Visual Abstract of Preliminary Results

Introduction

This research project investigates how three different embedding fusion mechanisms affect BERT's behavior. Here are specific models in this project with different embedding.

1.Base model where token embeddings are fused with word and positional embeddings by simple element-wise addition.
2.Conv2D model that uses a small convolutional kernel to combine embedding components in both sequence and hidden size dimensions.
3.FFT model that uses FFT-based circular convolution to fuse embeddings in hidden size dimension.

Conv2D and FFT model are different implementations of initial embedding convolution. This idea is inspired by Tony Plate's work using convolution of hyper-dimensional vectors to estimate sentence similarity, and it achieves good accuracy without machine learning models. [Plate, 1990] Therefore, it is worthwhile to explore the role of various embedding structures in deep learning NLP models, as well as its contribution and potential in representation learning.

Analysis techniques and preliminary results including:
(1) Training performance (accuracy and loss over epochs)
(2) Feature dynamics during training across layers and epochs levels
(3) Initial embedding space structure via t-SNE visualizations, token cosine similarity heatmaps, and statistic data visualization.
(4) Probing tasks presenting accuracy of simple linear regression probing classifiers on intermediate layer representations, with TF-IBF baseline, CLS hidden states extracted from trained models and random initialized models.

For preliminary experiments, only SST2 dataset are used. All models are trained under 32 batch size, 2e-5 learning rate under different epochs, considering variations of model learning patterns.

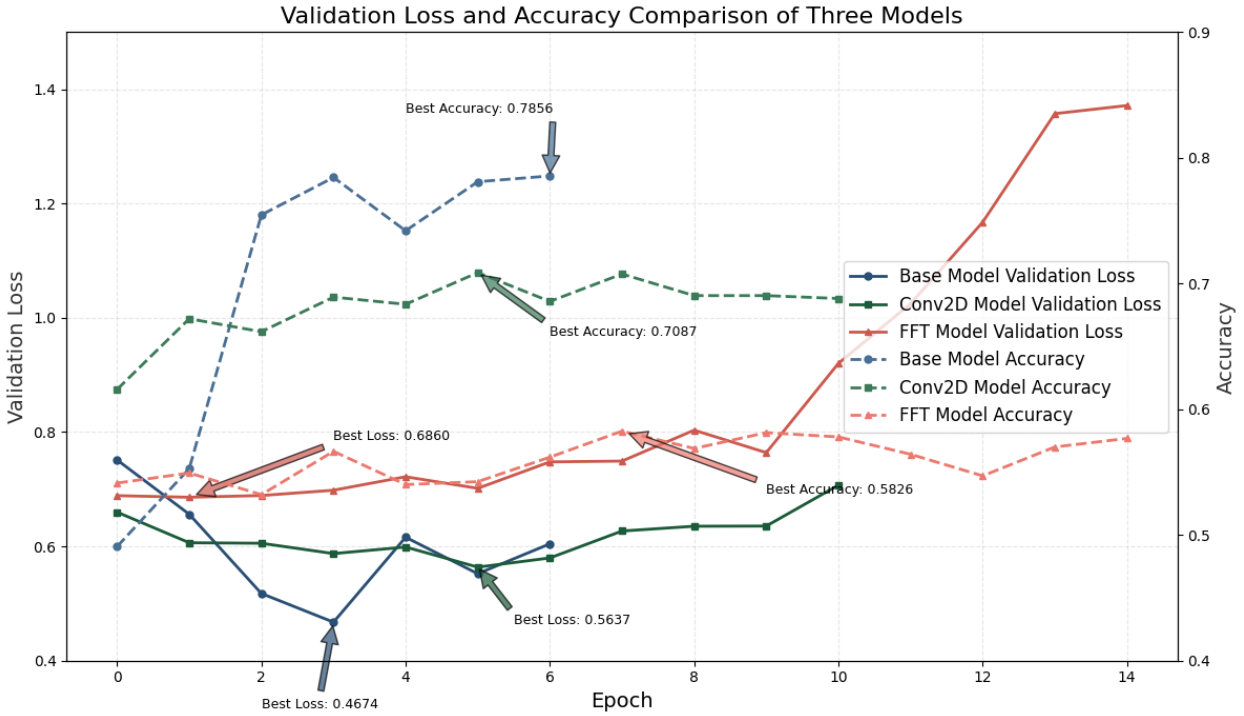
Training Performance Comparison

Overall Accuracy & Loss (Figure1) :

(1) Base model: Fast learning in early stages with rapidly boosted accuracy in first 3 epochs, then steps into overfitting, highly possible because of simple dataset. It indicates good learning capacity overall.
(2) Conv2D model: It shows a stable but slower learning process compared to base model according to its gradually increasing accuracy and decreasing validation loss. However, its best accuracy is almost 10 percent lower than base model.

(3) FFT model: Relatively stagnant learning process with lower accuracy and higher validation loss. This is possibly related to the whole-picture property of FFT embedding that will be mentioned below.

Figure1. Performance metrics for 3 models.



Feature Dynamics During Training

This section visualized the representations of CLS hidden states in a 2D plane for each model, at various layers and training epochs, as feature dynamics in [Hsu et al., 2024].
In the scatter plots below, each point represents an input token's representation (we color-code two different classes of inputs as orange vs blue for clarity), plotted by the first two principal components (PC1 and PC2). This gives a feature dynamics map of how separable or structured the representations are as training progresses. Additionally, feature dynamics in this section reflects semantic structure in learning processes of different models, and examination of pure structure patterns learned in these processes can be achieved by using other cluster techniques, providing insights and evidence of model structure learning and distinguishing ability, which further research can implement.

(1) Base Model Feature Evolution (Figure 2)

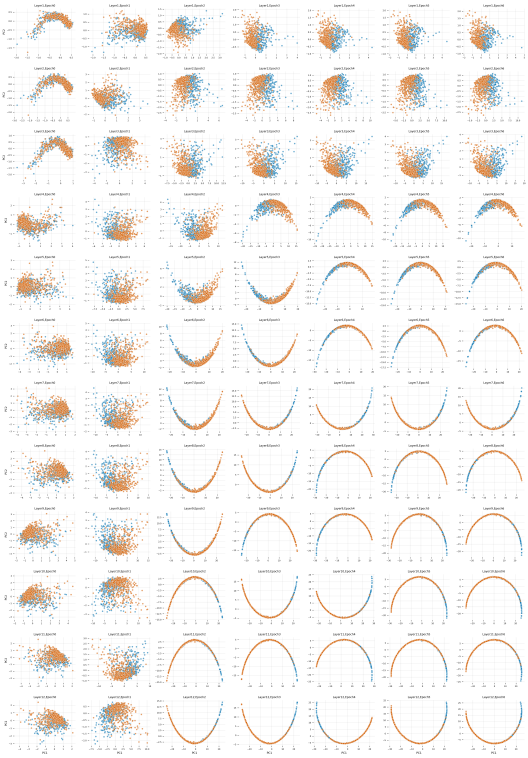


Figure 2: Base feature dynamics.

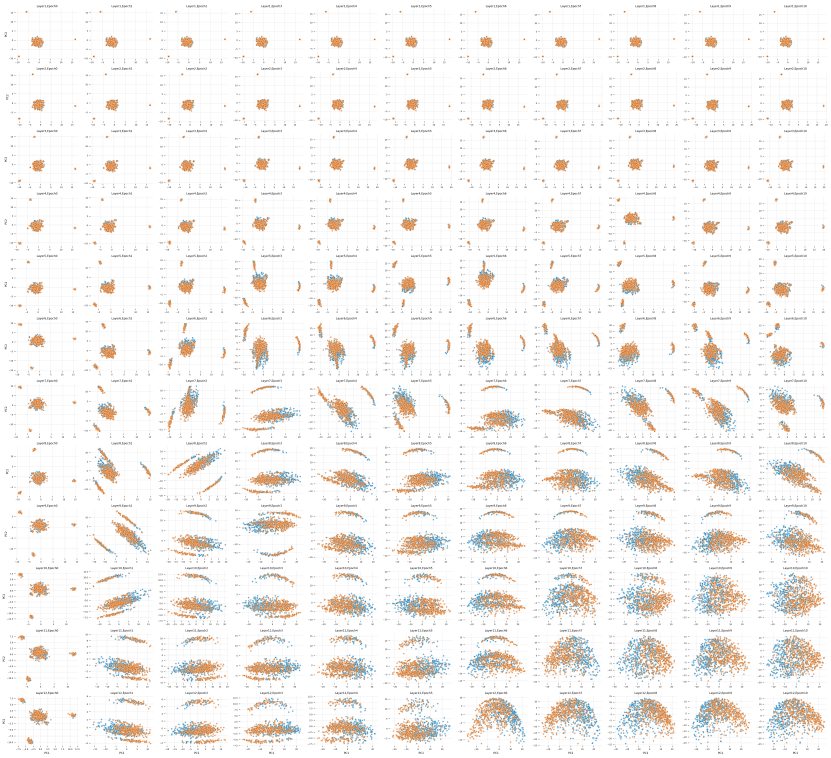


Figure 3: Conv2D feature dynamics

Feature dynamic process of base model follows a normal and classic pattern: initialization, fast fitting with separation, overfitting to collapse into semi-circle manifold. To be more specific, base model's feature dynamics show rapid early development (Epoch 0-1), quickly forming separable clusters in upper layers corresponding to peak performance (Epoch 3). However, continued training leads to representation over-compression or collapse (Epoch 4-6). This results in reduced class separation in later layers, correlating with its subsequent performance decline.

(2) Conv2D Model Feature Evolution (Figure 3)
Conv2D model feature pattern demonstrates an inverse process, where data points form 4 clusters first, then they gradually blend with each other as their layer and epoch goes deeper. The previous 4 cluster are composed of one main cluster and three tiny clusters with no semantic separation observed. Instead, in deeper layers and later epochs (From Layer 10 and Epoch 6), clusters are fully mixed and semantically separate data scatters evenly distributed are shown, indicating totally different learning pattern from base model.

(3) FFT Model Feature Evolution (Figure 4)
The FFT model's representation learning was delayed and initially ineffective, but given sufficient training, it eventually organized features to separate classes, though it didn't shown better performance in SST2 compared to other two models. But interestingly, it shows intermediate collapse, as semi-circle pattern shows up From Layer 7, Epoch 3 to Layer 12, Epoch 6, which contradicts with later collapse pattern in other standard Bert model, such as base model. The potential reason for intermediate collapse still remains unknown, waiting for further exploration.

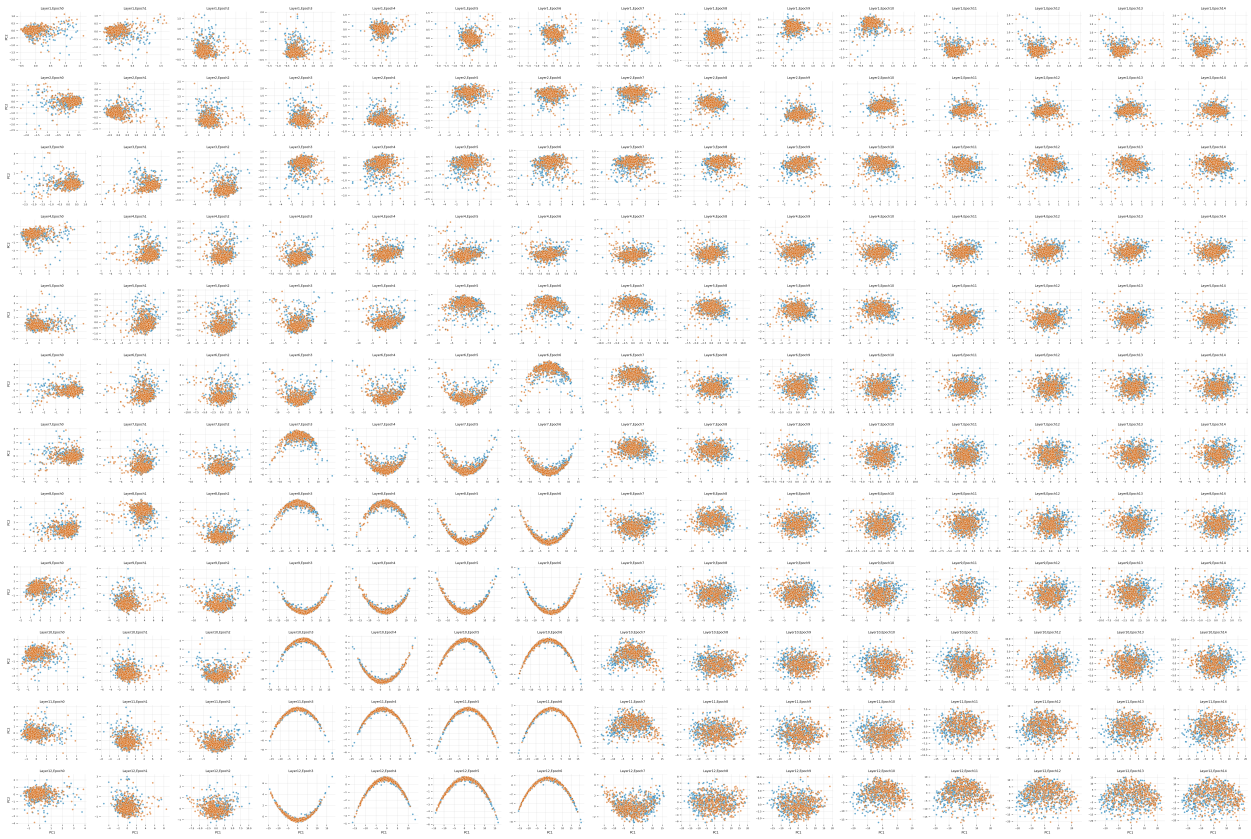


Figure 4. Feature dynamics in the FFT model

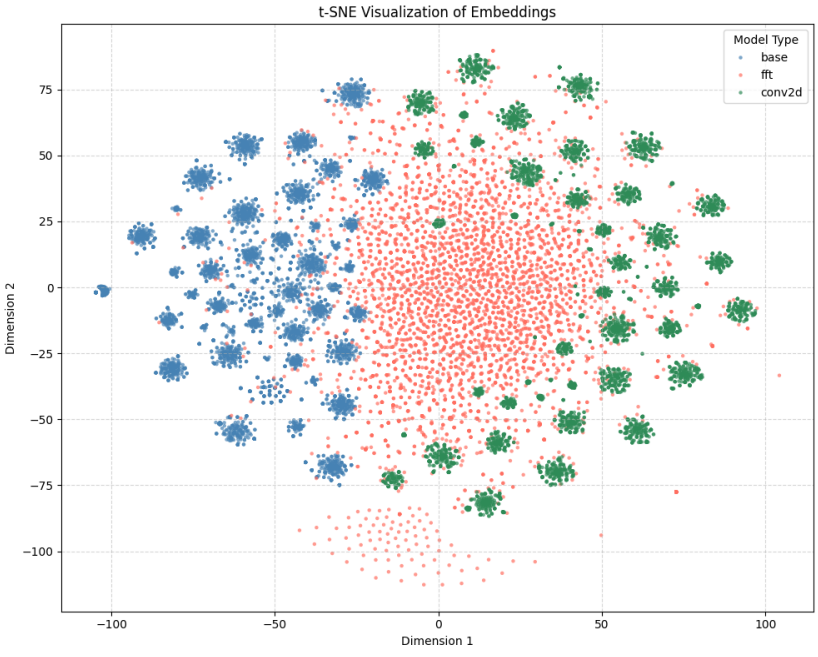


Figure 5: t-SNE Embeddings Visualization

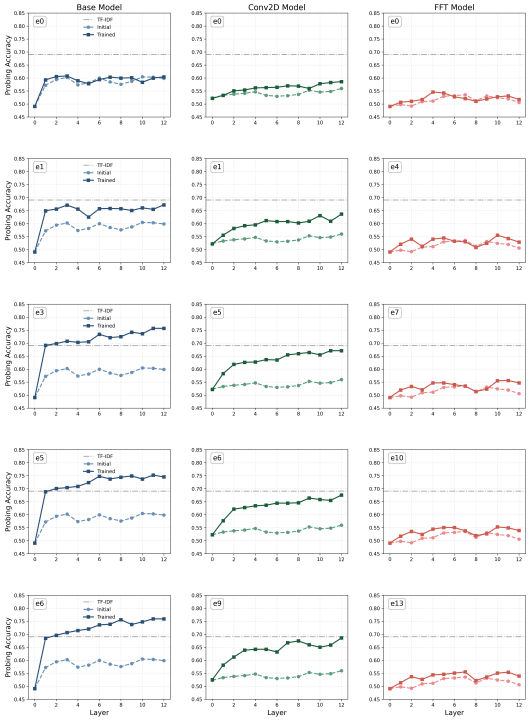


Figure 6: Probing Results of five stages

Embedding Distribution Properties

t-SNE Distribution Pattern

Figure 5 demonstrates three largely different embedding pattern caused by three embedding mechanism in two dimensional space.(blue for Base, green for Conv2D and red for FFT) Original base model exhibits multiple embedding clusters, suggesting the simple element-wise additive embedding forms locally focused initial input, which is more compatible with Transformer architecture design. On the contrary, while Conv2D model initial embedding shares the same cluster pattern, these clusters apparently gather in a semicircle shape on the right side of the plane, which resonates the rotation found in position embedding in previous studies. [Wennberg et al., 2024] Finally, considering FFT model's initial embeddings, they are almost evenly sprawled among 2D plane in a single dot form, indicating little local information exists after word and position input being compressed by FFT mechanism.

Cosine Similarity of Tokens

Figure 7 compares cosine similarities of the same token in three different models. FFT model heatmap is the coldest among them, with no similarity between different tokens, except small amount of points arranged in a regular horizontal and vertical pattern, which is speculated to be positional information.[Godey et al., 2024] In contrast, Conv2D model heatmap with the lightest color is the warmest overall, suggesting all token embedding vectors are the most anisotropic. Base model heatmap is a combination of these two models above, showing a balance between word information and position information.

Probing Internal Representations

Figure 6 illustrates probing results of five representative training stage: Unstructured, Entering structure, Maximum separation, Morphological changes and Late homogenization. The definitive epochs of each model corresponding to these stages are decided by both their unique training metrics change and feature dynamics pattern.

(1) Base model: Probes show rapid early task encoding in upper layers, but probe accuracy decreases over time in lower/middle layers, concentrating predictive power only in the final layers.

(2) Conv2D model: Conv2D probes reveal a stable, uniform learning progression where probe accuracy gradually increases up the layers throughout training, indicating hierarchical feature learning without loss of signal.

(3) FFT model: The FFT model's probing, while slightly exceeding random initialization results, remains overall unsatisfactory. Building on earlier embedding visualization analysis, this performance is likely attributed to the FFT embedding mechanism introducing frequency mismatch between word and position information.

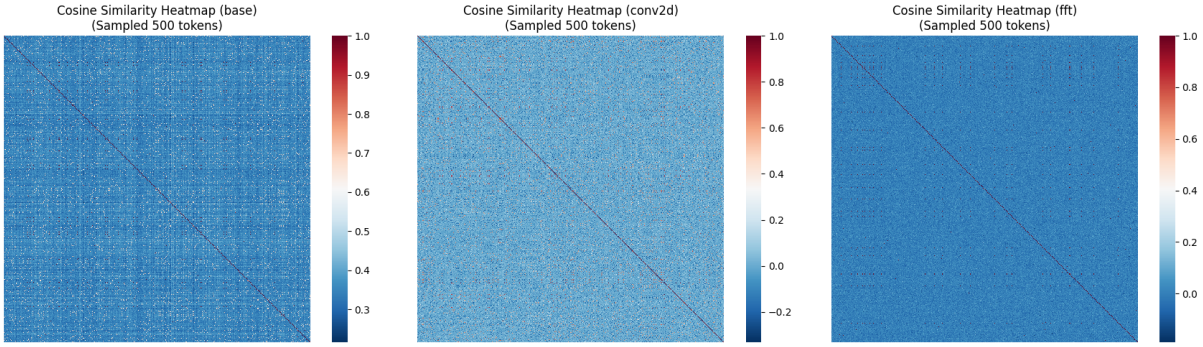


Figure 7.Cosine Similarity Heatmaps of Three Models

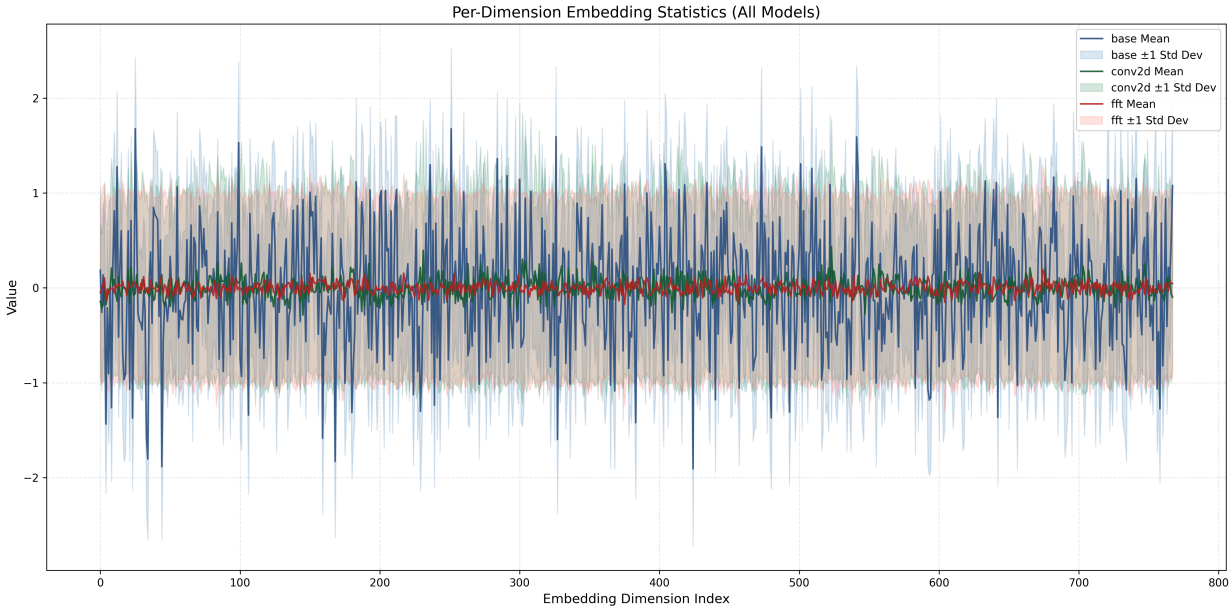


Figure 8.Value of embeddings among 768 dimensions

Key Conclusions and Future Plans

Key Conclusions

- Simple element-wise addition embedding mechanism in Base model provides meaningful embedding vectors, various in value (Figure 8) but relatively anisotropic. (Figure 5 & Figure 7)
- Convolutional embedding mechanism with convolution kernel creates a more more expressive representation space. We hypothesize that the Conv2D kernel functions as a learned optimal filter for combining word and position information. Initially random, this filter learns over training to preferentially pass through more meaningful local token-position combination patterns, potentially contributing to preventing updates in unstable directions. This hypothesis is supported by the Conv2D probing results, which demonstrate a stable, incremental increase in task-relevant accuracy from lower to upper layers.
- We can learn from FFT embedding results that they lack directional specificity and their components tend towards numerical homogeneity. We hypothesize that this is because

spectral element-wise multiplication in FFT embedding may results in mutual cancellation of magnitudes. However, it may in turn prevents FFT model from collapsing in the end, as shown in the feature dynamics.

Future Plans

- Theoretic foundations and explanation to help understand inner mechanisms better.
- Experiments testing the hypothesis of FFT mutual cancellation of magnitudes.
- More ablation experiments and attention analysis.
- Better modification of FFT embedding mechanism to solve potential frequency mismatch problems (mutual cancellation of magnitudes) to avoid drawbacks while amplify advantages.
- Generalization on larger datasets to test the validity of preliminary results.

Reference

1. Godey, N., de la Clergerie, É., & Sagot, B. (2024). *Anisotropy is inherent to self-attention in transformers.* arXiv preprint arXiv:2401.12143.

2. Hsu, A. R., Cherapanamjeri, Y., Park, B., Naumann, T., Odisho, A., & Yu, B. (2024). *Diagnosing Transformers: Illuminating Feature Spaces for Clinical Decision-Making*. In International Conference on Learning Representations (ICLR).

3. Plate, T. A. (1994). *Distributed representations and nested compositional structure*. University of Toronto, Department of Computer Science.