Skyformer: Remodel Self-Attention with Gaussian Kernel and Nyström Method

Yifan Chen* [1] Qi Zeng* [2] Heng Ji [2] Yun Yang [1]

[1]Department of Statistics [2]Department of Computer Science

NEURAL INFORMATION PROCESSING SYSTEMS

Background

Transformers are hard to train.

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{p}}\right)\boldsymbol{V} := \boldsymbol{D}^{-1}\boldsymbol{A}\boldsymbol{V} \tag{1}$$

- **Quadratic** time and space complexity in the self-attention mechanism, which makes **long sequence** processing resource-consuming,
- Small perturbations in parameter updates tend to be amplified, resulting in significant disturbances in the model output.

Kernel methods may be the answer to both challenges. To justify the introduction of kernel methods, we first build up the connections between Self-attention and **Gaussian Kernels**:

- The un-normalized attention score matrix can be formed via basic matrix operations on an empirical Gaussian kernel matrix,
- The form of Gaussian kernels has the natural interpretation of assigning "attention" to different tokens,
- Gaussian kernels automatically perform the normalization similar to how softmax does.

More details are provided in the next block, in which we propose "Kernelized Attention".

Kernelized Attention

- Kernelized Attention replaces the softmax structure with a Gaussian kernel

$$\text{Kernelized-Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \boldsymbol{C}\boldsymbol{V} := \kappa\left(\frac{\boldsymbol{Q}}{p^{1/4}},\frac{\boldsymbol{K}}{p^{1/4}}\right)\boldsymbol{V} \tag{2}$$

  where $\kappa(\boldsymbol{q}_i,\boldsymbol{q}_j) := \exp\left(-\|\boldsymbol{q}_i-\boldsymbol{q}_j\|^2/2\right)$ is the standard Gaussian kernel function,
- It can be rewritten in terms of the un-normalized attention score matrix as

$$\text{Kernelized-Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \boldsymbol{D}_Q^{-1/2}\cdot\boldsymbol{A}\cdot\boldsymbol{D}_K^{-1/2}, \tag{3}$$

  where $\boldsymbol{D}_Q$ (resp. $\boldsymbol{D}_K$) is a diagonal matrix with elements $(\boldsymbol{D}_Q)_{ii} = \exp\left(\frac{\|\boldsymbol{q}_i\|^2}{\sqrt{p}}\right), \forall i\in[n]$.

Comparing the two equations, we can check the un-normalized attention score matrix $\boldsymbol{A}$ is involved, and $\boldsymbol{D}_Q, \boldsymbol{D}_K$ perform the normalization similar to $\boldsymbol{D}$ in self-attention.

Attempts in past literature

The attention score matrix is known to exhibit a very fast rate of singular value decay, similar to that of an empirical kernel matrix. This near singular property motivates many low-rank attention approximation methods to skillfully leverage the computation techniques in kernel methods.

- Linformer [Wang et al., 2020] compresses the size of the key and value matrix with random projections based on the Johnson–Lindenstrauss transform;
- Reformer [Kitaev et al., 2020] applies locality-sensitive hashing (LSH) [Har-Peled et al., 2012] to simplify the computation of the attention score matrix, which is recently used in kernel density estimation [Charikar and Siminelakis, 2017, Backurs et al., 2019];
- Performer [Choromanski et al., 2020] projects both query and key matrix through random Fourier features [Rahimi et al., 2007], heavily exploiting Bochner Theorem for stationary kernels.

Nyström methods

We write Nyström methods in a sketching form as follows:

- Approximate $\boldsymbol{K}$ by $\widetilde{\boldsymbol{K}} = \boldsymbol{K}\boldsymbol{S}(\boldsymbol{S}^T\boldsymbol{K}\boldsymbol{S})^\dagger\boldsymbol{S}^T\boldsymbol{K}$;
- For classical Nyström methods, $\boldsymbol{S}$ is an n-by-d (uniform) sampling matrix;
- The complexity is reduced to $\mathcal{O}(nd^2)$, when $d \ll n$;
- A fun fact: we can check Nyström methods by letting $\boldsymbol{S} = \boldsymbol{I}$.

An implicit requirement for the matrix $\boldsymbol{K}$ to approximate is, $\boldsymbol{K}$ should be positive semidefinite (PSD). However, in general $\boldsymbol{A}$ is not PSD since $\boldsymbol{A}$ is not even symmetric.

Skyformer: a modified Nyström method

To resolve the issues mentioned above, we propose Skyformer:

- Complete the matrix into a PSD matrix $\bar{\boldsymbol{C}}$

$$\bar{\boldsymbol{C}} := \phi\left(\begin{pmatrix}\boldsymbol{Q}\\ \boldsymbol{K}\end{pmatrix},\begin{pmatrix}\boldsymbol{Q}\\ \boldsymbol{K}\end{pmatrix}\right), \quad \left(\boldsymbol{C} := \phi(\boldsymbol{Q},\boldsymbol{K}) = (\boldsymbol{I},\boldsymbol{0})\bar{\boldsymbol{C}}(\boldsymbol{0},\boldsymbol{I})^T\right) \tag{4}$$

- Approximate $\bar{\boldsymbol{C}}$ with $\tilde{\bar{\boldsymbol{C}}}$ through

$$\tilde{\bar{\boldsymbol{C}}} = \bar{\boldsymbol{C}}\boldsymbol{S}(\boldsymbol{S}^T\bar{\boldsymbol{C}}\boldsymbol{S})^\dagger\boldsymbol{S}^T\bar{\boldsymbol{C}} \tag{5}$$

- The final approximation will be given as

$$\tilde{\boldsymbol{C}} := (\boldsymbol{I},\boldsymbol{0})\tilde{\bar{\boldsymbol{C}}}(\boldsymbol{0},\boldsymbol{I})^T \tag{6}$$

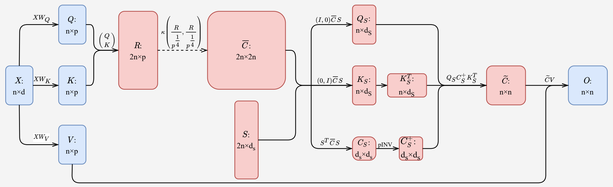

Figure 1. An overview of the proposed attention approximation method. This method approximates the original softmax matrix with a low dimensional randomized sketch of the complete symmetric matrix. Each box represents a matrix with the box size representing matrix size. The red boxes are matrices introduced by the proposed method besides the vanilla self-attention matrices in blue boxes. The dash line means that during computation, instead of the full $\bar{\boldsymbol{C}}$ we only compute the selected columns of $\bar{\boldsymbol{A}}$ with a sampling matrix $\boldsymbol{S}$.

Error analysis of Skyformer

An implicit advantage of using Kernelized Attention is that we can leverage the existing conclusions for kernel methods to analyze the theoretical properties of the model.

In the paper we state a high probability bound (Theorem 2) on the size d of the sub-sampling matrix used in Skyformer for the kernelized attention score matrix $\boldsymbol{C}$ by the following theorem. This theorem implies the time and space complexity of our proposed approximation depends on the statistical dimension $d_{stat} := \text{Tr}\left(\bar{\boldsymbol{C}}(\bar{\boldsymbol{C}}+\lambda\boldsymbol{I})^{-1}\right)$. If we directly use the conclusion from Gaussian kernels, d_{stat} should be $\widetilde{\mathcal{O}}(1)$ (complexity modulo poly-log term) [Yang et al., 2017] due to the exponential eigenvalue decay rate of Gaussian kernels, which is comparable to the complexity of most other efficient attention.

Empirical results

Empirical Approximation Evaluation

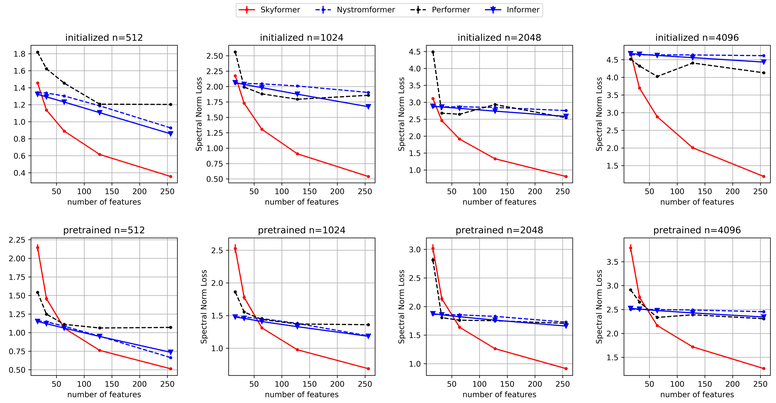

Figure 2. Spectral norm results with different sequence lengths under different W_Q, W_K, W_V settings, either from initialized or pretrained BERT models. All methods are approximating the original self-attention output. Y axis: Lower spectral loss means better approximation. X axis: Higher d (number of features) means visiting more elements in the original matrix and bringing more computation costs. The label "Skyformer" here means that we use the algorithm behind Skyformer, to approximate the raw attention score matrix A in self-attention. In this experiment, "Skyformer" also needs to first approximate A, and then approximate D, as Performer does.

Classification accuracy (%) on LRA benchmark.

Model	Text	ListOps	Retrieval	Pathfinder	Image	AVG.
Self-Attention	61.95	38.37	80.69	65.26	40.57	57.37
Kernelized Attention	60.22	38.78	81.77	70.73	41.29	58.56
Nystromformer	64.83	38.51	80.52	69.48	41.30	58.93
Linformer	58.93	37.45	78.19	60.93	37.96	54.69
Informer	62.64	32.53	77.57	57.83	38.10	53.73
Performer	64.19	38.02	80.04	66.30	41.43	58.00
Reformer	62.93	37.68	78.99	66.49	48.87	58.99
BigBird	63.86	39.25	80.28	68.72	43.16	59.05
Skyformer	64.70	38.69	82.06	70.73	40.77	**59.39**

References and supplementary materials

Omitted due to space limit. Please refer to https://openreview.net/forum?id=pZCYG7gjkKz.

yifanc10@illinois.edu