FNet: Mixing Tokens with Fourier Transforms

Google Research

정영석

Contents

Introduction

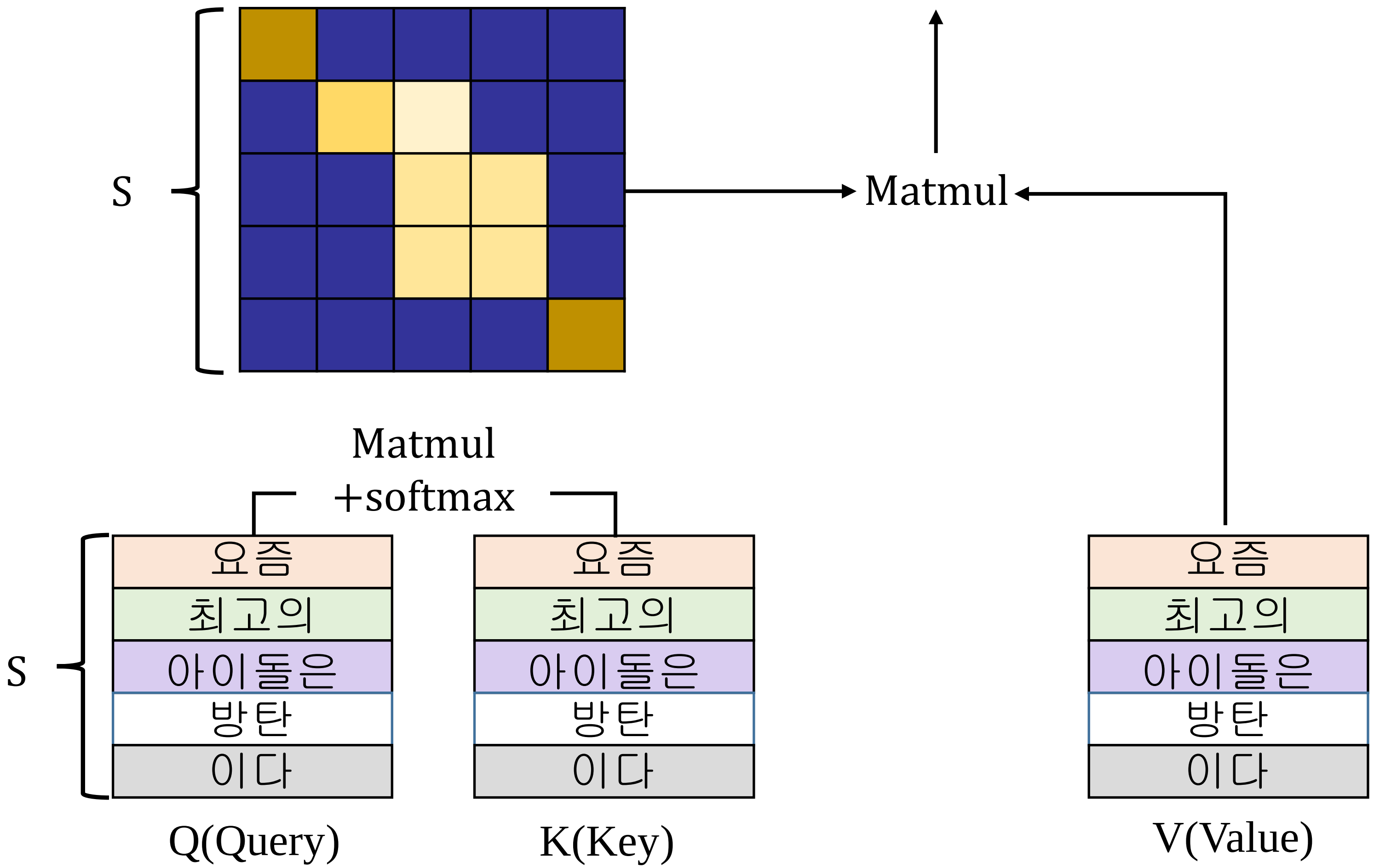
▪ Transformer

- At the heart of the Transformer is a self-attention mechanism.
    - An inductive bias that connects each token in the input through a relevance weighted basis of every other token.
    - Each hidden unit is represented in the basis of the hidden units of the previous layer.

"Self-Attention"

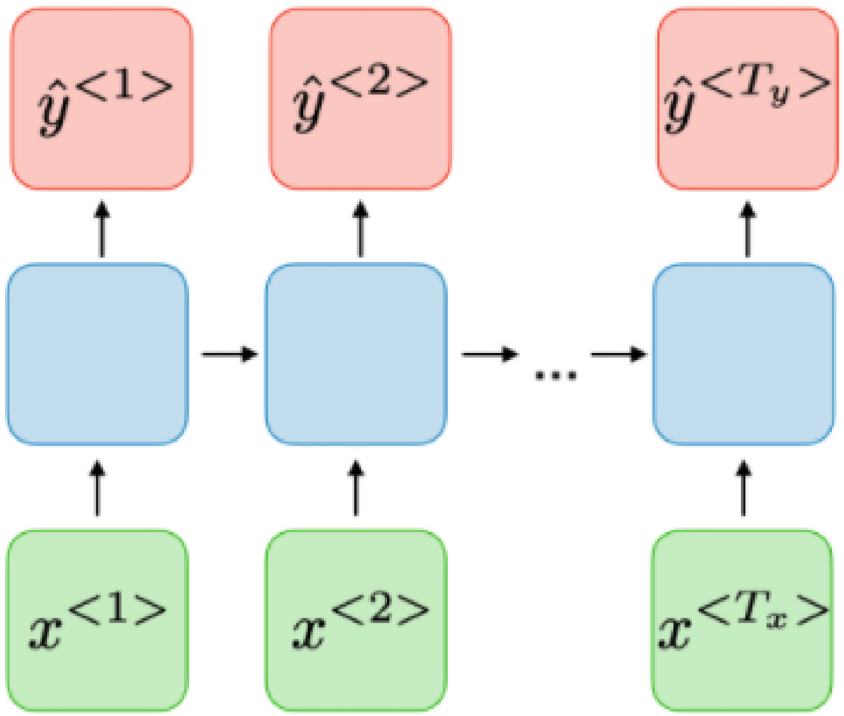
Introduction

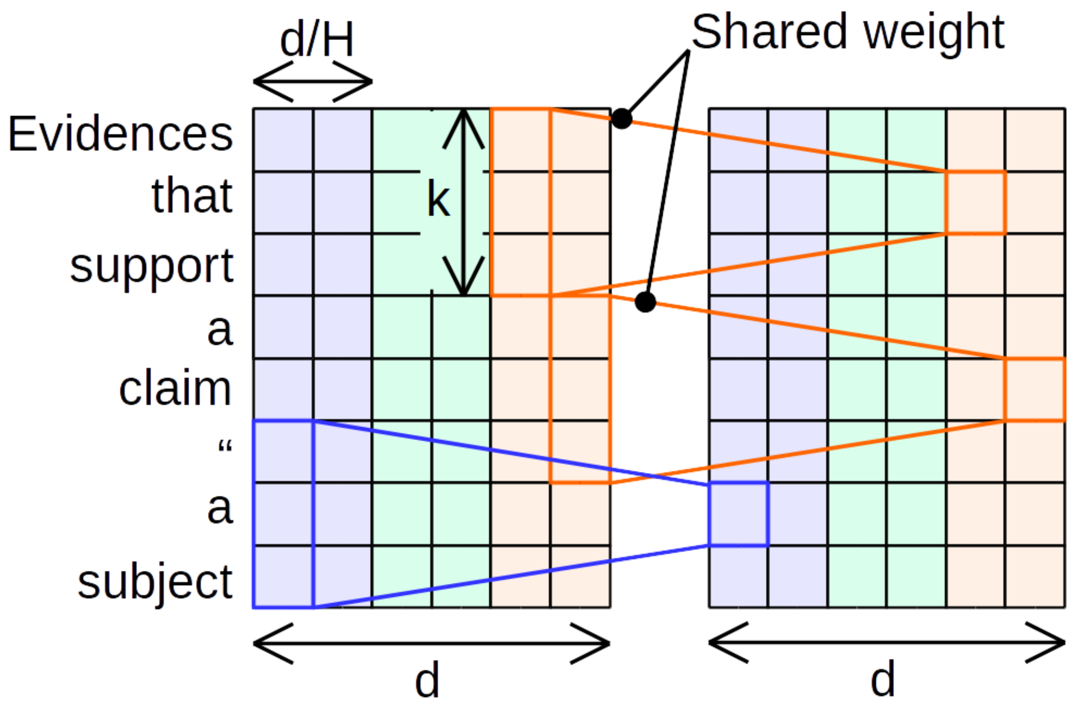
- **Self-Attention**



Introduction

- **Comparison with RNN and CNN**





Introduction

▪ Motivation of the FNet

- We investigate whether simpler mixing mechanisms can wholly replace the relatively complicated attention layers

  ✓ The standard self-attention mechanism (Vaswaniet al., 2017) has a quadratic time and memory bottleneck with respect to sequence length.

  ✓ This limits its applicability in text tasks involving **long range dependencies**, character-level modelling, speech processing, image and video processing

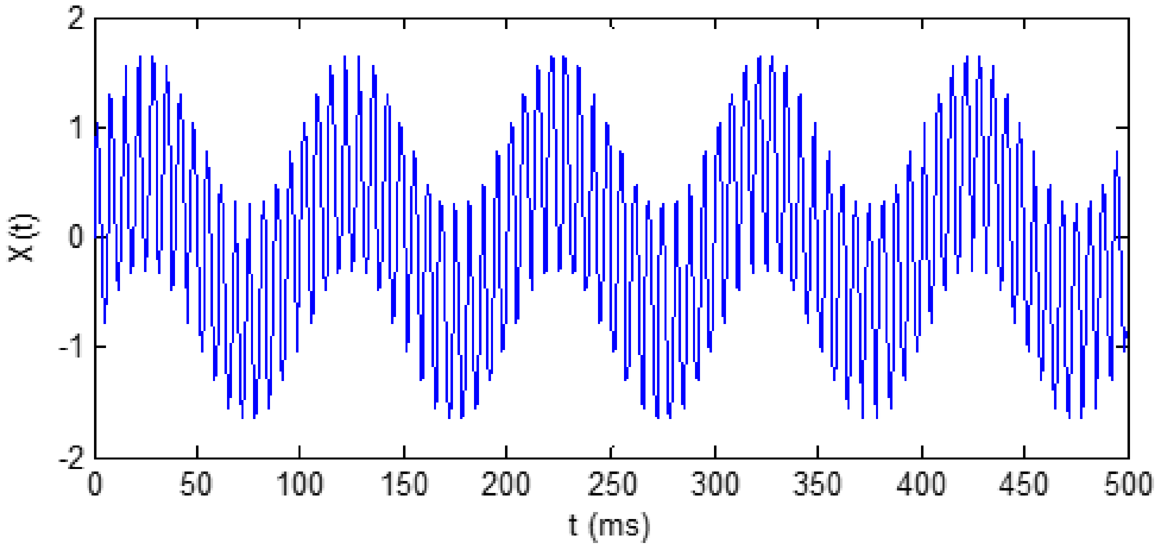
➢ Fourier Transform

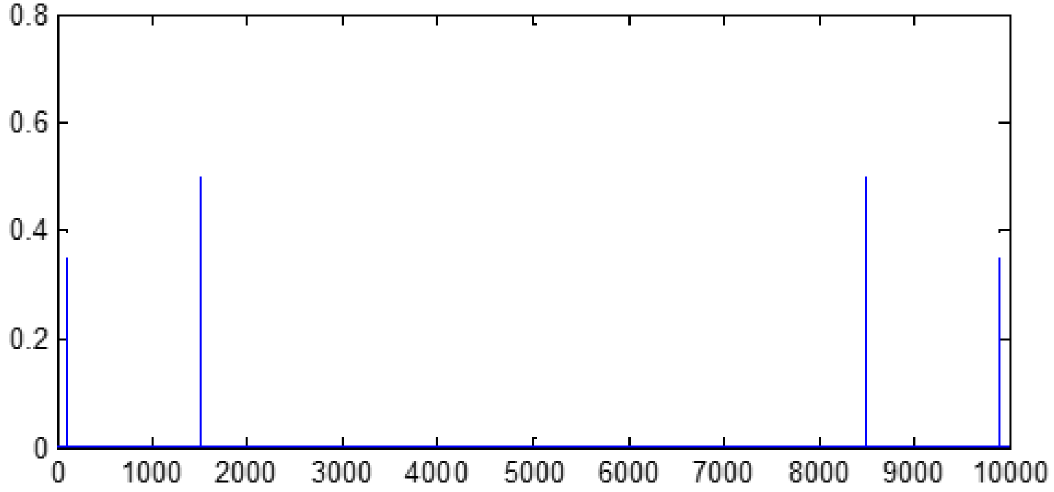
Introduction

▪ DFT (Discrete Fourier Transform)

- The Fourier Transform decomposes a function into its constituent frequencies.

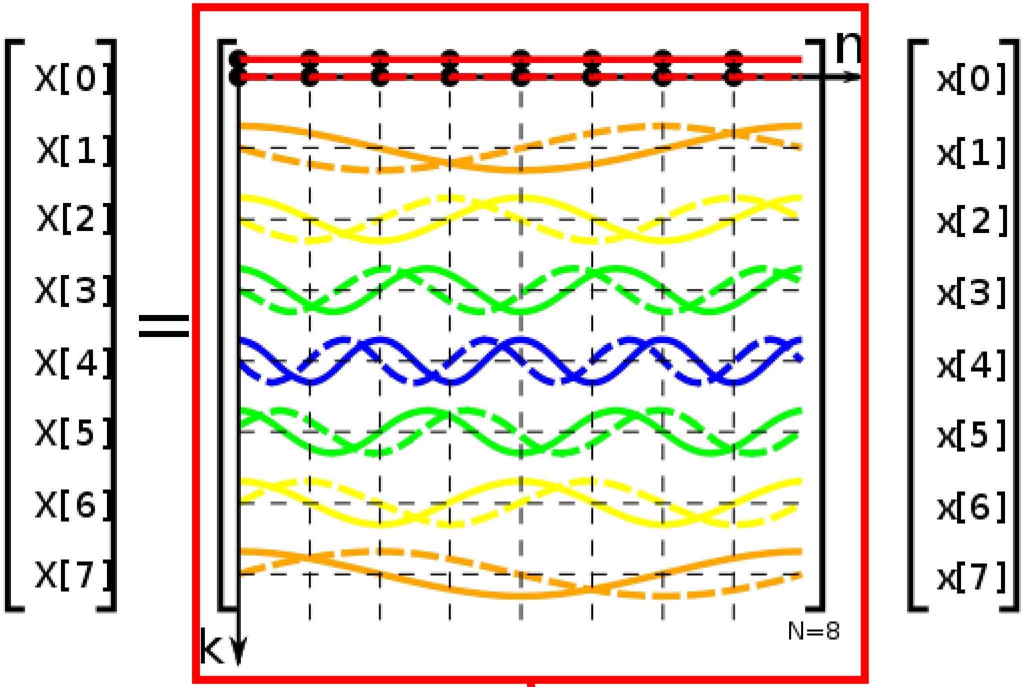
$$X_k = \sum_{n=0}^{N-1} x_n e^{-\frac{2\pi i}{N}nk}, 0 \leq k \leq N-1$$





Introduction

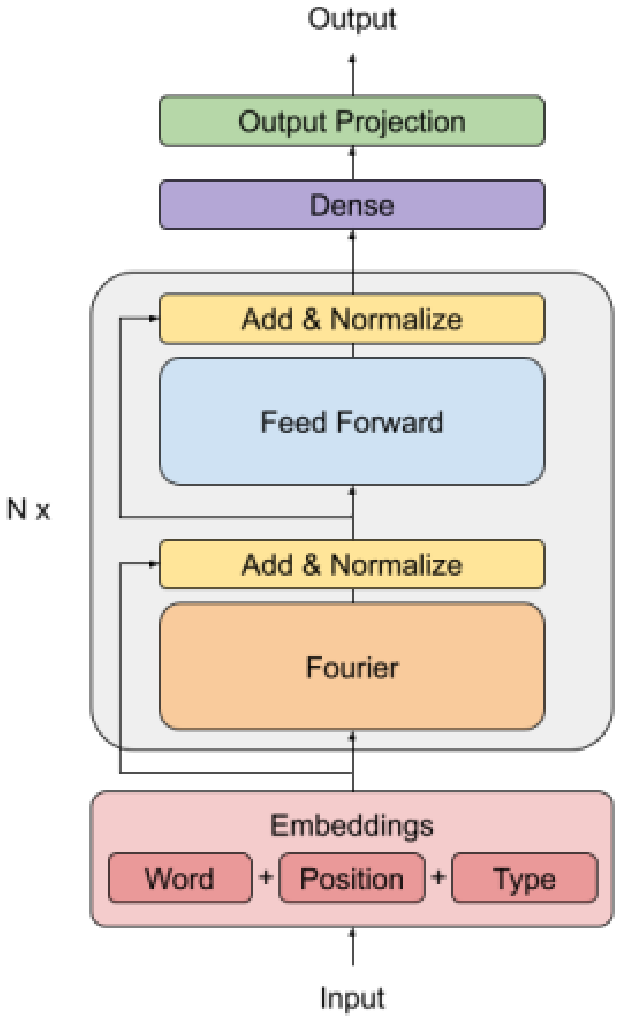
- DFT (Discrete Fourier Transform)



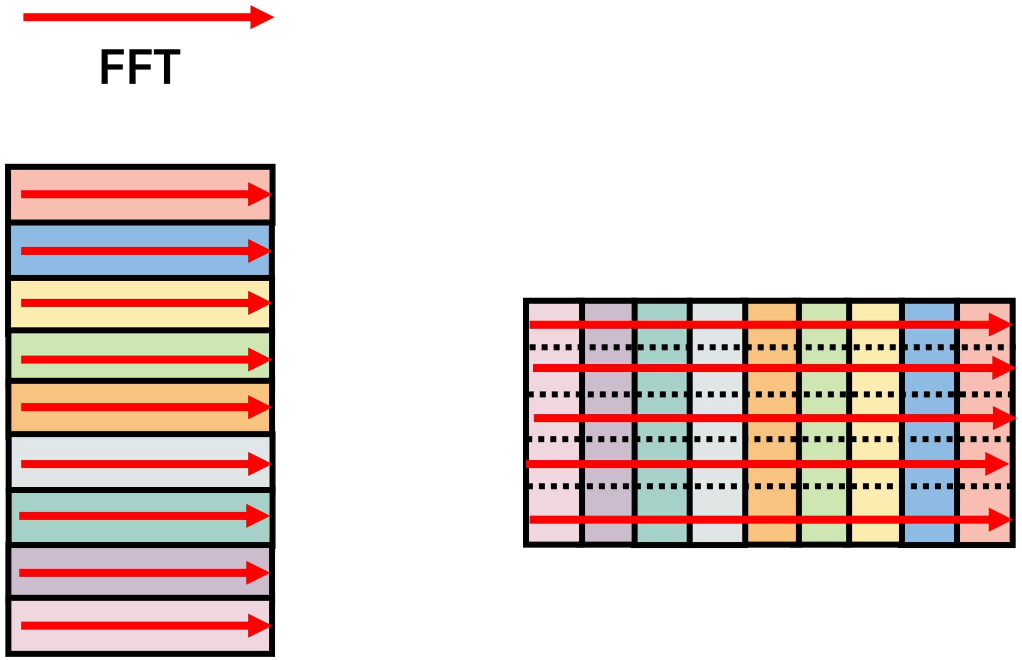
$$W = \frac{1}{\sqrt{N}} \begin{bmatrix} 1 & 1 & 1 & 1 & \cdots & 1 \\ 1 & \omega & \omega^2 & \omega^3 & \cdots & \omega^{N-1} \\ 1 & \omega^2 & \omega^4 & \omega^6 & \cdots & \omega^{2(N-1)} \\ 1 & \omega^3 & \omega^6 & \omega^9 & \cdots & \omega^{3(N-1)} \\ \vdots & \vdots & \vdots & \vdots & \ddots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

Method

- FNet Architecture



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



Result

▪ Comparison with other type model

- **Other type models**

✓ BERT-Base: a Transformer model.

✓ FNet encoder: we replace every self-attention sublayer with a Fourier sublayer

✓ Linear encoder: we replace each self-attention sublayer with a two learnable, dense, linear sublayers, one applied to the hidden dimension, and one applied to the sequence dimension.

✓ Random encoder: we replace each self-attention sublayer with a two constant random matrices, one applied to the hidden dimension, and one applied to the sequence dimension.

✓ Feed Forward-only (FF-only) encoder: we remove the self-attention sublayer from the Transformer layers; this leaves a ResNet with only feed-forward layers and no token mixing

Results

▪ Experiment Result

- MLM and NSP loss & Training speed of the models

Model	Loss			Accuracy	
	Total	MLM	NSP	MLM	NSP
BERT-Base	**1.76**	**1.48**	**0.28**	**0.68**	**0.86**
Linear-Base	2.12	1.78	0.35	0.62	0.83
FNet-Base	2.45	2.06	0.40	0.58	0.80
Random-Base	5.02	4.48	0.55	0.26	0.70
FF-only-Base	7.54	6.85	0.69	0.13	0.50
FNet-Hybrid-Base	2.13	1.79	0.34	0.63	0.84
BERT-Large	**1.49**	**1.23**	**0.25**	**0.72**	**0.88**
Linear-Large	1.91	1.60	0.31	0.65	0.85
FNet-Large	2.11	1.75	0.36	0.63	0.82

Model	GPU (64)	TPU (256)
BERT-Base	161	41
Linear-Base	28 (5.7x)	23 (1.8x)
FNet-Base	24 (6.9x)	21 (2.0x)
Random-Base	26 (6.1x)	21 (2.0x)
FF-only-Base	**21 (7.8x)**	**20 (2.0x)**
FNet-hybrid-Base	28 (5.7x)	22 (1.8x)
BERT-Large	FAIL	89
Linear-Large	FAIL	51 (1.8x)
FNet-Large	**70**	**44 (2.0x)**

* **Fnet-hybrid** : replace the final two Fourier sublayers of FNet with self-attention sublayers.

Results

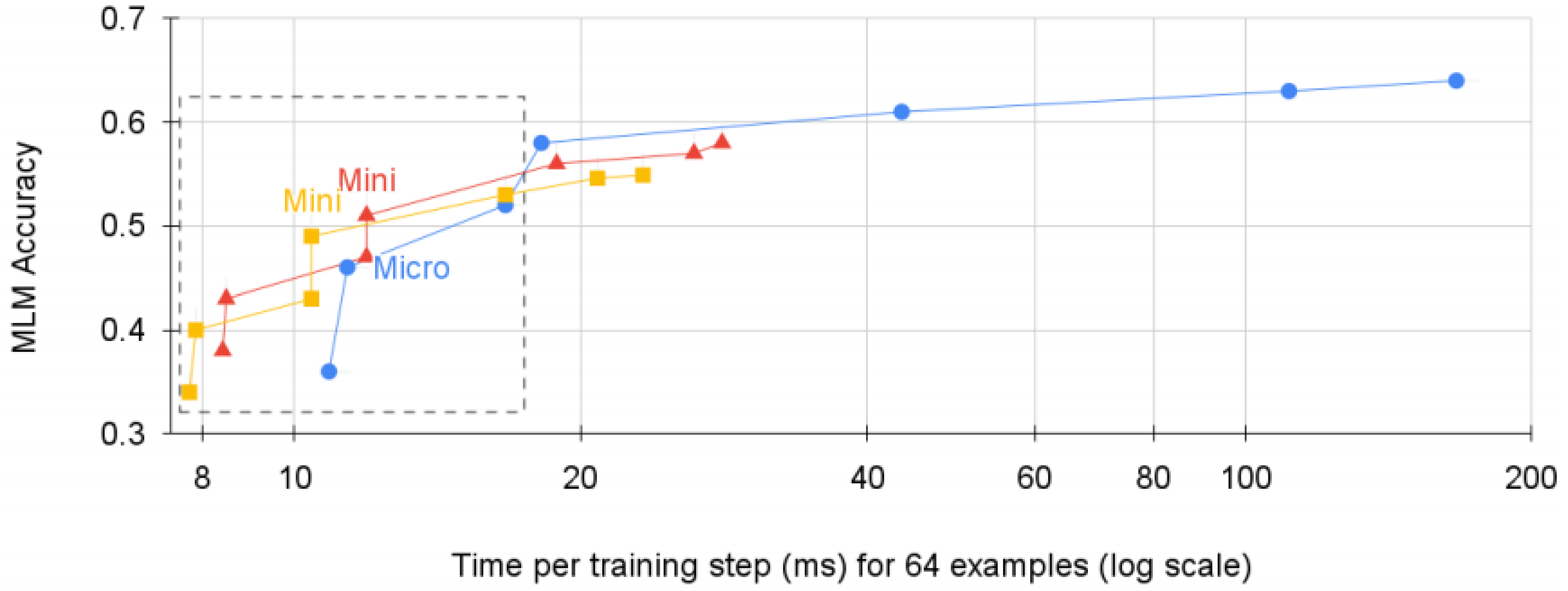
▪ Experiment Result

- GLUE (Text classification tasks) fine-tuning

Model	MNLI (m/mm)	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE	Avg.
BERT-Base	**84/81**	**87**	**91**	93	73	**89**	**83**	**69**	**83.3**
Linear-Base	74/75	84	80	94	67	67	83	69	77.0
FNet-Base	72/73	83	80	**95**	69	79	76	63	76.7
Random-Base	51/50	70	61	76	67	4	73	57	56.6
FF-only-Base	34/35	31	52	48	67	FAIL	73	54	49.3[4]
FNet-Hybrid-Base	78/79	85	88	94	**76**	86	79	60	80.6
BERT-Large	**88/88**	**88**	**92**	**95**	71	**88**	86	66	**84.7**
Linear-Large	35/36	84	80	79	67	24	73	60	59.8
FNet-Large	78/76	85	85	94	**78**	84	**88**	**69**	81.9

Results

- Speed and accuracy trade-offs



Results

- ## Long-Range Arena benchmark

Model	ListOps	Text	Retrieval	Image	Pathfinder	Path-X	Avg.
Transformer	36.06	61.54	59.67	41.51	80.38	FAIL	55.83
Linear	33.75	53.35	58.95	41.04	83.69	FAIL	54.16
FNet	35.33	65.11	59.61	38.67	77.80	FAIL	55.30

Seq. length	512	1024	2048	4096	8192	16386
	GPU					
Transformer	26	11	4	FAIL	FAIL	
Linear	49 (1.9x)	23 (2.0x)	11 (2.6x)	4	FAIL	
FNet (FFT)	**60 (2.3x)**	**30 (2.7x)**	**16 (3.9x)**	**8**	**4**	
Performer	32 (1.3x)	19 (1.6x)	10 (2.3x)	5	2	
	TPU					
Transformer	43	16	5	1	FAIL	FAIL
Linear	78 (1.8x)	**62 (3.8x)**	**28 (5.7x)**	**12 (9.9x)**	4	FAIL
FNet (matmul)	**92 (2.1x)**	61 (3.8x)	26 (5.4x)	11 (8.8x)	4	1
FNet (FFT)	36 (0.8x)	25 (1.5x)	13 (2.7x)	7 (5.4x)	3	1
Performer	59 (1.4x)	42 (2.6x)	23 (4.6x)	**12 (9.9x)**	**6**	**3**