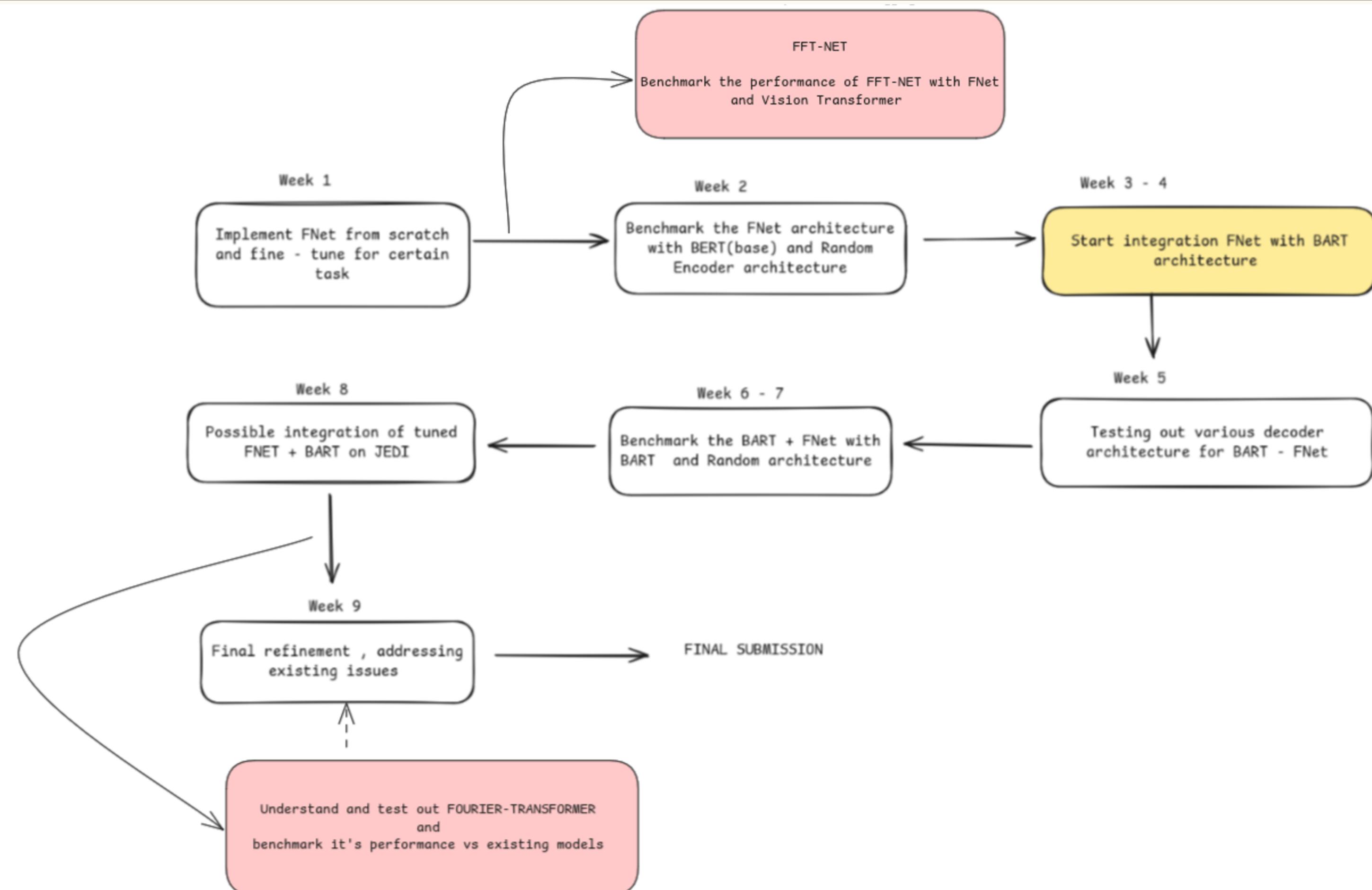
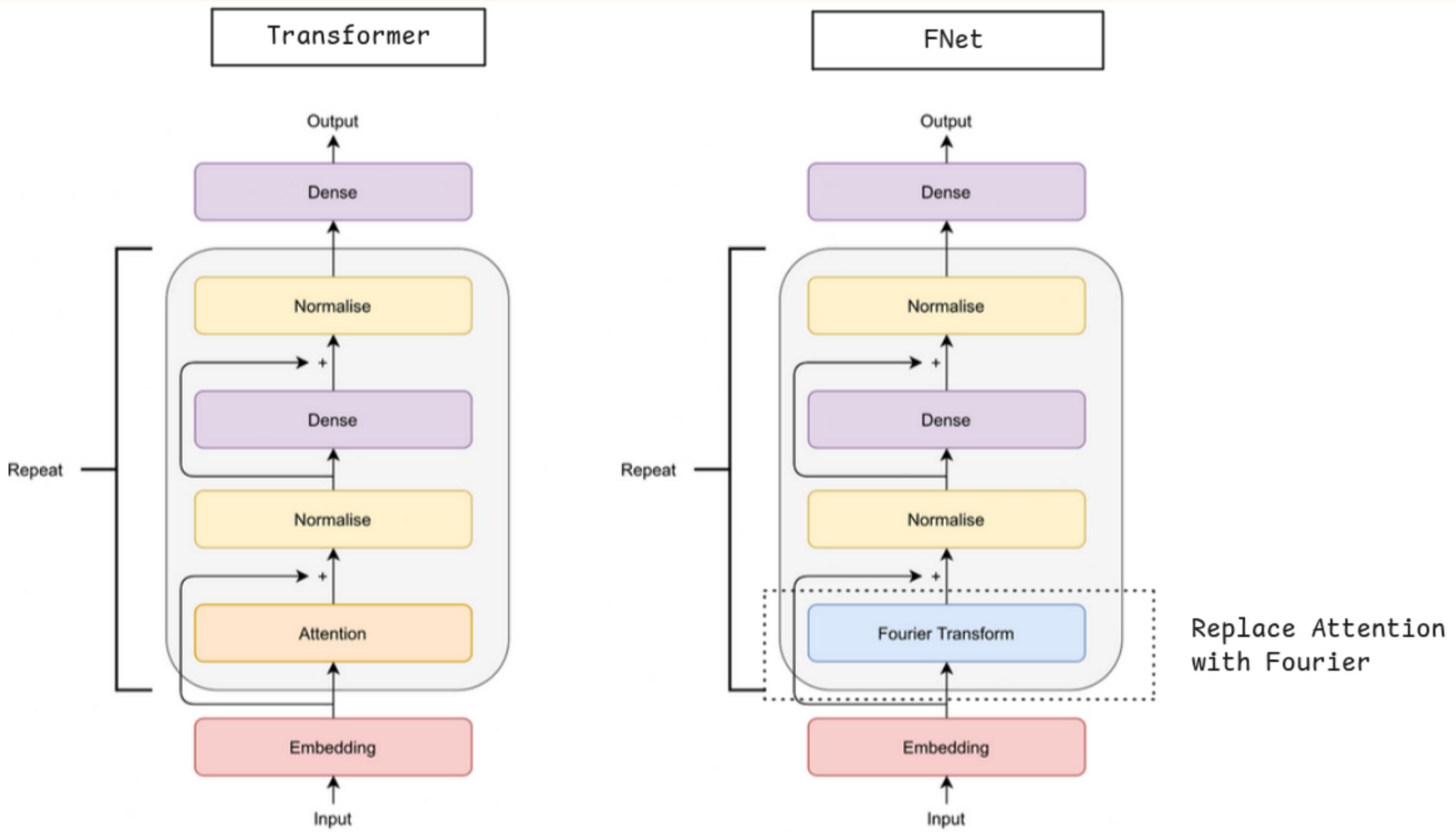


# FNET + BART : IS FOURIER ALL YOU NEED ?

Samkit Jain , Aryan Garg , Aniruth Suresh



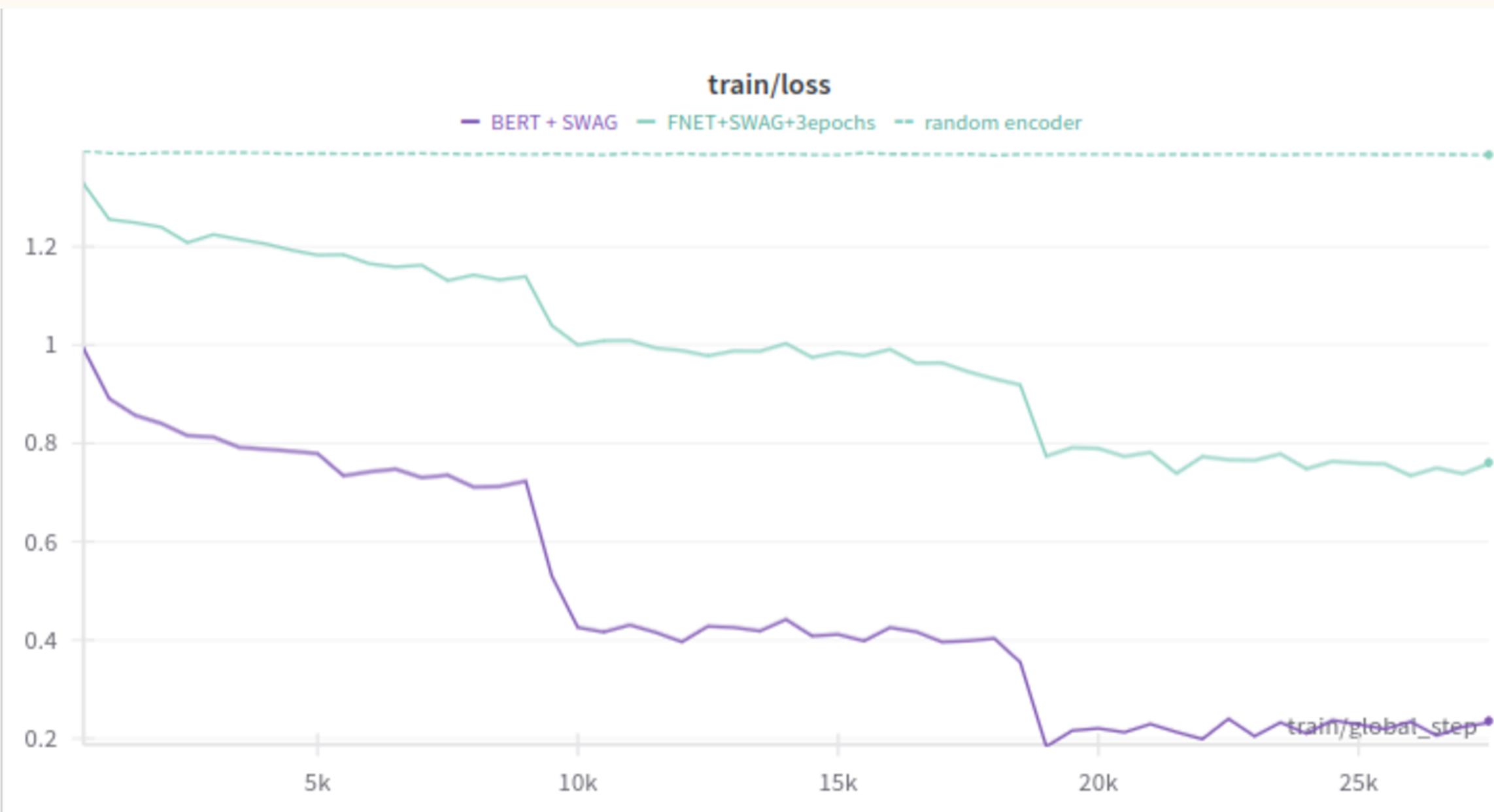
$$X(f) = \int_{-\infty}^{\infty} x(t)e^{-j2\pi ft} dt$$



F-Net swaps the standard, computationally heavy self-attention layers found in Transformers.

Uses DFT as the core operation for mixing of information. Capable of performing this mixing without the need for learnable parameters.

Reduces computational complexity from  $O(n^*n)$  to  $O(n \log n)$



Name (2 visualized)	Runtime
BERT + SWAG	2h 4m 15s
FNET+SWAG+3epochs	46m 11s

3 models :

1. Fnet + SWAG
2. BERT + SWAG
3. Random (replaces attention by fixed matrices)

## OBSERVATIONS :

1. Clearly from the above graph , we observe that we can't just replace attention by any random fixed matrix .
2. This clearly signifies the importance using **Fourier Transform** .

**Remember** : FT has no learnable parameters !

```
# Example test case
context = "The weather was getting colder and the leaves were falling from the trees."
choices = [
    "She decided to wear a light summer dress.",
    "He put on a heavy winter coat.",
    "They went to the beach to enjoy the sun.",
    "The sun was shining brightly in the sky."
]

predicted_index = predict(model, tokenizer, context, choices)
print(f"Predicted choice: {choices[predicted_index]}")
```

```
(fart) aniruth.suresh@gnode010:~$ python3 check.py
Evaluating: 100%
Accuracy: 0.7800
F1 Score: 0.7799
Predicted choice: He put on a heavy winter coat.
(fart) aniruth.suresh@gnode010:~$
```

Given a context and set of four options , F-Net predicted the most appropriate one which around 78% accuracy and F<sub>1</sub> -score !

The screenshot shows the arXiv preprint page for the paper "The FFT Strikes Again: An Efficient Alternative to Self-Attention". The page header includes the arXiv logo, a search bar, and navigation links for Help and Ad. The main content area displays the title, authors, abstract, and a summary of the paper's contributions. The abstract discusses the quadratic complexity of self-attention and how FFTNet addresses this by mapping inputs into the frequency domain, using Parseval's theorem to model long-range dependencies, and combining local windowing with a global FFT branch.

arXiv > cs > arXiv:2502.18394

Computer Science > Machine Learning

(Submitted on 25 Feb 2025 (v1), last revised 16 Mar 2025 (this version, v5))

**The FFT Strikes Again: An Efficient Alternative to Self-Attention**

Jacob Fein-Ashley, Rajgopal Kannan, Viktor Prasanna

Conventional self-attention mechanisms exhibit quadratic complexity in sequence length, making them challenging to scale for long inputs. We present FFTNet, an adaptive spectral filtering framework that uses the Fast Fourier Transform (FFT) to achieve global token mixing in  $\mathcal{O}(n \log n)$  time. By mapping inputs into the frequency domain, FFTNet exploits orthogonality and energy preservation-guaranteed by Parseval's theorem-to efficiently model long-range dependencies. Our main theoretical contributions include 1) An adaptive spectral filter that highlights salient frequency components, 2) A hybrid scheme combining local windowing with a global FFT branch, 3) Nonlinear feature transformations applied in both the frequency and token domains. Experiments on Long Range Arena and ImageNet validate our theoretical insights and demonstrate superior performance over fixed Fourier-based and standard attention models.

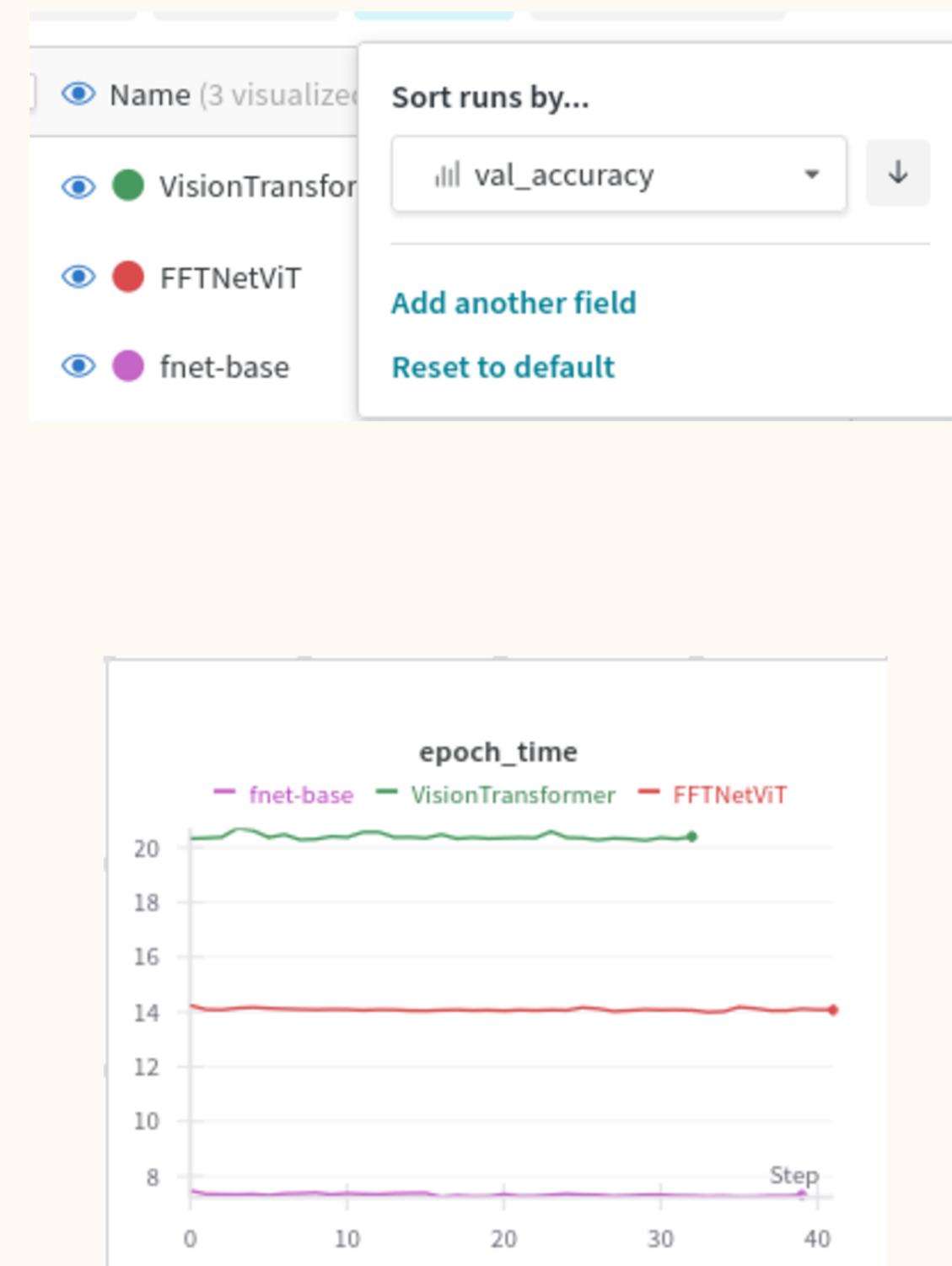
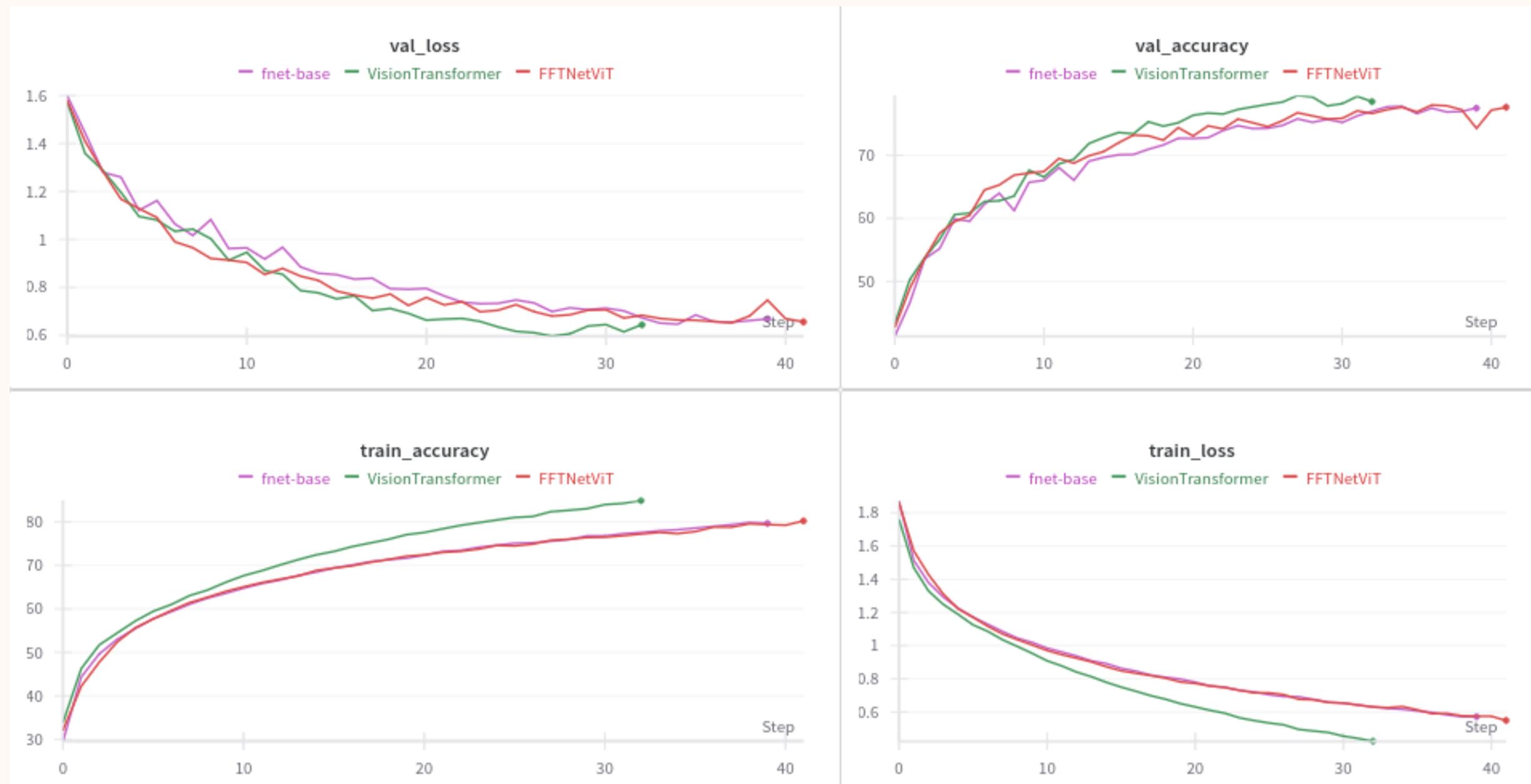
Paper published on **16<sup>th</sup> March 2025** that builds on the idea of F-Net.

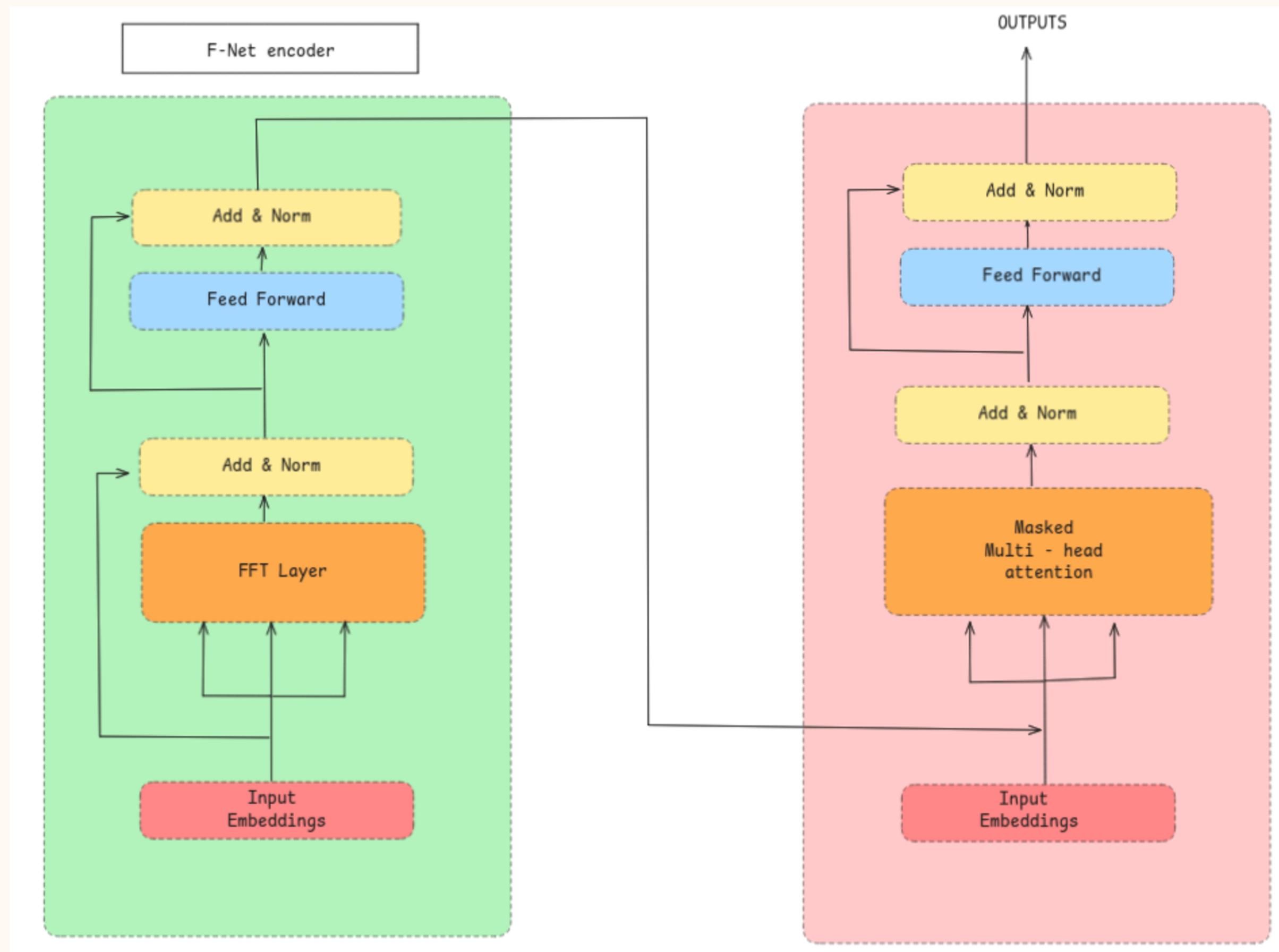
It introduces a learnable, context dependent filter in the frequency domain to dynamically emphasize or attenuate required frequency components, enhancing its performance over the fixed parameters of F-Net.

It also applies a non-linear activation function (modReLU) directly to the complex FT coefficients after filtering, resulting in better representation of token dependencies.

Even though this adds some computational overhead compared to the base F-Net, the complexity is still about  $O(n \log n)$ . Better benchmarks on datasets are also observed on finetuning.

# Task : CIFAR - 10 classification





FFT mixing incorporated in BART by replacing the attention heads by FFT Layers. (currently , we have tested replacing it in encoder . We further plan to extrapolate it to decoder stage as well) .



Model	Time (5 visualized)	Runtime
bart-base-sst2	~13m 33s	13m 33s
base-fnet-sst2	~22m 52s	22m 52s

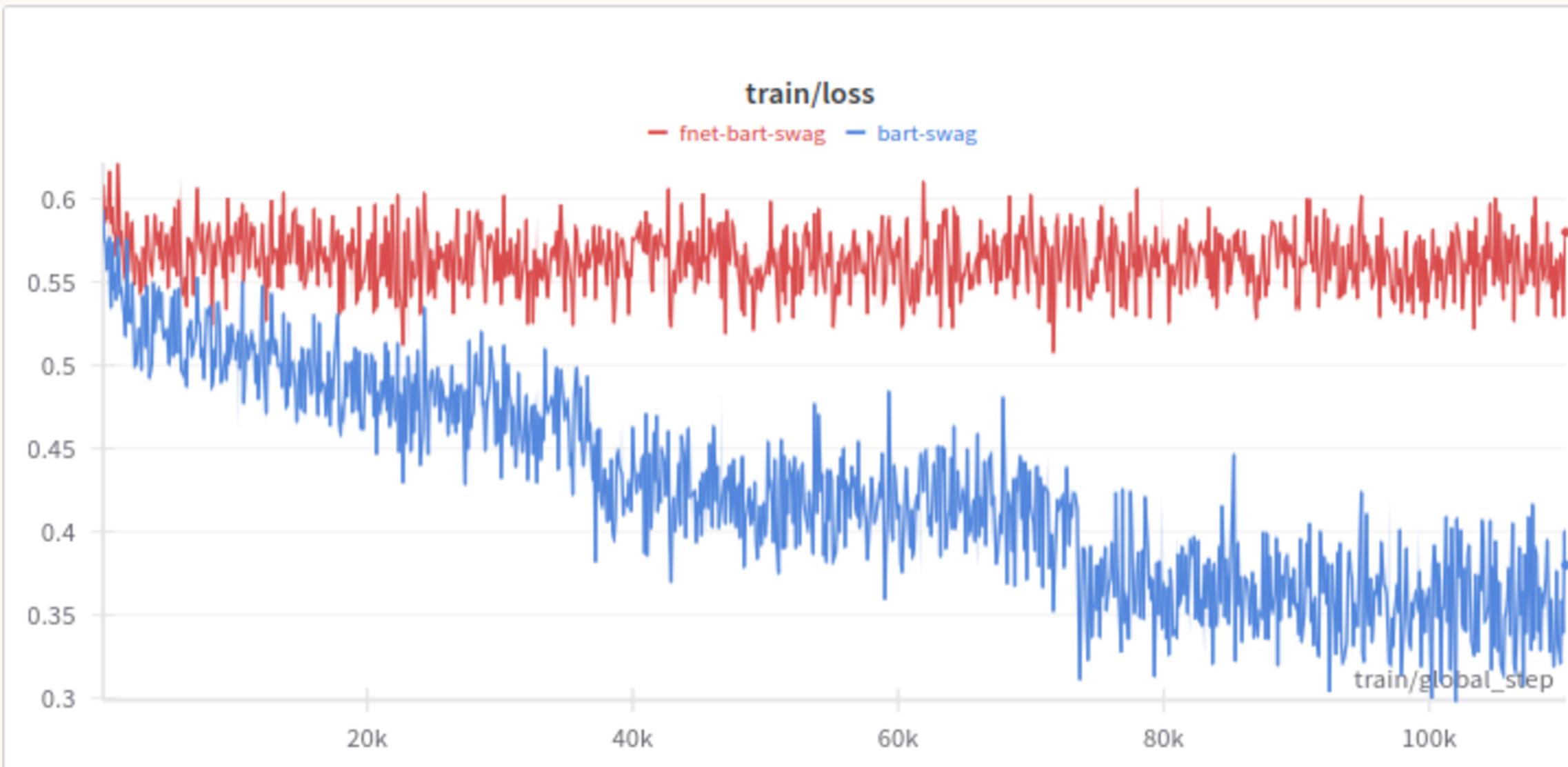
We train on the **SST-2 Dataset** which performs well for both regular BART and BART with FNet layers .

Time - Accuracy Tradeoff

<b>Model</b>	<b>Eval Accuracy</b>	<b>Eval Loss</b>	<b>Total time taken</b>
Base BART	92.231%	0.40153	22m 52s
BART + FNet	83.37%	0.559	<b>13m 33s</b>

```
warnings.warn(
{'eval_loss': 0.4015326499938965, 'eval_accuracy': 0.9231651376146789, 'eval_runtime': 1.4848, 'eval_samples_per_second': 587.272, 'eval_steps_per_second': 73.409, 'epoch': 3.0}
{'train_runtime': 1430.5598, 'train_samples_per_second': 141.236, 'train_steps_per_second': 17.655, 'train_loss': 0.21926027940411502, 'epoch': 3.0}
100%|██████████| 109/109 [00:01<00:00, 75.65it/s]
{'eval_loss': 0.35456112027168274, 'eval_accuracy': 0.9277522935779816, 'eval_runtime': 1.4543, 'eval_samples_per_second': 599.621, 'eval_steps_per_second': 74.953, 'epoch': 3.0}
```

```
warnings.warn(
{'eval_loss': 0.5593359470367432, 'eval_accuracy': 0.8337155963302753, 'eval_runtime': 1.301, 'eval_samples_per_second': 670.257, 'eval_steps_per_second': 83.782, 'epoch': 3.0}
{'train_runtime': 1249.9381, 'train_samples_per_second': 161.646, 'train_steps_per_second': 20.207, 'train_loss': 0.38505867073455396, 'epoch': 3.0}
100%|██████████| 109/109 [00:01<00:00, 86.53it/s]
{'eval_loss': 0.5593359470367432, 'eval_accuracy': 0.8337155963302753, 'eval_runtime': 1.2711, 'eval_samples_per_second': 685.997, 'eval_steps_per_second': 85.75, 'epoch': 3.0}
```

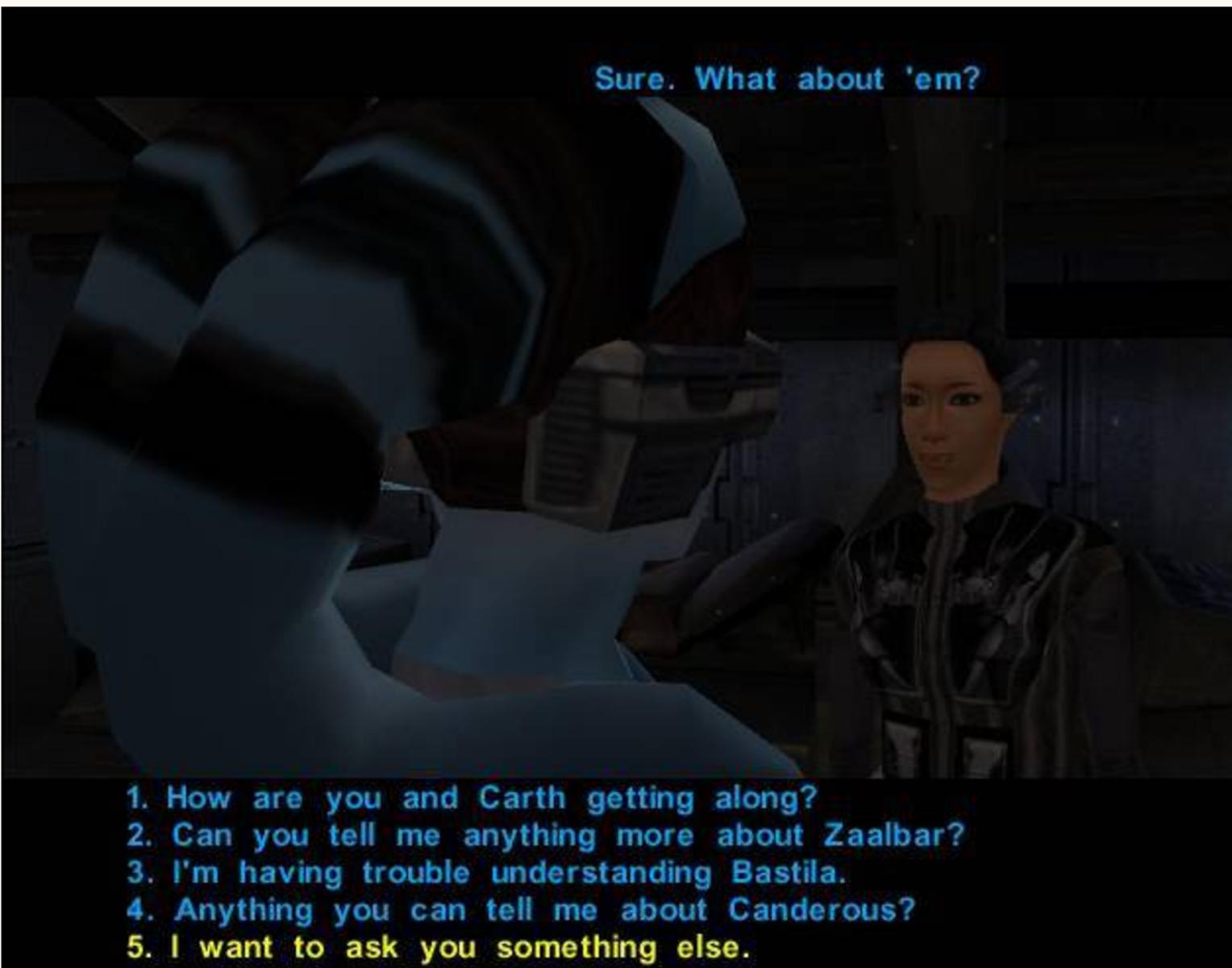


**Reasons for poor performance :**  
SWAG depends heavily on understanding context and making inferences, which often benefits from attention. By replacing the self-attention layer with Fourier Transforms the model may be **losing important semantic information** required for accuracy.

**CLAIM :** Implementing F-Net on BART encoder seems to perform well on simple tasks like sst-2 , MNIST and CIFAR but fails to generalize to heavy complex task like SWAG which requires context !!

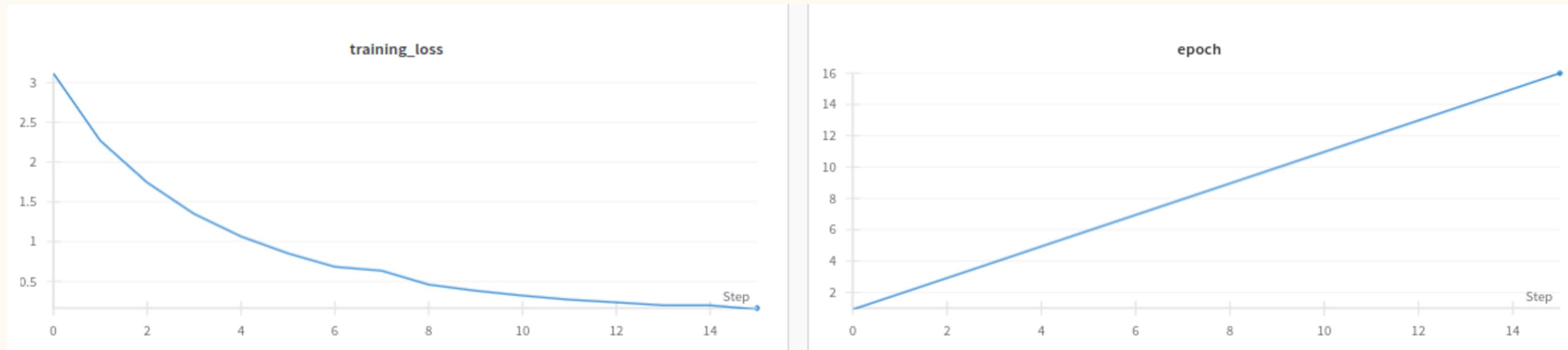
Not a simple one - to - one fixed conversations => it depends on player choices, character stats, and game states

## Star Wars: "Knights of the Old Republic" (KOTOR)



Key idea : Represent "Dialogue as a graph"

1. **Nodes** = individual dialogue utterances and **Edges** = transitions between utterances, which are determined by the game state
2. Similar dialogue nodes are grouped using clustering algorithms (A basic threshold F<sub>1</sub> score based algo is implemented).
3. Graph is linearized
4. During training, one utterance is masked at a time within this sequence. The model is asked to predict the masked line given the other lines in the cluster and the current game state !



```
Average Precision: 0.8625
Average Recall: 0.8591
Average F1 Score: 0.8606
```

```
DialogRPT Score: 0.6154
Average DialogRPT Score: 0.5027941809351749
(fart) aniruth.suresh@gnode076:~/JEDI$
```

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

1.  $Q$  = current dialogue input
2.  $K, V$  = representation of the game state
3.  $QK^T$  = computes how well each element in the dialogue input (query) matches each element in the game state (similarity score )

1. Plan on implementing **adaptive filtering techniques** similar to that used in FFTNet to make use of complex information from FFT instead of just taking real part .
2. Setup and benchmark Fourier Transformer which uses spectral filtering using Fourier Transform .

The image shows a screenshot of an arXiv preprint page. The title is "Fourier Transformer: Fast Long Range Modeling by Removing Sequence Redundancy with FFT Operator". It was submitted on May 24, 2023. The authors listed are Ziwei He, Meng Yang, Minwei Feng, Jingcheng Yin, Xinbing Wang, Jingwen Leng, Zhouhan Lin. The abstract discusses the inefficiency of transformer models due to self-attention and introduces the Fourier Transformer, which uses an FFT operator to perform Discrete Cosine Transformation (DCT) to reduce computational costs while retaining the ability to inherit weights from pretrained models. The code is available at [https://url\(this https URL\)](https://url(this https URL)).

3. Plan to modify the decoder architecture of BART and analyze the performance .
4. Integrate the FNet BART models on JEDI and compare and analyze the results .

All active code, results, and run details are documented.  
(As of mid-submission, there are 6 active branches.)

The image shows a GitHub repository page for "AniruthSuresh/FART—INLP". The repository is described as "Repo to maintain all the codes for INLP Project". It has 1 contributor, 0 issues, 0 stars, and 0 forks. The repository URL is [https://github.com/AniruthSuresh/FART--INLP](#). A note below states: "Repo to maintain all the codes for INLP Project. Contribute to AniruthSuresh/FART---INLP development by creating an account on GitHub." There is also a GitHub icon.

# THANK YOU

FOR YOUR **ATTENTION:)**