EECS 6322 Week 7, Paper 2
Rethinking Attention With Performers

Shivani Sheth (Student #: 218011783)
shivs29@yorku.ca

I. Summary and contributions

The paper proposes a novel architecture known as Performers, which are faster alternatives to Transformers. These architectures have linear time and space complexities and are able to estimate the regular (softmax) full rank attention Transformers with almost the same accuracy. The performers use Fast Attention Via positive Orthogonal Random features (FAVOR+) method to map the features in a given Transformer to a different dimension and estimate the features in linear time as compared to the quadratic time taken by the transformers. The proposed architecture also provides a low estimation variance, uniform convergence, an unbiased estimation of the attention matrix, and is fully compatible with the regular Transformers.

The regular transformers include a trainable attention approach that capture the dependencies between each input token in the input sequence. This approach led to a very high computational cost for large, or even medium size sequences of the input tokens. Thus, the existing transformer architecture restricted its use on applications with limited computational resources, even for medium length input tokens. The performer aims to resolve this drawback of the transformer by estimating the softmax of the attention matrix in linear time, thus reducing the resource requirements for large input sequences.

The proposed architecture introduces the FAVOR+ method that uses positive orthogonal random features to estimate the Gaussian and Softmax kernels of the original transformer mechanism. The Fast Attention (FA) in the FAVOR+ model is obtained by substituting the original attention model by the Generalized Kernelizable Attention. The original bidirectional attention model is given by: $Att_{\leftrightarrow}(Q,K,V) = D^{-1}AV, A = exp(QK^T/\sqrt{d}), D = diag(A1_L)$, *where* 1_L *is a vector of length L consisting of all 1's,* $diag()$ *is a diagonal matrix, and Q, K, V are the Queries, Keys, Values respectively. The time and space complexity of this model is* $O(L^2d)$ *and* $O(L^2 + Ld)$ *respectively. Similarly, the Generalized Kernelizable Attention model is given by* $\widehat{Att_{\leftrightarrow}}(Q,K,V) = \widehat{D}^{-1}(Q'((K')^TV)),\ \widehat{D} = diag(Q'\ ((K')^T\ 1_L))$ *which has a time and space complexity of* $O(Lrd)$ *and* $O(Lr + Ld + rd)$ *respectively. Hence, we see that the latter significantly reduces both the time and space complexity, allowing for larger input sequences to be processed with limited computational resources.*

The kernalizable model computes each element in the attention matrix by $A(i,j) = K(q_i^T, k_j^T)$, *where* $K(x,y) = E[\phi(x)^T\phi(y)]$. *This* ϕ *function maps the input elements to a different dimension where the original kernels can be computed linearly. To approximate the softmax kernels for attention in the performers architecture, the authors have defined the* ϕ *function as* $\phi(x) = \frac{h(x)}{\sqrt{m}}(f_1(\omega_1^Tx), \cdots, f_1(\omega_m^Tx), \cdots, f_l(\omega_1^Tx), \cdots, f_l(\omega_m^Tx))$. *By substituting the values of* $h(x) = exp(\frac{||x||^2}{2}), l = 2, f_1 = sin, f_2 = cos$, *we get the modified Softmax Kernel as* $\widehat{SM}_m^{trig}(x,y) = exp(\frac{||x||^2}{2})\ K_{gauss}(x,y)\ exp(\frac{||y||^2}{2})$. *However, this modified Softmax Kernel can produce feature maps with negative dimension values that result in unstable behaviour, especially when the kernel scores are close to 0. Hence, the Softmax Kernel is modified to only accommodate the positive feature maps, which is given by:* $\widehat{SM}_m^{hyp+} = E_{\omega\sim\mathcal{N}(0,I_d)}\ [exp(\omega^Tx - \frac{||x||^2}{2})\ exp(\omega^Ty - \frac{||y||^2}{2})] = \Lambda E_{\omega\sim\mathcal{N}(0,I_d)}\ cosh(\omega^Tz)$, *where cosh is a hyperbolic cosine and* $\Lambda = exp(-\frac{||x||^2+||y||^2}{2})$. *Further, the model's variance of the Gaussian or softmax kernel can be reduced by selecting the random samples* $\omega_1, \cdots, \omega_m$ *that are exactly orthogonal.*

The performers were written in Jax and implemented on the pre-existing Transformers by replacing its attention component. It was compared against baselines such as Reformers and Linformers and was pre-trained on the PG-19 dataset. The Performer performs significantly well, both in forward and backward pass in terms of time complexity against the transformers. It obtains a nearly linear time complexity and a sub-quadratic memory consumption. The orthogonal and positive features also result in a significantly lower Mean Squared Error of the model as compared to the IID and trigonometric sin/ cos features. The backward compatibility of the Performers model can result in a high accuracy in datasets such as the LM1B dataset upon fine-tuning, and the PG-19 dataset with a positive softmax with feature redrawing. The Performer-RELU also outperforms Transformers on the TrEMBL dataset in both unidirectional and bidirectional models having 36 layers.

Thus, the performer significantly reduces the space and time complexity over the regular Transformers. It performs an unbiased estimation of the Softmax Kernel of the Transformer and hence is able to perform equally well with better computational costs. The model can be applied to several domains in the real-world which were earlier restricted by computational limitations over large-sequence inputs such as Backward Compatibility in Transformers, Attention Beyond

Transformers, Biology and Medicine, Research on Transformers, and Environment.

II. Strengths

The Performer model significantly reduces the time and space complexities as compared to the Transformer. Since the Transformer is widely used in several domains, the Performer advances the research in those domains by lowering the computational requirements for large input data. It also performs equally well as compared to Transformers and can be extended to other scalable kernel methods.

III. Weaknesses

There does not exist a general Performer model that can be applied (with comparable accuracy) to different datasets. Currently, different versions of the Performer need to be implemented with different datasets, often also requiring additional procedures such as fine tuning or feature redrawing.

IV. Correctness

The claims and empirical methodology of the Performers architecture are correct and have been supplemented with in-depth theoretical proofs. They have also been tested to have comparable accuracies as compared to Transformers. The results also show a significant decrease in the time and space complexities of the Performer model due to the estimation of the Softmax Kernel in the regular Transformers.

V. Clarity

The paper is clear and very well written with significant theoretical grounding along with experimental results. The appendix also elaborates on the important components of the paper and supplements the understanding of the reader. The difference in the results have also been illustrated on various components such as time complexity in the backward and forward pass, and mean square error with different feature selection of the model that make the comparison easier to understand.

VI. Relation to prior work

Prior work aimed to estimate the attention matrix of the Transformer, in order to reduce the quadratic computational cost by methods such as sparsity, pooling-based compression, clustering/ convolutional techniques, attention mechanism to attend local neighborhoods, and local sensitive hashing. Other previous work also includes dense attention matrices that consist of low-rank kernels substituting the softmax non-linear activation. These methods are mainly based on the kernels that produce explicit representations as dot-products of the finite and positive feature vectors. They aim to propose simpler attention mechanisms by having additional constraints, such as identical key, query pairs and sparse attention using extra layers, instead of approximating the regular attention as proposed by the Performers.

Other methods to reduce the Transformer's computational complexity also include shared attention weights and reverse residual layers that restrict the applications to short-sequence problems because of the insufficient approximations of the attention mechanism. The truncated back-propagation methodology also fails to capture the long-range input sequence dependencies as it approximates the attention mechanism based on back-propagation inside a localized window.

In contrast, the Performers architecture is proved to be capable of accurately estimating the regular (softmax) full-rank attention in linear space and time complexity and it also does not rely on any priors such as low-ranking or sparsity. With a little fine-tuning, they are also fully compatible with the regular transformers, and provide an unbiased estimation of the attention matrix, strong theoretical guarantees, lower variance and uniform convergence. These architectures use the Fast Attention Via positive Orthogonal Random features (FAVOR+) method (as proposed in the paper) to closely approximate the attention mechanism.

VII. Reproducibility

The model has been implemented on JAX and enough details in terms of theoretical grounding, modifications from the regular Transformer, and code implementation have been provided to reproduce the major details of the work proposed by the authors.

VIII. Additional comments

The paper could further include more information on which datasets or features in a dataset would be best suited for the different versions of the model.