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

Yifan Chen*, Qi Zeng*, Heng Ji, Yun Yang

University of Illinois at Urbana-Champaign

October 18, 2021

Background

Transformers are expensive to train

Quadratic Complexity

- Quadratic time and space complexity in the self-attention mechanism
- Transformers cannot support long sequence processing and large batch size

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

Training Instability

- Small perturbations in parameter updates tend to be amplified, resulting in significant disturbances in the model output

Background

Kernel methods may be the answer to both challenges

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

Method

Kernelized Attention

- Kernelized Attention replaces the softmax structure with a Gaussian kernel

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

- It can be rewritten in terms of the un-normalized attention score matrix as

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

Method

Skyformer: a modified Nyström method

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

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

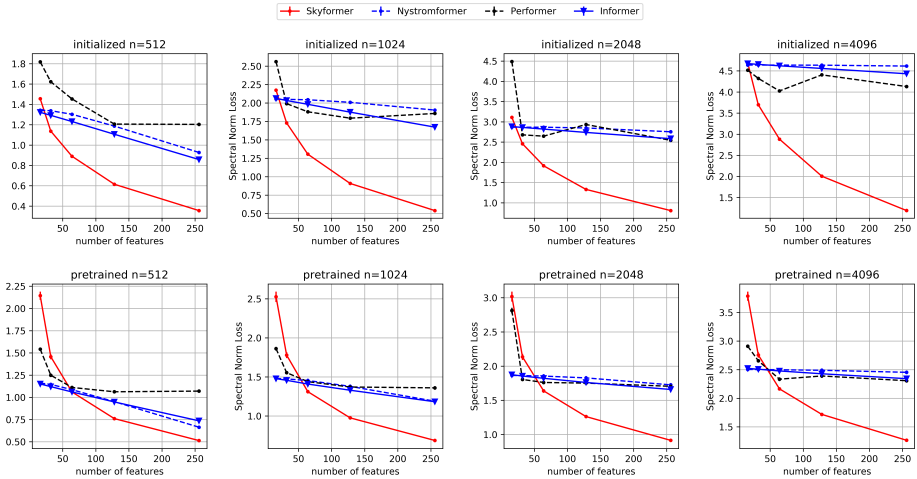
- Approximate $\bar{\boldsymbol{B}}$ with $\tilde{\bar{\boldsymbol{B}}}$ through

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

- The final approximation will be given as

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

Results: Empirical Approximation Evaluation



Results: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**

Conclusion

- We revisit the intrinsic connection between self-attention and kernel methods, and explore a new kernel-based structure to stabilize the training of Transformers
- We approximate the Kernelized Attention via low dimensional randomized sketches by adapting the Nyström method to a non-PSD matrix.
- We conduct extensive experiments showing that Skyformer achieves comparable performance to the original self-attention with fewer computational costs.