W'_K\right)\right|.$$

Also, with the property of $\ell_{2,\infty}$-norm,

$$\max_{i\in[N]} \left\|f_V\left(Y^{(i)}; \widetilde{W}_V\right) - f_V\left(Y^{(i)}; \widetilde{W}'_V\right)\right\|_{2,\infty} = \max_{i\in[N],t\in[M]} \left\|f_V\left(y_t^{(i)}; \widetilde{W}_V\right) - f_V\left(y_t^{(i)}; \widetilde{W}'_V\right)\right\|.$$

Now, we let $\mathcal{C}_K$ (a set of W_K) be the ε_K-cover for $\mathcal{F}_K$ over inputs $\left\{\left(y_t^{(i)}, q_\tau^{(i)}\right)\right\}_{t\in[M],\tau\in[T]}^{i\in[N]}$ of size

$$\mathcal{N}_\infty \left(\mathcal{F}_K; \varepsilon_K; \left\{\left(y_t^{(i)}, q_\tau^{(i)}\right)\right\}_{t\in[M],\tau\in[T]}^{i\in[N]}\right).$$

Also, let $\mathcal{C}_V$ (a set of $\widetilde{W}_V$) be the ε_V-cover for $\mathcal{F}_V$ over inputs $\left\{y_t^{(i)}\right\}_{t\in[M]}^{i\in[N]}$ of size

$$\mathcal{N}_\infty \left(\mathcal{F}_V; \varepsilon_V; \left\{y_t^{(i)}\right\}_{t\in[M]}^{i\in[N]}; \|\cdot\|_2\right).$$

We now construct the cover for $\mathcal{F}_{\text{hop}}$. First Set

$$\mathcal{C}_{\text{hop}} = \left\{f_{\text{hop}}\left(Y^{(i)}, q_\tau^{(i)}; W'_K \widetilde{W}'_V\right)_{\tau\in[T]}^{i\in[N]} : W'_K \in \mathcal{C}_K, \widetilde{W}'_V \in \mathcal{C}_V\right\}.$$

Then for any $W_K \in \mathcal{W}_K, \widetilde{W}'_V \in \widetilde{\mathcal{W}}_V$, there exists $W'_K \in \mathcal{C}_{\text{hop}}, \widetilde{W}'_V \in \mathcal{C}_V$ (using Lemma C.7):

$$\left\| f_{\text{hop}}(Y, q_\tau; W_K, \widetilde{W}_V) - f_{\text{hop}}(Y, q_\tau; W'_K, \widetilde{W}'_V) \right\| \leq 2B_V B_Y \varepsilon_K + \varepsilon_V.$$

The size of the cover $\mathcal{C}_{\text{hop}}$ we have constructed is,

$$\begin{aligned} & \log |\mathcal{C}_{\text{hop}}| \\ = & \log |\mathcal{C}_K| + \log |\mathcal{C}_V| \\ = & \log \mathcal{N}_\infty \left(\mathcal{F}_K; \varepsilon_K; \left\{ \left(y_t^{(i)}, q_\tau^{(i)} \right) \right\}_{t\in[M],\tau\in[T]}^{i\in[N]} \right) + \log \mathcal{N}_\infty \left(\mathcal{F}_V; \varepsilon_V; \left\{ y_t^{(i)} \right\}_{t\in[M]}^{i\in[N]}; \|\cdot\|_2 \right), \end{aligned}$$

where $\varepsilon = 2\varepsilon_{\text{Score}} + \varepsilon_K$ for $\mathcal{C}_{\text{hop}}$. □

Next, we introduce a useful lemma for completing the proof of Theorem 3.4.

Lemma C.9. (Lemma A.8 of (Edelman et al., 2022)) For $\alpha_i, \beta_i \geq 0$, the solution to the following optimization

$$\min_{x_1,\ldots,x_n} \sum_{i=1}^n \frac{\alpha_i}{x_i^2} \quad \text{subject to} \sum_{i=1}^n \beta_i x_i = C,$$

is γ^3/C^2 and is achieved at $x_i = C/\gamma \left(\alpha_i/\beta_i\right)^{1/3}$ where $\gamma = \sum_{i=1}^n \alpha_i^{1/3} \beta_i^{2/3}$.

Proof of Lemma C.9. The proof follows by a standard Lagrangian analysis. Let $f(x)$ be the objective function and $g(x)$ be the constraint function. With Lagrange multiplier, we have

$$\boldsymbol{\nabla} f(x) - \lambda \boldsymbol{\nabla} g(x) = 0.$$

By plugging in f and g we have

$$x_i = -\left(\frac{2\alpha_i}{\lambda \beta_i} \right)^{\frac{1}{3}}, \tag{C.5}$$

for all $i \in [n]$. In addition, we get λ by plugging (C.5) into the constraint g:

$$\lambda = \frac{\sum_{i=1}^n (2a_i)^{\frac{1}{3}} \beta_i^{\frac{2}{3}}}{C}.$$

□

To bound the covering number of $\mathcal{F}_K, \mathcal{F}_V$, we introduce the covering number bound for a linear function class:

Lemma C.10 (Covering Number Bound for Linear Function Class, Lemma 4.6 of (Edelman et al., 2022))**.** Let $\mathcal{W} : \left\{ W \in \mathbb{R}^{d_1 \times d_2} : \left\| W^\top \right\|_{2,1} \leq B_W \right\}$, and consider the function class $\mathcal{F} : \{x \mapsto Wx : W \in \mathcal{W}\}$. For any $\varepsilon > 0$ and $x^{(1)}, \ldots, x^{(N)} \in \mathbb{R}^{d_2}$ satisfying $\forall i \in [N], \left\| x^{(i)} \right\| \leq B_X$,

$$\log \mathcal{N}_\infty \left(\mathcal{F}; \varepsilon; x^{(1)}, \ldots, x^{(N)}; \|\cdot\|_2 \right) \lesssim \frac{\left(B_X B_W\right)^2}{\varepsilon^2} \log\left(d_1 N\right).$$

We now obtain the covering number bound of a Modern Hopfield Model explicitly by bounding the two function classes $\mathcal{F}_V$ and $\mathcal{F}_K$ using Lemma C.10.

Lemma C.11 (Covering Number Bound of Outlier-Efficient Hopfield Layer)**.**

$$\begin{aligned}&\log \mathcal{N}_\infty \left(\mathcal{F}_{\text{hop}} ; \varepsilon; \left\{ \left(Y^{(i)}, Q^{(i)} \right) \right\}^{i\in[N]} ; \|\cdot\|_{2,\infty} \right) \\ &\leq \varepsilon^{-2} \left(\left(4B_V^2 B_Y^2 \left(B_{\beta K}^{2,1} \right)^2 \log\left(dNM\right) \right)^{\frac{1}{3}} + \left(\left(B_V^{2,1} \right)^2 \log(dNM) \right)^{\frac{1}{3}} \right)^3 ,\end{aligned}$$

where $B_{\beta K}^{2,1} = \beta B_K^{2,1}$.

Proof of Lemma C.11. First observe that since $\|q_\tau\| \leq 1$ (Assumption 3.1-(A1)), we have

$$\left| \beta y_t^\mathsf{T} W_K q_\tau - \beta y_t^\mathsf{T} W_K' q_\tau \right| \leq \left\| \beta y_t^\mathsf{T} W_K - \beta y_t^\mathsf{T} W_K' \right\| .$$

We define the right hand side as (taking transpose to make it a column vector):

$$\widehat{\mathcal{F}}_K := \left\{ y_t \mapsto W_{\beta K}^\top y_t : \|W_{\beta K}\|_{2,1} \leq B_{\beta K}^{2,1} \right\},$$

where $W_{\beta K} := \beta W_K$ and $B_{\beta K}^{2,1} = \beta B_K^{2,1}$.

Since the covering number of $\mathcal{F}_K$ is at most the covering number of $\widehat{\mathcal{F}}_K$, instead of discussing the covering number bound of $\mathcal{F}_K$, we focus on $\widehat{\mathcal{F}}_K$.

Now, by Lemma C.10, Assumption 3.1-(A3) and Assumption 3.1-(A4) we have

$$\begin{aligned}\log \mathcal{N}_\infty \left(\mathcal{F}_K ; \varepsilon_K ; \left\{ \left(y_t^{(i)}, q_\tau^{(i)} \right) \right\}_{t\in[M],\tau\in[T]}^{i\in[N]} \right) &\leq \log \mathcal{N}_\infty \left(\widehat{\mathcal{F}}_K ; \varepsilon_K ; \left\{ \left(y_t^{(i)} \right) \right\}_{t\in[M]}^{i\in[N]} \right) \\ &\lesssim \frac{\left(B_{\beta K}^{2,1} \right)^2}{\varepsilon_K^2} \log\left(dNM\right) ,\end{aligned} \tag{C.6}$$

and

$$\log \mathcal{N}_\infty \left(\mathcal{F}_V ; \varepsilon_V ; \left\{ y_t^{(i)} \right\}_{t\in[M]}^{i\in[N]} ; \|\cdot\|_2 \right) \lesssim \frac{\left(B_V^{2,1} \right)^2}{\varepsilon_V^2} \log\left(dNM\right) . \tag{C.7}$$

Next, we find the optimal ε_K and ε_V to minimize the sum of (C.6) and (C.7), subject to

$$2B_V B_Y \varepsilon_K + \varepsilon_V = \varepsilon.$$

By Lemma C.9, the optimal bound is

$$\begin{aligned}&\log \mathcal{N}_\infty \left(\mathcal{F}_{\text{hop}} ; \varepsilon; \left\{ \left(Y^{(i)}, Q^{(i)} \right) \right\}^{i\in[N]} ; \|\cdot\|_{2,\infty} \right) \\ &\leq \log \mathcal{N}_\infty \left(\mathcal{F}_{\text{hop}} ; \varepsilon; \left\{ \left(Y^{(i)}, q_\tau^{(i)} \right) \right\}_{\tau\in[T]}^{i\in[N]} ; \|\cdot\|_2 \right) \\ &\leq \varepsilon^{-2} \left(\left(4B_V^2 B_Y^2 \left(B_{\beta K}^{2,1} \right)^2 \log\left(dNM\right) \right)^{\frac{1}{3}} + \left(\left(B_V^{2,1} \right)^2 \log(dNM) \right)^{\frac{1}{3}} \right)^3 .\end{aligned}$$

□

With the covering number bound, we arrive at the norm-based generalization bound Theorem 3.4 through Lemma C.4.

D. Additional Numerical Experiments

D.1. Supplemental Experimental Results (Figure 3 and Figure 4)

We conducted in-depth case studies on the BERT model. In Figure 3, we focus on the outlier performance in selected layers. The figure shows that outliers become stronger in deeper layers of the vanilla model, corroborating insights from Bondarenko et al. (2021). However, `OutEffHop` maintains a consistent maximum infinity norm $\|\mathbf{x}\|_\infty$ across all layers, demonstrating its effectiveness in controlling outliers. In Figure 4, we delve into the maximum infinity norm $\|\mathbf{x}\|_\infty$ within the 10th layer's various tensor components. These tensors are after attention layer, the first residual layernorm after attention, and the first, second FFN layers. As mentioned in (Bondarenko et al., 2023), FFN layers indeed increase the outliers heavily along the training process in vanilla attention. In contrast, `OutEffHop` suppresses the outliers growing in both FFN layers. The effectiveness of `OutEffHop` is due to its built-in no-operation (no-op) pattern which defaults queries to this pattern when updates are unnecessary. This eliminates the need to learn outlier values in FFN layers to direct attention weights toward specific tokens for a no-op. Additionally, we observe that the first residual LayerNorm after the attention mechanism tends to amplify outliers. This observation is also mentioned in Wei et al. (2022)'s finding. Furthermore, we note that the outliers of `OutEffHop` are larger than those of the vanilla model. Our method, `OutEffHop`, focusing solely on the attention mechanism, offers evidence of its effectiveness in mitigating outliers within our approach.

D.2. Verifying Theoretical Results

We also verify our theoretical findings following the settings in (Hu et al., 2023).

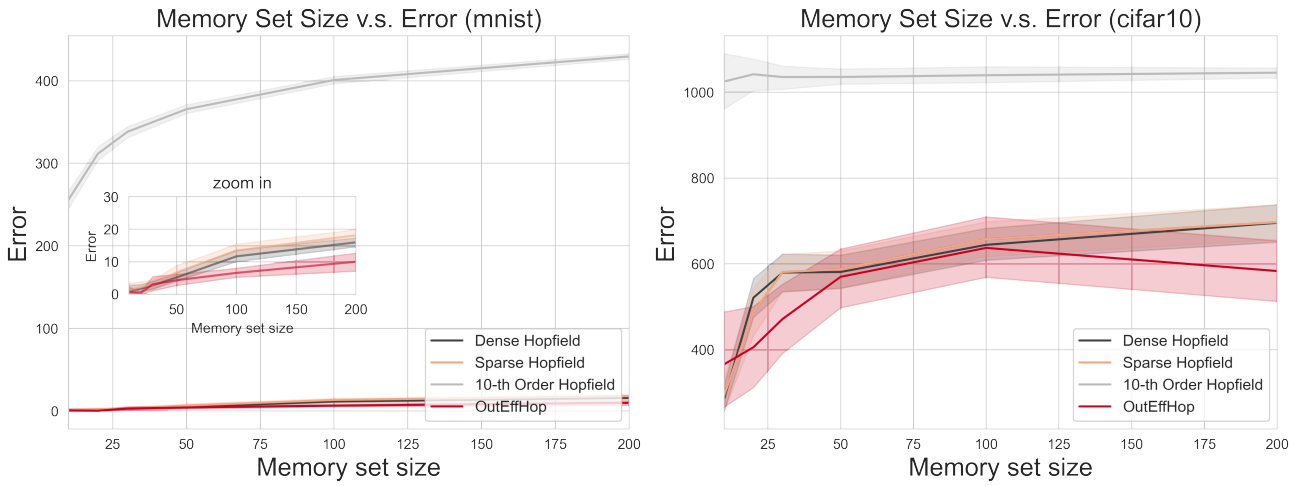


Figure 5. **Memory Capacity.** Our extensive evaluation of memory capacity across various Hopfield Networks, including Vanilla Modern Hopfield, Sparse Hopfield, 10th Order Hopfield, and our `OutEffHop`, is conducted on two image datasets: MNIST and CIFAR10. We observe that `OutEffHop` outperforms its baselines, especially when the memory set size is large.

Memory Capacity. For the memory capacity, we compare our Outlier-Efficient Modern Hopfield Model (`OutEffHop`) with Dense (Softmax) (Ramsauer et al., 2020), Sparse (Hu et al., 2023) and 10th order polynomial Hopfield model (Krotov and Hopfield, 2016) on MNIST (LeCun et al., 1998) (high sparsity) and CIFAR10 (Krizhevsky et al., 2009) (low sparsity) datasets. For all Hopfield models, we set $\beta = 1$. As shown in Figure 5, `OutEffHop` outperforms its baselines, especially when the memory set size is large.

Noise-Robustness. For the robustness against noise queries, we inject Gaussian noises varying variances (σ) into the images. The results, as shown in Figure 6, show that `OutEffHop` excels when the signal-to-noise ratio in patterns is low.

Faster Convergence. We numerically analyse the convergence of `OutEffHop`, Dense and Sparse Hopfield model by evaluating their loss and accuracy in two different datasets. We use the Vision Transfromer (Dosovitskiy et al., 2020) (ViT) as the backbone and then replace the attention layer with different Hopfield layers. The hyperparameters used in our experiment are listed in Table 3. As shown in Figure 7, our model surpasses its original counterpart across all datasets.

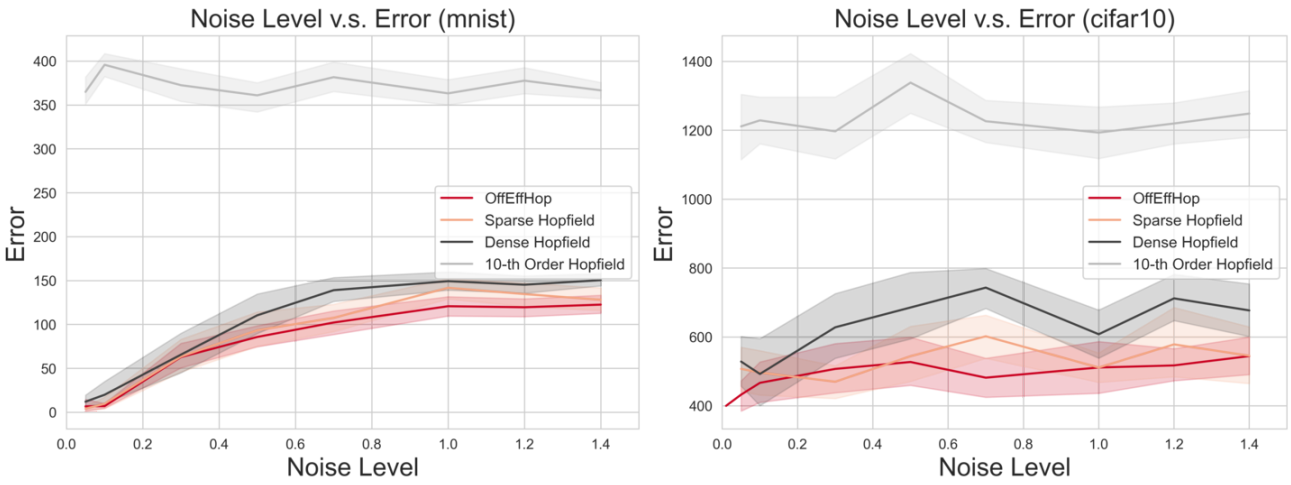


Figure 6. **Noise-Robustness.** Our extensive evaluation of noise robustness across various Hopfield Networks, including Vanilla Modern Hopfield, Sparse Hopfield, 10th Order Hopfield, and our `OutEffHop`, is conducted on two image datasets: MNIST and CIFAR10. The results show that as the noise level rises, the impact of `OutEffHop` on the error rate is minimal.

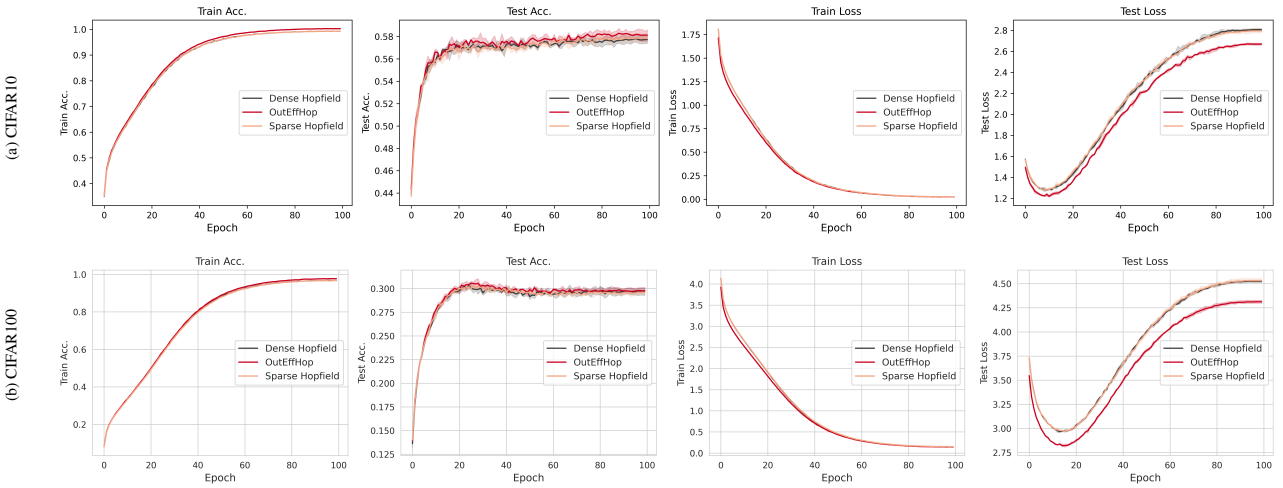


Figure 7. **Faster Convergence.** Our extensive evaluation of faster covergence across various Hopfield Networks, including Vanilla Modern Hopfield, Sparse Hopfield, and our `OutEffHop`, is conducted on two image datasets: CIFAR10 and CIFAR100. The results show that `OutEffHop` has faster convergence than baselines.

Table 3. Hyperparameter used in the fast convergence task.

parameter	**values**
learning rate	$1e-4$
embedding dimension	512
Feed forward dimension	1024
Dropout	0.3
activation function	GELU
Epoch	100
Batch size	512
Model optimizer	Adam
Patch size	32

D.3. Computational Cost Comparison

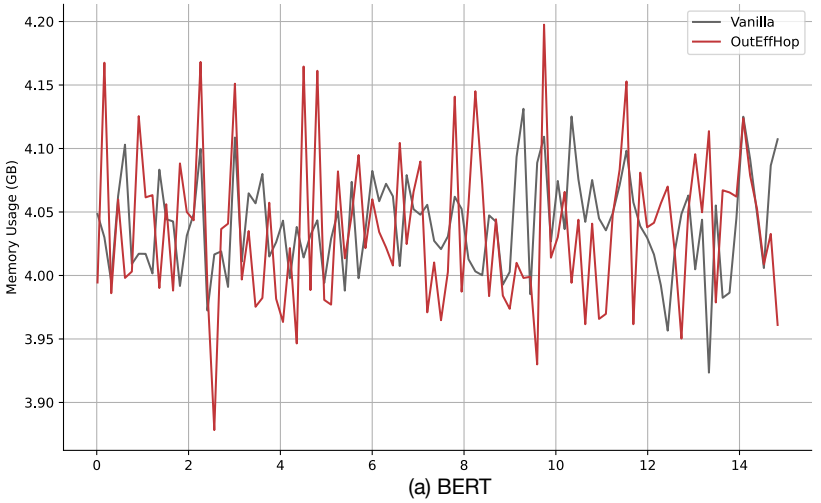
We compare the computational resources of four different models against the vanilla $\mathrm{Softmax}$ and `OutEffHop`, as detailed in Table 4. We measure the pre-training records of all four models. Memory usage for OPT, BERT, and ViT is monitored using Wandb [11], while for STanHop, it is tracked via system logs[12]. The model sizes for this experiment match those described in section 4.1. Our experimental setup used a Slurm system with two 80G A100 GPUs and a 24-core Intel(R) Xeon(R) Gold 6338 CPU at 2.00GHz. We also provide the wandb diagram of the system memory usage in Figure 8.

Table 4. The computational resource comparison of vanilla $\mathrm{Softmax}$ and `OutEffHop` in 4 models. We compare the Time and average of the Memory RAM usage in the model pre-training periods.

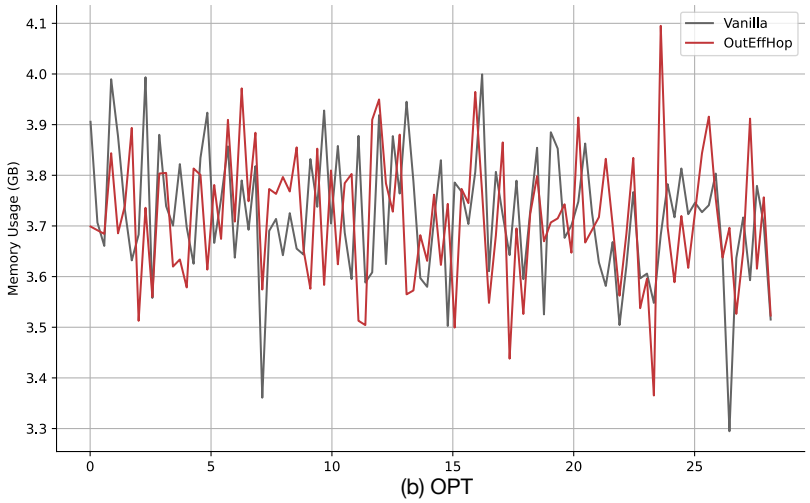
Model	Method	Memory Usage (Gb)
ViT	Vanilla	47.47
	OutEffHop	49.69
ERT	Vanilla	7.56
	OutEffHop	7.20
OPT	Vanilla	3.75
	OutEffHop	3.75
STN	Vanilla	5.30
	OutEffHop	5.28

[11] https://wandb.ai/

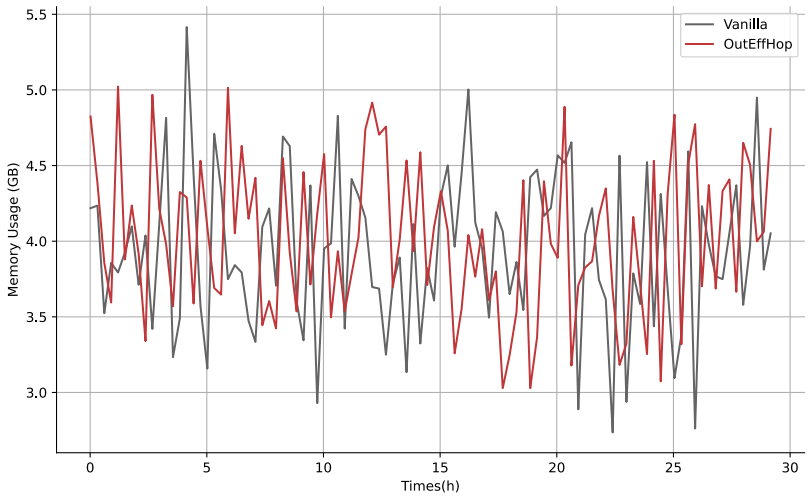
[12] We thank the authors of (Reneau et al., 2023) for their helpful comments on this part.



(a) BERT



(b) OPT



(c) ViT

Figure 8. The computational resource comparison between Vanilla Softmax and `OuTEffHop` involves measuring RAM usage via Wandb in a system equipped with 180G RAM under the Slurm system.