FNet + BART – Is **Fourier Transform** all you need ?

Intro to Natural Language Processing (Spring 2025)

February 17, 2025

Project Outline

NYPD

Aniruth Suresh (2022102055)
Aryan Garg (2022102074)
Samkit Jain (2022102062)

February 17 ,2024

INTERNATIONAL INSTITUTE OF
INFORMATION TECHNOLOGY
HYDERABAD

1 Introduction

Natural Language Processing (NLP) has been dominated by transformer-based architectures, but their computational complexity remains a challenge. The **FNet** [1] architecture, which replaces **self-attention with Fourier Transform**, offers a promising alternative with reduced computational overhead. This project aims to implement and benchmark FNet on various NLP tasks, comparing its performance across different datasets and evaluation metrics.

Furthermore, we explore an untested application: **integrating FNet within the BART model**. We evaluate its effectiveness on **JEDI: Justifiable End-dialogue Driven Interaction for NPC Entities in Role-Playing Games**, a dataset designed for modeling conversational agents in RPG environments.

2 Motivation

As **Electrical and Computer Engineering (ECE)** undergraduates, we have encountered Fourier Transforms extensively throughout our coursework. Inspired by its applications in signal processing, we were eager to investigate its role in NLP. The recent development of FNet, which utilizes Fourier Transforms instead of traditional attention mechanisms, aligns with our background.

Unlike self-attention, which has a computational complexity of $O(n^2)$, FNet leverages the Fa**st Fourier Transform (FFT)**, reducing the complexity to $O(n \log n)$. This significant improvement in efficiency arises because FFT decomposes the input sequence into a sum of sinusoidal components, allowing for faster global mixing of tokens compared to the quadratic overhead of self-attention.

3 Overview of FNet

FNet [1] is a transformer-based model that replaces the traditional self-attention mechanism with a token-mixing operation using the Fourier Transform. This modification significantly reduces computational complexity while maintaining competitive performance in various NLP tasks.

3.1 Fourier Transform and the Vandermonde Matrix

The Discrete Fourier Transform (DFT) can be represented as a matrix multiplication:

$$X_k = \sum_{n=0}^{N-1} x_n e^{-2\pi ink/N} \tag{1}$$

where X_k is the transformed sequence, and W, known as the DFT matrix is a **Vandermonde matrix** defined as:

$$W_{nk} = \frac{e^{-2\pi ink/N}}{\sqrt{N}} \tag{2}$$

where N is the sequence length, and i is the imaginary unit. The DFT matrix performs **global token mixing** in a single step, leveraging the Fourier basis to transform inputs efficiently. This avoids the need for explicit pairwise token interactions, unlike self-attention.

3.2 Why This Replacement Works ?

The authors justify replacing self-attention with Fourier Transforms by highlighting that the Discrete Fourier Transform (DFT) matrix enables **global mixing of token representations in the frequency domain**, which reduces computational complexity from the quadratic $O(n^2)$ of self-attention to $O(n \log n)$ via the Fast Fourier Transform (FFT), making the process significantly faster.

4 Literature Survey

4.1 BERT

BERT [3] is a transformer-based model designed to process text **bidirectionally**, capturing contextual information more effectively than previous unidirectional models. It uses **multiple layers of self-attention mechanisms** that enable the model to learn deep semantic relationships between words in a sentence.

BERT is pre-trained on vast amounts of textual data and can then be fine-tuned for various Natural Language Processing (NLP) tasks, such as question answering, sentiment analysis, named entity recognition, and text classification. This pre-training and fine-tuning approach allows BERT to achieve state-of-the-art results on numerous NLP benchmarks without training the entire architecture over and over again.

- **Pre-Training**:
  - **Masked Language Model (MLM)**: Randomly masks tokens in the input sentence and trains the model to predict the missing words, enabling bidirectional understanding.
  - **Next Sentence Prediction (NSP)**: Helps in understanding relationships between sentences, making it useful for tasks like question answering.

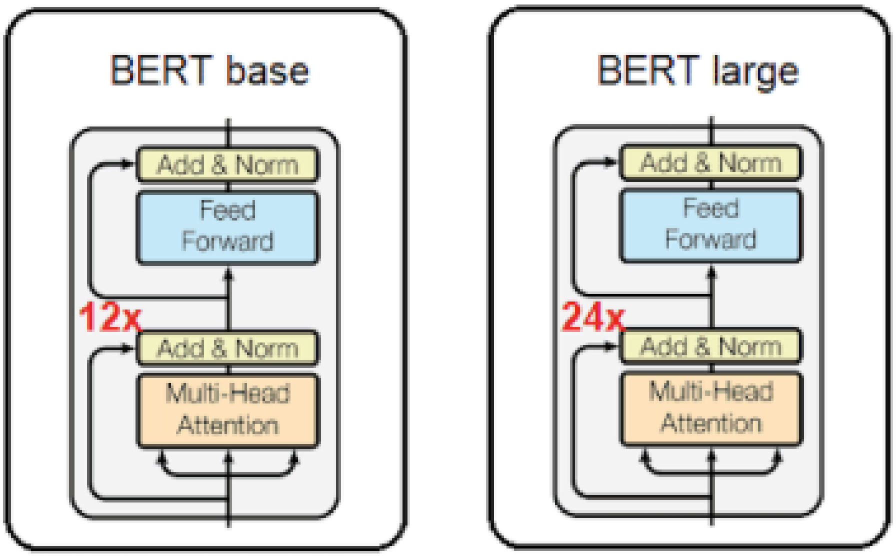

Figure 1: BERT Architecture

While BERT demonstrates state-of-the-art performance in various NLP benchmarks, its self-attention mechanism, scales **quadratically** with input size, posing computational challenges, particularly for long sequences.

4.2 BART

Unlike BERT and FNET, which primarily function as encoder-only models, BART extends the transformer framework by making use of **both an encoder and a decoder**, making it suitable for text generation tasks. BART [4] essentially makes use of the decoder half of the transformer architecture after the BERT encoders for generating text.
The primary characteristics of BART include:

- **Denoising Autoencoder Approach**: Trained by corrupting text (e.g., through token deletion, sentence shuffling) and learning to reconstruct the original input.

- **Bidirectionality and Autoregressiveness**: Combines bidirectional encoding, as in BERT, with an autoregressive decoder, making it a more generalized framework.

5 Project Outline

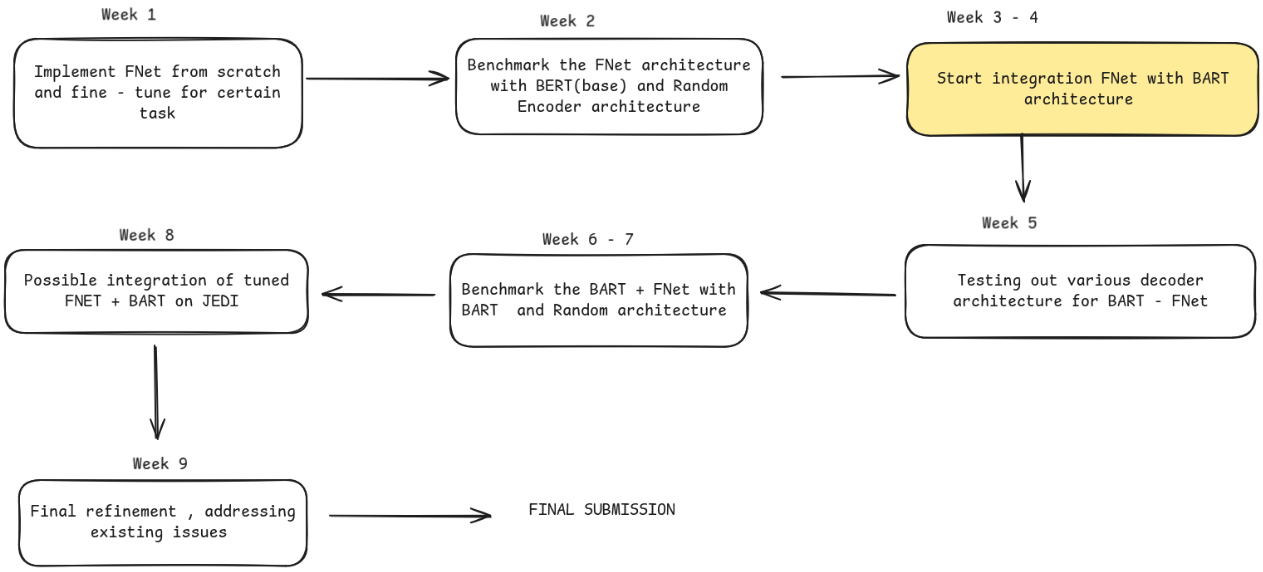
Week 1
Implement FNet from scratch and fine - tune for certain task
Week 2
Benchmark the FNet architecture with BERT(base) and Random Encoder architecture
Week 3 - 4
Start integration FNet with BART architecture
Week 5
Testing out various decoder architecture for BART - FNet
Week 6 - 7
Benchmark the BART + FNet with BART and Random architecture
Week 8
Possible integration of tuned FNET + BART on JEDI
Week 9
Final refinement , addressing existing issues
FINAL SUBMISSION

Figure 2: Project Plan

6 Project Objective

6.1 Overview

Transformer-based models like BART rely on self-attention mechanisms to capture contextual relationships, but these layers scale quadratically with sequence length, leading to high computational demands. Recent work on FNET demonstrates that **replacing attention with parameter-free Fourier Transforms** can achieve comparable performance to vanilla transformers on classification and generation tasks, with significant gains in training speed (up to 80% faster on GPUs).

6.2 Methodology

- Architectural Modifications:
  - Replace BART's **self-attention sub-layers with FNET's Fourier Transform layers**, which compute global token interactions via Fast Fourier Transforms (FFTs).
  - Investigate replacing cross-attention in the decoder: BART's decoder requires cross-attention to encoder outputs. We will test whether **Fourier or inverse Fourier transforms** can approximate this cross-modality interaction.
  - Preserve BART's encoder-decoder structure, positional embeddings, and feed-forward layers to retain its generative capabilities.
- Training and Evaluation:
  - Train the modified BART-FNET model on standard benchmarks, e.g., **GLUE and SWAG datasets**
  - Compare performance against baseline BART using metrics like Accuracy, F1-Score and MCC **(Matthews Correlation Coefficient)**
  - Quantify efficiency gains via **GPU memory usage, training time, and FLOPs**.

6.3 Further Exploration

JEDI [5] explores how large language models (LLMs) can enhance dynamic, context-sensitive dialogue in story-driven RPGs, using *Star Wars: Knights of the Old Republic* (KOTOR) as a case study. Unlike traditional dialogue systems that rely on pre-written scripts—lacking adaptability—the **JEDI project fine-tunes BART on graph-structured dialogue data and game state information** to improve the responsiveness and immersion of in-game conversations.

Using the FNET + BART model, we will benchmark its performance on the **KOTOR dataset**, evaluating it using metrics such as BLEU, ROUGE-1, and ROUGE-2 to assess the quality of generated dialogue responses in comparison to ground-truth references.

7 Datasets and Metrics explored

Dataset	**Description**	**Usage**	**Metrics**
1. GLUE – CoLA	Corpus of Linguistic Acceptability (Warstadt et al., 2018) consists of English acceptability judgments drawn from books and journal articles on linguistic theory.	Each example is a sequence of words annotated with whether it is a grammatical English sentence. BERT and FNET both papers have this.	Matthews Correlation Coefficient (MCC): For the formula, please refer to here. The value of this coefficient lies in the range [-1, 1], and a value of 0 indicates random output.
2. SWAG	The SWAG (Situations With Adversarial Generations) dataset is a multiple-choice commonsense reasoning benchmark. It consists of context sentences and four possible continuations, where the goal is to select the most plausible one.	Evaluates a model's ability to understand and generate commonsense-based continuations. Used in NLP benchmarks for natural language inference and commonsense reasoning tasks.	Accuracy and F1 score.
3. SST-2	The SST-2 dataset is a binary sentiment classification dataset derived from the Stanford Sentiment Treebank (SST). It consists of sentences from movie reviews, where each sentence is labeled as either positive (1) or negative (0).	Binary sentiment classification, where the task is to classify whether a sentence from a movie review is positive or negative.	Accuracy and F1 score. Used in BERT and FNET both.

Table 1: Dataset Descriptions, Usage, and Metrics

References

[1] J. Lee-Thorp, J. Ainslie, I. Eckstein, and S. Ontanon, *FNet: Mixing Tokens with Fourier Transforms* [Online]. Available: https://arxiv.org/abs/2105.03824

[2] N. Akoury, Q. Yang, and M. Iyyer, A Framework for Exploring Player Perceptions of LLM-Generated Dialogue in Commercial Video Games, in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

[3] I. Turc, M.-W. Chang, K. Lee, and K. Toutanova, *Well-Read Students Learn Better: On the Importance of Pre-training Compact Models*. [Online]. Available at https://arxiv.org/abs/1810.04805

[4] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, *ART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension*, Facebook AI, [Online]. Available at https://arxiv.org/abs/2106.01621

[5] W. Chan, O. Abul-Hassan, and S. Sun, *EDI: Justifiable End-dialogue Driven Interaction for NPC Entities in Role-Playing Games*, Stanford CS224N Custom Project, [Online]. Available at https://web.stanford.edu/class/cs224n/final-reports/256911920.pdf

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, *Attention Is All You Need*, in *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*, 2017. [Online]. Available: https://arxiv.org/abs/1706.03762