

# Deep Learning for Time Series

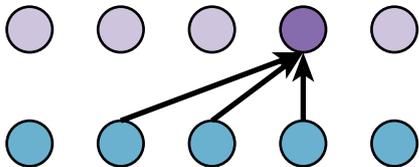
Session 3: Attention-based architectures

Romain Tavenard

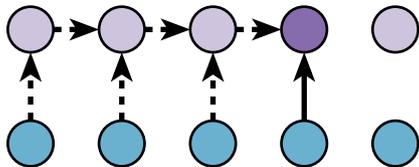
# Intro

---

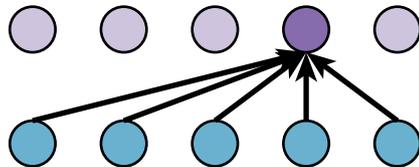
Causal Conv.



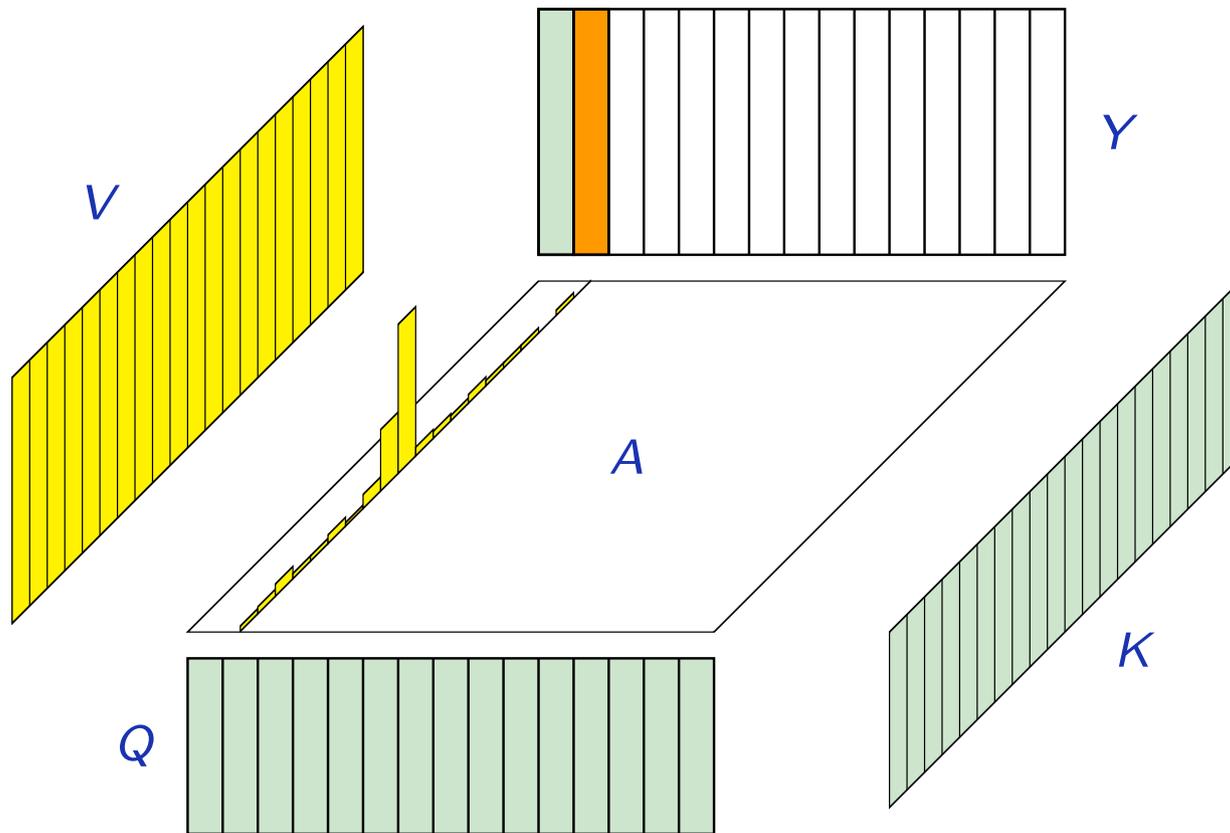
RNN



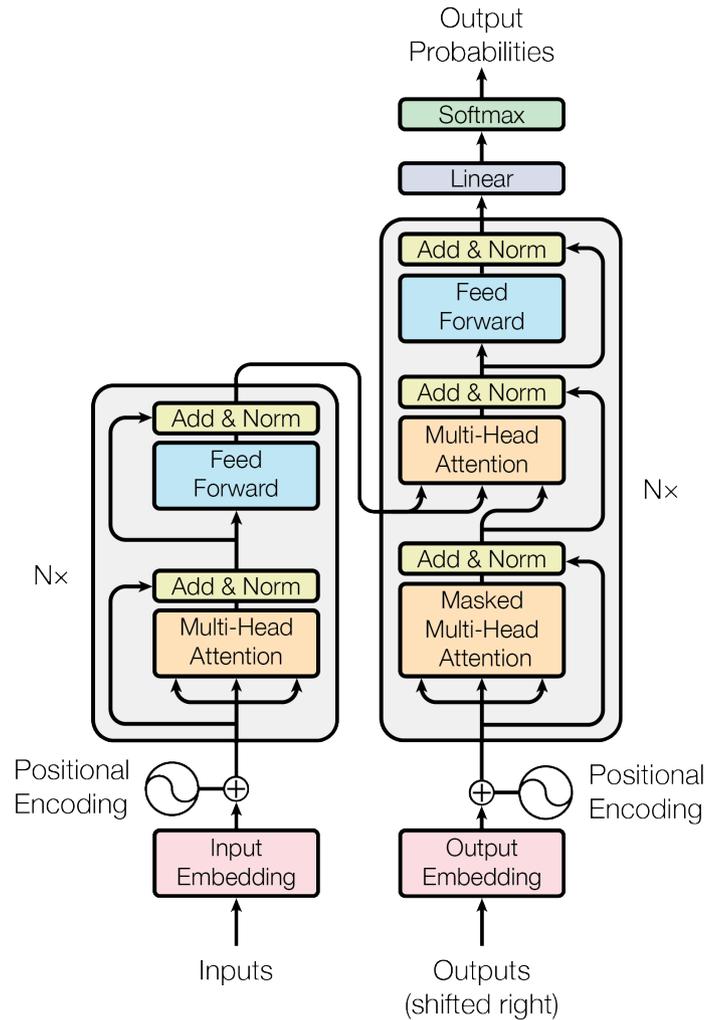
Self-Attention



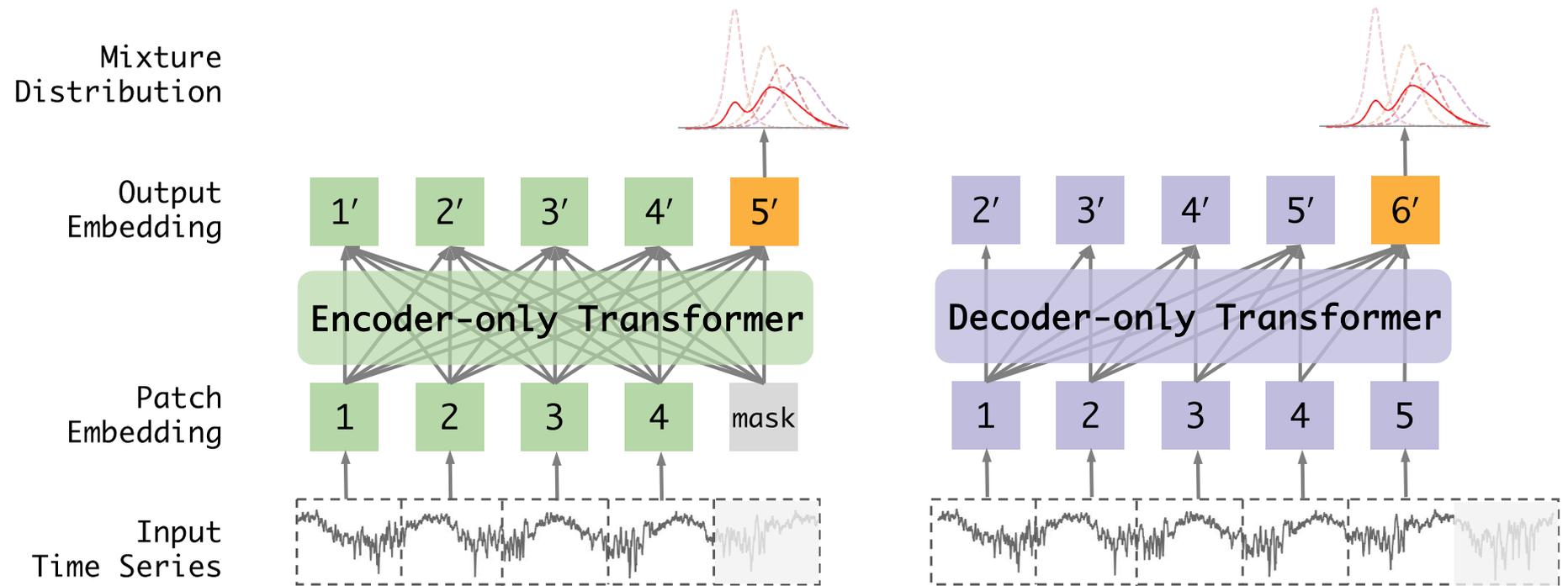
$$A_i = \text{softmax} \left( \frac{QK^\top}{\sqrt{d_k}} \right) V \quad Y_i = V^\top A_i$$



Source: Deep Learning Course, François Fleuret



Source: "Attention is all you need" (NIPS'17)



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

## Are Transformers Effective for Time Series Forecasting?

Ailing Zeng<sup>1\*</sup>, Muxi Chen<sup>1\*</sup>, Lei Zhang<sup>2</sup>, Qiang Xu<sup>1</sup>

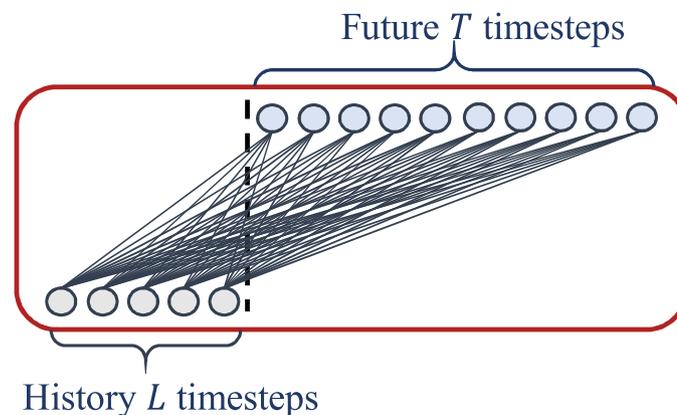
<sup>1</sup>The Chinese University of Hong Kong

<sup>2</sup>International 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)

## Are Transformers Effective for Time Series Forecasting?

Ailing Zeng<sup>1\*</sup>, Muxi Chen<sup>1\*</sup>, Lei Zhang<sup>2</sup>, Qiang Xu<sup>1</sup>

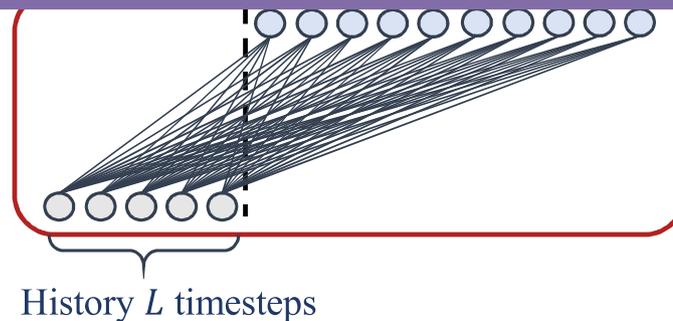
<sup>1</sup>The Chinese University of Hong Kong

<sup>2</sup>International Digital Economy Academy (IDEA)

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

- Comparison of transformer-based models

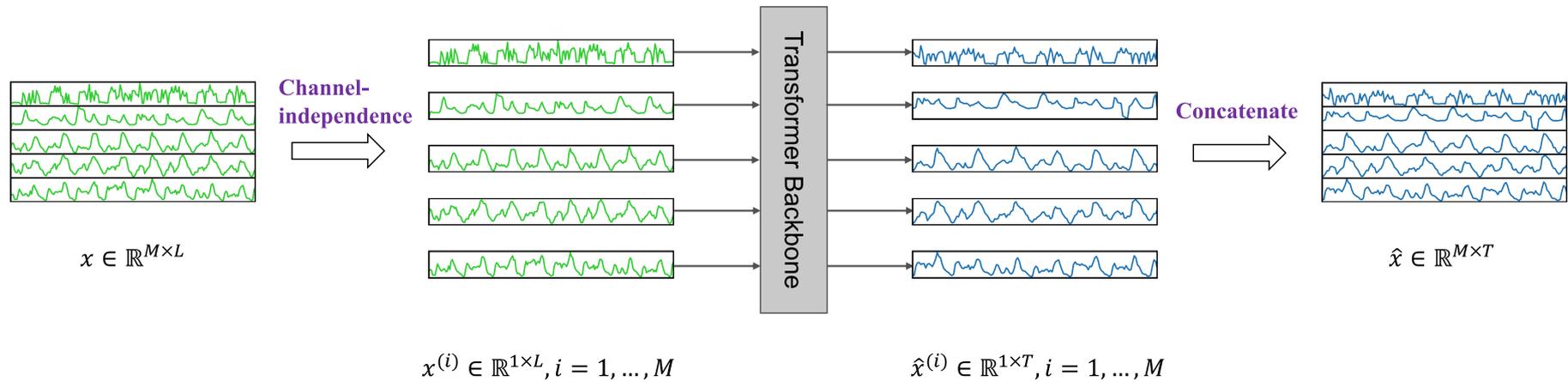
“LTSF-Linear surprisingly outperforms existing sophisticated Transformer-based LTSF models in all cases, and often by a large margin”



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

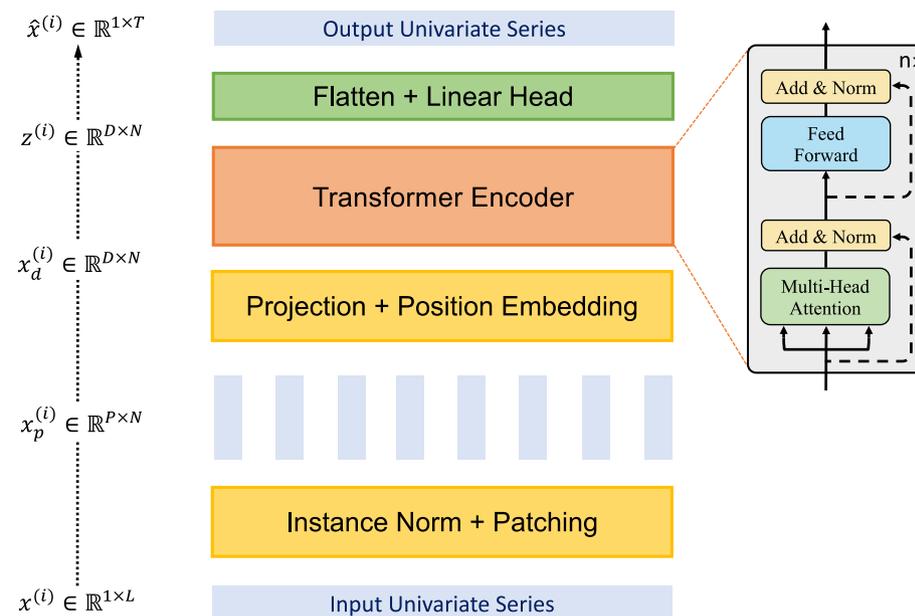
- Previous methods:
  1. each patch = 1 time point or some extracted information
  2. all variates are considered together
- PatchTST:
  1. patch = subseries
  2. channel-independence  
(bonus: much easier for transfer learning)
  3. proper preprocessing (RevIN)

## Channel independence



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

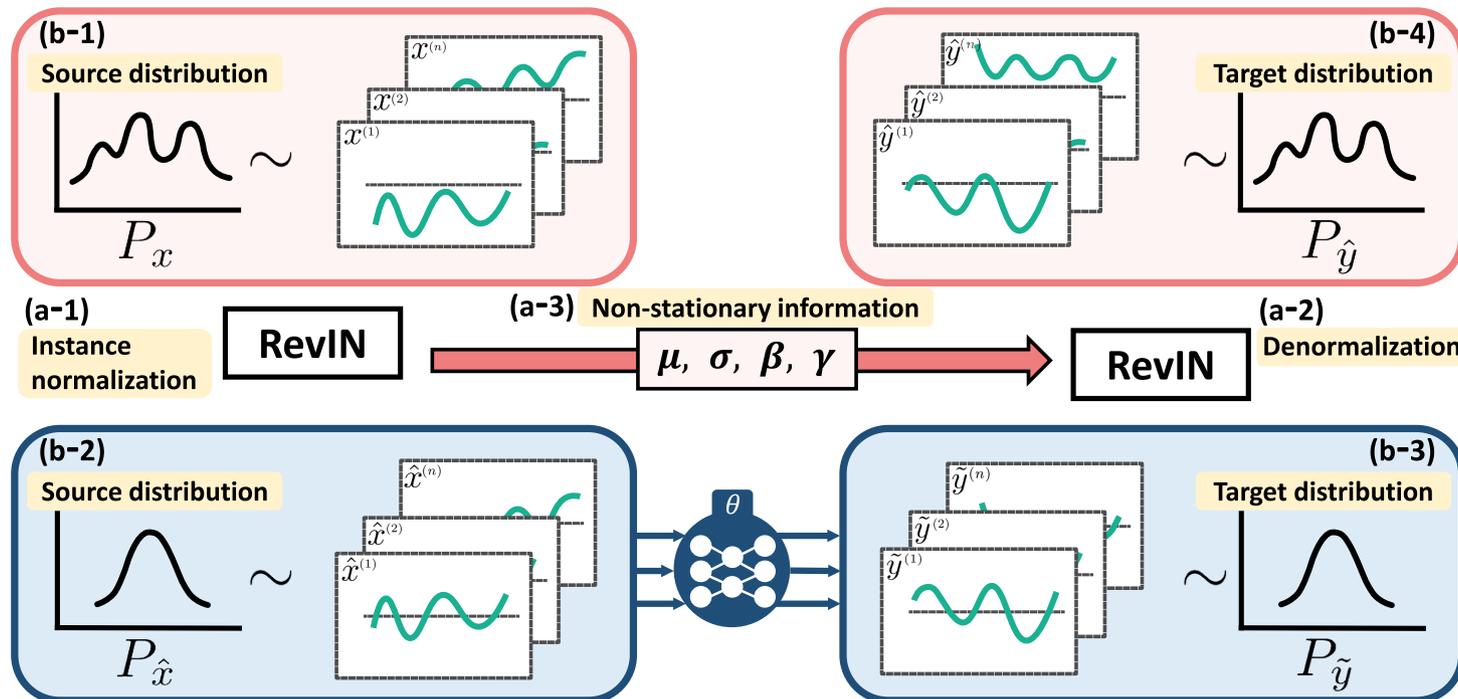
## Transformer backbone



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

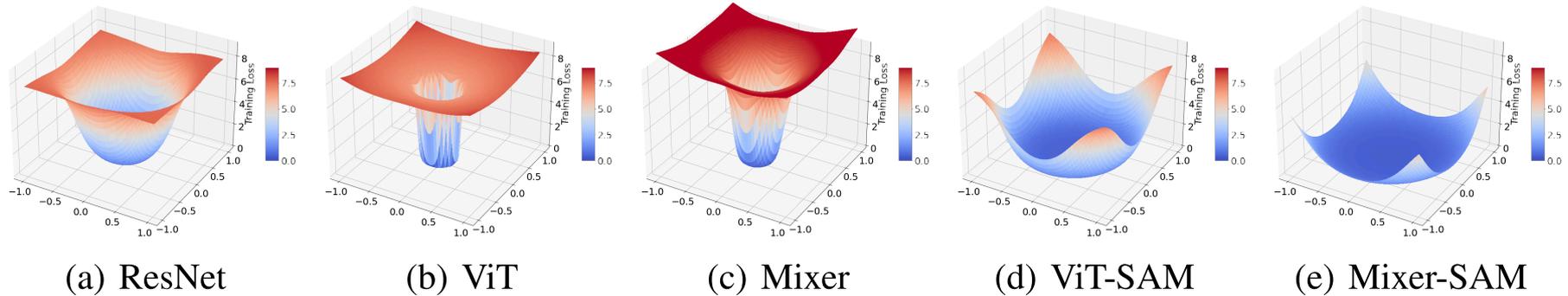
## Preprocessing (RevIN)

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



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

- Transformers have a sharp loss landscape:

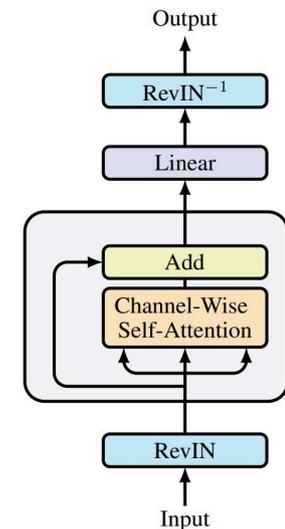
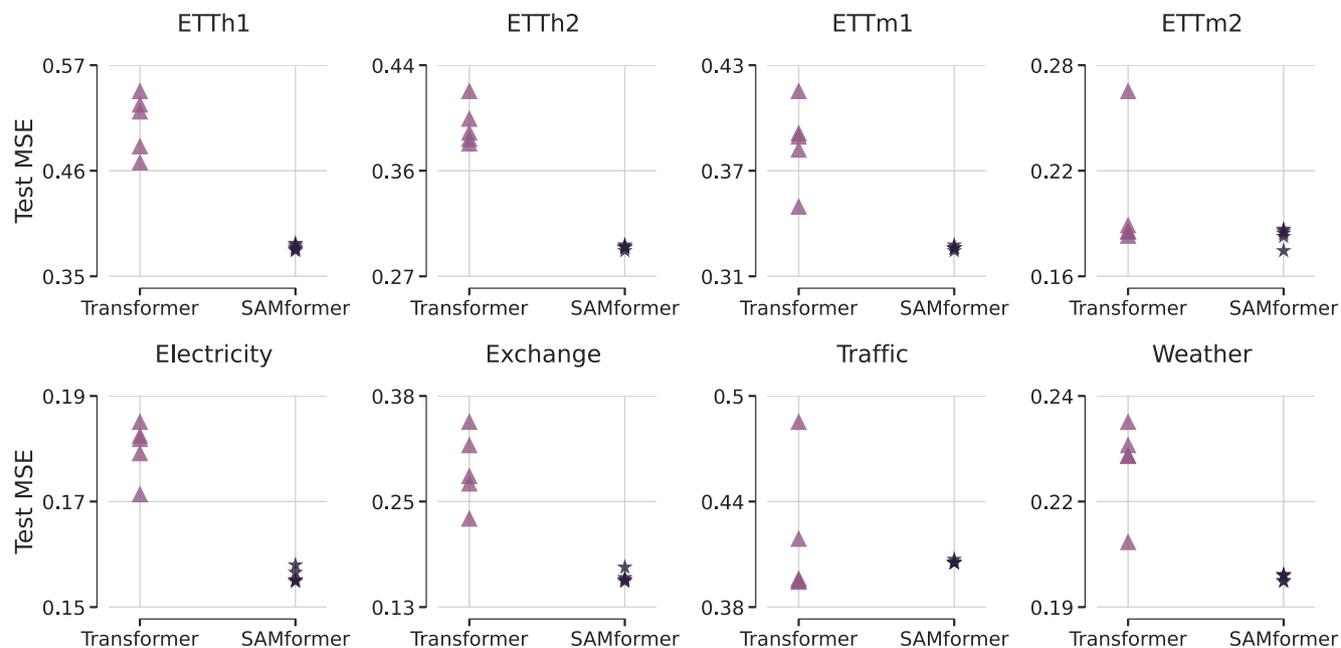


Source: “When Vision Transformers Outperform ResNets without Pre-training or Strong Data Augmentations” (ICLR’22)

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

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

- 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)