

Deep Learning for Time Series

Session 3: Attention-based architectures

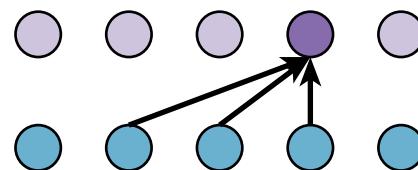
Romain Tavenard

Intro

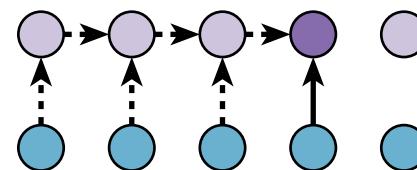
Self-attention vs RNNs vs ConvNets

Intro

Causal Conv.



RNN



Self-Attention

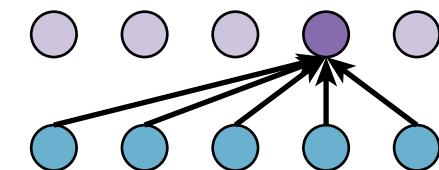
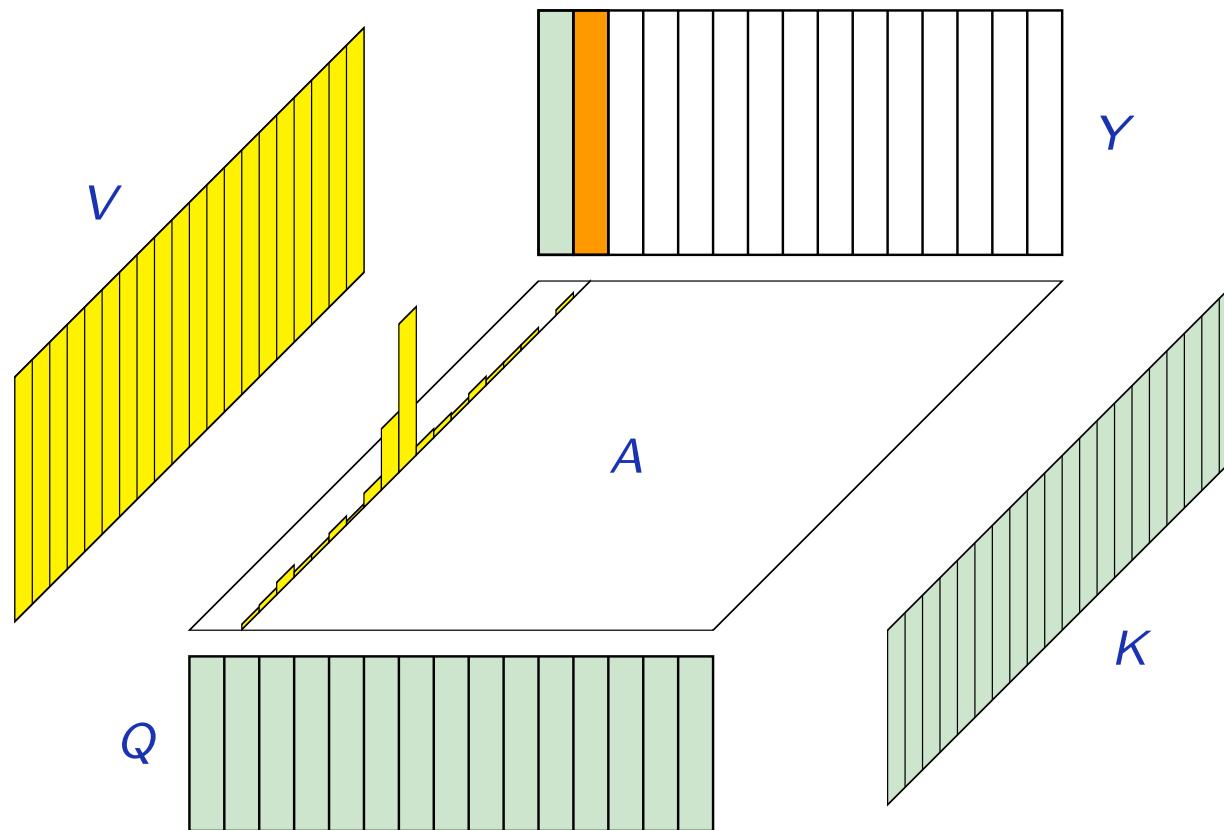


Illustration of self-attention

Intro

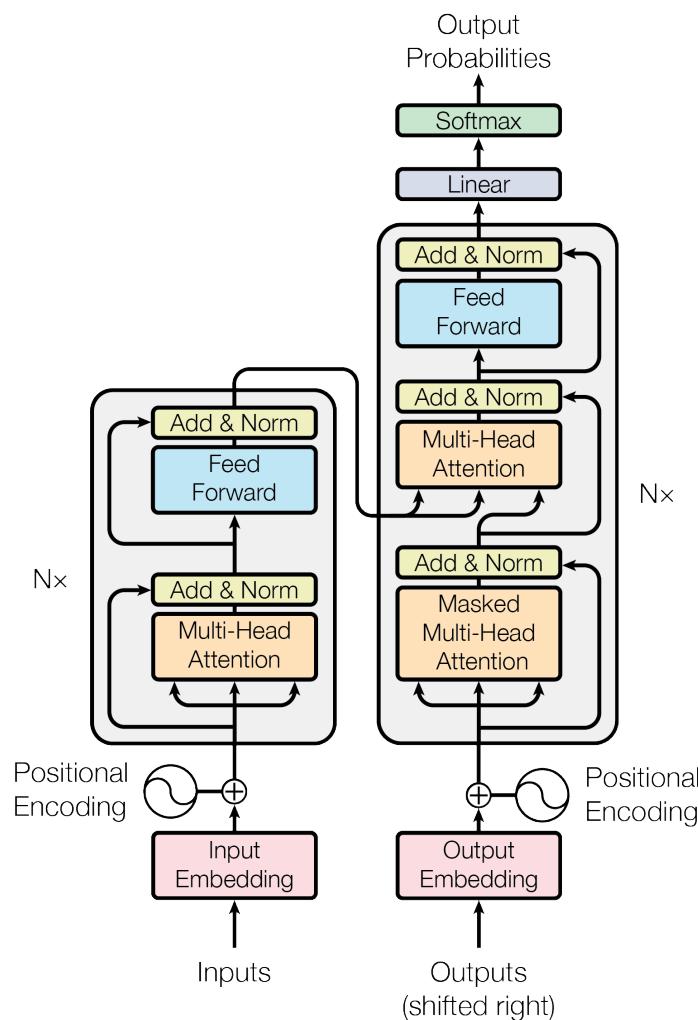
$$A_i = \text{softmax} \left(\frac{QK^\top}{\sqrt{d_k}} \right) V$$
$$Y_i = V^\top A_i$$



Source: Deep Learning Course, François Fleuret

Basic transformer architecture

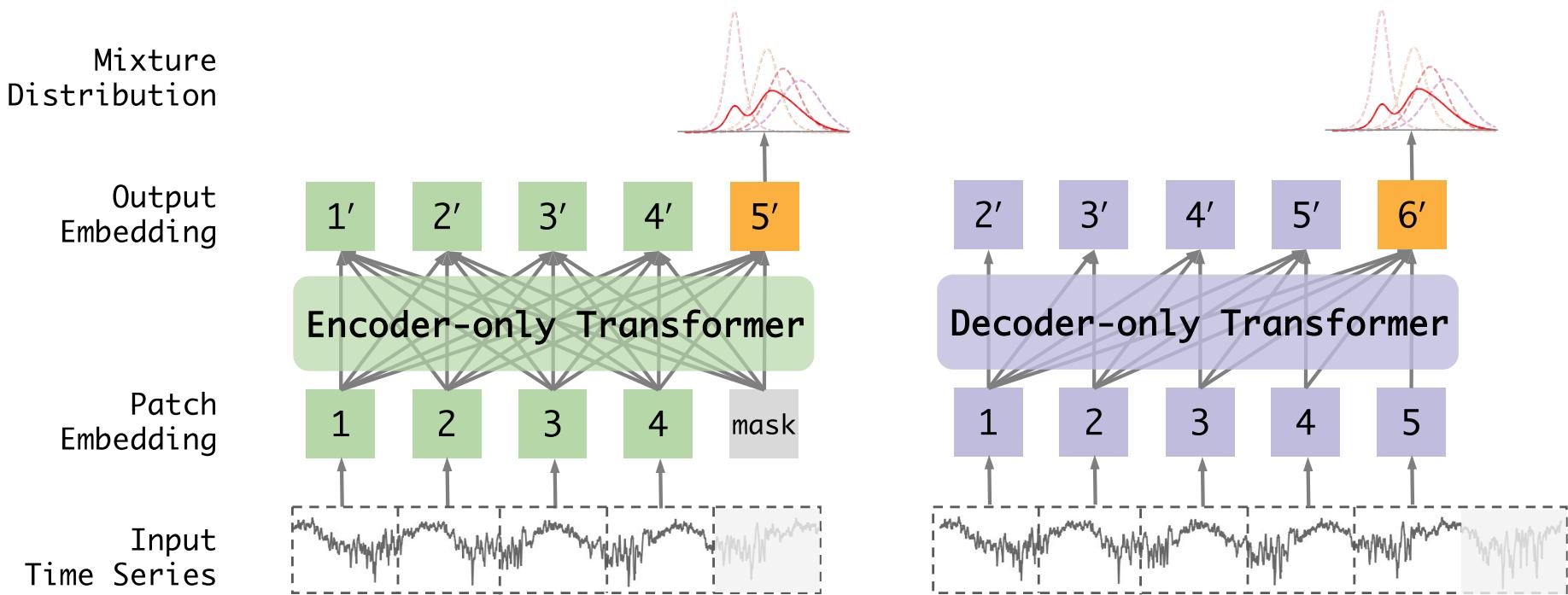
Intro



Source: “Attention is all you need” (NIPS’17)

Encoder-only vs Decoder-only for forecasting

Intro



Source: “Towards Neural Scaling Laws for Time Series Foundation Models” (ICLR’25)

- Encoder-only (bi-directional attention) is more versatile (can be applied to different tasks)
- Decoder-only is more suitable for forecasting

Failures of transformers in forecasting

Intro

Are Transformers Effective for Time Series Forecasting?

Ailing Zeng^{1*}, Muxi Chen^{1*}, Lei Zhang², Qiang Xu¹

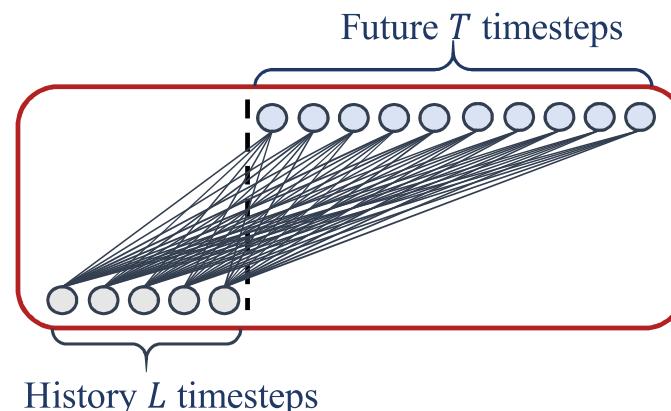
¹The Chinese University of Hong Kong

²International Digital Economy Academy (IDEA)

{alzeng, mxchen21, qxu}@cse.cuhk.edu.hk

{leizhang}@idea.edu.cn

- Comparison between simple linear predictor and SOTA transformers (at the time)



Source: “Are Transformers Effective for Time Series Forecasting?” (AAAI’23)

Failures of transformers in forecasting

Intro

Are Transformers Effective for Time Series Forecasting?

Ailing Zeng^{1*}, Muxi Chen^{1*}, Lei Zhang², Qiang Xu¹

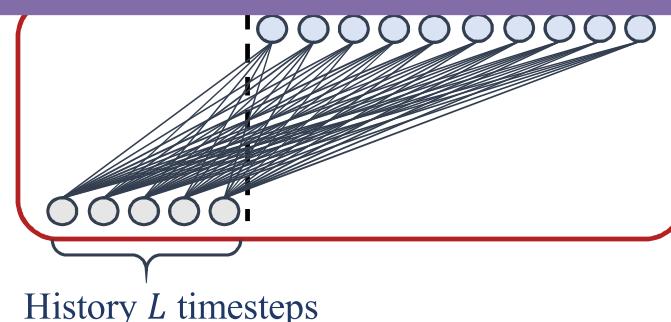
¹The Chinese University of Hong Kong

²International Digital Economy Academy (IDEA)

{alzeng, mxchen21, qxu}@cse.cuhk.edu.hk

- Compared to existing sophisticated Transformer-based LTSF models in all cases, and often by a large margin”

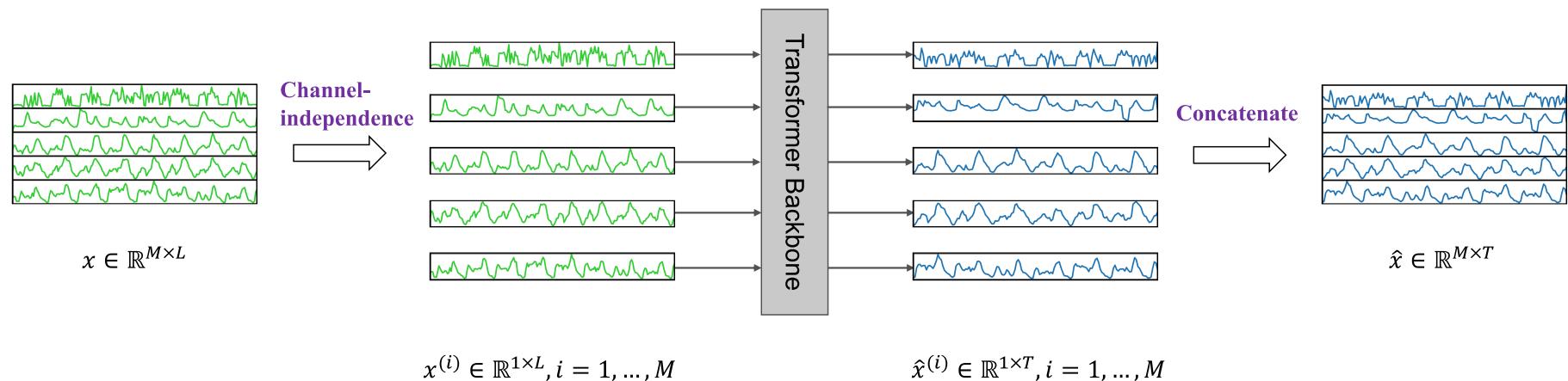
TA



Source: “Are Transformers Effective for Time Series Forecasting?” (AAAI’23)

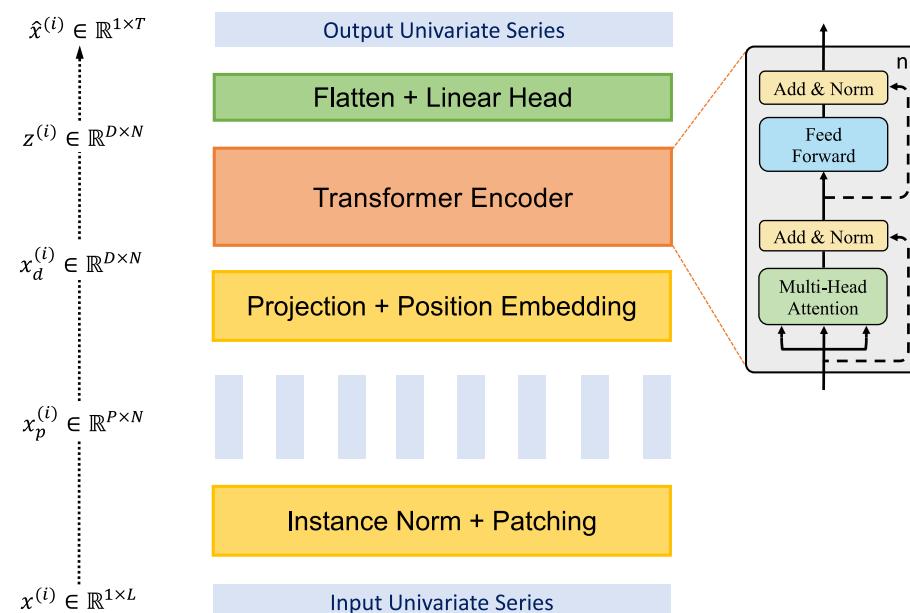
- Previous methods:
 1. each patch = 1 time point or some extracted information
 2. all variates are considered together
- PatchTST:
 1. patch = subseries
 2. channel-independence
(bonus: much easier for transfer learning)
 3. proper preprocessing (RevIN)

Channel independence



Source: “A Time Series is Worth 64 Words: Long-term Forecasting with Transformers” (ICLR’23)

Transformer backbone



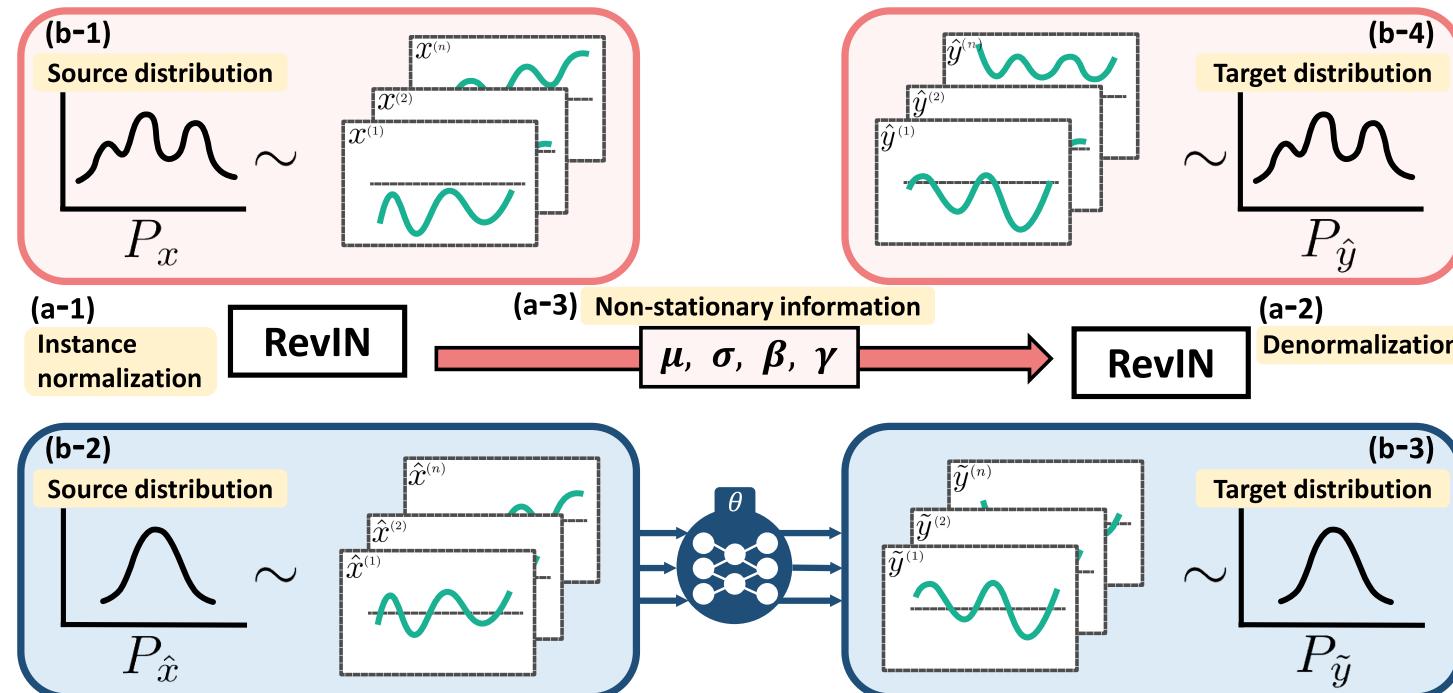
Source: “A Time Series is Worth 64 Words: Long-term Forecasting with Transformers” (ICLR’23)

PatchTST: a transformer that works well

Intro

Preprocessing (RevIN)

$$\hat{x}_{d,t}^{(i)} = \gamma_d \cdot \frac{x_{d,t}^{(i)} - \hat{\mu}_d^{(i)}}{\hat{\sigma}_d^{(i)}} + \beta_d \quad \text{and reverse before forecasting}$$

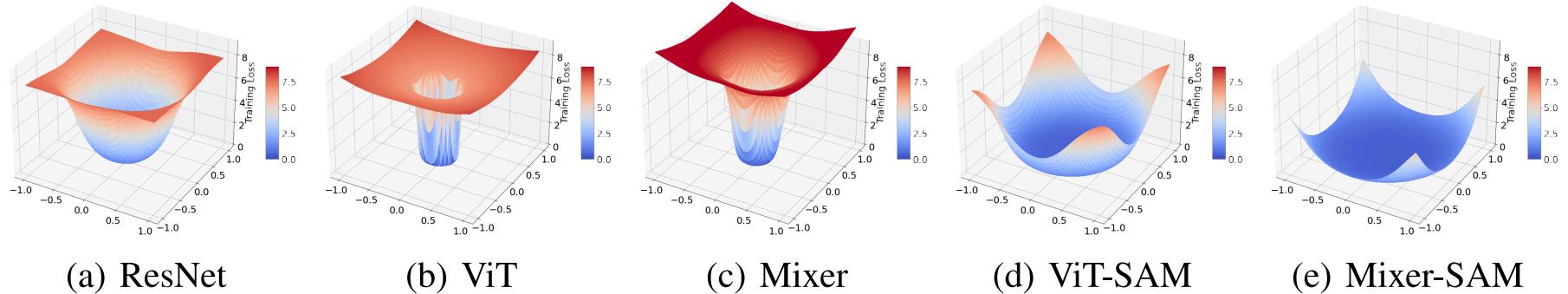


Source: “Reversible Instance Normalization for Accurate Time-Series Forecasting against Distribution Shift” (ICLR’22)

More insights on transformer failures

Intro

- Transformers have a sharp loss landscape:



Source: “When Vision Transformers Outperform ResNets without Pre-training or Strong Data Augmentations” (ICLR’22)

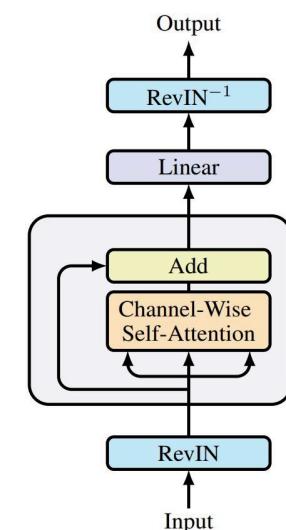
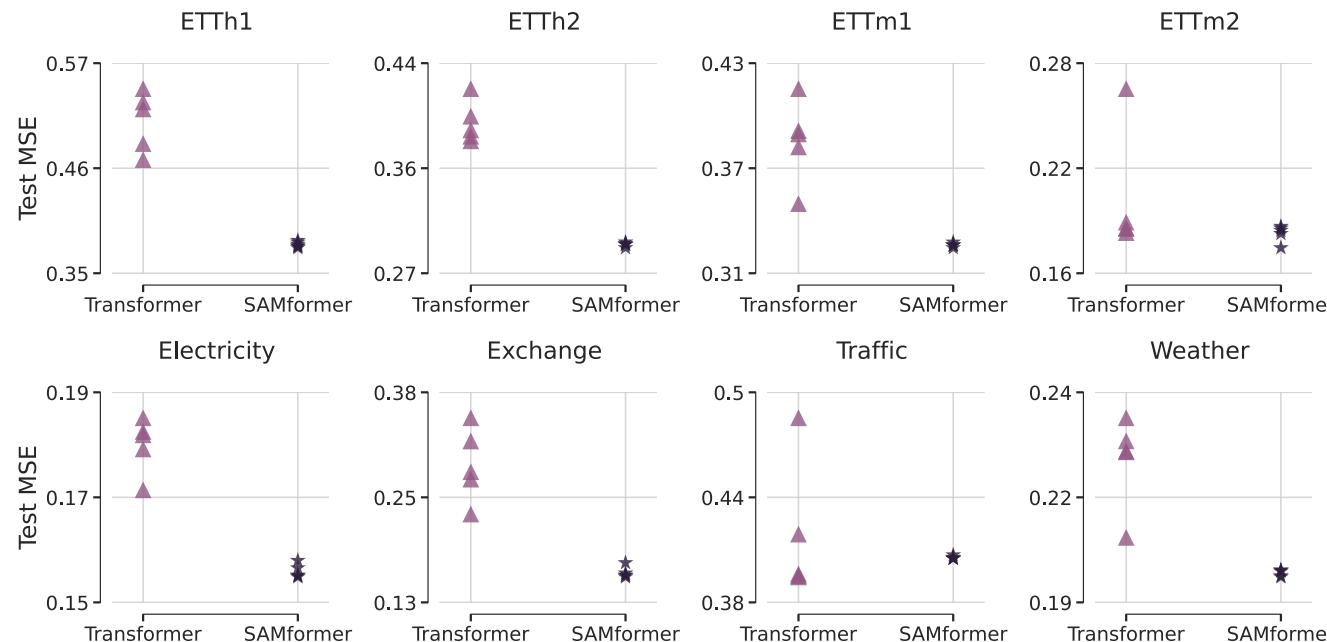
- Train with SAM (Sharpness-Aware Minimization) to smooth the landscape:

$$\mathcal{L}_{\text{SAM}} = \max_{\|\varepsilon\|_2 \leq \rho} \mathcal{L}(w + \varepsilon)$$

More insights on transformer failures

Intro

- SAMformer: a forecasting Transformer trained with SAM



Source: “SAMformer: [...]” (ICML’24)

- Transformers can work well for time series forecasting
- But need proper design choices:
 1. patch-based input
 2. channel-independence
 3. proper preprocessing (RevIN)
 4. smooth optimization (SAM)