Linformer & Performer

Gholamreza Dar - Amirhassan Amirmahani

Linformer

Linformer

Linformer: Self-Attention with Linear Complexity

Sinong Wang, Belinda Li, Madian Khabsa, Han Fang, Hao Ma

Facebook AI, USA

Transformer Complexity : $O(n^2)$

“ the representation of each token is updated by visiting all the other tokens that are present in the previous layer.”

- This is essential for retaining long-term information
- provides Transformers with the edge over recurrent models on long sequences

Model Architecture	Complexity per Layer	Sequential Operation
Recurrent	$O(n)$	$O(n)$
Transformer, (Vaswani et al., 2017)	$O(n^2)$	$O(1)$
Sparse Tansformer, (Child et al., 2019)	$O(n\sqrt{n})$	$O(1)$
Reformer, (Kitaev et al., 2020)	$O(n \log(n))$	$O(\log(n))$
Linformer	$O(n)$	$O(1)$

Table 1: Per-layer time complexity and minimum number of sequential operations as a function of sequence length (n) for various architectures.

Why Linformer?

- Training and deploying large Transformer models can be **COSTLY**
- The key efficiency bottleneck in standard Transformer models is the model's **self-attention mechanism**.

Answering "can Transformer models be optimized to avoid this quadratic operation, or is this operation required to maintain strong performance?"

Prior Work : introducing sparsity into attention layers

by having each token attend to only a subset of tokens in the whole sequence. This reduces the overall complexity of the attention mechanism to $O(n\sqrt{n})$

2% drop with only 20% speed up

Prior Work : Reformer

used locally-sensitive hashing (LSH) to reduce the self-attention complexity to O(nlog(n)).

efficiency gains only appear on sequences with length>2048

the Reformer's multi-round hashing approach actually increases the number of sequential operations, which further undermines their final efficiency gains.

Main idea of the paper

inspired by the key observation that **self-attention is low rank**.

“attention dot product matrix can always be closely approximated by a matrix of much lower **rank**.”

Which can be obtained by calculating the singular value decomposition (SVD) of the key and value matrices

https://en.wikipedia.org/wiki/Singular_value_decomposition

Rank

In linear algebra, the rank of a matrix A is the dimension of the vector space generated (or spanned) by its columns.

-wikipedia

the number of linearly independent rows or columns

$$\begin{bmatrix} 1 & -1 \\ 1 & -1 \\ 0 & 0 \\ 2 & -2 \end{bmatrix}$$

Rank : 1

$$\begin{bmatrix} 1 & 0 & 1 \\ -2 & -3 & 1 \\ 3 & 3 & 0 \end{bmatrix}$$

Rank : 2

The Observation

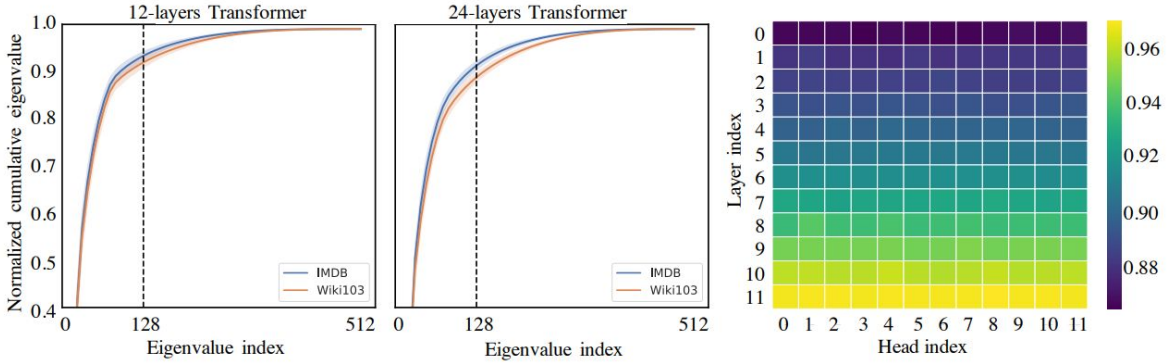


Figure 1: Left two figures are spectrum analysis of the self-attention matrix in pretrained transformer model (Liu et al., 2019) with $n = 512$. The Y-axis is the normalized cumulative singular value of context mapping matrix P, and the X-axis the index of largest eigenvalue. The results are based on both RoBERTa-base and large model in two public datasets: Wiki103 and IMDB. The right figure plots the heatmap of normalized cumulative eigenvalue at the 128-th largest eigenvalue across different layers and heads in Wiki103 data.

Benchmark results

- Linformer is faster but usually less performant (in some cases outperforms transformer)
- “the performance of Linformer model is mainly determined by the projected dimension k instead of the ratio n/k. “

n	Model	SST-2	IMDB	QNLI	QQP	Average
512	Liu et al. (2019), RoBERTa-base	93.1	94.1	90.9	**90.9**	92.25
	Linformer, 128	92.4	94.0	90.4	90.2	91.75
	Linformer, 128, shared kv	**93.4**	93.4	90.3	90.3	91.85
	Linformer, 128, shared kv, layer	93.2	93.8	90.1	90.2	91.83
	Linformer, 256	93.2	94.0	90.6	90.5	92.08
	Linformer, 256, shared kv	93.3	93.6	90.6	90.6	92.03
	Linformer, 256, shared kv, layer	93.1	94.1	**91.2**	90.8	**92.30**
512	Devlin et al. (2019), BERT-base	92.7	93.5	91.8	89.6	91.90
	Sanh et al. (2019), Distilled BERT	91.3	92.8	89.2	88.5	90.45
1024	Linformer, 256	93.0	93.8	90.4	90.4	91.90
	Linformer, 256, shared kv	93.0	93.6	90.3	90.4	91.83
	Linformer, 256, shared kv, layer	93.2	**94.2**	90.8	90.5	92.18

Table 2: Dev set results on benchmark natural language understanding tasks. The RoBERTa-base model here is pretrained with same corpus as BERT.

Inference-time Efficiency Results

- benchmarking inference speed and memory on a 16GB Tesla V100 GPU card
- Left table shows time saved compared to a regular Transformer
- Left table shows memory saved compared to a regular Transformer

length n	projected dimensions k				
	128	256	512	1024	2048
512	1.5x	1.3x	-	-	-
1024	1.7x	1.6x	1.3x	-	-
2048	2.6x	2.4x	2.1x	1.3x	-
4096	3.4x	3.2x	2.8x	2.2x	1.3x
8192	5.5x	5.0x	4.4x	3.5x	2.1x
16384	8.6x	7.8x	7.0x	5.6x	3.3x
32768	13x	12x	11x	8.8x	5.0x
65536	20x	18x	16x	14x	7.9x

length n	projected dimensions k				
	128	256	512	1024	2048
512	1.7x	1.5x	-	-	-
1024	3.0x	2.9x	1.8x	-	-
2048	6.1x	5.6x	3.6x	2.0x	-
4096	14x	13x	8.3x	4.3x	2.3x
8192	28x	26x	17x	8.5x	4.5x
16384	56x	48x	32x	16x	8x
32768	56x	48x	36x	18x	16x
65536	60x	52x	40x	20x	18x

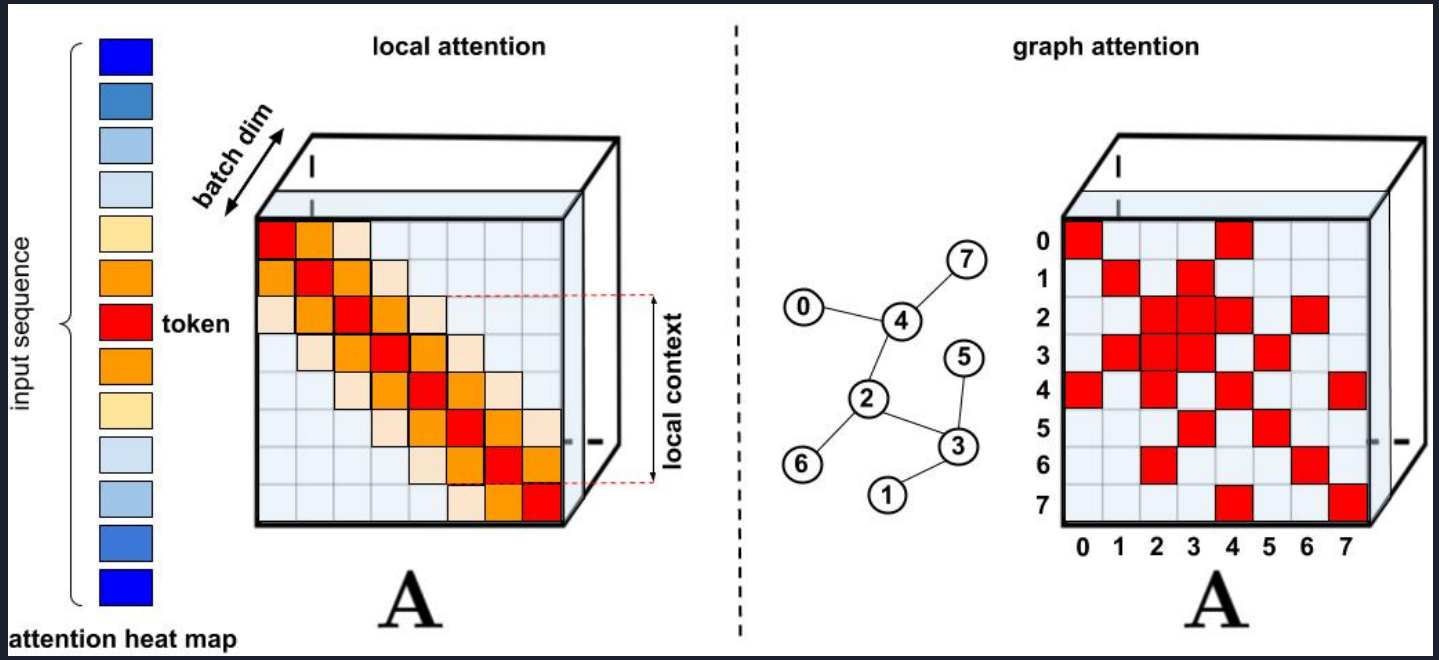
Table 3: Inference-time efficiency improvements of the Linformer over the Transformer, across various projected dimensions k and sequence lengths n. Left table shows time saved. Right table shows memory saved.

Broader Impact

- Can improve working with images drastically
- Can be more environment friendly due to less power consumption
- However may have potential ethical problems (with great power comes great responsibility)

Performer

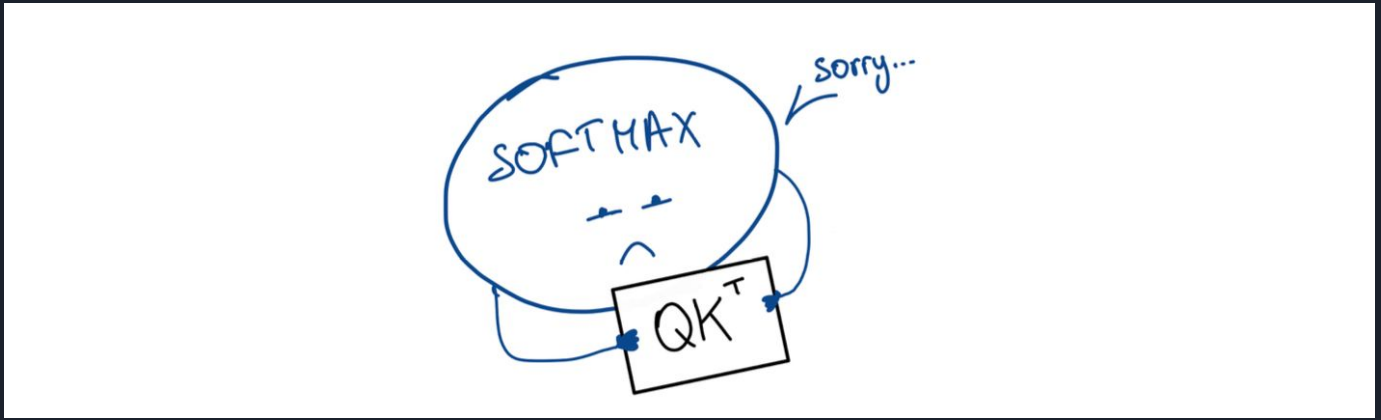
sparse attention



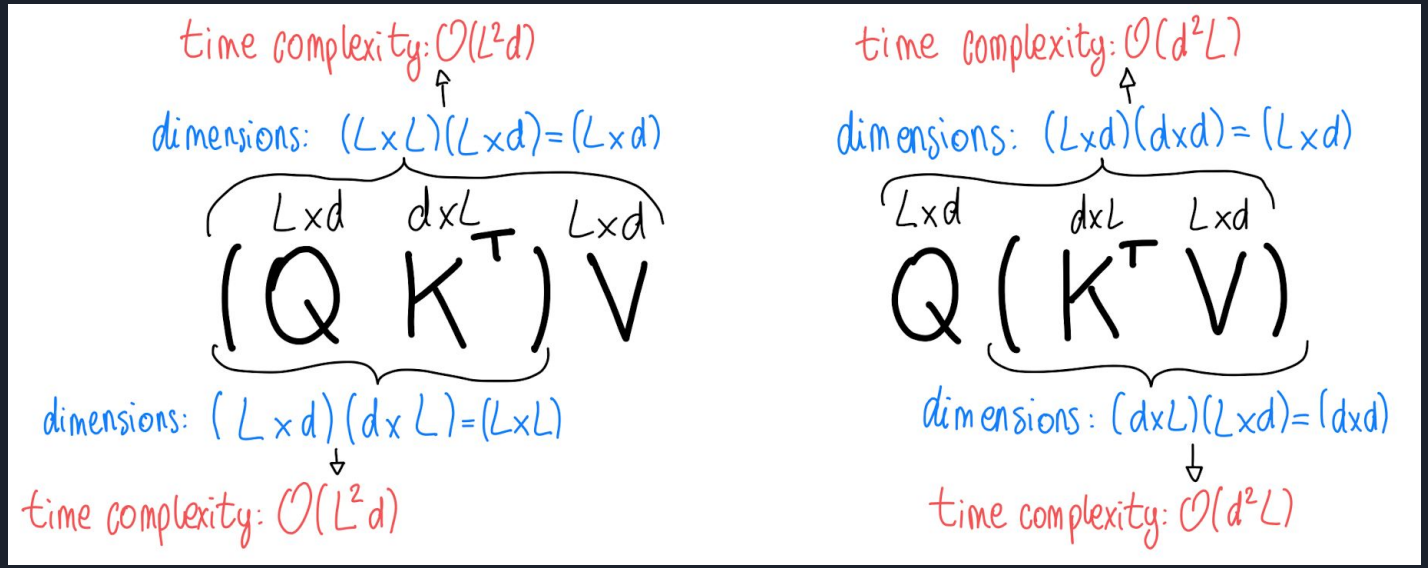
sparse attention limitations

1. They require efficient sparse-matrix multiplication operations, which are not available on all accelerators
2. they usually do not provide rigorous theoretical guarantees for their representation power
3. they are optimized primarily for Transformer models and generative pre-training
4. they usually stack more attention layers to compensate for sparse representations, making them difficult to use with other pre-trained models, thus requiring retraining and significant energy consumption

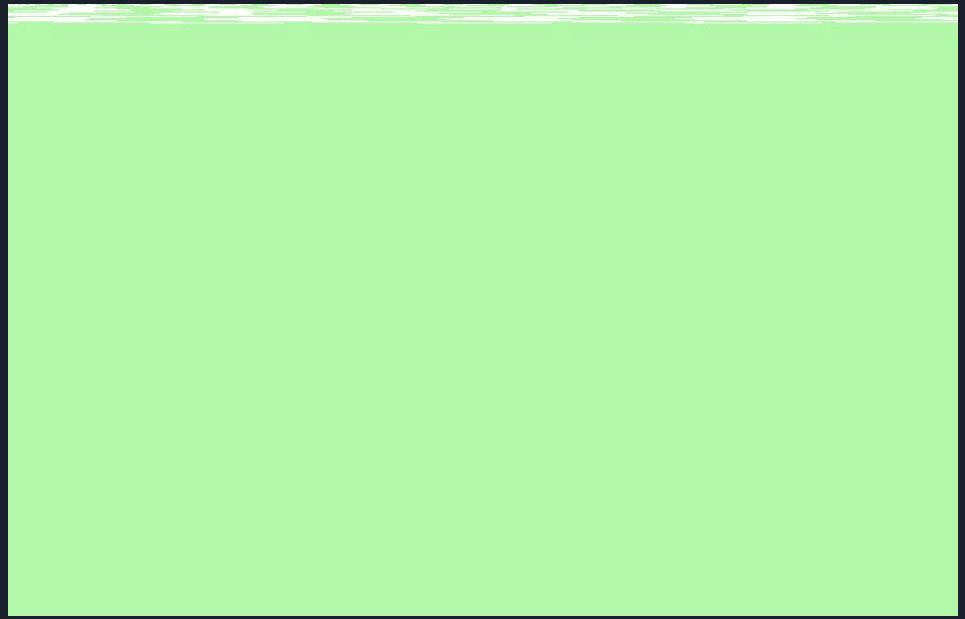
sparse attention limitations



Performer work style



Performer work style



Performer work style

$$\text{softmax}\left(\frac{\mathbf{QK}^\top}{\sqrt{d}}\right) = \mathbf{D}^{-1}\mathbf{A}$$

$$\text{with:} \quad \mathbf{A} = \exp\left(\mathbf{QK}^\top/\sqrt{d}\right), \quad \mathbf{D} = \text{diag}(\mathbf{A}\mathbf{1}_L)$$

$$\mathbf{A}(i,j) = \text{K}(\mathbf{q}_i, \mathbf{k}_j) = \exp(\mathbf{q}_i \mathbf{k}_j^\top) = \phi(\mathbf{q}_i)^\top \phi(\mathbf{k}_j)$$

Performer work style

$$\phi(\mathbf{x}) = \frac{h(\mathbf{x})}{\sqrt{m}}(f_1(w_1^\top \mathbf{x}), ..., f_1(w_m^\top \mathbf{x}), ..., f_l(w_1^\top \mathbf{x}), ..., f_l(w_m^\top \mathbf{x}))$$

$$h(\mathbf{x}) = 1, l = 2, f_1 = \sin, f_2 = \cos, \mathcal{D} = \mathcal{N}(0, \mathbf{I}_d)$$

$$h(\mathbf{x}) = \exp\left(\frac{||\mathbf{x}||^2}{2}\right)$$

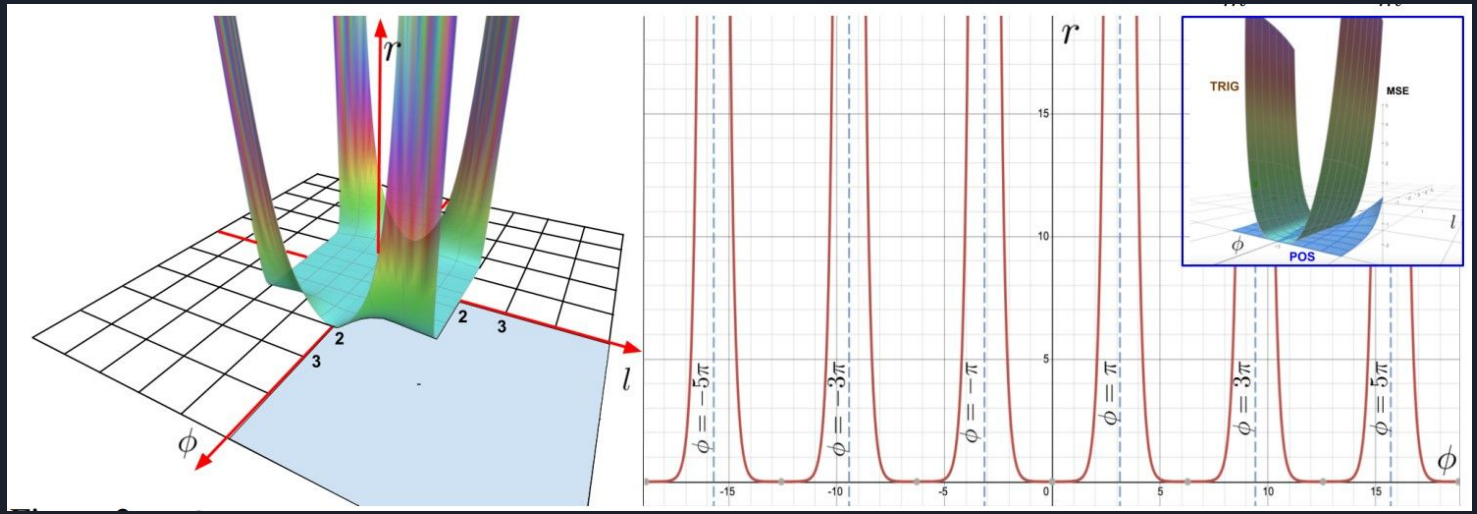
Performer work style

$$\mathrm{K_{SM}}(\mathbf{x},\mathbf{y}) = \exp(\mathbf{x}^\top \mathbf{y}) = \mathbb{E}_{w\sim\mathcal{N}(0,\mathbf{I}_d)}[\exp(w^\top \mathbf{x} - \frac{\|\mathbf{x}\|^2}{2})\exp(w^\top \|\mathbf{y}\| - \frac{\mathbf{y}^2}{2})]$$

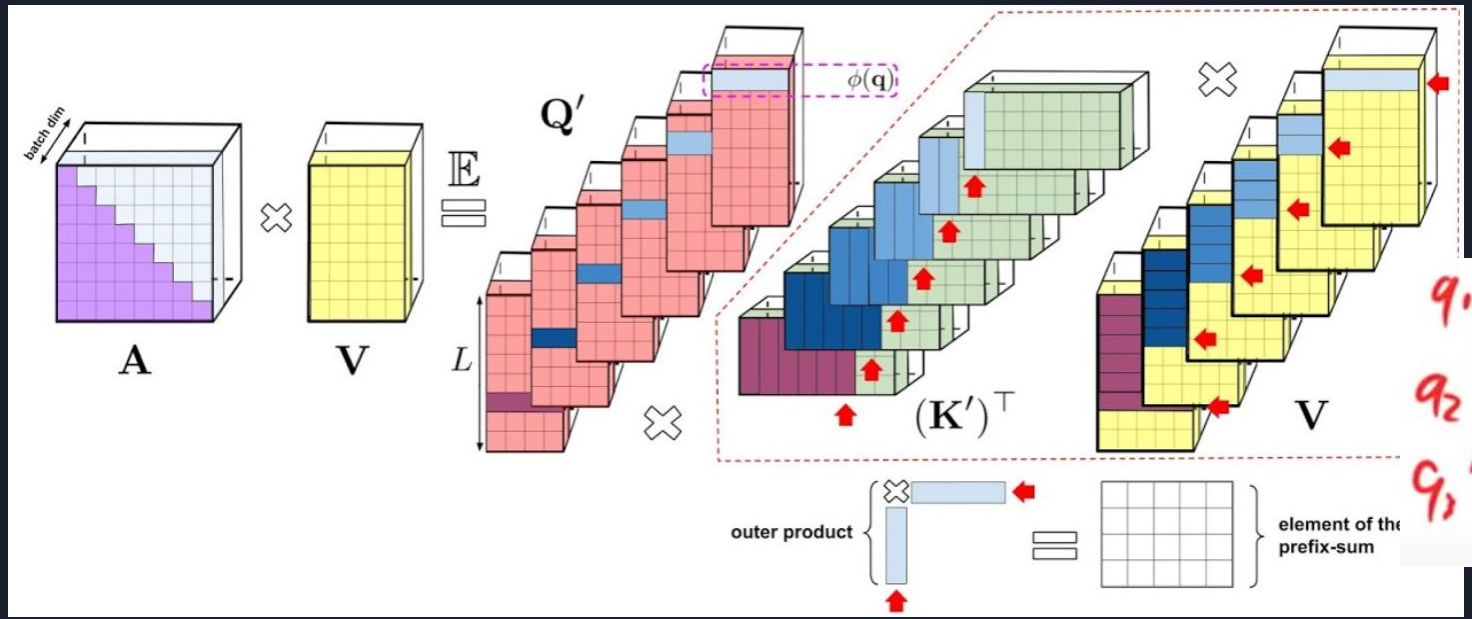
$$= \mathbb{E}_{w\sim\mathcal{N}(0,\mathbf{I}_d)}[\exp(-\frac{\|\mathbf{x}\|^2}{2})\exp(-\frac{\|\mathbf{y}\|^2}{2})\exp(w^\top \mathbf{x})\exp(w^\top \mathbf{y})$$

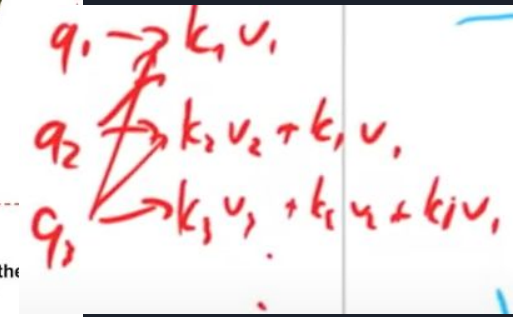
$$h(\mathbf{x}) = \exp(-\frac{||\mathbf{x}||^2}{2}), l = 1, f_1 = \exp, \mathcal{D} = \mathcal{N}(0, \mathbf{I}_d)$$

Problem

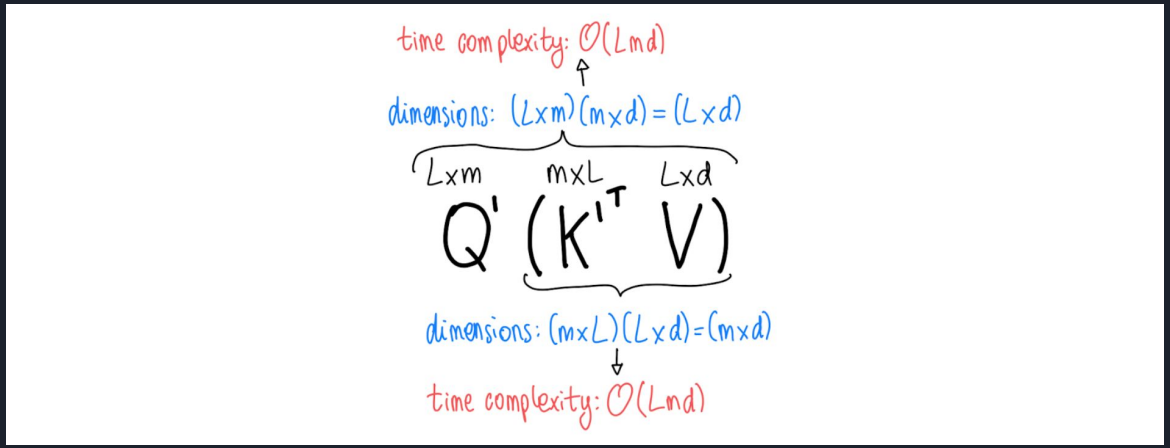

r
l
ϕ
2
3
2
3
r
15
10
5
ϕ = −5π
ϕ = −3π
ϕ = −π
ϕ = π
ϕ = 3π
ϕ = 5π
-15
-10
-5
0
5
10
15
ϕ
TRIG
MSE
POS
ϕ
l

Causal Attention via prefix sums

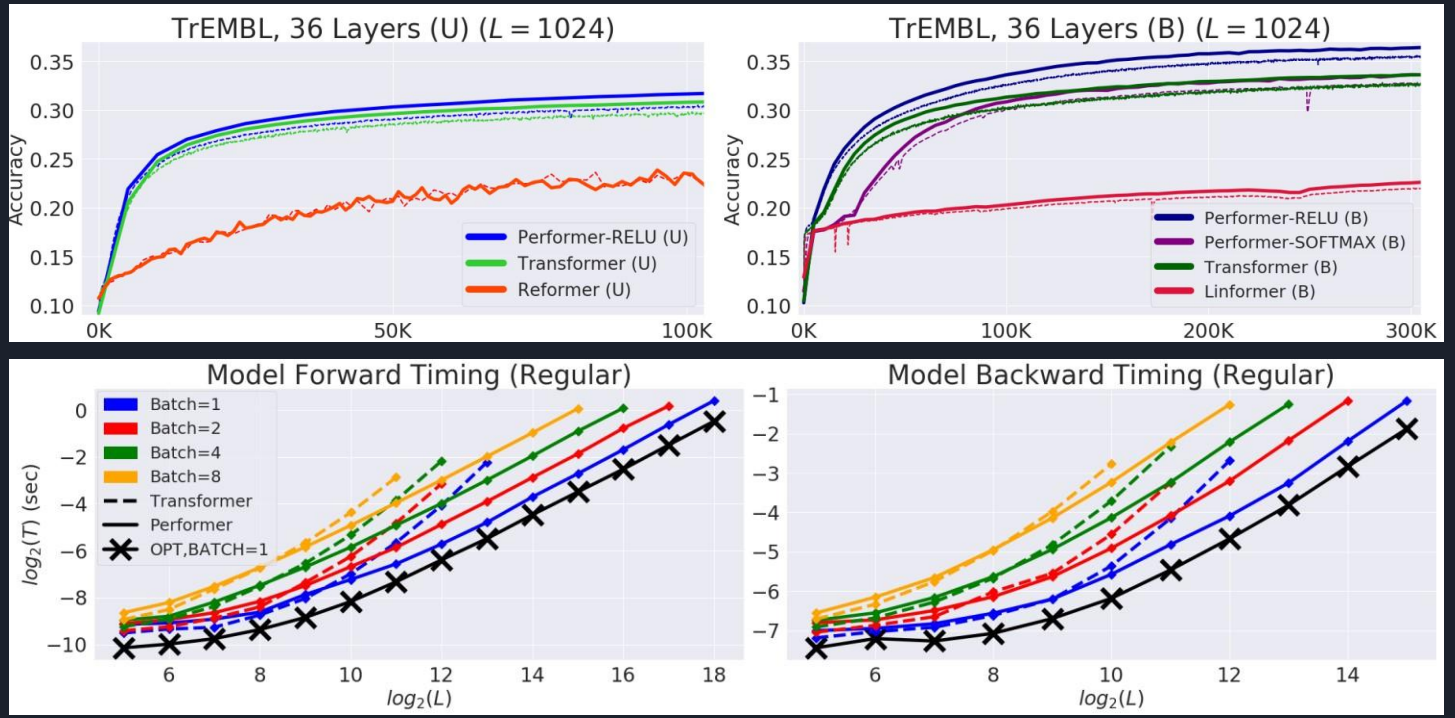




Finally :)


time complexity: O(Lmd)
dimensions: (Lxm)(mxd) = (Lxd)
Lxm
mxL
Lxd
Q' (K'^T V)
dimensions: (mxL)(Lxd) = (mxd)
time complexity: O(Lmd)

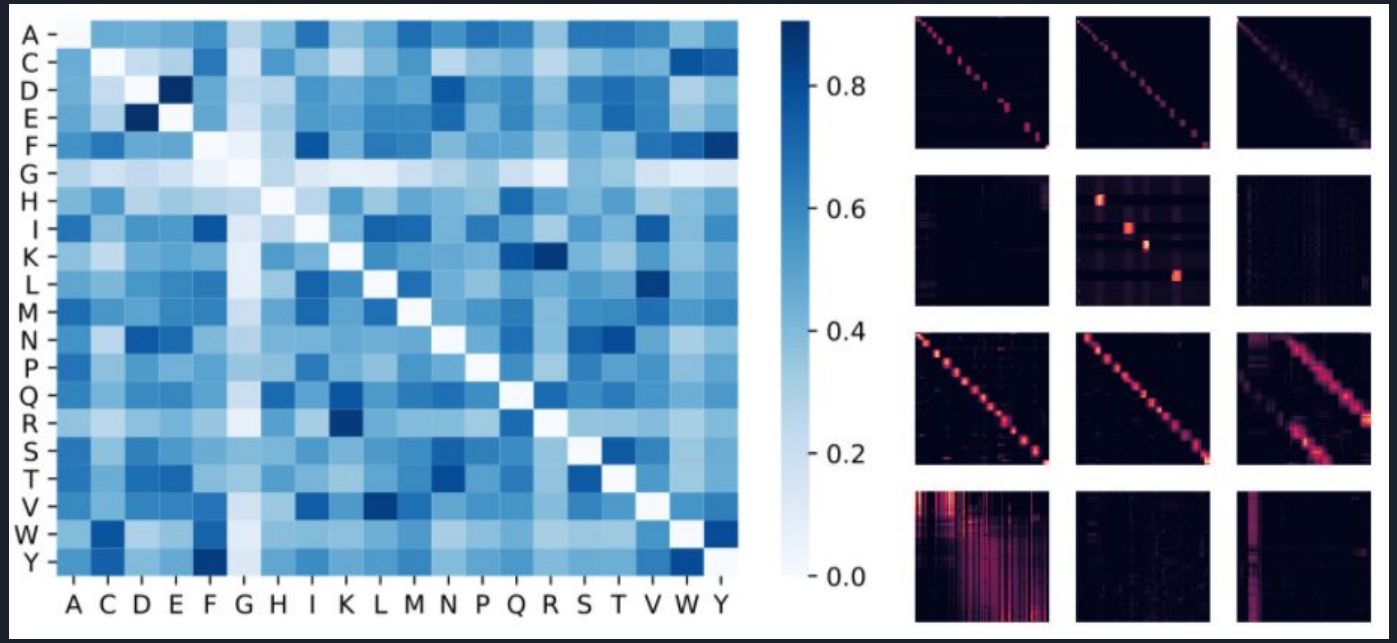
Comparisons



BROADER IMPACT

1. Biology and Medicine
2. Environment
3. Research on Transformers
4. Backward Compatibility
5. Attention Beyond Transformers

Applications



Resources

Linformer: Self-Attention with Linear Complexity (paper review)

Singular_value_decomposition

Linformer paper review YT video

Linformer: Self-Attention with Linear Complexity - arxiv

From Transformers to Performers: Approximating Attention

Rethinking Attention with Performers

Rethinking Attention with Performers - arxiv

Rethinking Attention with Performers (Paper Explained)

Thanks

hasanmahani08@gmail.com

rezadar1378@gmail.com