Outlier-Efficient Hopfield Layers for Large Transformer-Based Models

Jerry Yao-Chieh Hu*, Pei-Hsuan Chang*, Haozheng Luo*,
Hong-Yu Chen, Weijian Li, Wei-Po Wang, Han Liu

Northwestern University, Computer Science

https://arxiv.org/abs/2404.03828

ICML 2024

Northwestern University

Introduction

Problem: Transformer-based models encounter outlier-inefficient problems.
Proposal: Outlier-Efficient Modern Hopfield Model (termed `OutEffHop`), a quantization-friendly modern Hopfield model.

- Serves as an outlier-efficient alternative for vanilla attention
- Retains and improves the desirable properties of modern Hopfield model, such as fix-point convergence and exponential memory capacity
- Outperforms across 3 large transformer-based and 1 Hopfield-based models

Motivation: Outliers and Hopfield Model

No-Update Situation: When an input X is informative enough, the attention mechanism (attention + residual) tends to behave like a identity map (Bondarenko et al., 2023). $\rightarrow$ A should be a zero matrix

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

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

- Low-information tokens attract large attention probability to achieve no-update
- Causes wide range of $QK^\intercal$, namely no-op outliers

Modern Hopfield Model: An alternative to attention mechanism in Transformer. This is an energy-based model with:

- The retrieval dynamics mirrors attention mechanism and leads to its compatibility with deep learning architecture
- Fix-point convergence and exponential memory capacity

Methodology: No-op Classification Mechanism

We introduce an extra dimension on query and memory patterns for outlier classification, and propose a novel outlier-efficient Hopfield model.

Add an extra dimension on query and memory for outlier classification

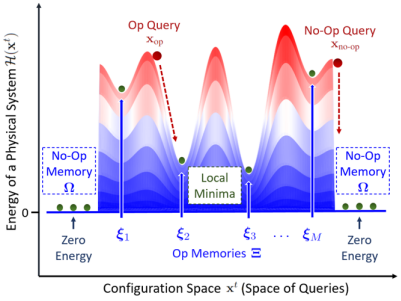
$$\mathbf{x} \leftarrow (x_1, \ldots, x_d, 0), \quad \xi_\mu \leftarrow (\xi_1^\mu, \cdots, \xi_d^\mu, \omega)$$

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

The no-op classification function

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

The outlier-efficient Hopfield model has zero energy point as no-op memory



If the input is a no-op query, it will converge to a no-op memory Ω instead of a low-similarity memory pattern.

Outlier-Efficient Modern Hopfield Model

From no-op classification mechanism, we find the corresponding Hopfield energy and retrieval dynamics:

$$\mathcal{H}(x) = -\mathrm{lse}_1(\beta, \Xi^\intercal x) + \frac{1}{2}\langle x, x\rangle + \text{Const.},\ \mathrm{lse}_1(\beta, \Xi^x) := \beta^{-1}\log\Big(\sum_{\mu=1}^{M}\exp(\beta\langle \xi_\mu, \mathbf{x}\rangle) + 1\Big)$$

$$\mathcal{T}_{\mathsf{OutEff}}(x_t) := \Xi\mathrm{Softmax}_1(\beta\Xi^\intercal x_t) = x_{t+1},\ \mathrm{Softmax}_1 := \exp(z)/\Big(\sum_{\mu=1}^{M}\exp(z_\mu) + 1\Big)$$

1. Our model has a zero-energy point associated with $\exp\{\beta\langle\Omega, x\rangle\} = \exp\{0\}$.
2. When $\mathcal{T}_{\mathsf{OutEff}}$ is applied only once, the retrieval dynamics is equivalent to outlier-efficient attention (Miller, 2023).

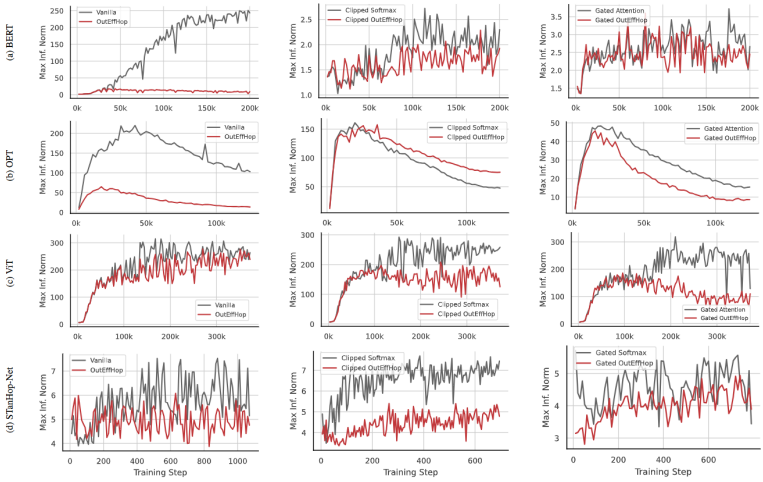
Experimental Studies: Outlier-Efficiency and Quantization Results

Compare `OutEffHop` with the vanilla attention and their combination with `Clipped_Softmax` and `Gated_Attention`.

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	

Results: `OutEffHop` reduce 22+% in average kurtosis and 26+% in $\|x\|_\infty$ across four models, and the performance of models maintain the same after quantization.

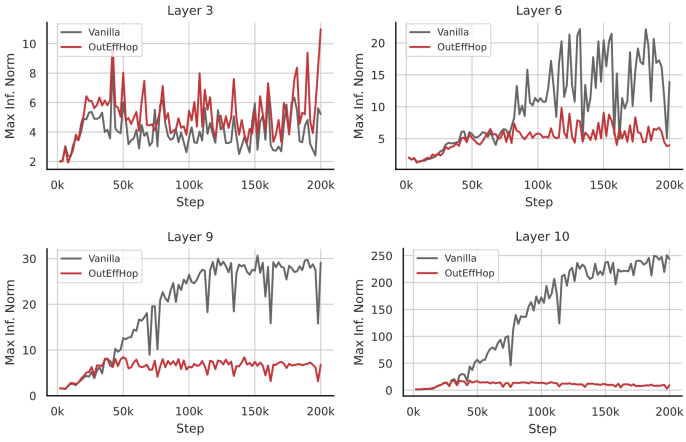
Experimental Studies: The Changes of $\|x\|_\infty$ during Pretraing



Results: In vanilla attention, the outliers grow during the pretraining stage, while OutEffHop delivers significant reductions of $\|x\|_\infty$ and improves `Clipped_Softmax` and `Gated_Attention`.

Experimental Studies: Outlier Performance in Selected Layers.

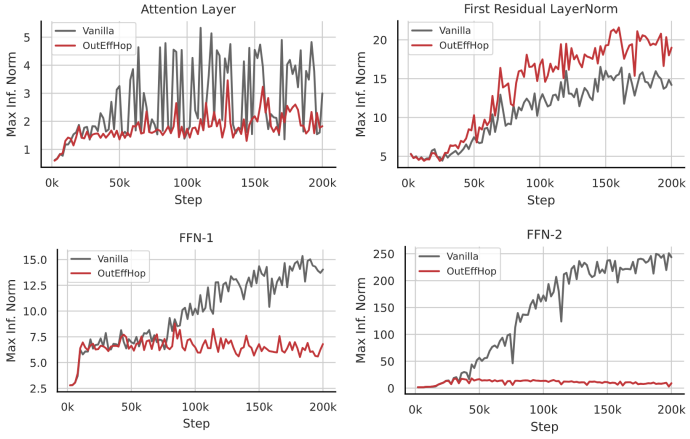
The trend of $\|x\|_\infty$ of FFN output in layers 3, 6, 9, and 10 of BERT model.



Results: Outliers become stronger in deeper layers of the vanilla model, but OutEffHop maintains a consistent value of $\|x\|_\infty$ across all layers.

Experimental Studies: Outlier for different tensor components.

The trend of $\|x\|_\infty$ of the tensors after attention layer, the first residual layernorm after attention, and the first, second FFN layers within layer 10 of BERT.



Results: OutEffHop suppresses the outliers growing in both FFN layers.

Experimental Studies: Case Study on STanHop-Net.

We test our method on STanHop-Net, a Hopfield-based time series model, and compare it with common Hopfiled layers.

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

Results: `OutEffHop` achieves top-tier outlier-efficiency in 25 out of 30 evaluated scenarios, ranking either first or second with marginal sacrifice of model performance.

Summary

- Outlier-Efficient Modern Hopfield Model
  - Manages outliers in large transformer-based models
  - Strong physics intuition and efficient memory retrievals
- No-op classification mechanism
  - Maps all no-op patterns into Ω
  - Prevents no-op patterns from contributing to retrieval output
- Theoretical enhancements
  - Improves fixed point convergence and exponential memory capacity
  - Provides model-based interpretation for $Softmax_1$ attention (Miller, 2023)
- Empirical performance of `OutEffHop`
  - $\sim$22+% reduction in average kurtosis, $\sim$26+% reduction in $\|x\|_\infty$, maintain performance after quantization
  - Top two in outlier efficiency in most settings compared to 3 variants of Hopfield-based time series model

Thank You!

Jerry Yao-Chieh Hu*, Pei-Hsuan Chang*, Haozheng Luo*,
Hong-Yu Chen, Weijian Li, Wei-Po Wang, Han Liu

jhu@u.northwestern.edu
b09202022@ntu.edu.tw
robinluo2022@u.northwestern.edu
hong-yuchen2029@u.northwestern.edu
weijianli@u.northwestern.edu
b09202009@ntu.edu.tw
hanliu@northwestern.edu
http://magics.cs.northwestern.edu