Can we make a homegrown GPT?

Hyeonjin Jo*, Jaewoo Park*, and Chaerin Sim‡
Instructor: Jongeun Lee †

Department of CSE*, School of New UNIStars‡, Department of EE†
Ulsan National Institute of Science and Technology (UNIST), Ulsan, South Korea

ULSAN NATIONAL INSTITUTE OF SCIENCE AND TECHNOLOGY 2007

Intelligent **C**omputing & **C**odesign **L**aboratory

Introduction

Keywords: Natural Language Processing, Multi-head Attention, Fourier Transform, Field Programmable Gate Array, Hardware-Software Co-Design

In recent years, **Transformer** has achieved an outstanding performance in Natural Language Processing, such as ChatGPT. This remarkable outcome is attributed to ***Multi-head attention***, which consists of multiple projection layers and matrix multiplications.

However, the quadratic complexity of computation and memory toward sequence length has led to poor adaptability to deep learning accelerators.

Hence, finding more efficient attention with lower complexity has emerged as a new research trend.

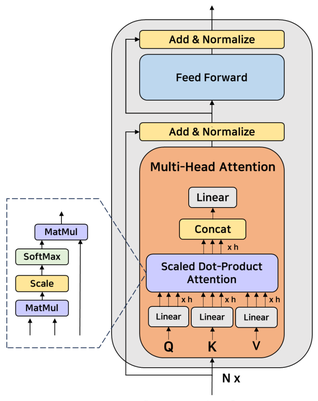


Figure 1. Transformer Architecture

Fnet and Fourier Transform

The **Fnet** replaces the Multi-head attention with a ***Fourier attention***, which performs a two-dimensional ***discrete Fourier transform (DFT)*** on input vectors.

By Leveraging the efficient Fast Fourier transform (FFT) on GPUs and eliminating parameters for Fourier attention, Fnet outperforms transformer with up to 5x speedup on training with comparable model accuracy.

The DFT converts a sequence into a complex-valued sequence in the frequency domain. To efficiently compute a N-length DFT, it can be represented as a matrix form, a ***DFT matrix***.

$$y = \Re\left(\mathcal{F}_{\text{seq}}\left(\mathcal{F}_{\text{h}}(x)\right)\right)$$

$\mathcal{F}_{\text{h}}$: *1D DFT along the hidden dimension*
$\mathcal{F}_{\text{seq}}$: *1D DFT along the sequence dimension*

Equation 1. Fourier Attention

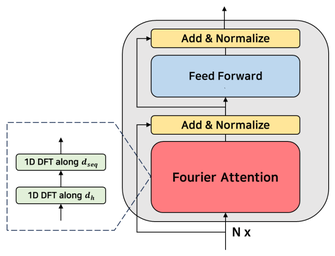


Figure 2. Fnet Architecture

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

Equation 2. Discrete Fourier Transform

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

where $w = e^{-2\pi i/N}$

Equation 3. Discrete Fourier Transform matrix

	Operations	Parameters
Multi-head Attention	$2n^2 d_h + 4nd_h^2$	$3nd_h$
Fourier Attention(DFT)	$n^2 d_h + nd_h^2$	0
Fourier Attention(FFT)	$nd_h \log n + nd_h \log d_h$	0

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

Table 1. Analysis of Computational Complexity

Motivation & Challenges

Motivation

- Inference Fnet on systolic array-based edge devices

Challenges

Most deep neural network accelerators are designed to

- Optimize General Matrix Multiplication or Convolution operations, not DFT
- Support Low-bit integer arithmetic ⬌ DFT requires complex arithmetic

Our Approach: Mapping 2D DFT

Decompose Complex Arithmetic

- Convert complex arithmetic to integer arithmetic

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

: Each Elements

- Rearrange the sequence of elements

$$\begin{bmatrix} a_{11} & -b_{11} & a_{12} & -b_{12} \\ b_{11} & a_{11} & b_{12} & a_{12} \\ a_{21} & -b_{21} & a_{22} & -b_{22} \\ b_{21} & a_{21} & b_{22} & a_{22} \end{bmatrix} \mapsto \begin{bmatrix} a_{11} & a_{12} & -b_{11} & -b_{12} \\ a_{21} & a_{22} & -b_{21} & -b_{22} \\ b_{11} & b_{12} & a_{11} & a_{12} \\ b_{21} & b_{22} & a_{21} & a_{22} \end{bmatrix}$$

: Real Part : Imaginary Part

Increasing the computation by **a quadruple**

Optimize DFT Matrix Multiplications

- Utilize the known information about input and output

$$y = \Re\left(\mathcal{F}_{\text{seq}}\left(\mathcal{F}_{\text{h}}(x)\right)\right) \quad where, input\ x \in \mathbb{R}\ and\ output\ y \in \mathbb{R}$$

$\mathcal{F}_{\text{h}}$: *1D DFT along the hidden dimension*
$\mathcal{F}_{\text{seq}}$: *1D DFT along the sequence dimension*

$$\begin{bmatrix} W_{Re}^d & -W_{Im}^d \\ W_{Im}^d & W_{Re}^d \end{bmatrix} \cdot \begin{bmatrix} A_{Re}{}^T \\ O \end{bmatrix} = \begin{bmatrix} W_{Re}^d A_{Re}{}^T \\ W_{Im}^d A_{Re}{}^T \end{bmatrix} \Bigg| \begin{bmatrix} W_{Re}^s & -W_{Im}^s \\ W_{Im}^s & W_{Re}^s \end{bmatrix} \cdot \begin{bmatrix} A_{Re} W_{Re}^d{}^T \\ A_{Re} W_{Im}^d{}^T \end{bmatrix} = \begin{bmatrix} W_{Re}^s A_{Re} W_{Re}^d{}^T - W_{Im}^s A_{Re} W_{Im}^d{}^T \\ W_{Im}^s A_{Re} W_{Re}^d{}^T + W_{Re}^s A_{Re} W_{Im}^d{}^T \end{bmatrix}$$

$\mathcal{F}_{\text{h}}(x)$ $\qquad$ $\mathcal{F}_{\text{seq}}\left(\mathcal{F}_{\text{h}}(x)\right)$

Reducing the computation by **a half**

Experiment I: VTA & FPGA

Both Multi-head attention and our Fourier attention are compiled using TVM, an open-source, end-to-end optimizing compiler for machine learning.

The Inference latency of each layer is measured on the default configuration of VTA, an open-source, configurable deep learning accelerator, using Xilinx PYNQ-Z2 FPGA.

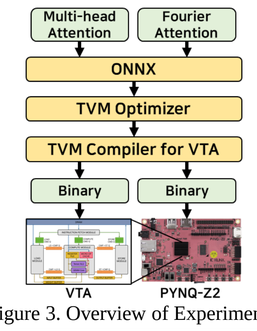


Figure 3. Overview of Experiment I

Experiment II: Gemmini & FireSim

Our optimized Fourier attention, non-optimized Fourier attention, and I-BERT, which is an integer-only transformer, were converted into binaries through the workload generator, FireMarshal.
The clock cycles spent on each sublayer of models are evaluated on Gemmini using FireSim. Gemmini is a full-stack generator of deep learning accelerator. FireSim is an FPGA-accelerated Cycle-exact simulation tool on Amazon EC2 F1 instance.

	Component	Configuration
CPU	Processor	Rocket
	Core	Single Core
Gemmini	Spatial Array	16 × 16
	Scratchpad	256 KB
	Accumulator	64 KB
	Dataflow	Weight-stationary

Table 2. Gemmini accelerator System-on-a-Chip Configuration

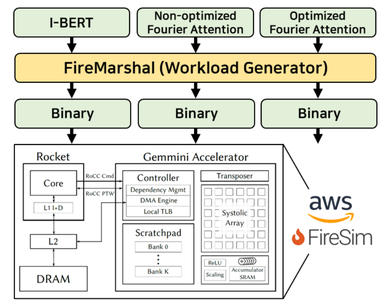
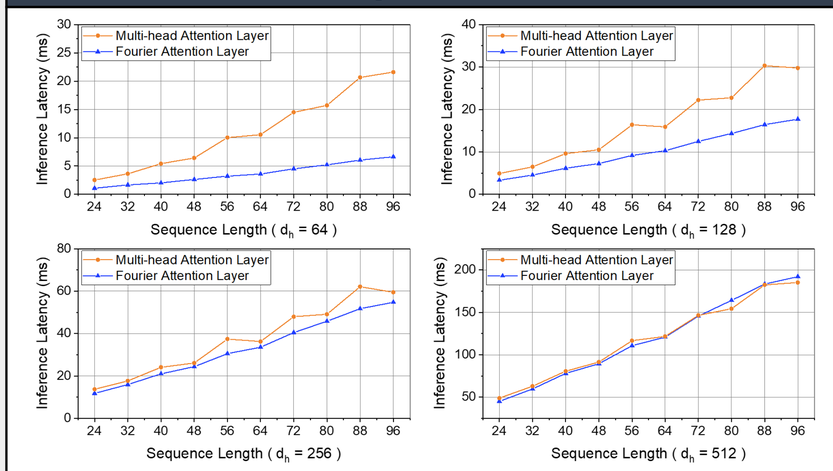
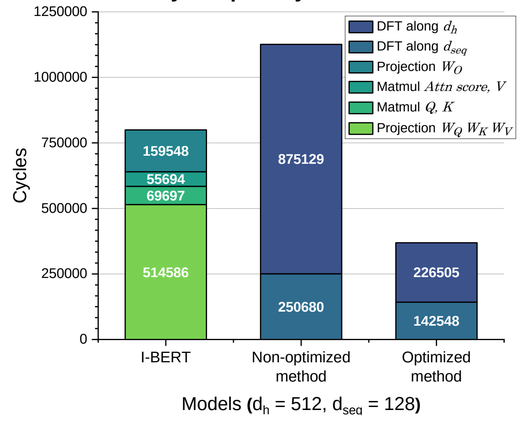


Figure 4. Overview of Experiment II

Experimental Results



Up to: **3.41 x**
Average: **1.42 x**
times improvement



vs I-BERT: **2.16 x**
vs Non-optimized: **3.05 x**
times improvement

Achievement & Future Plans

Achievement

- Publication and Oral Presentation at an international conference, ISOCC 2023
  Title: Accelerating Transformers with Fourier-Based Attention for Efficient On-Device Inference
  Authors: Hyeonjin Jo, Chaerin Sim, Jaewoo Park and Jongeun Lee

Future Plans

- Publish an extended paper with an efficient, end-to-end mapping of DFT
- Release implementations of our Fourier attention as an open-source

Conclusion

We propose an efficient mapping of DFT for edge devices

- **Experiment I** – Our approach enhances the inference latency by up to 70.7% and by 29.6% on average compared to the Multi-head attention
- **Experiment II** – Our optimized method reduces the total clock cycles by 53.8% compared to I-BERT and by 67.2% compared to the non-optimized approach