COLUMBIA | ENGINEERING
The Fu Foundation School of Engineering and Applied Science

Machine Learning and High Dimensional Data Mining (IEORE661)

The Unreasonable Effectiveness of Structured Random Orthogonal Matrices: Rethinking Attention with Hadamard-Rademacher Performers

Authors :
Raphaël ADDA (rea2157)
Sheldon ALLEN (sja14)

Professor :
KRZYSZTOF CHOROMANSKI

December 17, 2021

ABSTRACT:

In this report, we build on previous code and results (Koker, 2020) to show the basic Performer model's accuracy and compute efficiency in comparison to its extended version, using both Fourier (trigonometric) and positive random feature mappings. Specifically, we present the variance and compute times of the Hadamard-Rademacher (HR) extension compared to i.i.d. Gaussian and orthogonal Gaussian random feature maps, shedding some light on the accuracy and time trade-off between Performer models.

Contents

1 Introduction

Transformer model architecture introduced by Vaswani et al., 2017 underlies many of the deep learning advancements in fields such as Natural Language Processing, image and video recognition/classification, bioinformatics, etc. Specifically, the Transformer's parallelizable attention mechanism has largely replaced the sequential processing which the previous state-of-the-art deep learning approaches to sequence-to-sequence modeling – such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory networks (LSTMs) – relied on. The disadvantages of sequential processing are many: Processing a single token involves processing a vector of context information meant to capture all relevant relationships among the prior tokens. In addition to the excessive memory usage and training time implied by sequential processing, the long-range dependencies involved in these methods' backwards propagation are further prone to the vanishing or exploding gradients problem. The self-attention mechanism of the Transformer voids these issues by incorporating ("attending to") all tokens (or all prior tokens in the case of causal attention) at each training step, but with positional embeddings to retain information about the token locations. With that said, since the introduction of scaled dot-product attention in the Transformer architecture, practitioners and researchers have struggled with its lack of scalability: "Attention is all you need" but you also need computation time and memory space that scale quadratically with the sequence length (or number of input tokens). In its most basic form, the Performer architecture introduced by Choromanski et al. 2021 approximates the regular softmax full-rank attention but in time and space that scale only linearly with sequence length, without making assumptions about priors (e.g., sparsifying). This basic Performer entails orthogonally projecting the query and key matrices of regular attention onto a matrix (randomized feature map), where the number of rows is less than the sequence length, which dimensionality reduction already provides some compute efficiency gains. As an extension to the basic Performer architecture, projecting onto a structured random orthogonal matrix such as the Hadamard-Rademacher matrix further enhances both the accuracy and complexity advantages when the number of features is large enough. This improvement is based on the observation from Choromanski et al., 2017 that more structured matrices (such as the so-called "SD-product matrices") provide computational efficiency in addition to improved MSE. This benefit is due to the O(n log n) time in which Walsh-Hadamard transform calculates matrix-vector products (where the matrix is a Hadamard). What's more, the calculation also requires reduced memory space as the Rademacher is only a diagonal matrix. In this paper, we investigate the

accuracy and time requirement of the basic Performer architecture in comparison to the version of the Performer model that uses the Hadamard-Rademacher matrices in the random feature map. The rest of this paper is structured as follows: Having presented some of the most salient features of the Transformer model. In section 2, we discuss how the Performer's FAVOR+ mechanism differs from and improves on the original self-attention mechanism of the Transformer. We also describe two types of random orthogonal matrices, the Gaussian orthogonal and the more structured SD-product matrices, of which the Hadamard-Rademacher ("HR") is an example. In section 3, we provide some implementation details of using FAVOR+ with the HR matrix, including the pseudo-code for the Fast-Walsh-Hadamard transform and for the process to generate blocks of the HR matrix for projection. In section 4, we compare and discuss the accuracy and compute time results for the three approaches (IID Gaussian vs. orthogonal Gaussian vs. Hadamard Rademacher), using random Fourier features (sine and cosine) and using positive features. In the last section, we conclude with final observations and areas ripe for future study.

2 Further Introduction and Background

2.1 Attention Mechanisms

2.1.1 Regular Attention

Vaswani et al., 2017, introduced the concept of dot-product attention: it consists of the mapping of the three matrices, Q, K, V of $\mathcal{R}^{L\times d}$ query, key and value (L is the input sequence length and d the hidden dimension). The following dot-product attention is:

$$\text{Att}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d}})V$$

and $\text{Att}(Q, K, V)$ is in $\mathcal{R}^{L\times L}$.

2.1.2 FAVOR+ mechanism & positive orthogonal random features

Choromanski et al. 2020, introduced the concept of fast attention with orthogonal random features.
In the softmax operation, query Q and key K are substituted by Q' and K' in $\mathcal{R}^{L\times r}$ where $r \ll d$ using a gaussian orthogonal random features mapping function ϕ

in $\mathcal{R}^r$.
Thus:

$$\widehat{Att}(Q, K, V) = A^T V$$

with $A_{ij} = K(q_i^T, k_j^T)$, K a kernel function from $\mathcal{R}^d \times \mathcal{R}^d$ to $\mathcal{R}_+$ such that $K(x, y) = E[\phi(x)^T \phi(y)]$.

2.2 The family of Random Ortho-Matrices *(ROMS)*

Choromanski et al. 2018, proved the efficiency of random orthogonal embeddings for the two following types of random orthogonal matrices in several applications: Gaussian orthogonal matrices and SD-product matrices.

2.2.1 GAUSSIAN ORTHOGONAL MATRICES

A Gaussian orthogonal matrix G is defined as a random $\mathcal{R}^{n \times n}$ such that the rows marginally have multivariate Gaussian N (0, I) distribution and forms an orthogonal basis of $\mathcal{R}^n$. It can be obtained by drawing independently each element of the matrix using $\mathcal{N}(0, 1)$ distribution, extracting the matrix Q with QR decomposition (it is equivalent to apply Graham-Schmidt process on each row of the matrix), and finally independently scaling each row so they have mean 0 and variance 1.

2.2.2 SD-PRODUCT MATRICES WITH HADAMARD MATRIX

The SD-product matrix allows for similar result while gaining computational efficiency.
A SD orthogonal matrix is the result of the following product:

$$SD = \prod_{i=1}^{k} SD_i$$

where $D_i \in \mathcal{R}^{n \times n}$ are independent diagonal matrices and $S \in \mathcal{R}^{n \times n}$ is a particular kind of structured matrix with $|s_{ij}| = \frac{1}{\sqrt{n}}$ and orthogonal rows.
In our work, we will consider the H-Rademacher matrix, where we use as structured matrix S the Hadamard Matrix H defined as: $H_1 = (1)$ and $H_k = \frac{1}{\sqrt{2}} \begin{pmatrix} H_{k-1} & H_{k-1} \\ H_{k-1} & -H_{k-1} \end{pmatrix}$ and as D the diagonal matrix with independent 1 and -1 on the diagonal with probability $\frac{1}{2}$ (Rademacher distribution).
H_k is in $\mathcal{R}^{2^k \times 2^k}$

3 Adapation of SD-product matrices with Hadamard matrix to Performers

3.1 Implementation of Hadamard orthogonal projection matrix

A standard calculation of the product of the Hadamard matrix with the matrix D is to simply multiply the Hadamard Matrix H by the diagonal matrix D. The multiplication of a Hadamard matrix H_m and a vector of $\mathcal{R}^{2^m}$ is in 4^m.
However, the `FWHT` algorithm (fast Walsh–Hadamard transform) permits reduced complexity to $O(nlog(n))$, $n = 2^m$.
Indeed, it uses a **divide and conquer algorithm** to perform faster:

`Fast-Walsh-Hadamard-Transform`(v)
if length(v) == 1 **then return** v
end if
$v = (v_0, v_1)$ with length(v_0) == length(v_1) == length(v/2)
y_0 =`Fast-Walsh-Hadamard-Transform`($H_{m-1}v_0$)
y_1 =`Fast-Walsh-Hadamard-Transform`($H_{m-1}v_1$)
return $(y_0 + y_1, y_0 - y_1)$.

Using the complexity master theorem, it is clear that the complexity of this algorithm is $O(nlog(n))$ with $n = 2^m$.
To obtain a matrix product $H \times D_i$, we just need to perform this algorithm on each column of the matrix D.

3.2 FAVOR+ with Hadamard orthgonal projection

The `FAVOR+` algorithm approximates regular attention thanks to random feature mapping.
Choromanski et al. 2020 proved the efficiency of projecting the query and the key matrices using orthogonal random features.
Orthogonal matrices are obtained using Gaussian orthogonal matrices. However, Choromanski et al. 2018 showed that SD-product matrices, especially Hadamard matrices, can be at least as efficient as Gaussian matrices for orthogonal projection: This is why decided to implement the `FAVOR+` algorithm with a new option: computing the projection matrix using the Hadamard Process.

Here is the pseudo code of the algorithm to obtain Hadamard orthogonal projection matrix:

`Orthogonal-Hadamard-square`(d, k)
SD = matrix of size (d,d)
for i = 1 to k **do**
D =`Rademacher`(d)
SD += `Fast-Walsh-Hadamard-Transform`$(H_d D_i)$
return SD.

`Rademacher`(d) is the function to obtain diagonal matrix of size d with Rademacher ($Unif\{-1, 1\}$) elements on the diagonal, and D_i is the i-th column of D.

Here is the generalization of the Hadamard process to obtain non-square Hadamard matrices:

`Orthogonal-Hadamard`(m, d, k)
Let blocks be a matrix of $\mathcal{R}^{m \times d}$
num-squares = $\lfloor \frac{m}{d} \rfloor$
for l = 1 to num-squares **do**
$blocks_l$ = `Orthogonal-Hadamard-square`(d,k)
R = m - d $\times$ num$_s quares$ **if** R>0 **then**
Add the first R columns of `Orthogonal-Hadamard-square`(d,k) at the right of blocks

blocks /= $\sqrt{\frac{\text{num-squares + remainder}}{d}}$
return blocks

We added the Hadamard option on the official tensorflow implementation of the Performer on our github repository: github.com/Raphaeladda/Hadamard_Performers.git.

4 Experiments and Results

Choromanski et al. 2018 showed that Hadamard-Rademacher orthogonal projection permits similar results than Gaussian orthogonal projection while gaining computational efficiency in some applications. We tried our implementation on the FAVOR+ to determine if it was the case for this algorithm.
In order to assess our results, we adapted our Hadamard orthogonal projection to the numpy short implementation of the FAVOR+ algorithm by Teddy Koker: https://teddykoker.com/2020/11/performers/.

4.1 Comparison of MSE for the different methods

We first examined the MSE of the following three methods of approximating the attention matrix:

- IID Gaussian feature mapping
- Orthogonal Gaussian feature mapping
- Orthogonal Hadamard-Rademacher feature mapping

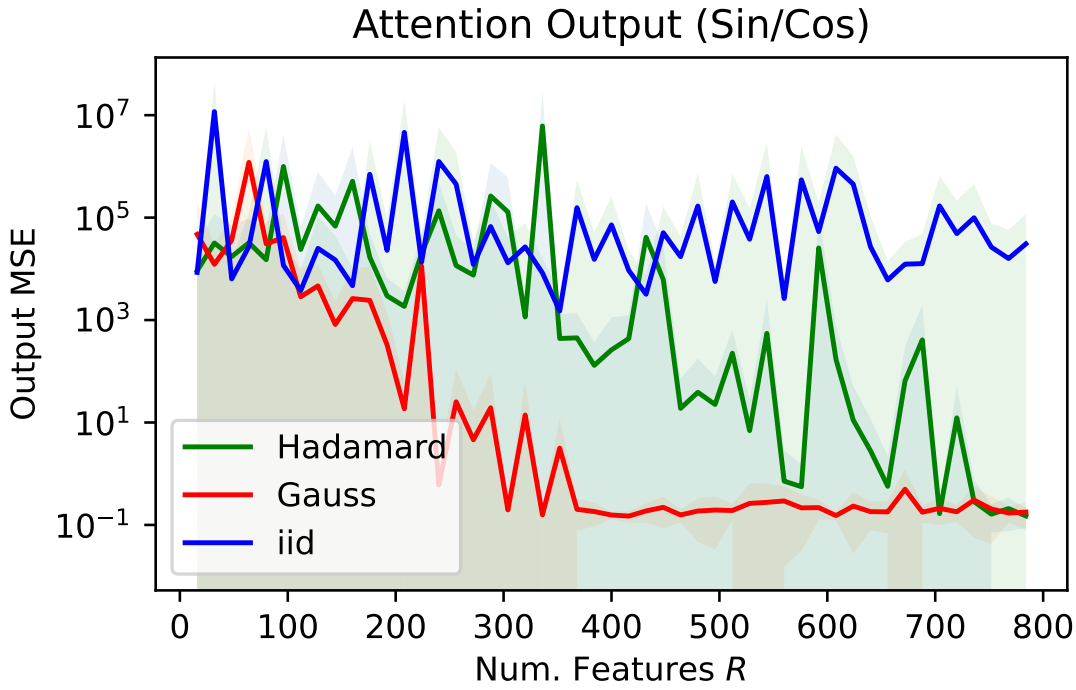

Figure 1: MSE of sin/cos attention approximation in log scale

The figure 1 above visualizes the MSE of each attention approximation, depending of the number of random features.
First, Hadamard and Gaussian feature mappings have an MSE much more lower than the regular iid feature mapping when the number of feature is not too low: it thus proves the efficiency of the orthogonal projection when using either the Hadamard method or the Gaussian one.
Moreover, the MSE for the Gaussian method appears lower than that of the Hadamard method.
However, we can detect on the graph that asymptotically, the MSE of the two methods will be the same, so for a high number of random features, both method are as efficient.

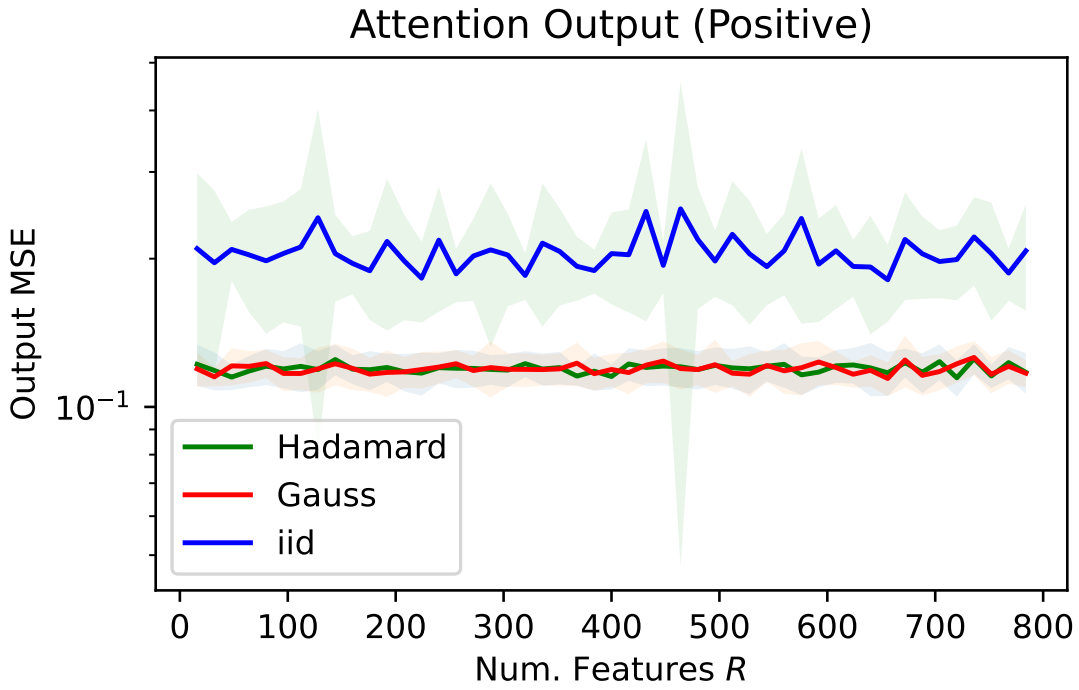

Figure 2: MSE of positive attention approximation in log scale with d = 16, l=1024

The figure 2 above visualizes the MSE of each attention approximation, depending of the number of random features.
Here, the difference of MSE between orthogonal and non-orthogonal methods is even more obvious.
However, unlike for the sin/cos attention approximation (2), the MSE of the Hadamard-Rademacher method is very low from the beginning: thus, for the positive attention, which is the attention approximation used in the performer (+ of the FAVOR+), it is not required to have a very high number of random features to equalize the efficiency of the Gaussian orthogonal projection method.

4.2 Comparison of computation times for the different methods

Here, we examined the computation times of the sin/cos and positive approximations of the attention using FAVOR+ with Hadamard and Gaussian orthogonal projections.

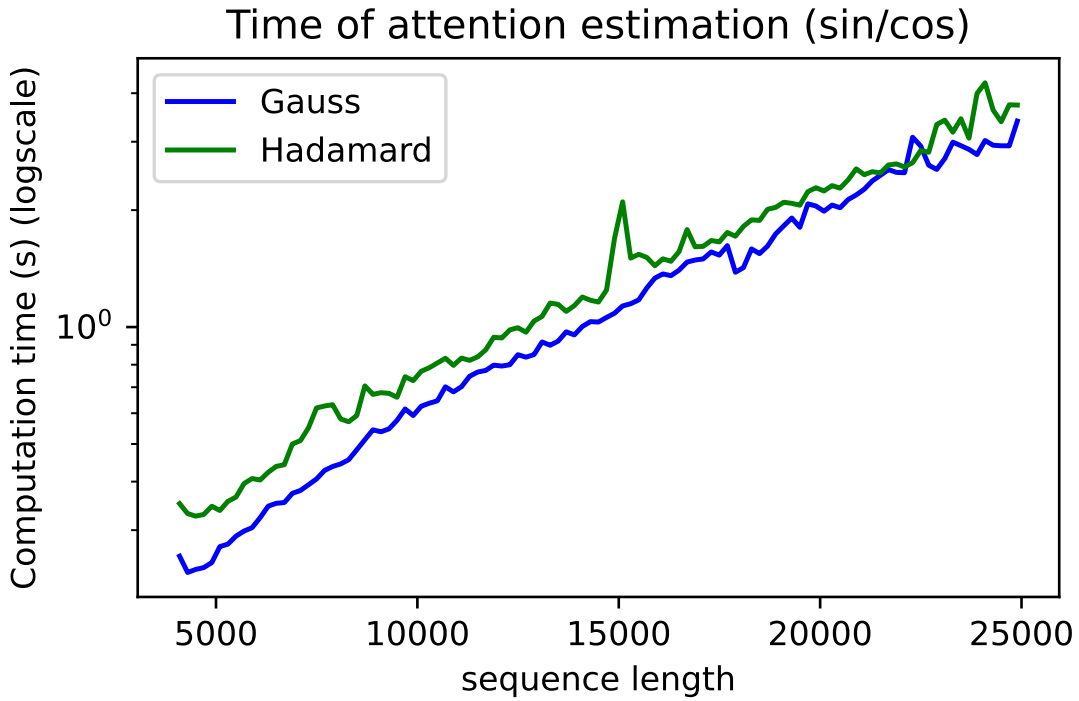

Figure 3: computation time of sin/cos attention approximation with d=16 and 300 random features

The figure 3 above visualizes the computation time of each approximation of the attention using orthogonal randoms features, depending of the input sequence length.
Here, we observe that Gauss and Hadamard method allow to compute approximate attention in an equivalent time. The Gaussian method has a slightly lower computation time compared to Hadamard method.

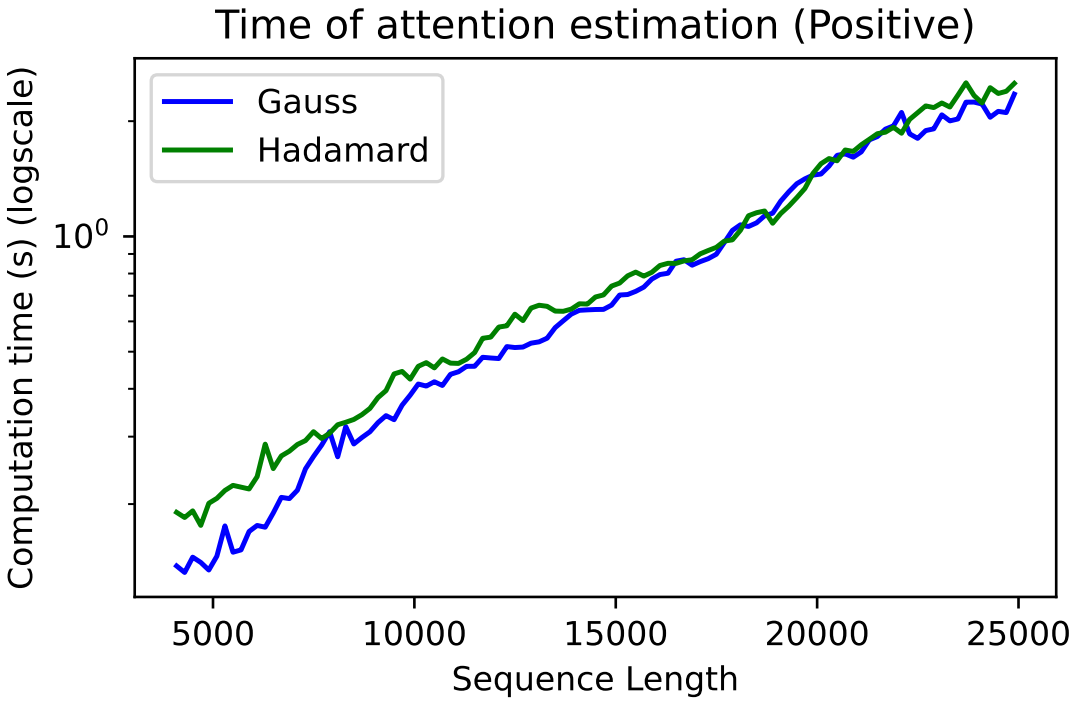

Figure 4: computation time of sin/cos attention approximation with d=16 and 300 random features

The figure 4 shows the same information than the figure 3 but for the positive attention.
Here, for long sequences, the computation tine is approximately the same.
Thus, once again, we conclude that for the sin/cos approximation of the attention, Hadamard projections give less efficient results than Gaussian projections, but for positive approximation, independent of whether we consider computation time or accuracy of the estimation of the attention, Hadamard projections provide almost as satisfying results as the Gaussian method.

5 Conclusion

As we observed above, the Hadamard-Rademacher projection matrix can be viewed as nearly equivalent to the orthogonal Gaussian matrix in terms of accuracy and compute time. Both of these orthogonal approaches are superior to the i.i.d. Gaussian projection matrix in these two measures, and therefore should be preferred in general. Between the two orthogonal matrix approaches (HR vs. Gaussian), the added complexity of the more structured HR approach requires an additional processing step to prepare the matrix. So without any clear compute or accuracy gains, practitioners might justifiably opt for the slightly simpler orthogonal Gaussian. However, further research would probably bear out the observations in Choromanski et al., 2017, that compute efficiency gains can be attained by using multiple blocks (especially, number of blocks, k = 3) in the construction of the Hadamard matrix – where we have in this study constructed the matrix from only a single block. Quite possibly the true efficiencies of the HR extension lie in this hyperparameter selection. With the code provided in the repo referenced above, this further investigation would be quite straightforward, though it had to be set outside the scope of this study due to time constraints.

6 Bibliography

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Advances in Neural Information Processing Systems 30, pp. 5998–6008. Curran Associates, Inc., 2017. URL http://papers. nips.cc/paper/7181-attention-is-all-you-need.pdf.

Krzysztof Marcin Choromanski, Mark Rowland, and Adrian Weller. The unreasonable effectiveness of structured random orthogonal embeddings. In Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA, pp. 219–228, 2017.

Krzysztof Choromanski, Mark Rowland, Wenyu Chen, and Adrian Weller. Unifying orthogonal Monte Carlo methods. In Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA, volume 97 of Proceedings of Machine Learning Research, pp. 1203–1212. PMLR, 2019. URL http://proceedings.mlr.press/v97/choromanski19a.html.

Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos. Rethinking Attention with Performers. ArXiv abs/2009.14794, pp. 1-38, 2021.

Koker, Teddy. “Performers: The Kernel Trick, Random Fourier Features, and Attention” https://teddykoker.com/2020/11/performers/.