# Accelerating Transformers with Fourier-Based Attention for Efficient On-Device Inference

Hyeonjin Jo[*], Chaerin Sim[‡], Jaewoo Park[*] and Jongeun Lee[†]
Department of CSE[*], School of New UNIStars[‡], Department of EE[†]
Ulsan National Institute of Science and Technology (UNIST), Ulsan, Korea
{zxzx1825, chaerinsim, hecate64, jlee}@unist.ac.kr

*Abstract*—**Multi-head attention based transformers have achieved significant success in various natural language processing applications. However, their quadratic computation complexity and low arithmetic intensity present challenges for inference acceleration. To address this issue, attention mechanisms based on Fourier transforms have been proposed. Nevertheless, the acceleration of complex arithmetic involved in Fourier transform on systolic array based edge devices remains unexplored. In this paper, we analyze the inference of transformers on VTA, a tensor accelerator designed for mobile devices, and propose an efficient mapping of Fourier based attention on VTA. Our experimental results demonstrate that on-device inference of Fourier based attention can improve inference latency up to 70.7% and 29.6% on average compared to Multi-head attention.**

***Keywords; Natural Language Processing, FPGA, Fourier Transform, Multi-head Attention***

## I. Introduction

In recent years, the transformer architecture [2] has proved its dominance across various NLP tasks. Though it was originally proposed for NLP tasks, it is now widespread in ML tasks of all sorts including image classification and generative modeling. At the heart of its outstanding performance is the novel *multi-head attention* mechanism, which consists of several projection layers and matrix multiplications between replicated inputs. However, the use of memory-intensive operations such as reshape and transpose operations makes the adaptability onto existing DNN accelerators extremely poor. Also, the algorithmic complexity of the multi-head attention is quadratic with respect to the sequence length, requiring hundreds of GPUs to train a single large language model.

To overcome the poor scalability of attention-based transformers, it became a research trend to find a better attention mechanism with a lower complexity compared to multi-head attention. The FNet architecture [1] replaces the multi-head attention with a Fourier attention layer, which performs 2D discrete Fourier transform (DFT) on the input vectors and returns only the real part. Due to the efficient fast Fourier transform (FFT) on GPUs, FNet outperforms transformers with 5x speedup on training with comparable model accuracy. But still, it remains a challenge to efficiently accelerate FNet on systolic array-based edge devices for the following reasons:

1) Most of the DNN accelerators on the edge target optimizing GEMM or convolution operations, not DFT.

2) DFT requires complex arithmetic, which contrasts with typical inference accelerators focusing on low-bit integer arithmetic.

In this paper, we evaluate the performance of muti-head attention based transformers using VTA. VTA [2] is an open-source, programmable DNN accelerator for edge devices. We then implement and evaluate Fourier attention on VTA, introducing novel techniques to avoid complex arithmetic and efficiently compute DFT. Experimental results demonstrate that the inference latency of Fourier attention outperforms multi-head attention up to 70.7% and 29.6% on average, depending on sequence length and hidden dimension size.

## II. Multi-Head Attention and Fourier Attention

The standard architecture of a transformer contains a multiple stack of encoder blocks. Fig. 1 represents the inner structure of a single encoder block. The original transformer proposed in [3] incorporates the multi-head attention (MHA) layer, where the input matrix is replicated as three identical queries, keys, and values pair followed by a linear projection and a scaled dot-product attention. Because of the parallel dataflows and heterogeneous operations, the computation of MHA is highly memory-bound.

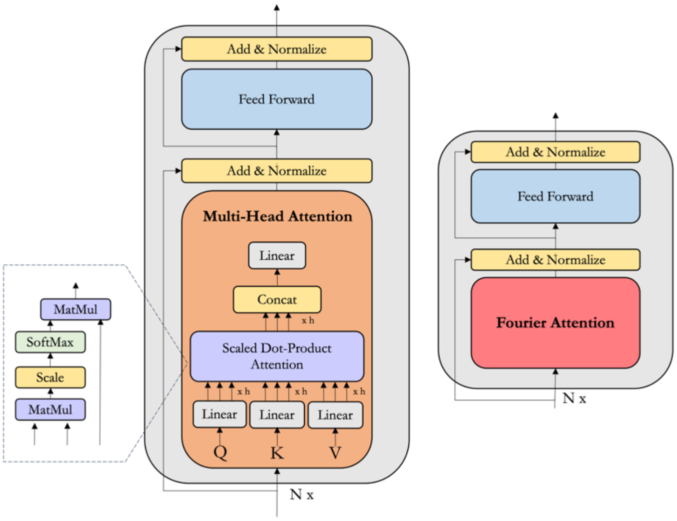

Figure 1. Comparison between multi-head attention and Fourier attention.

Compared to the MHA layer, the Fourier attention layer in [1] is surprisingly simple yet effective. The whole MHA layer is substituted to a 2D Fourier transform layer where only the real component of the output is passed to the next layer. The 2D discrete Fourier transform (DFT) is applied as a 1D DFT along the hidden dimension and another 1D DFT along the sequence dimension.

The Fourier attention is efficient compared to MHA in two ways. First, the model parameters of the linear projection layer can be eliminated. Secondly, the parallel dataflow is reduced to a single 2D Fourier transformation. The implementation of a Fourier attention layer can be done in two ways, a fast Fourier

This research was supported in part by the MSIT, Korea, under the ITRC support program (IITP-2021-0-02052) supervised by the IITP, and in part by the UIRP (Undergraduate Interdisciplinary Research Project) program of UNIST Innovative Education Center.

Transform (FFT) method and a matrix multiplication based method, namely the DFT method. The N-point FFT method has a lower computational complexity of $O(N \log N)$ whereas the DFT method can make use of GEMM operations and therefore could be optimized for inference accelerators. Table 1 summarizes the operation and model parameter cost for each case.

TABLE I. COMPUTATIONAL COMPLEXITY COMPARSION

| | Operations | Model Parameters |
|---|---|---|
| **MHA** | $2n^2d_h + 4nd_h^2$ | $3nd_h$ |
| **DFT-FA** | $n^2d_h + nd_h^2$ | 0 |
| **FFT-FA** | $nd_h \log n + nd_h \log d_h$ | 0 |

$n$ is the sequence length, $d_h$ is the model's hidden dimension.

## III. ACCELERATING DFT ON VTA

The discrete Fourier transform (DFT) converts a complex-valued sequence of length $N$ into a sequence of the same length as shown in (1).

$$X_k = \sum_{0}^{N-1} x_n \cdot e^{-\frac{2\pi i}{N}kn} \qquad (1)$$

To efficiently compute a length $N$ DFT, it can be converted into a matrix form (2), where $w$ is the $N$th root of identity $e^{-2\pi i/N}$.

$$W = \frac{1}{\sqrt{N}}\begin{bmatrix} 1 & 1 & 1 & 1 & \cdots & 1 \\ 1 & w & w^2 & w^3 & \cdots & w^{N-1} \\ 1 & w^2 & w^4 & w^6 & \cdots & w^{2(N-1)} \\ 1 & w^3 & w^6 & w^9 & \cdots & w^{3(N-1)} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & w^{N-1} & w^{2(N-1)} & w^{3(N-1)} & \cdots & w^{(N-1)(N-1)} \end{bmatrix} \qquad (2)$$

However, $W$ is a $N \times N$ complex matrix whereas most of the DNN accelerators do not support complex arithmetic natively. Using the fact that an arbitrary $N \times M$ complex matrix can be transformed into a $2N \times 2M$ real matrix as shown in (3), the DFT matrix can be transformed into a real matrix form.

$$\begin{bmatrix} a_{11} + ib_{11} & a_{12} + ib_{12} \\ a_{21} + ib_{21} & a_{22} + ib_{22} \end{bmatrix} \mapsto \begin{bmatrix} a_{11} & -b_{11} & a_{12} & -b_{12} \\ b_{11} & a_{11} & b_{12} & a_{12} \\ a_{21} & -b_{21} & a_{22} & -b_{22} \\ b_{21} & a_{21} & b_{22} & a_{22} \end{bmatrix} \qquad (3)$$

This transform enforces the total computation to increase by four times. In the case of Fourier attention, the imaginary part of the input vector is always zero, and only the real part of the output vector is passed to the next layer. Using this observation, we further optimize the DFT computation by removing redundant operations for the imaginary part of the input and output vectors. Fig. 2 illustrates the optimization of 2D DFT matrix for Fourier attention. By introducing this optimization, the computation is decreased by half.

## IV. EVALUATION

To evaluate the performance of our proposed optimization of Fourier attention on VTA, we synthesized the default configuration of VTA on a Xilinx PNYQ-Z2 board. The MHA layer and the Fourier attention layer were compiled by using the TVM compiler, quantized to 8-bits. In the case of Fourier attention, the DFT matrix was pre-computed and stored inside the VTA before runtime. The inference latency of different sequence lengths and hidden dimensions $d_h$ was measured for both cases. The results are presented in Fig. 3.

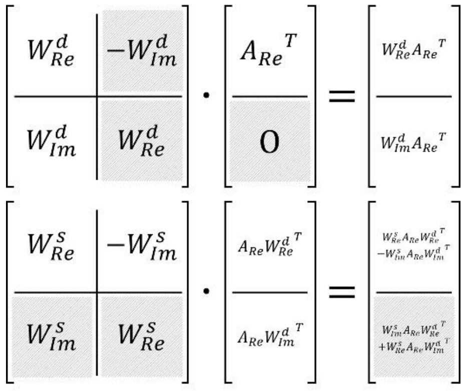

Figure 2. Optimization of 2D DFT matrix for Fourier attention. The gray area indicates redundant operations therefore can be ignored during computation.

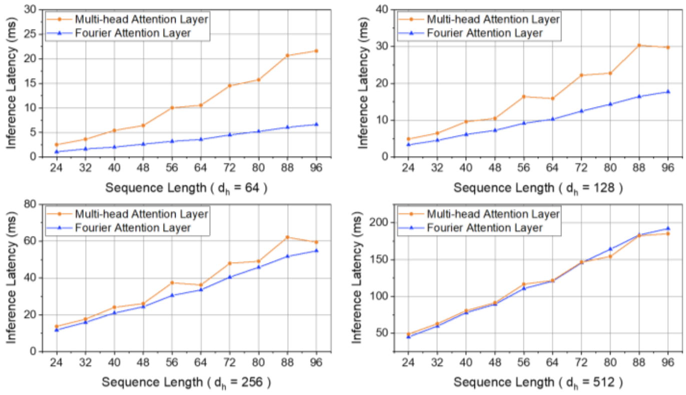

Figure 3. VTA Inference latency comparison of MHA and Fourier attention layer for various sequence lengths and number of hidden dimensions $d_h$.

Overall, we observe that Fourier attention has a better speedup for cases with lower $d_h$ and longer sequence length. There was no significant performance improvement for $d_h$ higher than 512. For $d_h$ of 64 and sequence length of 88 Fourier attention had a maximum improvement of 70.7%. Under cases when $d_h$ is smaller than 512 Fourier attention outperformed MHA by a geometric mean of 29.6%.

## V. CONCLUSION

This paper presents an efficient mapping of Fourier attention on VTA, a DNN accelerator for edge devices. To address the challenge of avoiding complex arithmetic, we propose a novel optimization of DFT matrices for Fourier attention. Our results demonstrate that Fourier attention can provide inference latency improvement up to 70.7% and 29.6% on average compared to MHA layers.

## REFERENCES

[1] James Lee-Thorp *et al.*, "FNet: Mixing tokens with fourier transforms." arXiv preprint arXiv:2105.03824 (2021).

[2] Thierry Moreau *et al.*, "A hardware–software blueprint for flexible deep learning specialization." IEEE Micro 39.5 (2019): 8-16.

[3] Ashish Vaswani *et al.*, "Attention is all you need." Advances in neural information processing systems 30 (2017).