A Time Series is Worth 64 Words: Long-term Forecasting with Transformers

Angela Cui, Vipin Gunda, James Kim, Derek Liu, Oliver Lopez

Cornell University - CS4872 Deep Learning

Cornell University

Introduction / Background / Motivation

- **Problem:** Traditional Transformer-based models struggle with long-term time series forecasting.
  - O(N²) complexity for sequence length N and limited ability to capture local semantic information and incorporate longer look-back windows
- **Original Contributions:**
  1. **Patching:** Divides time series into subseries-level patches to capture local semantic information while reducing sequence length and computation.
  2. **Channel Independence:** Each univariate time series is processed separately with shared weights
- **Goal:** Enhance long-term time series forecasting
  - Improve patching through 1D convolutions
  - Improve self-supervised model by randomizing the mask ratio instead of using set patches.

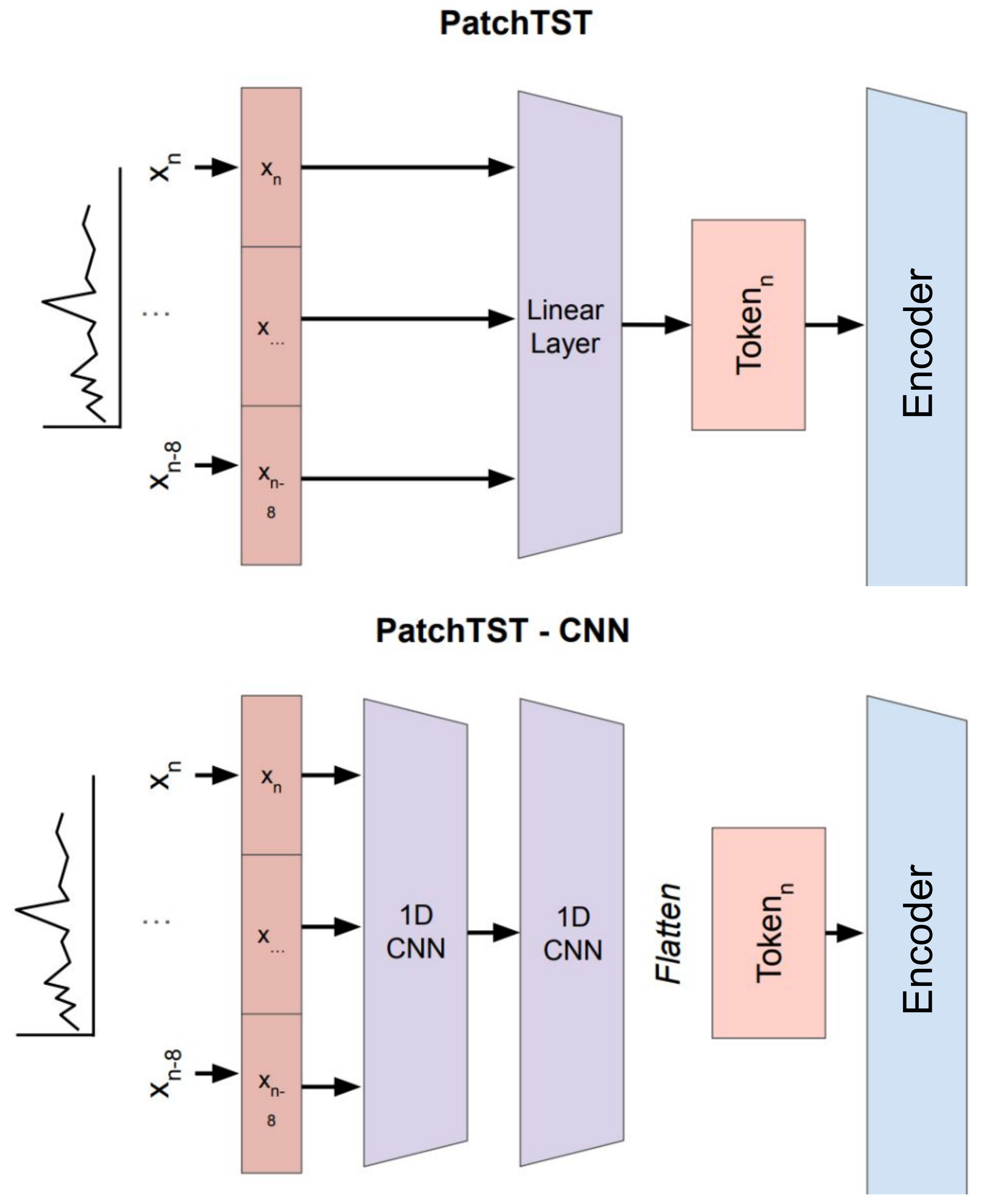
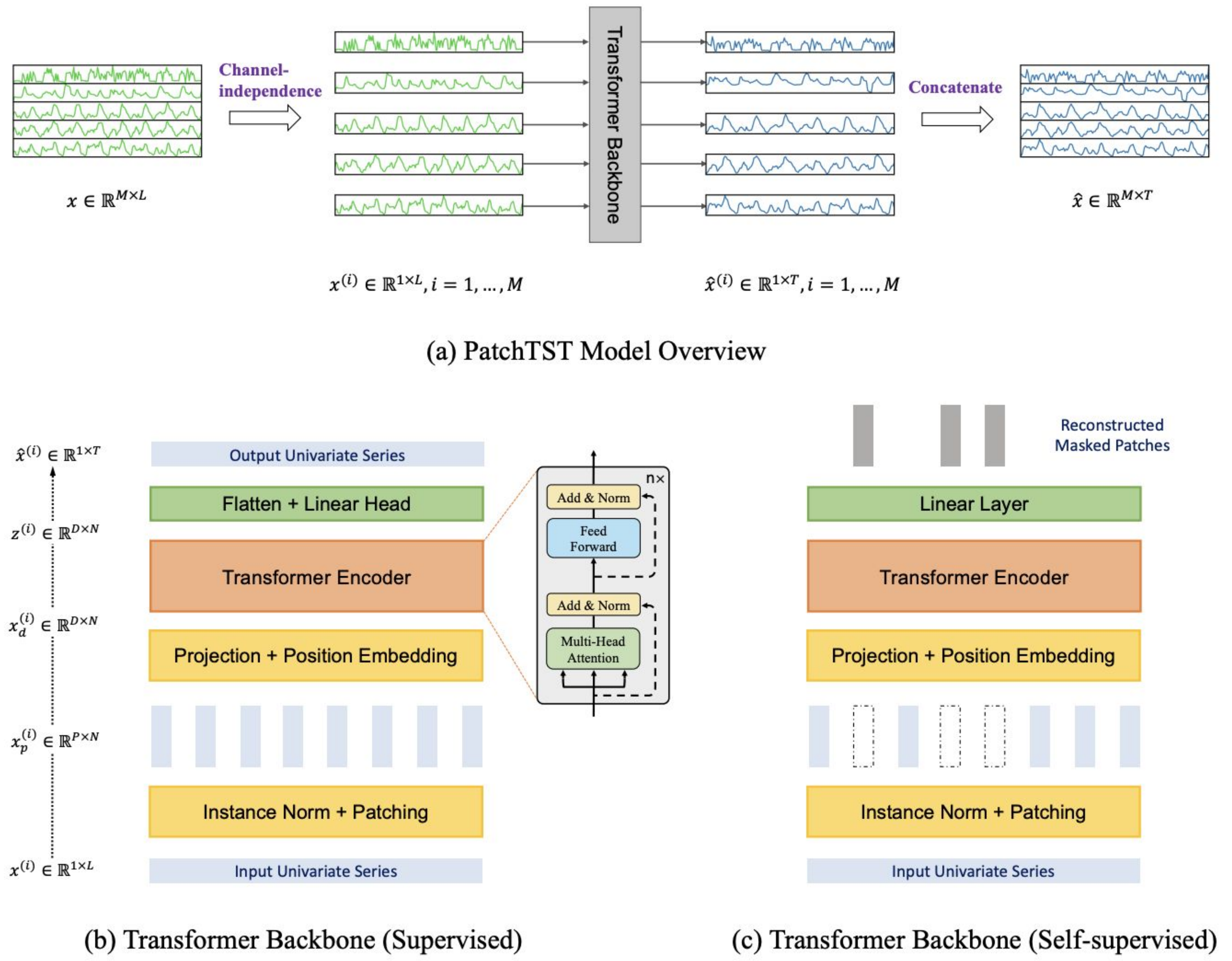


Figure 2: Top is original patch embedding, Bottom uses 1D CNN layers to extract features from a given patch

Methodology


Channel-independence
Transformer Backbone
Concatenate
$x \in \mathbb{R}^{M \times L}$
$x^{(i)} \in \mathbb{R}^{1 \times L}, i = 1, \dots, M$
$\hat{x}^{(i)} \in \mathbb{R}^{1 \times T}, i = 1, \dots, M$
$\hat{x} \in \mathbb{R}^{M \times T}$
$\hat{x}^{(i)} \in \mathbb{R}^{1 \times T}$
$z^{(i)} \in \mathbb{R}^{D \times N}$
$x_d^{(i)} \in \mathbb{R}^{D \times N}$
$x_p^{(i)} \in \mathbb{R}^{P \times N}$
$x^{(i)} \in \mathbb{R}^{1 \times L}$
Output Univariate Series
Flatten + Linear Head
Transformer Encoder
Projection + Position Embedding
Instance Norm + Patching
Input Univariate Series
n×
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Reconstructed Masked Patches
Linear Layer
Transformer Encoder
Projection + Position Embedding
Instance Norm + Patching
Input Univariate Series

(a) PatchTST Model Overview

(b) Transformer Backbone (Supervised)

(c) Transformer Backbone (Self-supervised)

Figure 1: PatchTST architecture. (a) Multivariate time series data is divided into different channels. They share the same Transformer backbone, but the forward processes are independent. (b) Each channel univariate series is passed through instance normalization operator and segmented into patches. These patches are used as Transformer input tokens, and the output channels are concatenated together. (c) Masked self-supervised representation learning with PatchTST where patches are randomly selected and set to zero. The model will reconstruct the masked patches.

Conclusions

- Randomized mask ratios do not yield significant improvements in long-term time series forecasting performance. In some cases, randomization slightly worsens results, though the differences are likely negligible.
- Augmenting the PatchTST architecture with 1D convolutional layers (as shown in Figure 2) does not provide meaningful performance gains, and is still prone to overfitting.
- These results suggest that the original patch-based linear embedding is already sufficiently expressive for the task, and further architectural complexity yields diminishing returns.

Results

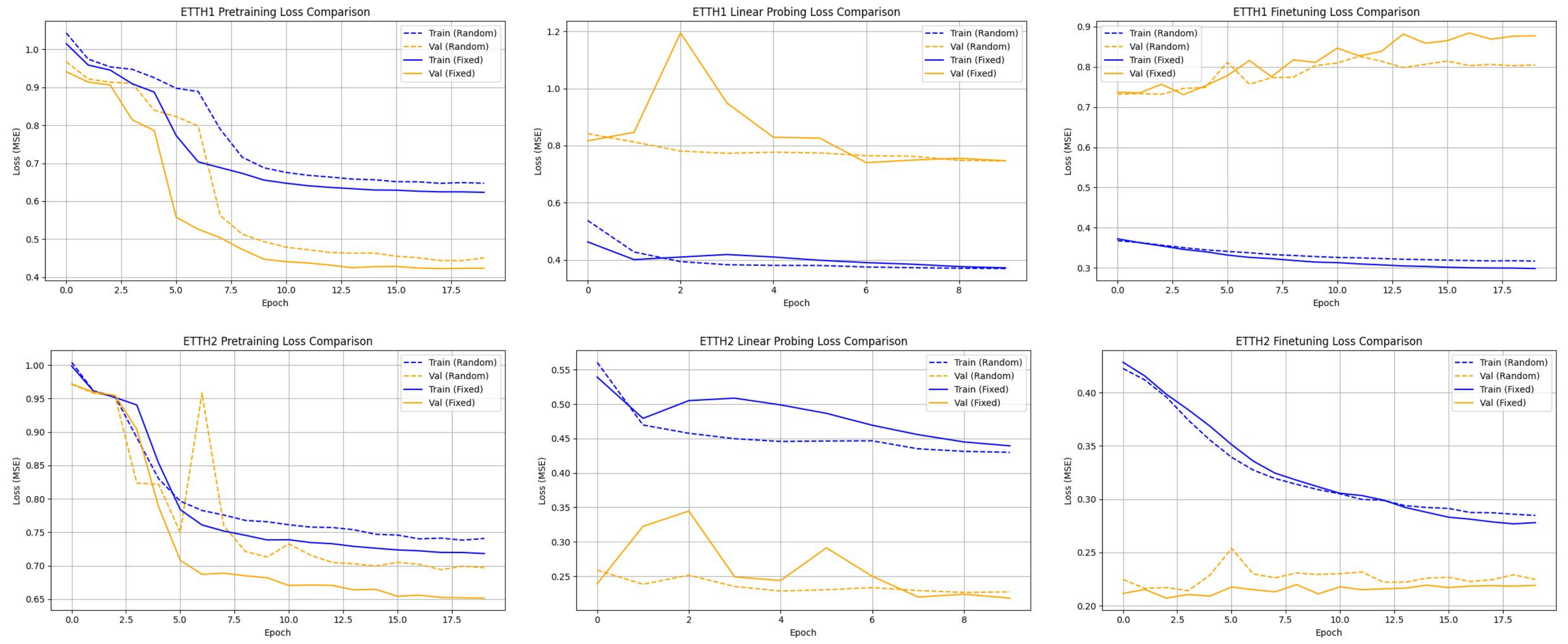


Figure 3: Results of pretraining, linear probing, and fine tuning across randomized vs. fixed mask ratios on the ETTH1 and ETTH2 datasets. Losses for both methods were very similar.

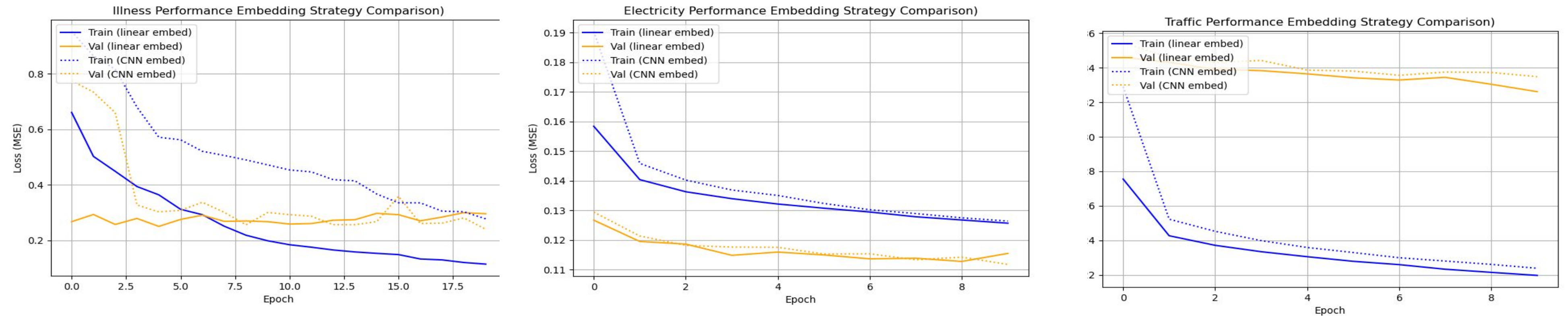


Figure 4: Results of supervised training with linear and convolutional embeddings. The latter shows marginal improvement, although both methods are prone to overfitting. This may be caused by the introduction of more more learnable parameters to a small experiment.

Future Work

- Diverse patch embedding mechanisms (mean, max, sum, dropout, EMA, etc.) with diverse time series datasets and full transformer architecture
- Different patch ratio “schedules” to see if the model can learn better representations by better controlling/varying how many patches are masked

References

[1] G. L. Asher, “Exploring Tokenization Techniques to Optimize Patch-Based Time-Series Transformers,” Computer Science Senior Theses, no. 47, Dartmouth College, 2024. [Online]. Available: https://digitalcommons.dartmouth.edu/cs_senior_theses/47

[2] Y. Nie, N. H. Nguyen, P. Sinthong, and J. Kalagnanam, “A Time Series is Worth 64 Words: Long-term Forecasting with Transformers,” in Proc. Int. Conf. Learn. Representations (ICLR), 2023.