OutEffHop: A Principled Outlier-Efficient Attention Layer from Dense Associative Memory Models

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

Abstract

We introduce a principled approach to Outlier-Efficient Attention Layers via associative memory models to reduce outlier emergence in large transformer-based model. Our main contribution is a novel associative memory model that facilitates outlier-efficient associative memory retrievals. This model subsumes the outlier-efficient attention mechanism ($\mathrm{Softmax}_1$) as a special case of its memory retrieval process. Methodologically, this enables the introduction of novel outlier-efficient Hopfield layers as powerful alternatives to traditional attention mechanisms, offering superior post-quantization performance. 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 `Clipped_Softmax` and `Gated_Attention`. Notably, our method achieves an average reduction of over 22% in average kurtosis and over 26% in the maximum infinity norm of model outputs across the four models, without sacrificing model performance after quantization.[1]

[*]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: Haozheng Luo <hluo@u.northwestern.edu>, Jerry Yao-Chieh Hu <jhu@u.northwestern.edu>, Pei-Hsuan Chang <b09202022@ntu.edu.tw>, Hong-Yu Chen <hong-yuchen2029@u.northwestern.edu>, 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]This paper presents a concise version of (Hu et al., 2024a).

1. Introduction

We tackle the outlier-inefficiency issue in large Transformer-based models by presenting a new principled approach to outlier-efficient attention layers, termed `OutEffHop`. This problem is of practical importance in the era of Large Foundation Models (Xu et al., 2024b; Zhou et al., 2024b; 2023; Wu et al., 2023; Bommasani et al., 2021; Ji et al., 2021; Ho et al., 2020; Brown et al., 2020; Floridi and Chiriatti, 2020).

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

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

We focus on the part of transformer right after attention

$$\mathrm{Output} = \mathrm{Residual}(X + A). \tag{1.1}$$

If the input X already contains sufficient information and does not require further feature extraction, the attention mechanism tends to behave like an identity map, outputting a zero matrix A. This is known as the *no-update situation*. A direct consequence is that the attention mechanism forces tokens with large values (as in V) to receive *close-to-zero* attention probability (as in $\mathrm{Softmax}(QK^\top)$), resulting in small-value tokens having large attention probabilities. Due to the normalization nature of the softmax function, this operation forces its input $QK^\top$ to span a wide range. This wide range is the fundamental source of outliers: some tokens must cause this "wide range" of $QK^\top$ (termed *outliers*). Since attention to these tokens behaves as a "no-op" (no operation), we call these "no-op" outliers. Furthermore, since the softmax function never reaches exactly zero, it always sends back a gradient signal, leading to the magnification of outliers during training (Bondarenko et al., 2023).

To address this, we draw motivation from recent progress in dense associative memory models (Wu et al., 2024a;b; Hu et al., 2024b;c; 2023; Chaudhry et al., 2023; Hoover et al., 2023; Krotov, 2023; Krotov and Hopfield, 2021; Ramsauer et al., 2020) and introduce `OutEffHop` to provide a principled understanding (theoretical guarantees and empirical evidences) of the outlier problem in transformer attention heads. This model-based understanding includes the $\mathrm{Softmax}_1$ activation (a quantization-robust alternative to the $\mathrm{Softmax}$ function in vanilla attention) as a special case.

2. Outlier-Efficient Hopfield Layers

The key idea is to add a ”no-op classification” dimension to the Hopfield energy function’s state space, identifying ”no-op” outliers as distinct patterns with no similarity to other memory patterns. Let $x \in \mathbb{R}^d$ represent the query patterns and $\Xi = [\xi^1, \cdots, \xi^M] \in \mathbb{R}^{d\times M}$ the M memory patterns. We extend their dimension such that x and ξ^μ become

$$x = (x_1, \ldots, x_d, 0), \quad \xi^\mu = (\xi^\mu_1, \cdots, \xi^\mu_d, \omega),$$

with an extra $\omega \in \mathbb{R}$. We set 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, then we introduce the following function:

$$\Lambda(\xi_\mu) = \begin{cases} (\xi^\mu_1, \cdots, \xi^\mu_d, 0) = \xi^\mu_{\text{op}} \in \mathbb{R}^{d+1}, & \text{if } \omega = 0, \\ (\underbrace{0, \cdots, 0}_{d}, C) = \Omega \in \mathbb{R}^{d+1}, & \text{if } \omega \neq 0 \end{cases}, \tag{2.1}$$

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.” We term the Λ function (2.1) the “no-op classification mechanism.” We introduce the outlier-efficient Modern Hopfield energy as:

$$\mathcal{H}(x) = -\text{lse}_1\left(\beta, \Xi^\mathsf{T} x\right) + \frac{1}{2}\langle x, x\rangle + \text{Const.}, \tag{2.2}$$

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

$$\text{lse}_1\left(\beta, \Xi^\mathsf{T} x\right) \tag{2.3}$$
$$:= \beta^{-1}\log\Big(\sum_{\mu=1}^M \exp\{\beta\langle\xi_\mu, x\rangle\} + \overbrace{\exp\{\beta\underbrace{\langle\Omega, x\rangle}_{=0}\}}^{(I)}\Big).$$

Here, (I) has an unique “*no-op memory class vector* $\Omega \in \mathbb{R}^d$” whose inner product with the query $x \in \mathbb{R}^d$ is zero: $\langle\Omega, x\rangle = 0$. Intuitively, by (2.1), Ω represents an outlier of the stored memory set $\Xi := [\xi_1, \cdots, \xi_M, \Omega]$, and $\langle\Omega, x\rangle = 0$ indicates it does not participate the retrieval process. Specifically, Hopfield energy (2.2) can be monotomically minimized by following memory retrieval dynamics:

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

$$\mathcal{T}_{\text{OutEff}}(x_t) := \Xi\text{Softmax}_1\left(\beta\Xi^\mathsf{T} x_t\right) = x_{t+1}, \tag{2.4}$$

monotonically minimizes the energy (2.2) over t.

Proof. Since (2.3) is concave by design, we prove this by standard CCCP derivation following (Hu et al., 2023). ☐

Remark 2.1 (1-Iteration $\mathcal{T}_{\text{OutEff}}$ is Softmax_1)**.** Due to the monotonic decreasing property of Lemma 2.1, for any given input query x, (2.4) retrieves a memory closest to it by approaching to the nearest local minimum of $\mathcal{H}$. Interestingly, when $\mathcal{T}_{\text{OutEff}}$ is applied only once, (2.4) is equivalent to an outlier-efficient attention (Miller, 2023).

Connection to Deep Learning By the connection with attention mechanism as shown above, Outlier-efficient Hopfield model is applicable to nowadays deep learning architectures. 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 (2.4) and multiplying with W_V (letting $V := KW_V$), we get: $Q^{\text{new}}W_V = \text{Softmax}_1(\beta QK^\mathsf{T})V$. We present the Outlier-Efficient Hopfield (`OutEffHop`) layer for deep learning:

$$\begin{aligned} Z &= \texttt{OutEffHop}(R, Y) \\ &= \text{Softmax}_1\left(\beta R W_Q W_K^\mathsf{T} Y^\mathsf{T}\right) Y W_K W_V, \end{aligned} \tag{2.5}$$

which takes R and Y as input, paired with weight matrices W_Q, W_K, and W_V. This attention-like layer is designed to be outlier-robust, i.e., it filters out *low-relevance* tokens in attention computation. Therefore, OutEffHop serves as a powerful alternative for quantization and compression, with strong theoretical foundations. Consequently, it offers a robust implementation for large foundation models, enabling more economical training without sacrificing performance.

Remark 2.2. Note that we only have to identify outlier when our model serves as associtative memory models. For using `OutEffHop` as attention-like layer like (2.5), 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 (2.4).

3. Experimental Studies

We conduct a series of experiments to validate the effectiveness of the Outlier-Efficient Attention Layers. Specifically, we benchmark our model against SOTA methods as outlined in (Bondarenko et al., 2023), employing 3 widely-used large transformer-based models and 1 Hopfield-based model.

3.1. Outlier Efficiency of `OutEffHop`

To evaluate the model’s resilience to outliers, we integrate `OutEffHop` into various architectures, including 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), by substituting the standard attention (Vaswani et al., 2017) and Hopfield layers (Hu et al., 2023; Ramsauer et al., 2020) with our module. We then train these models from scratch and evaluate them on the validation set. Each evaluation is conducted three times using different random seeds, with the average and standard deviation reported for each metric.

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	$\underline{14.210} \pm 0.003$	33.619 ± 0.001	$\mathbf{6.118} \pm 0.002$	$\mathbf{6.189} \pm 0.001$	
	Clipped OutEffHop	$\mathbf{11.839} \pm 0.001$	$\mathbf{30.107} \pm 0.001$	$\underline{6.133} \pm 0.000$	$\underline{6.199} \pm 0.001$	
	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	$\underline{32.777} \pm 0.000$	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	$\underline{21.542} \pm 0.000$	$\underline{13.302} \pm 0.001$	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	$\mathbf{15.677} \pm 0.000$	$\underline{16.236} \pm 0.074$	124.07m
	Gated OutEffHop	$\mathbf{11.449} \pm 0.000$	$\mathbf{7.568} \pm 0.000$	$\underline{15.751} \pm 0.000$	$\mathbf{16.148} \pm 0.005$	
ViT	Vanilla	37.104 ± 0.000	272.198 ± 0.000	$\underline{76.810} \pm 0.000$	74.935 ± 0.046	22.03m
	OutEffHop	31.601 ± 0.001	249.163 ± 0.000	76.788 ± 0.000	$\mathbf{76.313} \pm 0.012$	
	Clipped Softmax	33.868 ± 0.00	257.613 ± 0.00	76.612 ± 0.000	75.179 ± 0.013	
	Clipped OutEffHop	$\underline{24.642} \pm 0.000$	$\underline{196.199} \pm 0.001$	$\mathbf{76.871} \pm 0.001$	$\underline{76.083} \pm 0.007$	
	Gated Attention	45.145 ± 0.864	269.279 ± 1.426	69.922 ± 2.436	67.479 ± 1.447	22.04m
	Gated OutEffHop	$\mathbf{21.979} \pm 0.254$	$\mathbf{60.169} \pm 1.153$	74.089 ± 2.585	73.958 ± 3.126	
STanHop-Net	Vanilla	2.954 ± 0.063	5.048 ± 0.232	$\underline{0.360} \pm 0.008$	0.362 ± 0.000	35.13m
	OutEffHop	2.897 ± 0.011	4.565 ± 0.209	$\mathbf{0.360} \pm 0.004$	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	$\mathbf{4.145} \pm 0.23$	0.506 ± 0.05	0.517 ± 0.000	
	Gated Attention	$\underline{2.487} \pm 0.017$	4.277 ± 0.163	0.380 ± 0.006	0.375 ± 0.000	35.15m
	Gated OutEffHop	$\mathbf{2.459} \pm 0.041$	$\underline{4.240} \pm 0.155$	0.376 ± 0.007	0.367 ± 0.000	

Metrics. We report the *maximum infinity norm* $\|x\|_\infty$ and *average kurtosis* of the activation tensors $\mathbf{x}$ across all transformer layers as a metric of outliers. For BERT, we average the output tensors from the Feed-Forward Network (FFN) layer and Layer Normalization. Both are known for the presence of outliers, as confirmed by our experiments and previous studies (Bondarenko et al., 2023; Wei et al., 2022; Bondarenko et al., 2021). For OPT, ViT, and STanHop, we average over every output component in transformer layers. These metrics demonstrate strong correlations with model quantizability, reflecting robustness against outliers (Bondarenko et al., 2021; Shkolnik et al., 2020). Prior research (Dettmers et al., 2022; Wei et al., 2022; Bondarenko et al., 2021) highlight a substantial decline in model performance after quantization when outliers exist. Consequently, we record the models' performance both before and after quantization. For pre-quantization performance, we evaluate the *Perplexity Score* for BERT and OPT using **FP16** (16-bit floating-point), the *Top-1 Accuracy* for ViT using **FP32** (32-bit floating-point), and the *Mean Square Error* (MSE) for STanHop-Net. For post-quantization performance in **W8A8** (8-bit floating-point), we report the same metrics.

Datasets. We employ 4 real-world datasets for our evaluations: Bookcorpus (Zhu et al., 2015) and wiki40b/en (Guo et al., 2020) are for language models such as OPT and BERT; ImageNet-1k (Russakovsky et al., 2015) is for the vision model, i.e. ViT; and ETTh1 (Zhou et al., 2021) is for the time series model, i.e. STanHop-Net.

Models. Following Bondarenko et al. (2023), we evaluate our approach (`OutEffHop`) across four prominent models: two language models (BERT, OPT), one vision model (ViT), and one time series model (STanHop). For BERT, we utilize the BERT-base-uncased model, which contains 109 million parameters, and pretrain it using the masked language modeling (MLM) technique as outlined in (Devlin et al., 2019). The OPT model, equipped with 125 million parameters, is pretrained using causal language modeling (CLM). We configure the sequence lengths to 128 for BERT and 512 for OPT to enhance training efficiency. The ViT-S_16 variant,

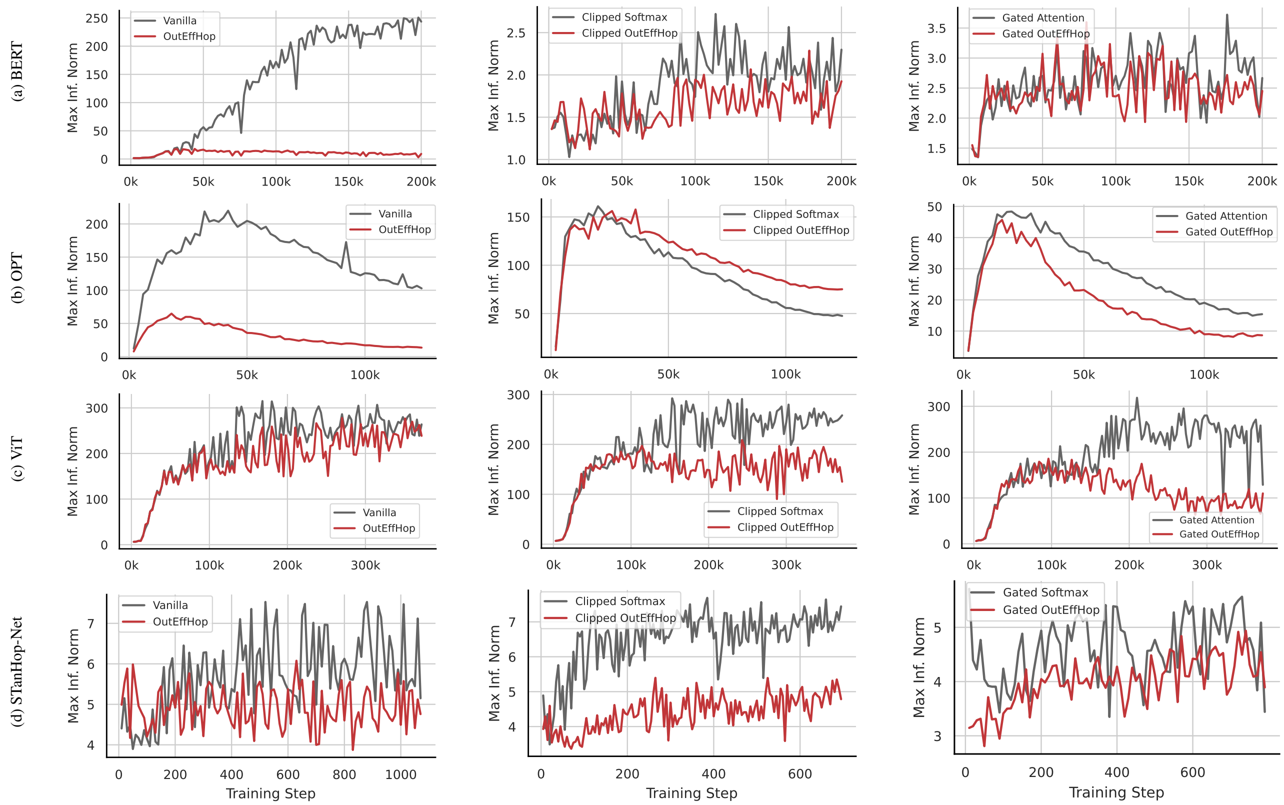


Figure 1. **The Impact of** `OutEffHop` **on Maximum Infinity Norm $\|x\|_\infty$ Changes During Pretraining of (a) BERT, (b) OPT, (c) ViT, and (d) STanHop-Net.** The plots, from left to right, compare `OutEffHop` with the vanilla attention baseline and their combination with `Clipped_Softmax` and `Gated_Attention` 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 $\|x\|_\infty$ compared to the vanilla attention and improves `Clipped_Softmax` and `Gated_Attention`.

with 22.03 million parameters, is pretrained using a conventional image classification task. Lastly, the STanHop-Net model, possessing 35.13 million parameters, is pretrained on a multivariate time series prediction task.

Results. In Table 1 and Figure 1, our findings reveal that `OutEffHop` matches the outlier reduction capabilities of `Clipped_Softmax` and `Gated_Attention`. When combined with these methods, `OutEffHop` further enhances their effectiveness, achieving an average reduction of approximately 22% in average kurtosis and 26% in maximum infinity norm across four tested models. An exception is the Clipped `OutEffHop` in the OPT model, which, as Bondarenko et al. (2023) suggests, does not perform well with the `Clipped_Softmax` method. Notably, `OutEffHop` lowers the maximum infinity norm during pre-training, particularly in layer 10 of BERT, ViT, and OPT models, and in layer 9 of the STanHop model, as shown in Figure 1. This underscores `OutEffHop`'s superiority in reducing outliers during pre-training compared to baseline methods, with significant enhancements particularly in the OPT model.

4. Conclusion and Discussion

We introduce the Outlier-Efficient Modern Hopfield Model to tackle the computational difficulties associated with outliers in large transformer-based models. This model not only improves the desirable properties of modern Hopfield networks, but also incorporates the `OutEffHop` layers as innovative deep learning components that enhance outlier reduction in large transformer architectures. Empirical evaluations show that `OutEffHop` achieves an average reduction of 22% in average kurtosis and 26% in maximum infinity norm across four different models.

Limitation and Future Work. The main limitation of `OutEffHop` is its inability to address outliers induced by LayerNorm, as indicated in the First Residual LayerNorm in Figure 3. Indeed, Wei et al. (2022) note that the origins of outliers in LayerNorm differ from those in the attention mechanisms we study. Future research focuses on integrating these outliers within the `OutEffHop` framework.

Impact Statement

We believe this methodology offers an opportunity to enhance the cores 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.

Acknowledgment

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

Mohammad Shahmeer Ahmad, Zan Ahmad Naeem, Mohamed Eltabakh, Mourad Ouzzani, and Nan Tang. Retclean: Retrieval-based data cleaning using foundation models and data lakes. *arXiv preprint arXiv:2303.16909*, 2023.

Josh Alman and Zhao Song. Fast attention requires bounded entries. In *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

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

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.

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.

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.

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.

Yelysei Bondarenko, Markus Nagel, and Tijmen Blankevoort. Understanding and overcoming the challenges of efficient transformer quantization, 2021.

Yelysei Bondarenko, Markus Nagel, and Tijmen Blankevoort. Quantizable transformers: Removing outliers by helping attention heads do nothing. *Advances in Neural Information Processing Systems (NeurIPS)*, 36, 2023.

Johannes Brandstetter. Blog post: Hopfield networks is all you need, 2021. 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.

Thomas F Burns. Semantically-correlated memories in a dense associative model. In *Forty-first International Conference on Machine Learning (ICML)*, 2024.

Thomas F Burns and Tomoki Fukai. Simplicial hopfield networks. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

Vivien Cabannes, Elvis Dohmatob, and Alberto Bietti. Scaling laws for associative memories. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

Hamza Chaudhry, Jacob Zavatone-Veth, Dmitry Krotov, and Cengiz Pehlevan. Long sequence hopfield memory. *Advances in Neural Information Processing Systems (NeurIPS)*, 36, 2023.

Jerry Chee, Yaohui Cai, Volodymyr Kuleshov, and Christopher M De Sa. Quip: 2-bit quantization of large language models with guarantees. *Advances in Neural Information Processing Systems*, 36, 2024.

Minshuo Chen, Song Mei, Jianqing Fan, and Mengdi Wang. An overview of diffusion models: Applications, guided generation, statistical rates and optimization. *arXiv preprint arXiv:2404.07771*, 2024.

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.

Tim Dettmers and Luke Zettlemoyer. The case for 4-bit precision: k-bit inference scaling laws. In *International Conference on Machine Learning*, pages 7750–7774. PMLR, 2023.

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.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36, 2024.

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.

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.

Zehao Dou, Minshuo Chen, Mengdi Wang, and Zhuoran Yang. Provable statistical rates for consistency diffusion models. *arXiv preprint arXiv:2406.16213*, 2024.

Luciano Floridi and Massimo Chiriatti. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30:681–694, 2020.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. Gptq: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*, 2022.

Hengyu Fu, Zhuoran Yang, Mengdi Wang, and Minshuo Chen. Unveil conditional diffusion models with classifier-free guidance: A sharp statistical theory. *arXiv preprint arXiv:2403.11968*, 2024.

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.

Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*, 2020.

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.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023.

Jiuxiang Gu, Yingyu Liang, Heshan Liu, Zhenmei Shi, Zhao Song, and Junze Yin. Conv-basis: A new paradigm for efficient attention inference and gradient computation in transformers. *arXiv preprint arXiv:2405.05219*, 2024a.

Jiuxiang Gu, Yingyu Liang, Zhenmei Shi, Zhao Song, and Chiwun Yang. Toward infinite-long prefix in transformer. *arXiv preprint arXiv:2406.14036*, 2024b.

Jiuxiang Gu, Yingyu Liang, Zhenmei Shi, Zhao Song, and Yufa Zhou. Tensor attention training: Provably efficient learning of higher-order transformers. *arXiv preprint arXiv:2405.16411*, 2024c.

Jiuxiang Gu, Yingyu Liang, Zhenmei Shi, Zhao Song, and Yufa Zhou. Unraveling the smoothness properties of diffusion models: A gaussian mixture perspective. *arXiv preprint arXiv:2405.16418*, 2024d.

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.

Yingqing Guo, Hui Yuan, Yukang Yang, Minshuo Chen, and Mengdi Wang. Gradient guidance for diffusion models: An optimization perspective. *arXiv preprint arXiv:2404.14743*, 2024.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

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.

John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982.

John J Hopfield. Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the national academy of sciences*, 81(10):3088–3092, 1984.

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.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

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.

Jerry Yao-Chieh Hu, Pei-Hsuan Chang, Robin Luo, Hong-Yu Chen, Weijian Li, Wei-Po Wang, and Han Liu. Outlier-efficient hopfield layers for large transformer-based models. In *Forty-first International Conference on Machine Learning (ICML)*, 2024a.

Jerry Yao-Chieh Hu, Bo-Yu Chen, Dennis Wu, Feng Ruan, and Han Liu. Nonparametric modern hopfield models. *arXiv preprint arXiv:2404.03900*, 2024b.

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)*, 2024c.

Jerry Yao-Chieh Hu, Maojiang Su, En-Jui Kuo, Zhao Song, and Han Liu. Computational limits of low-rank adaptation (lora) for transformer-based models. *arXiv preprint arXiv:2406.03136*, 2024d.

Jerry Yao-Chieh Hu, Weimin Wu, Zhuoru Li, Zhao Song, and Han Liu. On statistical rates and provably efficient criteria of latent diffusion transformers (dits). *arXiv preprint arXiv:2407.01079*, 2024e.

Xijie Huang, Zechun Liu, Shih-Yang Liu, and Kwang-Ting Cheng. Rolora: Fine-tuning rotated outlier-free llms for effective weight-activation quantization. *arXiv preprint arXiv:2407.08044*, 2024.

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.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Can Jin, Hongwu Peng, Shiyu Zhao, Zhenting Wang, Wujiang Xu, Ligong Han, Jiahui Zhao, Kai Zhong, Sanguthevar Rajasekaran, and Dimitris N Metaxas. Apeer: Automatic prompt engineering enhances large language model reranking. *arXiv preprint arXiv:2406.14449*, 2024.

johnowhitaker. Blog post: Exploring softmax1, or "community research for the win!", 2023. 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.

Sehoon Kim, Coleman Hooper, Amir Gholami, Zhen Dong, Xiuyu Li, Sheng Shen, Michael W Mahoney, and Kurt Keutzer. Squeezellm: Dense-and-sparse quantization. *arXiv preprint arXiv:2306.07629*, 2023.

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

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Dmitry Krotov. A new frontier for hopfield networks. *Nature Reviews Physics*, 5(7):366–367, 2023.

Dmitry Krotov and John J. Hopfield. Dense associative memory for pattern recognition. *CoRR*, 2016.

Dmitry Krotov and John J. Hopfield. Large associative memory problem in neurobiology and machine learning. In *International Conference on Learning Representations*, 2021.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. Bloom: A 176b-parameter open-access multilingual language model. 2023.

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.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021.

Hongkang Li, Meng Wang, Songtao Lu, Xiaodong Cui, and Pin-Yu Chen. How do nonlinear transformers learn and generalize in in-context learning? In *Forty-first International Conference on Machine Learning*, 2024a.

Hongkang Li, Meng Wang, Shuai Zhang, Sijia Liu, and Pin-Yu Chen. Learning on transformers is provable low-rank and sparse: A one-layer analysis. *arXiv preprint arXiv:2406.17167*, 2024b.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. *Proceedings of Machine Learning and Systems*, 6:87–100, 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

Mengyang Liu, Haozheng Luo, Leonard Thong, Yinghao Li, Chao Zhang, and Le Song. Sciannotate: A tool for integrating weak labeling sources for sequence labeling. *arXiv preprint arXiv:2208.10241*, 2022.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. Gpt understands, too. *AI Open*, 2023.

Haozheng Luo, Ningwei Liu, and Charles Feng. Question and answer classification with deep contextualized transformer. In *Advances in Information and Communication: Proceedings of the 2021 Future of Information and Communication Conference (FICC), Volume 2*, pages 453–461. Springer, 2021.

Haozheng Luo, Ruiyang Qin, Chenwei Xu, Guo Ye, and Zening Luo. Open-ended multi-modal relational reasoning for video question answering. In *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 363–369. IEEE, 2023a.

Haozheng Luo, Jiahao Yu, Wenxin Zhang, Jialong Li, Jerry Yao-Chieh Hu, Xingyu Xin, and Han Liu. Decoupled alignment for robust plug-and-play adaptation. *arXiv preprint arXiv:2406.01514*, 2024.

Yukui Luo, Nuo Xu, Hongwu Peng, Chenghong Wang, Shijin Duan, Kaleel Mahmood, Wujie Wen, Caiwen Ding, and Xiaolin Xu. Aq2pnn: Enabling two-party privacy-preserving deep neural network inference with adaptive quantization. In *Proceedings of the 56th Annual IEEE/ACM International Symposium on Microarchitecture*, pages 628–640, 2023b.

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.

Cristian Meo, Ksenia Sycheva, Anirudh Goyal, and Justin Dauwels. Bayesian-lora: Lora based parameter efficient fine-tuning using optimal quantization levels and rank values trough differentiable bayesian gates. *arXiv preprint arXiv:2406.13046*, 2024.

Evan Miller. Blog post: Attention is off by one, 2023. URL https://www.evanmiller.org/attention-is-off-by-one.html. Accessed: July 4, 2024.

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.

Zhenyu Pan, Haozheng Luo, Manling Li, and Han Liu. Chain-of-action: Faithful and multimodal question answering through large language models. *arXiv preprint arXiv:2403.17359*, 2024a.

Zhenyu Pan, Haozheng Luo, Manling Li, and Han Liu. Conv-coa: Improving open-domain question answering in large language models via conversational chain-of-action. *arXiv preprint arXiv:2405.17822*, 2024b.

Ethan Perez, Patrick Lewis, Wen-tau Yih, Kyunghyun Cho, and Douwe Kiela. Unsupervised question decomposition for question answering. *arXiv preprint arXiv:2002.09758*, 2020.

Ruiyang Qin, Haozheng Luo, Zheheng Fan, and Ziang Ren. Ibert: Idiom cloze-style reading comprehension with attention. *arXiv preprint arXiv:2112.02994*, 2021.

Ruiyang Qin, Dancheng Liu, Zheyu Yan, Zhaoxuan Tan, Zixuan Pan, Zhenge Jia, Meng Jiang, Ahmed Abbasi, Jinjun Xiong, and Yiyu Shi. Empirical guidelines for deploying llms onto resource-constrained edge devices. *arXiv preprint arXiv:2406.03777*, 2024.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

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.

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.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

Karan Samel, Zelin Zhao, Binghong Chen, Kuan Wang, Robin Luo, and Le Song. How to design sample and computationally efficient vqa models. *arXiv preprint arXiv:2103.11537*, 2021.

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.

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.

Moran Shkolnik, Brian Chmiel, Ron Banner, Gil Shomron, Yury Nahshan, Alex Bronstein, and Uri Weiser. Robust quantization: One model to rule them all, 2020.

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.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.

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.

Neng Wang, Hongyang Yang, and Christina Dan Wang. Fingpt: Instruction tuning benchmark for open-source large language models in financial datasets. *arXiv preprint arXiv:2310.04793*, 2023.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

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.

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.

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.

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.

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.

Weimin Wu, Jiayuan Fan, Tao Chen, Hancheng Ye, Bo Zhang, and Baopu Li. Instance-aware model ensemble with distillation for unsupervised domain adaptation. *arXiv preprint arXiv:2211.08106*, 2022.

Xinhua Wu, Haoyu He, Yanchao Wang, and Qi Wang. Pre-trained mobility transformer: A foundation model for human mobility. *arXiv preprint arXiv:2406.02578*, 2024c.

Yuchen Wu, Minshuo Chen, Zihao Li, Mengdi Wang, and Yuting Wei. Theoretical insights for diffusion guidance: A case study for gaussian mixture models. In *Forty-first International Conference on Machine Learning*, 2024d.

Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. Smoothquant: Accurate and efficient post-training quantization for large language models. In *International Conference on Machine Learning*, pages 38087–38099. PMLR, 2023.

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)*, 2024a.

Zhuoyan Xu, Zhenmei Shi, Junyi Wei, Fangzhou Mu, Yin Li, and Yingyu Liang. Towards few-shot adaptation of foundation models via multitask finetuning. *arXiv preprint arXiv:2402.15017*, 2024b.

Zhewei Yao, Reza Yazdani Aminabadi, Minjia Zhang, Xiaoxia Wu, Conglong Li, and Yuxiong He. Zeroquant: Efficient and affordable post-training quantization for large-scale transformers. *Advances in Neural Information Processing Systems*, 35:27168–27183, 2022.

Jiahao Yu, Xingwei Lin, and Xinyu Xing. Gptfuzzer: Red teaming large language models with auto-generated jailbreak prompts. *arXiv preprint arXiv:2309.10253*, 2023a.

Jiahao Yu, Haozheng Luo, Jerry Yao-Chieh Hu, Wenbo Guo, Han Liu, and Xinyu Xing. Enhancing jailbreak attack against large language models through silent tokens. *arXiv preprint arXiv:2405.20653*, 2024.

Zhongzhi Yu, Shang Wu, Yonggan Fu, Shunyao Zhang, and Yingyan Celine Lin. Hint-aug: Drawing hints from foundation vision transformers towards boosted few-shot parameter-efficient tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11102–11112, 2023b.

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.

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.

Yu Zhang, Mei Di, Haozheng Luo, Chenwei Xu, and Richard Tzong-Han Tsai. Smutf: Schema matching using generative tags and hybrid features. *arXiv preprint arXiv:2402.01685*, 2024a.

Yufeng Zhang, Fengzhuo Zhang, Zhuoran Yang, and Zhaoran Wang. What and how does in-context learning learn? bayesian model averaging, parameterization, and generalization. *arXiv preprint arXiv:2305.19420*, 2023.

Zhi Zhang, Weijian Li, and Han Liu. Multivariate time series forecasting by graph attention networks with theoretical guarantees. In *International Conference on Artificial Intelligence and Statistics*, pages 2845–2853. PMLR, 2024b.

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.

Yuqi Zhou, Sunhao Dai, Liang Pang, Gang Wang, Zhenhua Dong, Jun Xu, and Ji-Rong Wen. Source echo chamber: Exploring the escalation of source bias in user, data, and recommender system feedback loop. *arXiv preprint arXiv:2405.17998*, 2024a.

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.

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*, 2024b.

Fengbin Zhu, Wenqiang Lei, Chao Wang, Jianming Zheng, Soujanya Poria, and Tat-Seng Chua. Retrieving and reading: A comprehensive survey on open-domain question answering. *arXiv preprint arXiv:2101.00774*, 2021.

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.

Supplementary Material

- **Appendix A. Related Works**
- **Appendix B. Additional Numerical Experiments**

A. Related Works

Associative Memory Models for Deep Learning. The classical Hopfield models (Hopfield, 1984; 1982; Krotov and Hopfield, 2016) emulate the associative memory functions of the human brain, emphasizing the storage and retrieval of distinct memory patterns. Recent renewed interest in associative memory models can be attributed to (i) improvements in memory storage capabilities (Wu et al., 2024a; Chaudhry et al., 2023; Krotov and Hopfield, 2016; Demircigil et al., 2017), (ii) innovative architectural developments (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). Modern associative memory networks, or contemporary Hopfield models (Hu et al., 2024a;b; 2023; Wu et al., 2024b; Burns and Fukai, 2023; Brandstetter, 2021; Ramsauer et al., 2020), demonstrate advantageous properties such as rapid convergence and exponential memory capacity. These models create a link to Transformer architectures (Hu et al., 2024b; Gu et al., 2024c; Hu et al., 2023; Wu et al., 2024b; Cabannes et al., 2024; Bietti et al., 2023; Ramsauer et al., 2020), effectively acting as advanced extensions of attention mechanisms. As a result, their applications span various domains, including drug discovery (Schimunek et al., 2023), immunology (Widrich et al., 2020), time series forecasting (Wu et al., 2024b; Auer et al., 2023; Zhang et al., 2024b), tabular learning (Xu et al., 2024a), out-of-distribution detection (Hofmann et al., 2024), reinforcement learning (Paischer et al., 2022), and computer vision (Fürst et al., 2022). Our study advances this research direction by focusing on more efficient models. We believe this work is essential for steering future research towards a Hopfield-driven design paradigm, particularly for large-scale models.

Outlier-Efficient Methods. Quantization is a technique used to lessen the computational demands of expansive models via low-bit precision computing (Huang et al., 2024; Qin et al., 2024; Luo et al., 2023b; Horowitz, 2014; Tang and Kwan, 1993; Marchesi et al., 1993). Common quantization strategies, such as INT8 and INT4, reduce the models' weights and activations to 8-bit or 4-bit integers, respectively (Chee et al., 2024; Lin et al., 2024; Xiao et al., 2023; Kim et al., 2023; Dettmers and Zettlemoyer, 2023; Frantar et al., 2022; Wei et al., 2022; Yao et al., 2022; Dettmers et al., 2022; Zafrir et al., 2019; Bhandare et al., 2019; Junczys-Dowmunt et al., 2018). Nonetheless, the quantization efficacy of transformer-based models is often hampered by the presence of outliers, which lead to disproportionately large attention weights (Bondarenko et al., 2023; 2021). To address this, Wei et al. (2022) revise LayerNorm to facilitate the quantization of activation tensors devoid of outliers and introduced Token-Wise Clipping to optimize the clipping ranges for each token. Dettmers et al. (2022) apply varying degrees of precision to quantize outlier features and other features. Additionally, Meo et al. (2024) adopt a Bayesian perspective by employing a prior distribution on quantization levels, effectively helping in mitigating outliers. Despite these advancements, since outliers originate from the $\mathrm{Softmax}$ function, these methods do not tackle the root cause of the issue. In response, Bondarenko et al. (2023) develope `Clipped_Softmax` and `Gated_Attention`, which enforce the attention mechanism to produce exact zeros, thus addressing the source of outliers. Specifically, `Clipped_Softmax` expands the output range of the softmax function beyond (0,1), and `Gated_Attention` decides whether to retain or eliminate updates. However, these methods require hyperparameter tuning for optimal performance, with `Clipped_Softmax` showing suboptimal results in the OPT model and `Gated_Attention` adding extra training parameters. In our paper, we introduce a novel approach using the modern Hopfield model, which inherently supports outlier-efficient computation. Surprisingly, its retrieval dynamics include $\mathrm{Softmax}_1$ outlier-efficient attention as a specific instance[2]. Preliminary experimental findings (johnowhitaker, 2023) validate its efficacy in managing outliers. We anticipate our work illuminate the theoretical and methodological research into (Hopfield-based) large foundation models.

Transformer-Based Foundation Models. In recent years, foundation models achieve significant advancements within the field of artificial intelligence, concentrating on diverse key research areas such as reasoning (Zhou et al., 2024a; Pan et al., 2024a;b; Wang et al., 2022), question and answering (Zhu et al., 2021; Luo et al., 2021; Qin et al., 2021; Perez et al., 2020), safety (Luo et al., 2024; Yu et al., 2024; 2023a), prompting (Jin et al., 2024; Liu et al., 2023; Lester et al., 2021; Gao et al., 2020), multi-modality (Liu et al., 2024; Girdhar et al., 2023; Luo et al., 2023a; Samel et al., 2021), theory (Li et al.,

[2]For any $x \in \mathbb{R}^d$, $\mathrm{Softmax}_1(x)_i = \frac{\exp(x_i)}{1+\sum_j \exp(x_j)}$.

2024a;b; Gu et al., 2024d;b; Chen et al., 2024; Fu et al., 2024; Hu et al., 2024e; Dou et al., 2024; Guo et al., 2024; Wu et al., 2024d; Zhang et al., 2023), data cleaning (Zhang et al., 2024a; Ahmad et al., 2023; Liu et al., 2022) and parameter-efficient fine-tuning (PEFT) (Dettmers et al., 2024; Yu et al., 2023b; Wu et al., 2022; Hu et al., 2021). They hold a central position not only in machine learning but also across various scientific fields, prominently including natural language processing (Touvron et al., 2023a;b; Jiang et al., 2023; Le Scao et al., 2023; Floridi and Chiriatti, 2020; Brown et al., 2020), vision (Saharia et al., 2022; Ramesh et al., 2022; Dosovitskiy et al., 2020), finance (Wang et al., 2023; Wu et al., 2023), genomics (Zhou et al., 2024b; 2023; Ji et al., 2021), human mobility (Wu et al., 2024c) and many others.

Outlier Related Transformer Theories. Recent studies demonstrate the benefits of outlier removal from attention heads in large transformer-based foundation models. Alman and Song (2023) demonstrate that efficient transformers, including vanilla and tensor versions, require bounded attention weights through precise reduction methods. Hu et al. (2024c) indicate that efficient modern Hopfield models and their networks also require bounded query and key patterns for sub-quadratic time complexity using fine-grained reduction techniques. Additionally, Hu et al. (2024d) theoretically show that the existence of outliers hamper the efficiency and performance of LoRA fine-tuning. Further, Gu et al. (2024a;c); Alman and Song (2024); Gao et al. (2023) find that bounded weight matrices are essential for the efficient training of transformer-based models.

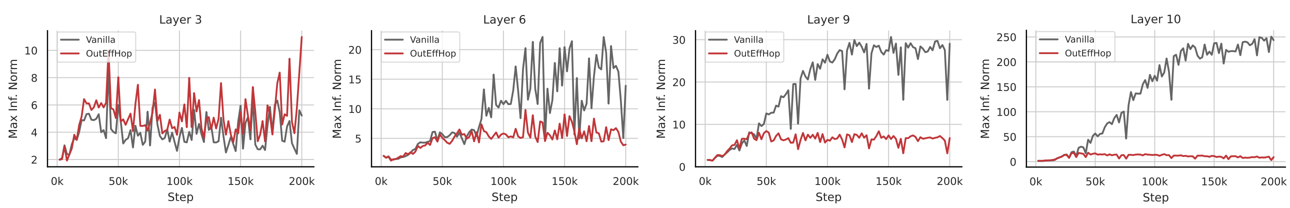


Figure 2. The trend of Feed-Forward Network (FFN) output maximum infinity norm values in layers 3, 6, 9, and 10 of a BERT encoder is analyzed using two softmax variations: `OutEffHop` (represented in red) and vanilla $\mathrm{Softmax}$ (in grey). The findings indicate that `OutEffHop` significantly reduces outliers in the model compared to the vanilla $\mathrm{Softmax}$.

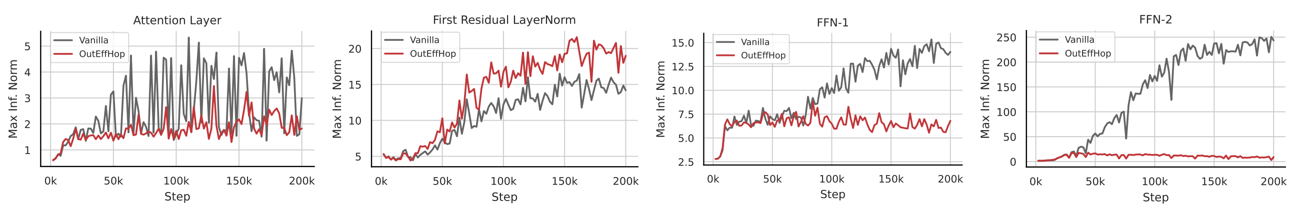


Figure 3. Maximum infinity norm $\|x\|_\infty$ for different tensor components within layer 10 of BERT. Our work is analysed using two softmax variations: `OutEffHop` (represented in red) and vanilla $\mathrm{Softmax}$ (in grey). We find `OutEffHop` suppresses the outliers growing in both FFN layers.

B. Additional Numerical Experiments

B.1. Supplemental Experimental Results (Figure 2 and Figure 3)

We conduct extensive case studies on the BERT model. In Figure 2, we analyze the outlier performance across different layers, observing an increase in outlier strength in the deeper layers of the standard model, consistent with the observations by Bondarenko et al. (2021). The `OutEffHop` model demonstrates robust control over the maximum infinity norm $\|x\|_\infty$ across all layers, highlighting its effective outlier management capabilities. In Figure 3, we assess the maximum infinity norm $\|x\|_\infty$ in the 10th layer's components—post-attention layer, initial residual LayerNorm following attention, and the first and second FFN layers. As noted by (Bondarenko et al., 2023), FFN layers significantly contribute to outlier amplification during training in standard attention models. In contrast, `OutEffHop` limits this growth in both FFN layers by employing a no-operation (no-op) mode that engages when updates are unnecessary, thus preventing the inadvertent learning of outlier values. Furthermore, the initial residual LayerNorm post-attention is observed to exacerbate outliers, a phenomenon also reported in Wei et al. (2022)'s research. Despite this, `OutEffHop`, primarily focusing on the attention mechanism, demonstrates effective reduction of outliers, showcasing its potential in our model.

B.2. Verifying Theoretical Results

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

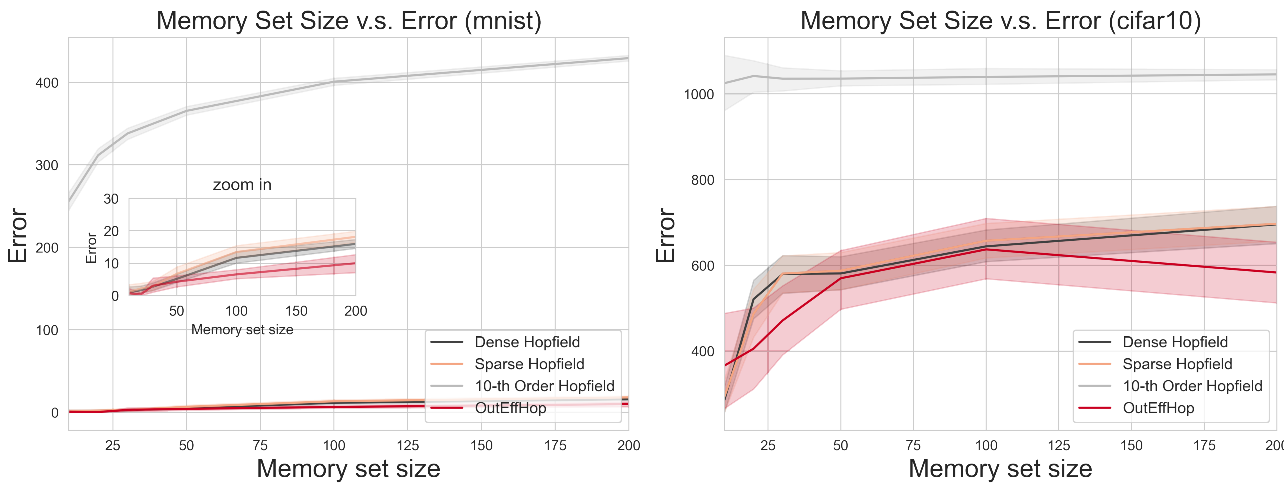


Figure 4. **Memory Capacity.** Our extensive evaluation of memory capacity across various Hopfield Networks, including Vanilla Modern Hopfield, Sparse Hopfield, 10th Order Hopfield, and our `OutEffHop`, is conducted on two image datasets: MNIST and CIFAR10. We observe that `OutEffHop` outperforms its baselines, especially when the memory set size is large.

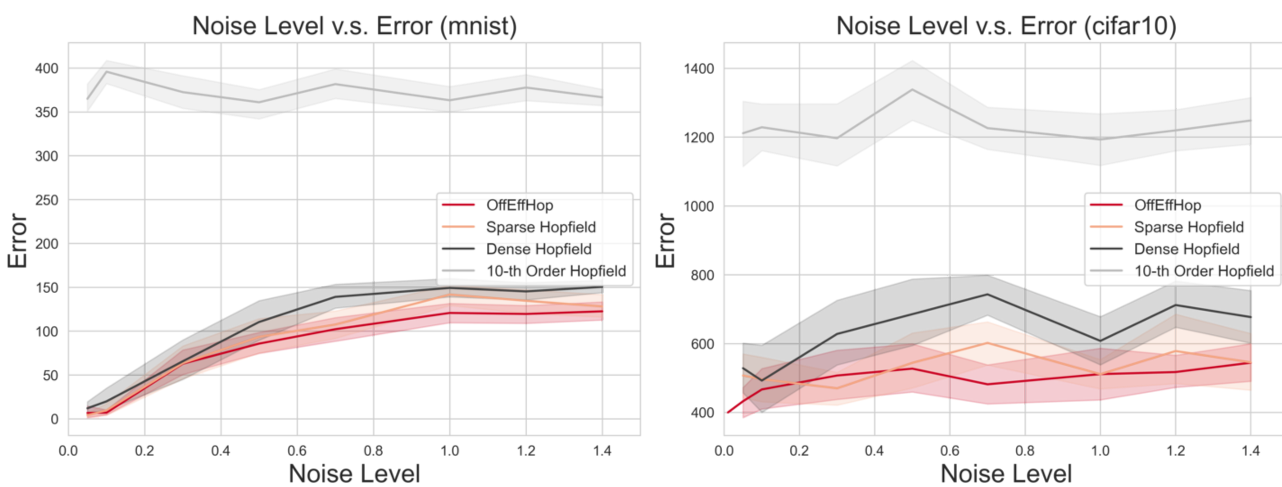


Figure 5. **Noise-Robustness.** Our extensive evaluation of noise robustness across various Hopfield Networks, including Vanilla Modern Hopfield, Sparse Hopfield, 10th Order Hopfield, and our `OutEffHop`, is conducted on two image datasets: MNIST and CIFAR10. The results show that as the noise level rises, the impact of `OutEffHop` on the error rate is minimal.

Memory Capacity. For memory capacity evaluation, we contrast our Outlier-Efficient Modern Hopfield Model (`OutEffHop`) with traditional Dense (Softmax) (Ramsauer et al., 2020), Sparse (Hu et al., 2023), and 10th order polynomial Hopfield models (Krotov and Hopfield, 2016) using the MNIST (LeCun et al., 1998) (high sparsity) and CIFAR10 (Krizhevsky et al., 2009) (low sparsity) datasets. In all Hopfield models, we employ a fixed $\beta = 1$. As depicted in Figure 4, `OutEffHop` surpasses its counterparts, particularly noticeable when the memory set size is extensive.

Noise-Robustness. For the robustness against noise queries, we inject Gaussian noises varying variances (σ) into the images. The results, as shown in Figure 5, show that `OutEffHop` excels when the signal-to-noise ratio in patterns is low.

Faster Convergence. We numerically analyze the convergence of `OutEffHop` alongside the Dense and Sparse Hopfield models by assessing their loss and accuracy on two distinct datasets. We employ the Vision Transformer (ViT) (Dosovitskiy et al., 2020) as the backbone architecture, replacing its attention layer with various Hopfield layers. The hyperparameters utilized in our experiments are detailed in Table 2. As illustrated in Figure 6, our model consistently outperforms its original counterpart across all datasets.

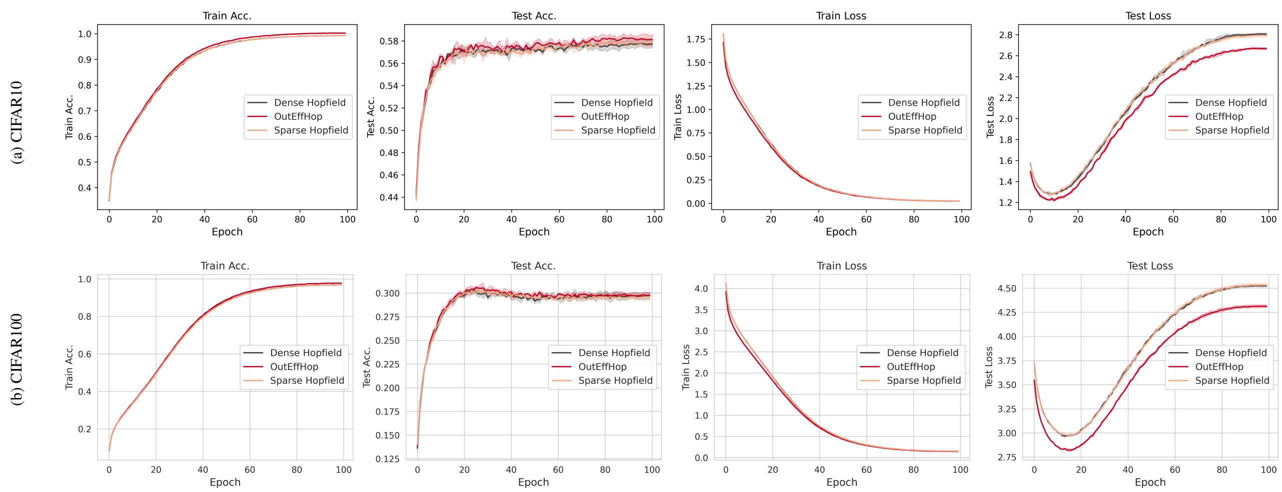


Figure 6. **Faster Convergence.** Our extensive evaluation of faster covergence across various Hopfield Networks, including Vanilla Modern Hopfield, Sparse Hopfield, and our `OutEffHop`, is conducted on two image datasets: CIFAR10 and CIFAR100. The results show that `OutEffHop` has faster convergence than baselines.

Table 2. 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

B.3. Computational Cost Comparison

We evaluate the computational resource utilization of four different models compared to the vanilla $\mathrm{Softmax}$ and `OutEffHop`, as outlined in Table 3. We document the pre-training metrics for all models. Memory usage for the OPT, BERT, and ViT models is monitored using Wandb[3], while for the STanHop model, it is tracked via system logs. The model configurations for this experiment are as described in Section 4.1. Our experimental infrastructure includes a Slurm system equipped with two 80G A100 GPUs and a 24-core Intel(R) Xeon(R) Gold 6338 CPU at 2.00GHz. Additionally, the Wandb diagram illustrating the system memory usage is presented in Figure 7.

Table 3. 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

B.4. `OutEffHop` Improves Hopfield-Centric Deep Learning Model: A Case Study on STanHop-Net

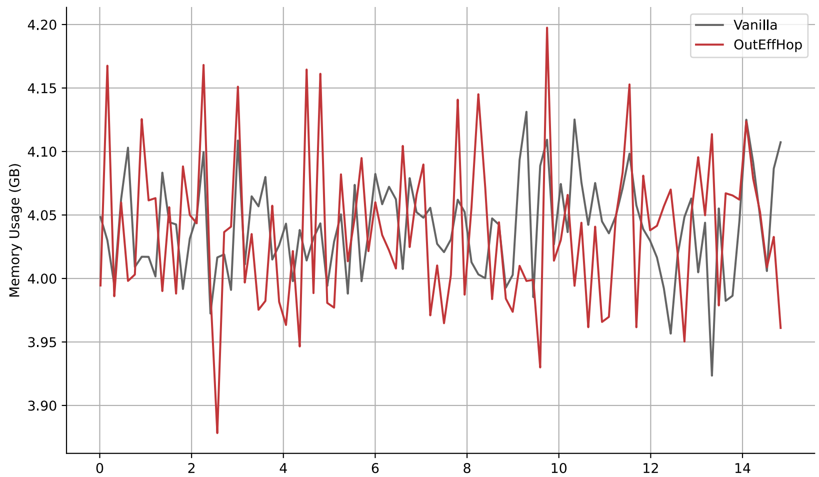
We also test our method on STanHop-Net (Wu et al., 2024b), a Hopfield-based time series prediction model. We compare our method with common modern Hopfield layers (Hu et al., 2023; Ramsauer et al., 2020).

Data. Following Wu et al. (2024b); Zhang et al. (2024b), we employ three realistic datasets for our multivariate time series prediction tasks: ETTh1 (Electricity Transformer Temperature-hourly), ETTm1 (Electricity Transformer Temperature-minutely), and WTH (Weather). These datasets are partitioned into training, validation, and test sets with a ratio of 14/5/5. For each dataset, we perform evaluations across a range of prediction horizons.

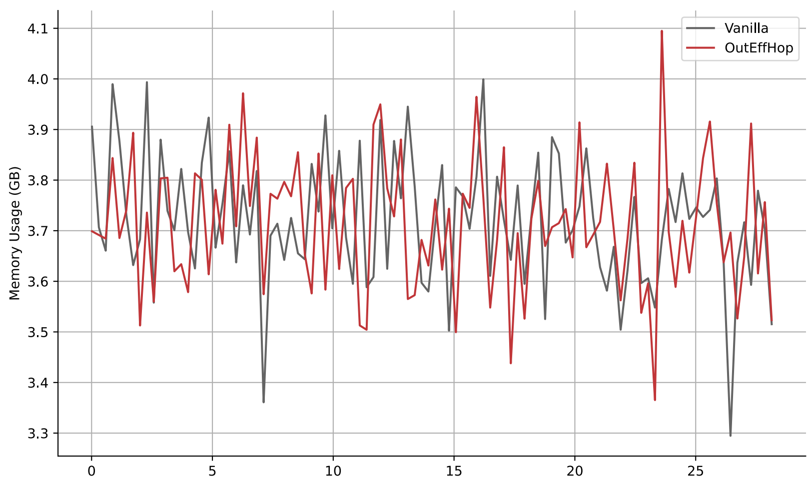
Metrics. To assess outlier efficiency, we employ the same metrics as used in previous experiments: the maximum infinity norm $\|x\|_\infty$ and *average kurtosis* across 12 decoder layers. For evaluating prediction accuracy, we utilize Mean Squared Error (MSE) and Mean Absolute Error (MAE). Each experiment is conducted ten times to ensure reliability, and the results reported are the averages of these runs.

Results. In Table 4, our findings highlight the efficacy of `OutEffHop` in augmenting the outlier efficiency of modern Hopfield network architectures. `OutEffHop` achieves significant enhancements in outlier efficiency with only a minor compromise in model performance. It secures top-tier outlier efficiency in 25 out of 30 evaluated scenarios, ranking either first or second in these assessments. Within the STanHop-Net framework, the `OutEffHop` model exhibits a noticeable improvement in outlier efficiency relative to the Vanilla and Sparse, Generalized Sparse Modern Hopfield Models. This includes reductions of 3% and 4% in $\|x\|_\infty$ and average kurtosis, respectively.

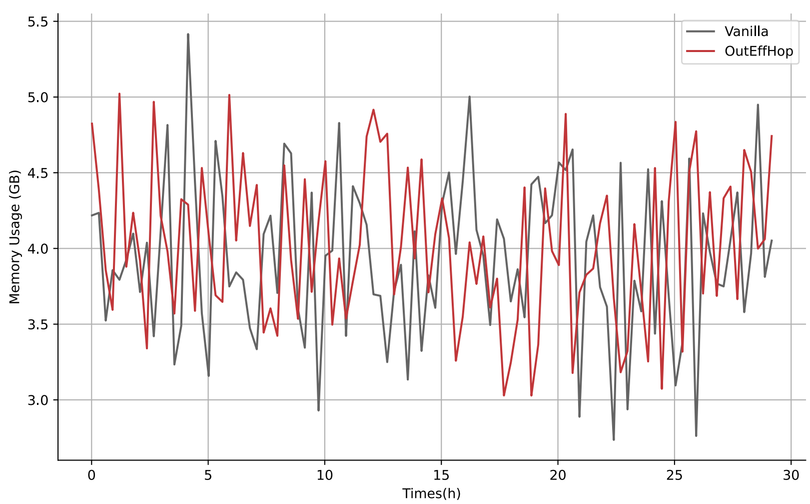
[3] https://wandb.ai/



(a) BERT



(b) OPT



(c) ViT

Figure 7. The computational resource comparison between Vanilla $\mathrm{Softmax}$ and `OuTEffHop` involves measuring RAM usage via Wandb in a system equipped with 180G RAM under the Slurm system.

Table 4. **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`			
Dataset	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