Evaluating Transformer-Based Models for Time Series Forecasting: A Comprehensive Study on Model Architecture

Deep Learning Final Project

National Yang Ming Chiao Tung University

Ya-Chun Chang*, Ming-Chih Lo*, Yong-En Tian*

Introduction

- Time series forecasting is ubiquitous but challenging in the real world.
- **Transformer**, which has achieved tremendous success in NLP and CV, is widely adopted and exhaustedly modified for time series forecasting.
- However, recent works revealed that simple **linear forecasters** could achieve comparable performance with Transformers, which questions the direction of architecture-oriented research on Transformer-based forecasters.

Problem Formulation

iTransformer

- The whole lookback series of individual variate is regarded as the variate token.
- LayNorm and FFN work in coordination for variate-centric representations.
- Multivariate correlations can explicitly captured by the self-attention mechanism.

Challenges and Our proposal

- The design of the FFN for learning temporal representations is overly simplistic.
- The architecture for handling the noise component in the series data shouldn't be overly complex.

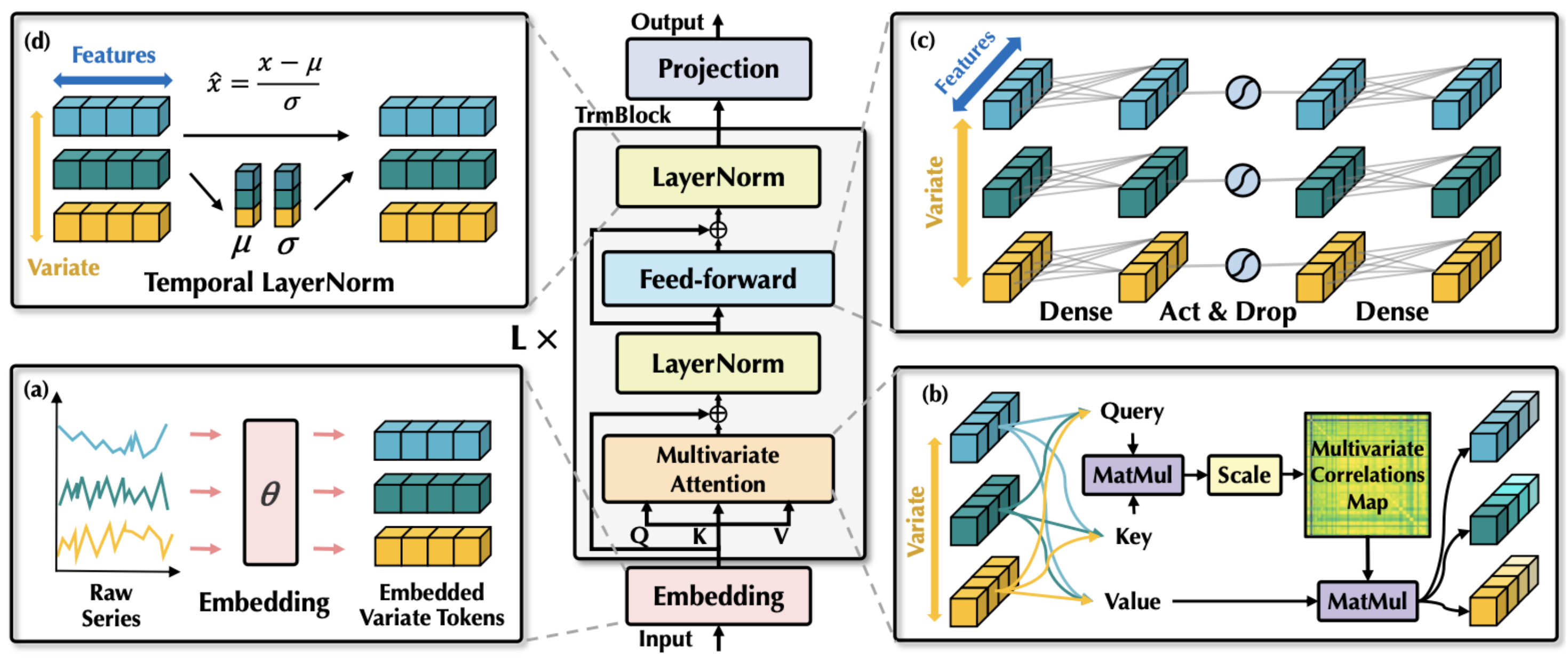

Figure 1: Overall structure of iTransformer

Methodology

iTransformer-FFN2.0

Expand the double-layer conv layer to a triple-layer conv layer.

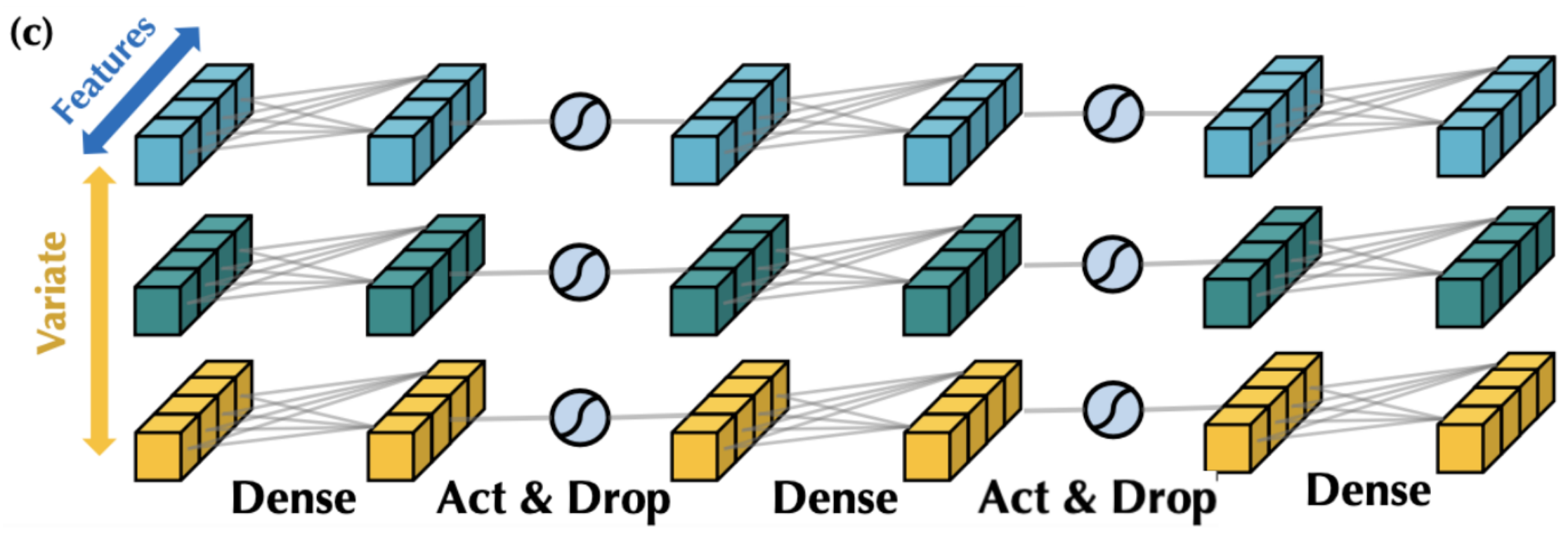

Figure 2: The overall concept of of iTransformer-FFN2.0.

Series Decomposition

Feed each series component into the appropriate model block.

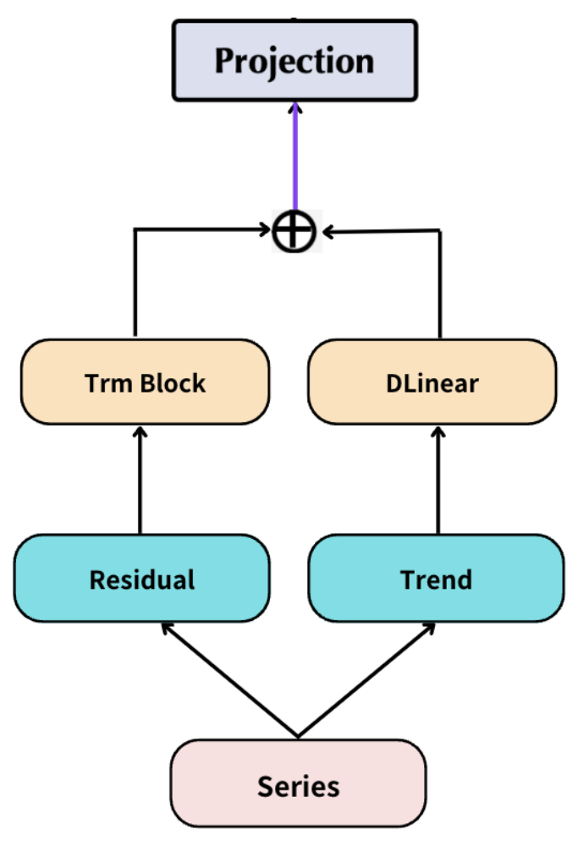

Figure 3: The overall concept of Series Decomposition.

iPatching

Replace conv layers with the encoder of PatchTST.

Dual Attention iTransformer (DAT)

Utilize both variable-wise and time-wise attention mechanisms to capture relationships across different variables and time steps.

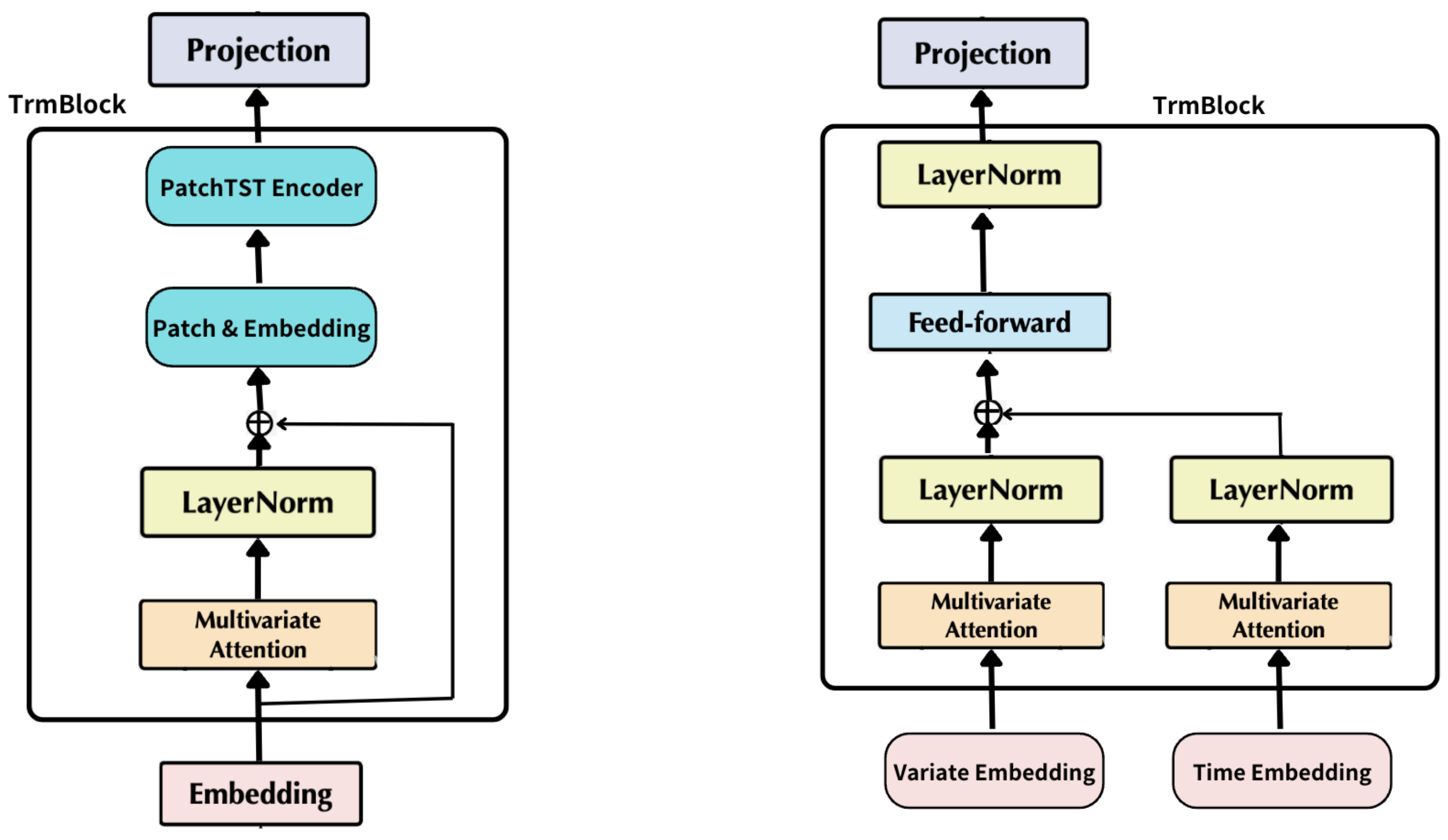

Figure 4: The overall concept of our approach. Left: iPatching. Right: DAT.

Experiment Results

Dataset	iTransformer/MSE	iPatching/MSE	Series Decomposition/MSE	iTransformer FFN2.0/MSE	iTransformer NLinear/MSE	DAT/MSE
Exchange	0.364	0.359	0.368	0.381	0.355	0.360
ETTm1	0.407	0.398	0.407	0.401	0.412	0.403
ECL	0.178	0.215	0.185	0.176	0.263	0.178
Traffic	0.428	0.511	0.433	0.432	0.603	0.426

Table 1: Comparison of MSE values for different models across various datasets.

Dataset	iTransformer/MAE	iPatching/MAE	Series Decomposition/MAE	iTransformer FFN2.0/MAE	iTransformer NLinear/MAE	DAT/MAE
Exchange	0.407	0.403	0.411	0.412	0.403	0.402
ETTm1	0.411	0.405	0.407	0.408	0.415	0.409
ECL	0.271	0.292	0.273	0.270	0.354	0.269
Traffic	0.282	0.294	0.290	0.291	0.405	0.288

Table 2: Comparison of MAE values for different models across various datasets.

- The modification of the temporal component (iPatching and DAT) in the model doesn't seem to provide significant assistance, as only the less variable datasets (Exchange and ETT) show a slight improvement.
- Series decomposition appears to be ineffective, yielding results similar to those of iTransformer, likely due to the inconsistent patterns in the data.
- Based on the table above, for datasets where the original iTransformer performed poorly, such as ETTm1, iTransformer-FFN2.0 showed a slight improvement. For other datasets where iTransformer already performed well, such as Traffic and ECL, iTransformer-FFN2.0 did not show a particularly noticeable difference.
- However, according to the following figure, we found that even in datasets like ECL, where iTransformer-FFN2.0 did not stand out in terms of MSE and MAE, it fits the trend fluctuations better.

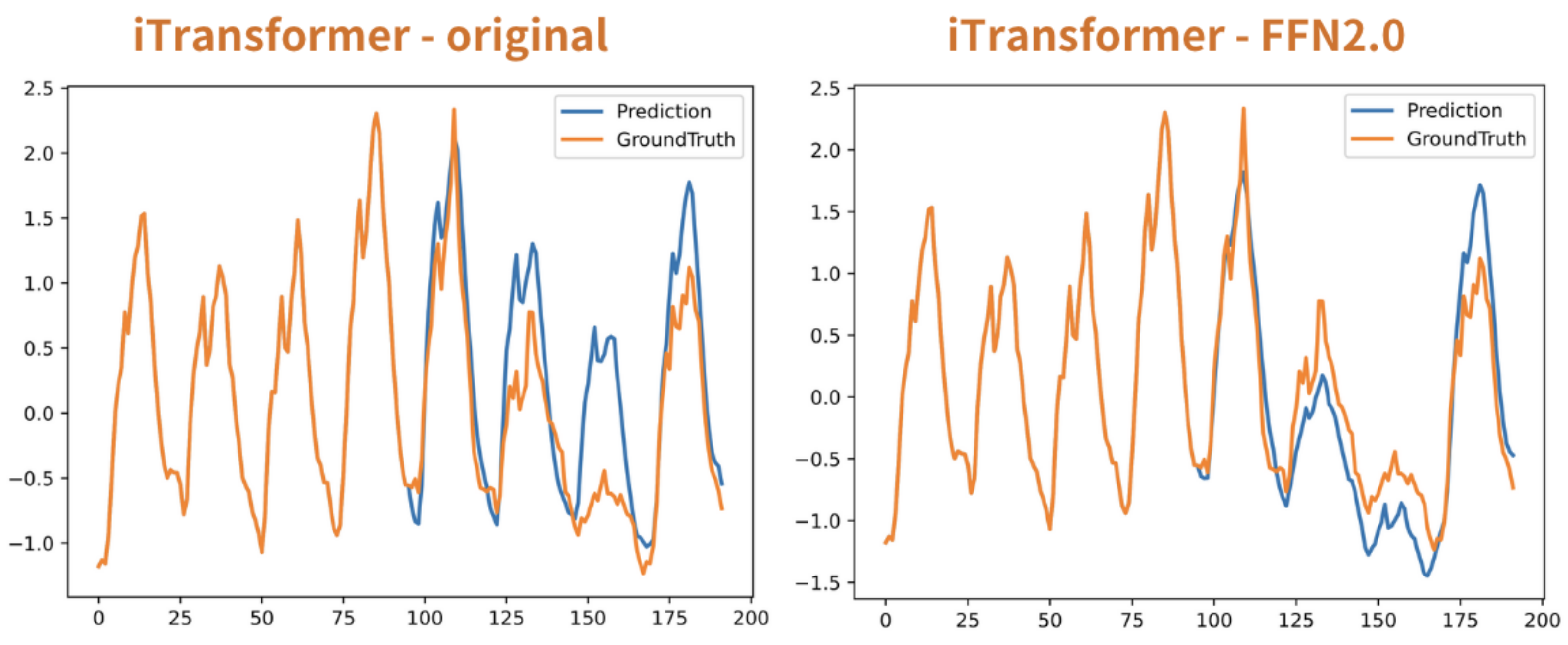

Figure 5: Visualization of input-96-predict-96 results on the ECL dataset.

Conclusion

- All the models appear to have similar performance. We observed that these minor differences might be resolved by simply adjusting the parameters.
- FFN should not be too complex when it is used together with attention mechanism. Original FFN is simple yet effective and sufficient. The FFN typically uses simple linear or convolutional layers in order to introduce non-linearity. Overcomplicating it can cause a loss of focus and diminish the transformer's original, most effective design.
- In this project, we experimented with various SOTA transformer-based models to explore their potential and conducted a comprehensive comparison of these different models.