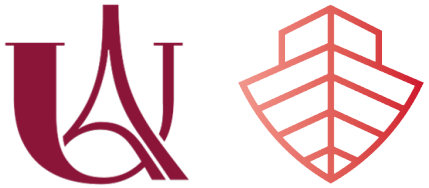
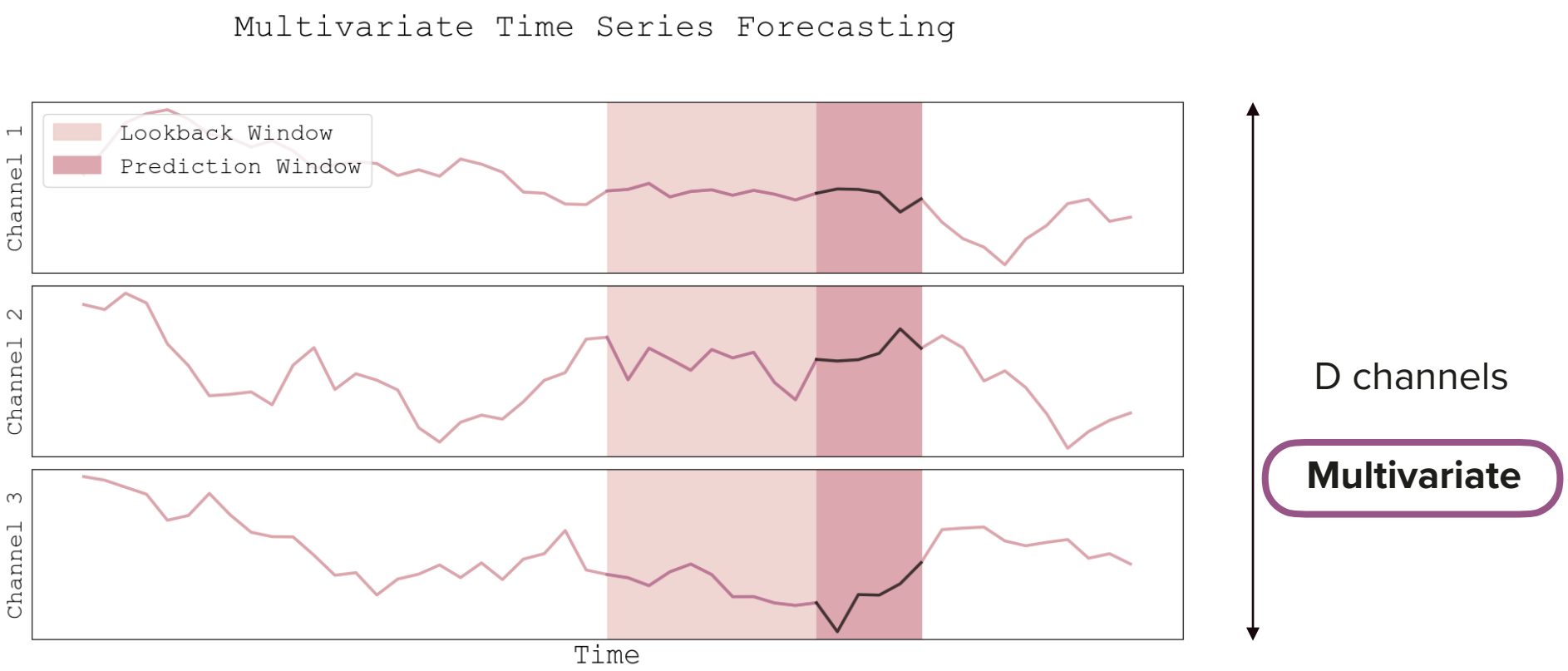
SAMformer : Unlocking the Potential of Transformers in Time-Series Forecasting

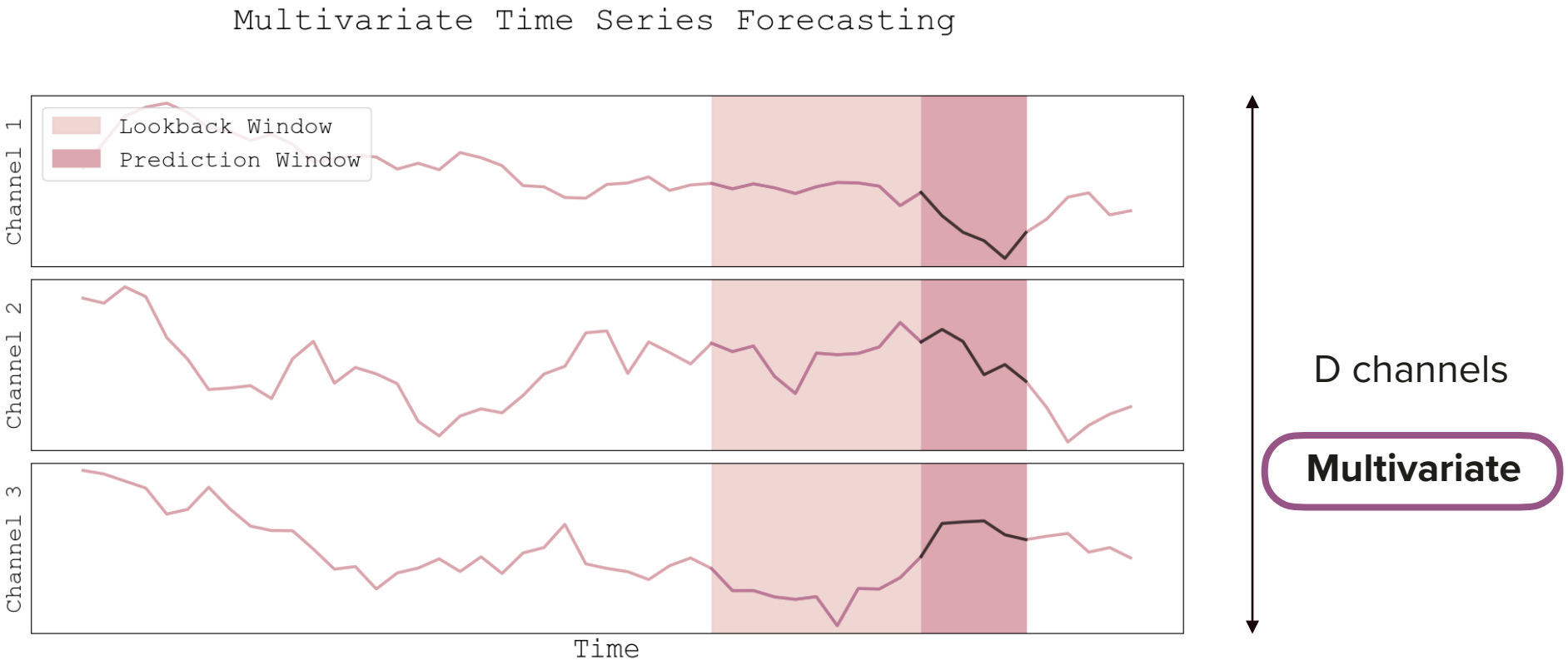
Romain ILBERT, Ambroise Odonnat, Vasilii Feofanov, Aladin Virmaux,
Giuseppe Paolo, Ievgen Redko, Themis Palpanas



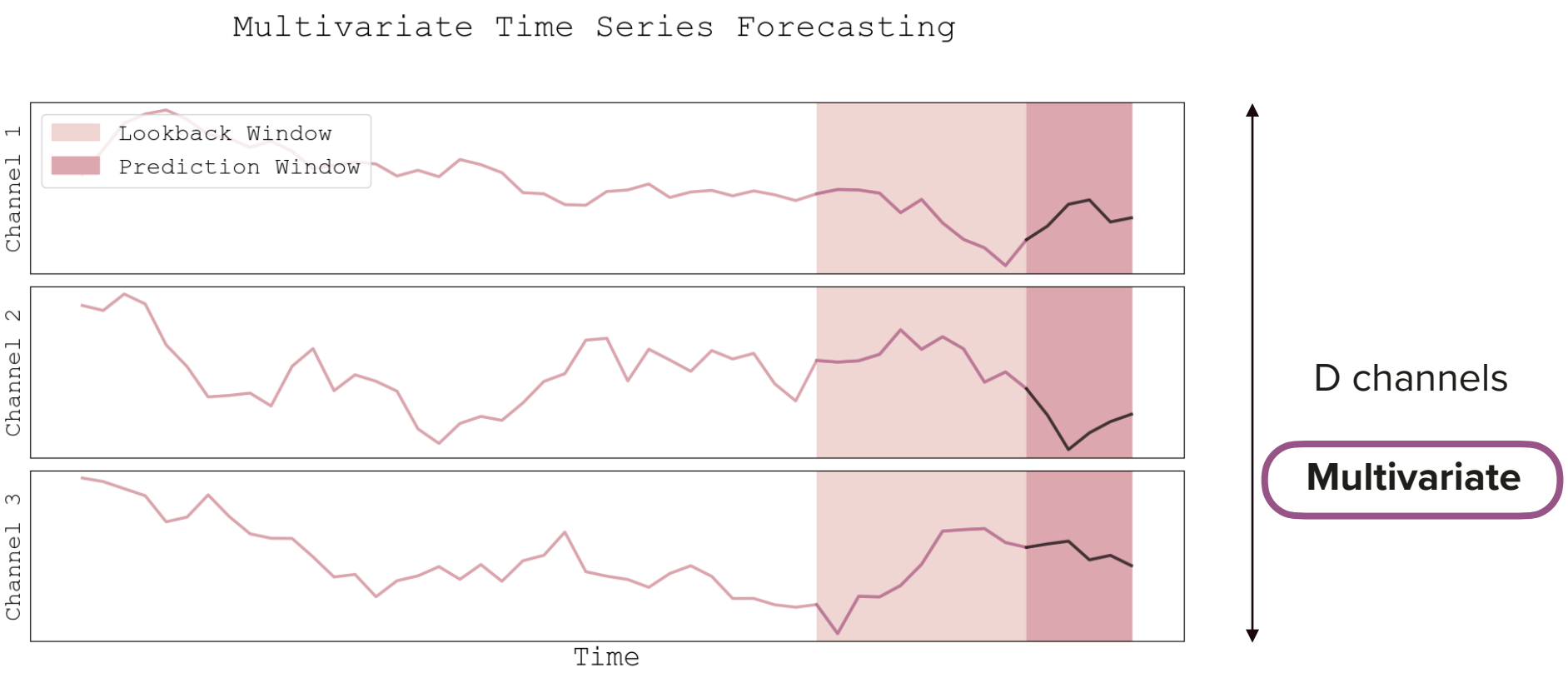
Time Series Forecasting : Problem Setup



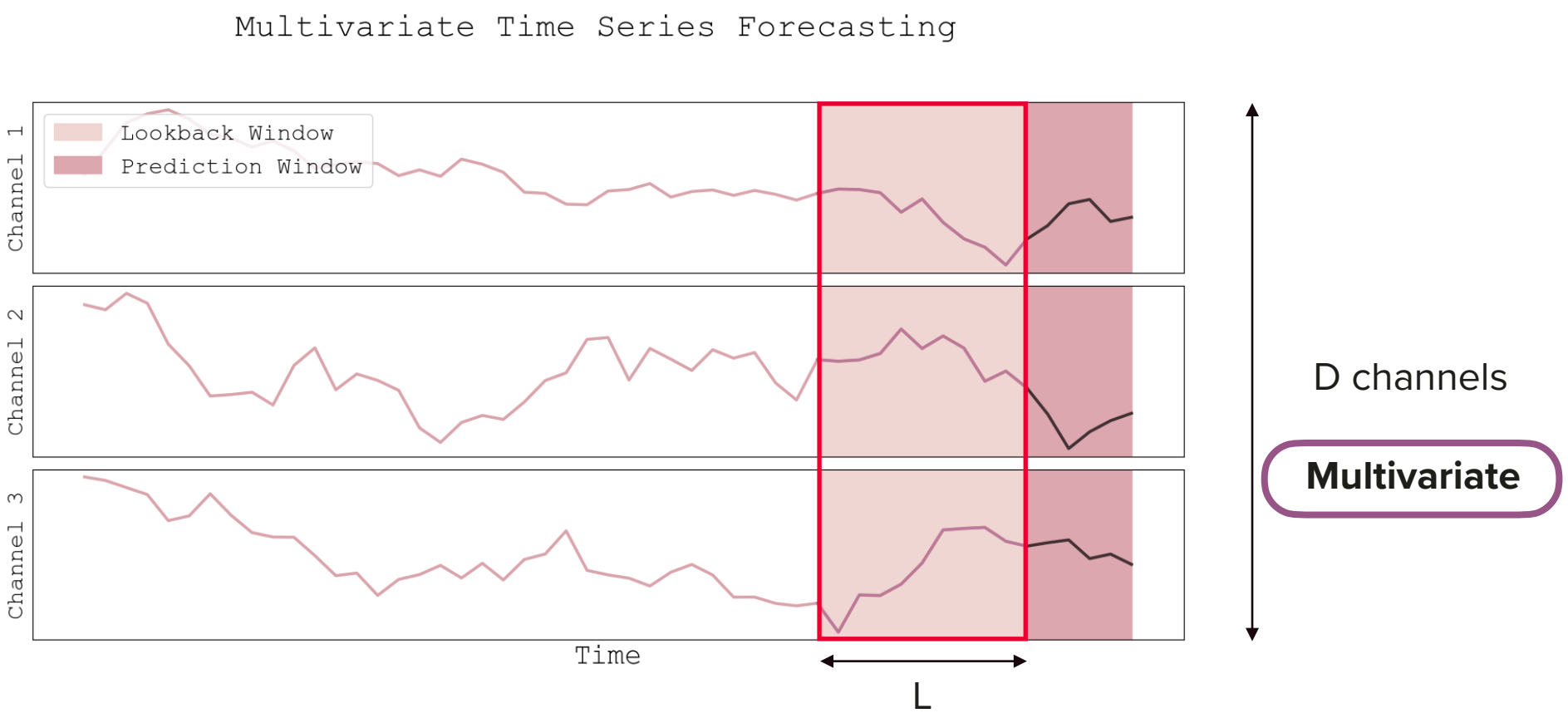
Time Series Forecasting : Problem Setup



Time Series Forecasting : Problem Setup

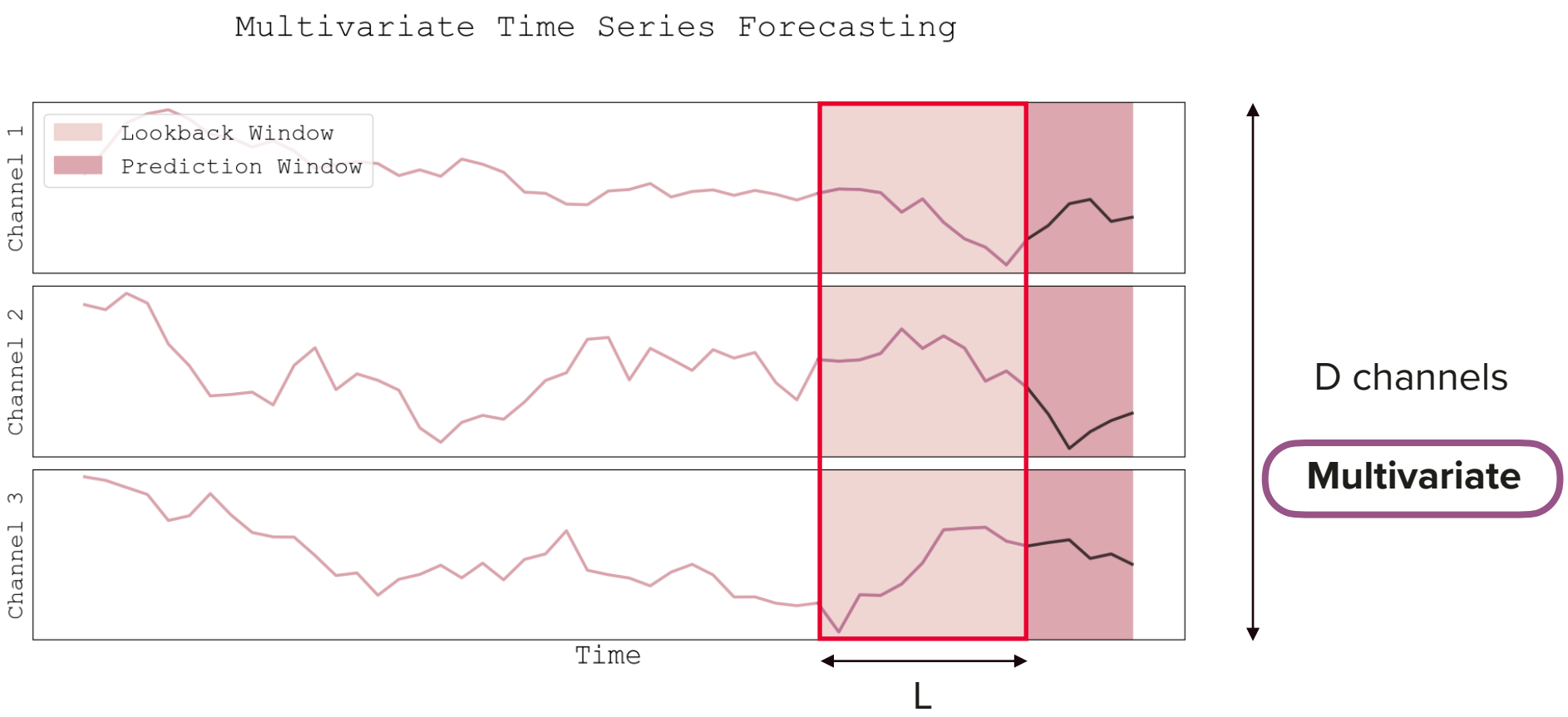


Time Series Forecasting : Problem Setup



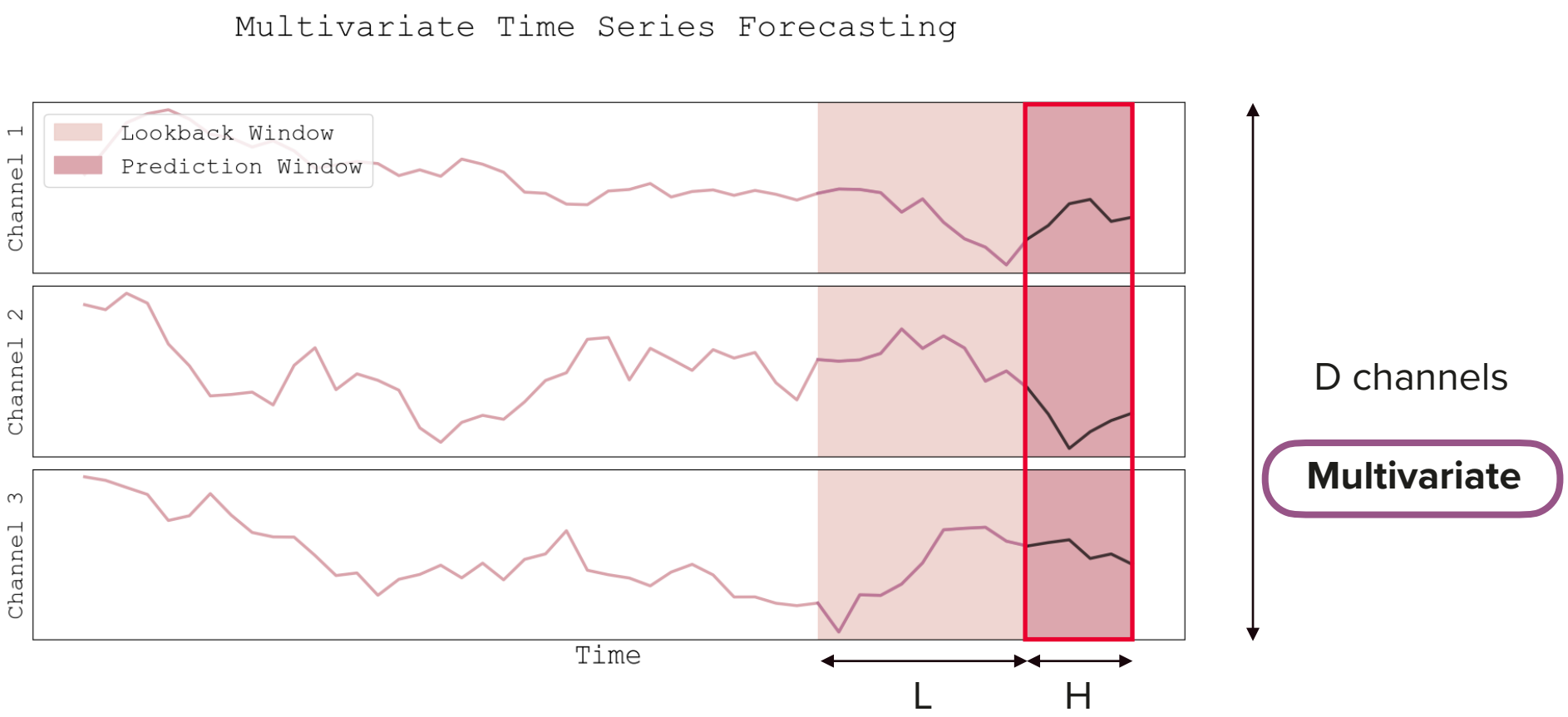
Inputs : $X \in \mathbb{R}^{D \times L}$

Time Series Forecasting : Problem Setup



Inputs : $X \in \mathbb{R}^{D \times L}$ $\xrightarrow{f}$

Time Series Forecasting : Problem Setup

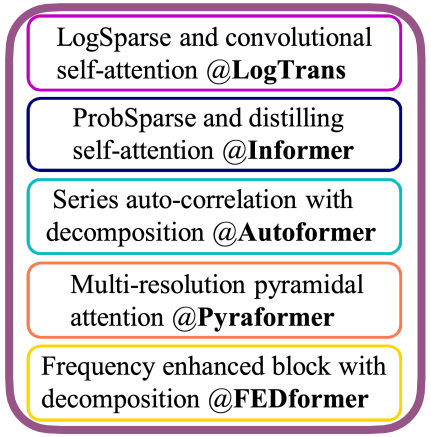


Inputs : $X \in \mathbb{R}^{D \times L}$ $\xrightarrow{f}$ Outputs : $Y \in \mathbb{R}^{D \times H}$, $H \in \{96,192,336,720\}$ **Long-Term**

Failure of Transformers in Time Series Forecasting

Main conclusions from *Are Transformers Effective for Time Series Forecasting* ?*

1. Transformer-based methods don't work well in forecasting

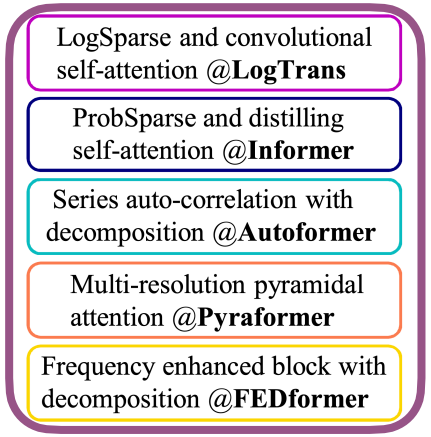


1. Are Transformers Effective for Time Series Forecasting ? Zeng et al. 2023.

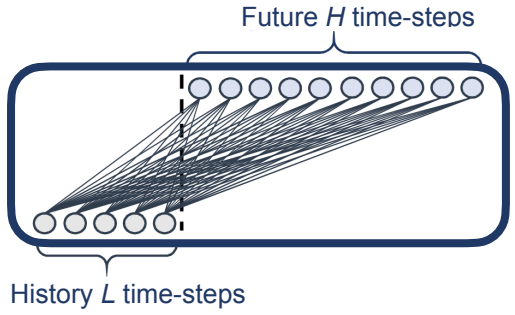
Failure of Transformers in Time Series Forecasting

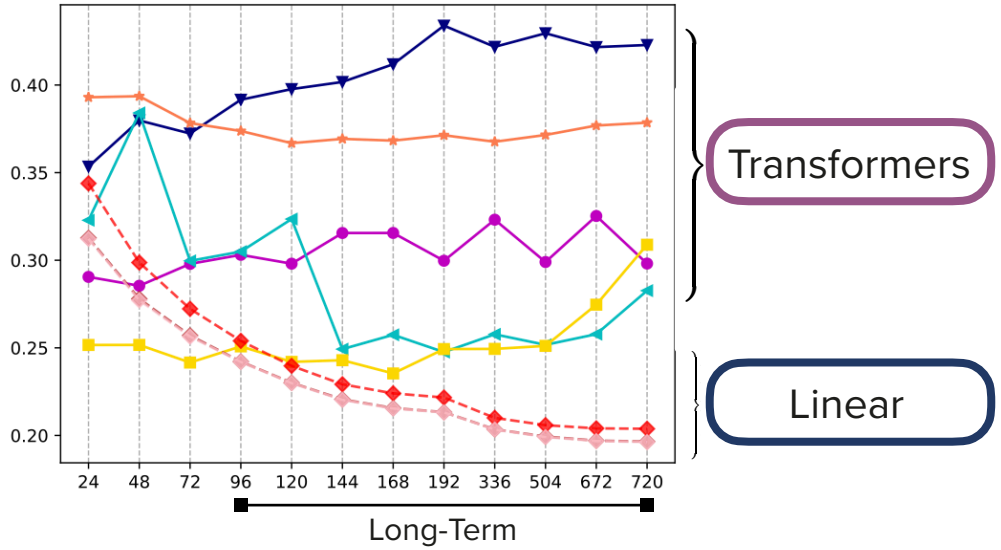
Main conclusions from *Are Transformers Effective for Time Series Forecasting* ?*

1. Transformer-based methods don't work well in forecasting
2. A simple Linear Model surpasses all of them in long-term forecasting



VS.



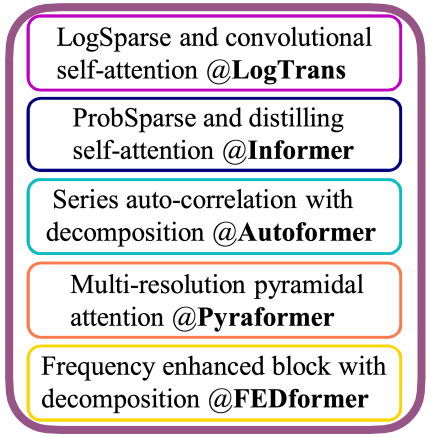


1. Are Transformers Effective for Time Series Forecasting ? Zeng et al. 2023.

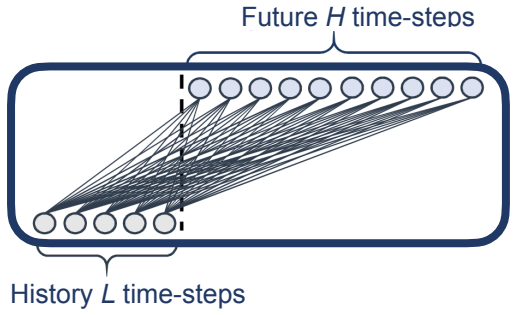
Failure of Transformers in Time Series Forecasting

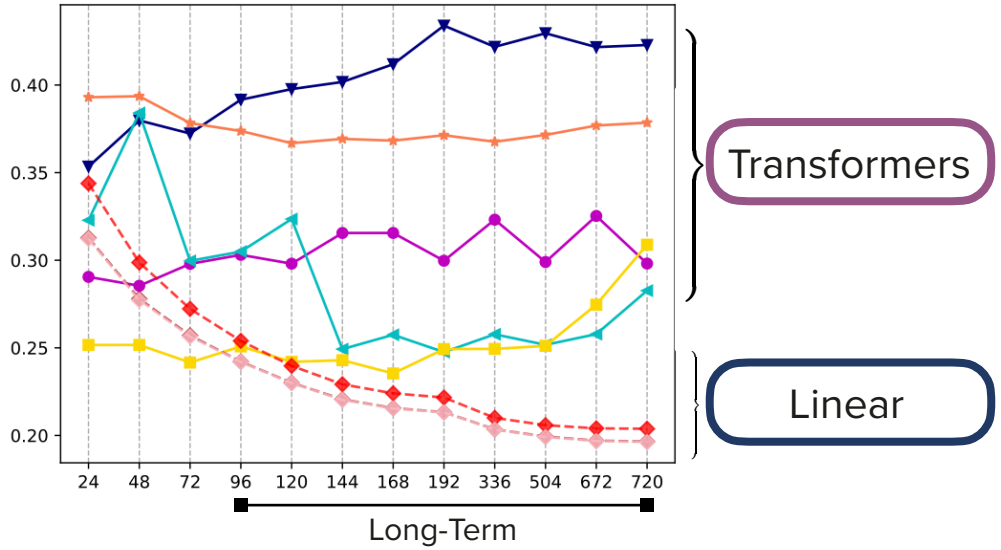
Main conclusions from *Are Transformers Effective for Time Series Forecasting* ?*

1. Transformer-based methods don't work well in forecasting
2. A simple Linear Model surpasses all of them in long-term forecasting



VS.





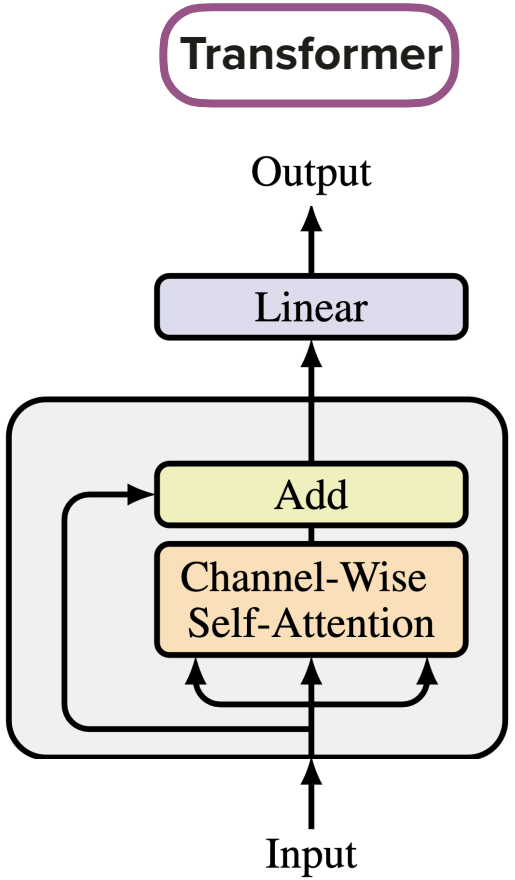
Yet Transformers dominate in NLP and vision… Why?

1. Are Transformers Effective for Time Series Forecasting ? Zeng et al. 2023.

Failure of Transformers : A simple forecasting example

Context :

- Consider a simple forecasting problem $\mathbf{Y} = \mathbf{X}\mathbf{W}_{\text{toy}} + \epsilon$ with L=512, H=96 and D=7



Failure of Transformers : A simple forecasting example

Context :

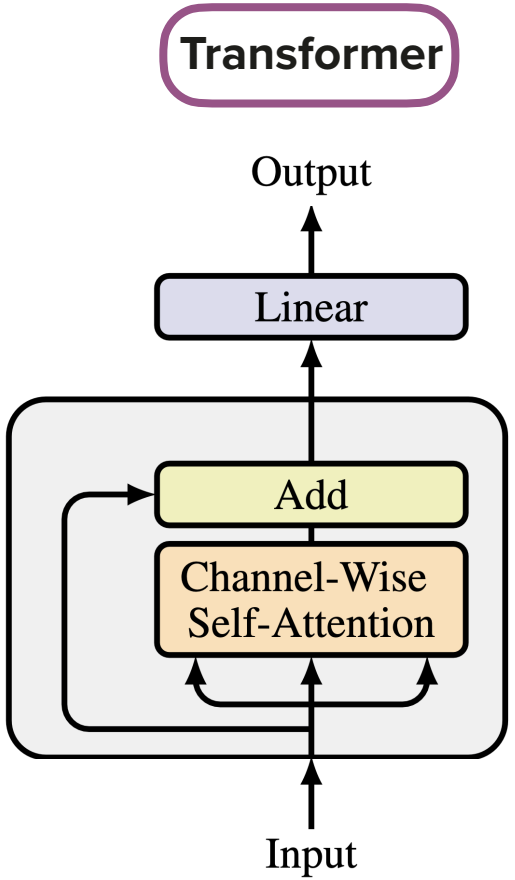
- Consider a simple forecasting problem $\mathbf{Y} = \mathbf{X}\mathbf{W}_{\text{toy}} + \epsilon$ with L=512, H=96 and D=7

Why channel-wise attention ? :

$$f(\mathbf{X}) = [\mathbf{X} + \text{attention}(\mathbf{X})]\mathbf{W}$$

$$\text{attention}(\mathbf{X}) = A(\mathbf{X})\mathbf{X}\mathbf{W}_V\mathbf{W}_O$$

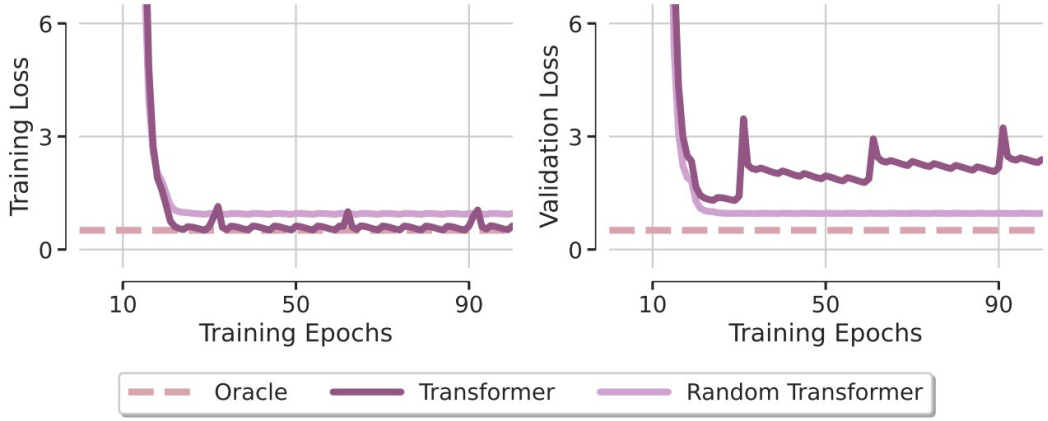
$$A(\mathbf{X}) \in \mathbb{R}^{D \times D}$$



Failure of Transformers : A simple forecasting example

Conclusions :

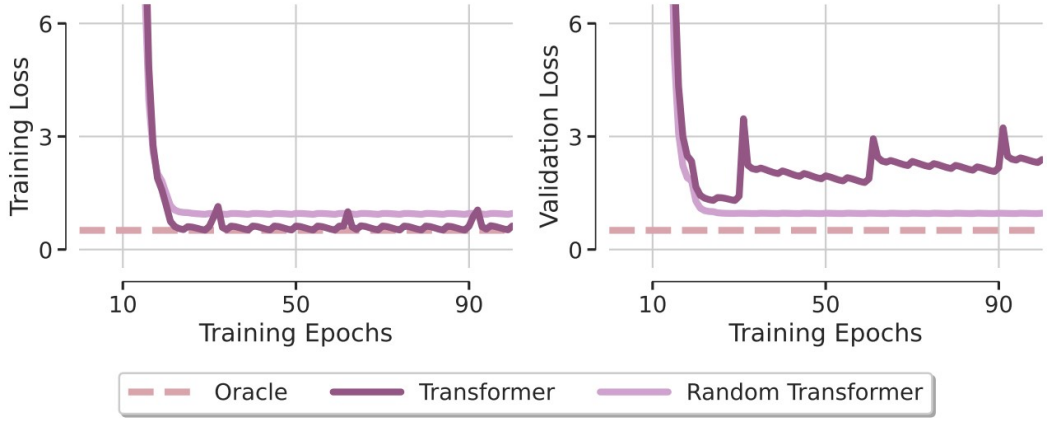
1. Our Transformer severely overfits…
2. … and works better if we freeze the attention…
3. … because the attention get stuck in a bad local minimum and does not move afterwards



Failure of Transformers : A simple forecasting example

Conclusions :

1. Our Transformer severely overfits…
2. … and works better if we freeze the attention…
3. … because the attention get stuck in a bad local minimum and does not move afterwards

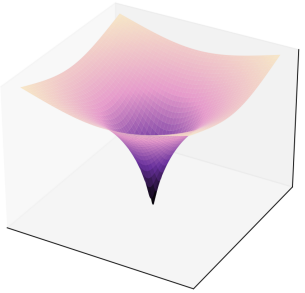


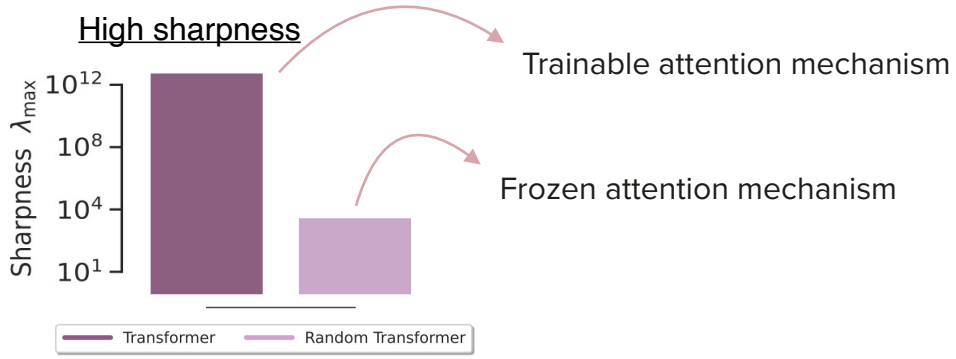
Pathological behavior suggesting sharp local minima

Failure of Transformers : A simple forecasting example

Why transformers fail?

1. Transformers have a sharp loss landscape

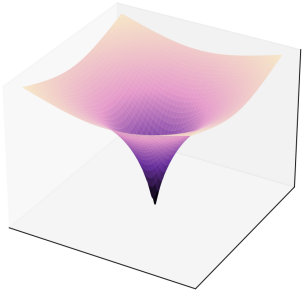


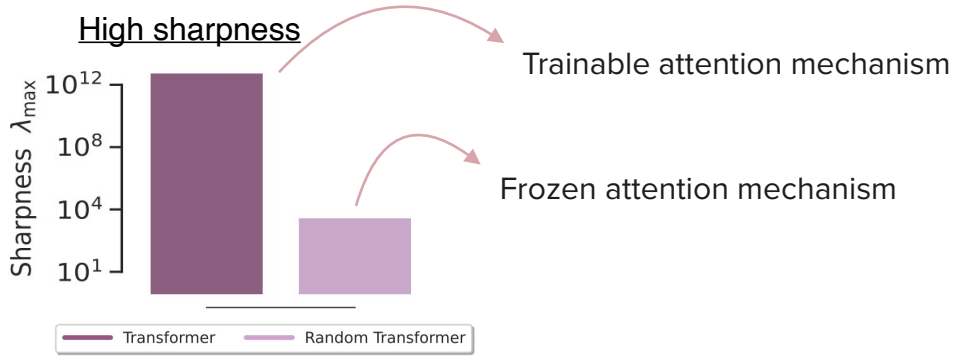


Failure of Transformers : A simple forecasting example

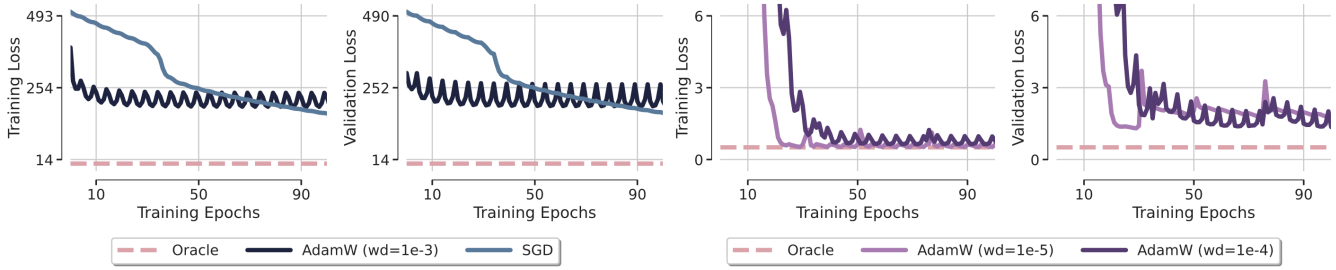
Why transformers fail?

1. Transformers have a sharp loss landscape





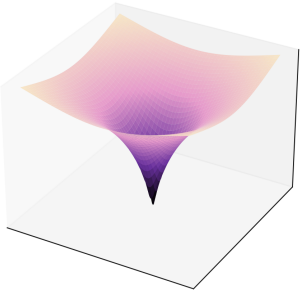
2. And no change in the optimizer helps to solve this

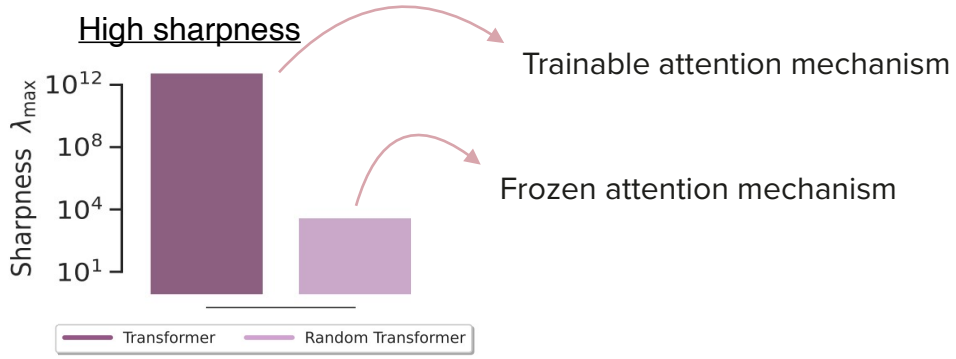


Failure of Transformers : A simple forecasting example

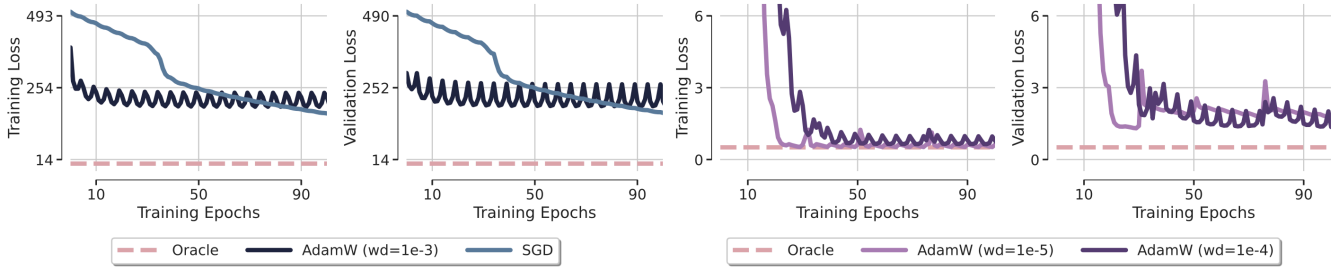
Why transformers fail?

1. Transformers have a sharp loss landscape





2. And no change in the optimizer helps to solve this



3. Well-known in NLP and vision (Chen et al., 2022, Zhai et al. 2023), ignored in TS Forecasting

Failure of Transformers : Solution

Sharpness - Aware Minimization (Foret et al. 2021, Chen et al. 2022)

- Smooths the loss landscape => flatter, more generalizable local minima

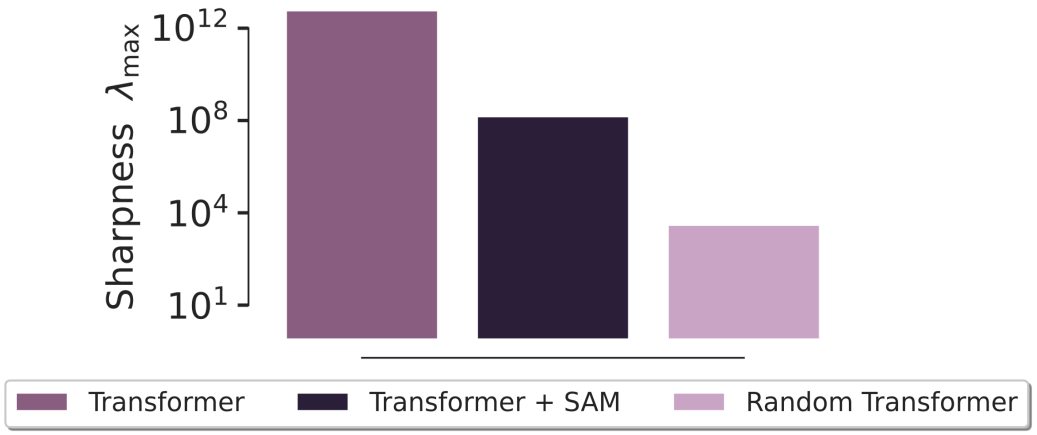
$$\mathcal{L}^{\text{SAM}}_{\text{train}}(\omega) = \max_{\|\epsilon\|<\rho} \mathcal{L}_{\text{train}}(\omega + \epsilon)$$

Failure of Transformers : Solution

Sharpness - Aware Minimization (Foret et al. 2021, Chen et al. 2022)

- Smooths the loss landscape => flatter, more generalizable local minima

$$\mathscr{L}^{\text{SAM}}_{\text{train}}(\omega) = \max_{\|\epsilon\|<\rho} \mathscr{L}_{\text{train}}(\omega + \epsilon)$$

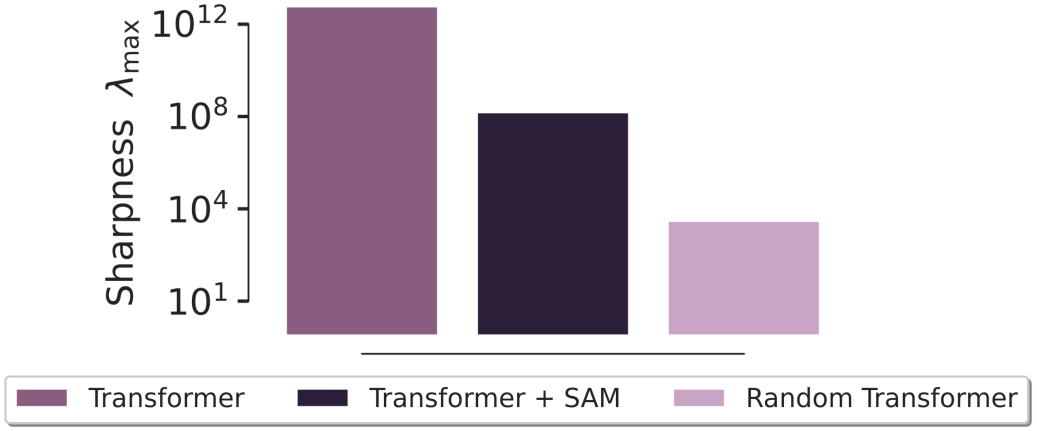


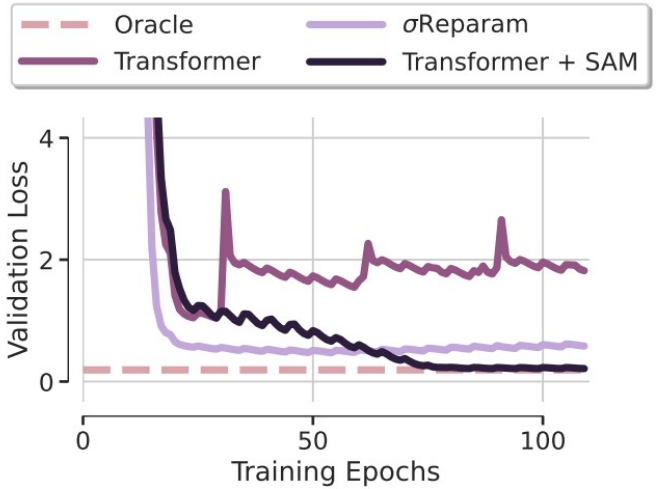
Failure of Transformers : Solution

Sharpness - Aware Minimization (Foret et al. 2021, Chen et al. 2022)

- Smooths the loss landscape => flatter, more generalizable local minima

$$\mathcal{L}_{\text{train}}^{\text{SAM}}(\omega) = \max_{\|\epsilon\|<\rho} \mathcal{L}_{\text{train}}(\omega + \epsilon)$$



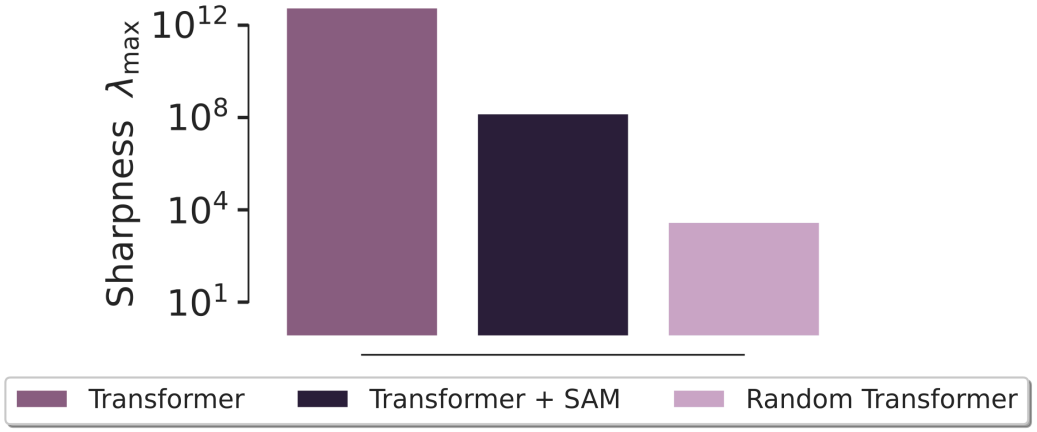


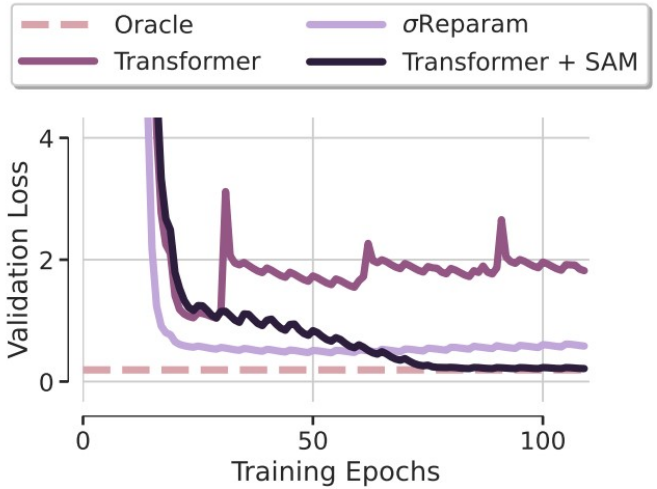
Failure of Transformers : Solution

Sharpness - Aware Minimization (Foret et al. 2021, Chen et al. 2022)

- Smooths the loss landscape => flatter, more generalizable local minima

$$\mathscr{L}^{\text{SAM}}_{\text{train}}(\omega) = \max_{\|\epsilon\|<\rho} \mathscr{L}_{\text{train}}(\omega + \epsilon)$$





SAM = Desired Solution

Congrats, you now know how to solve a linear forecasting problem with transformers!

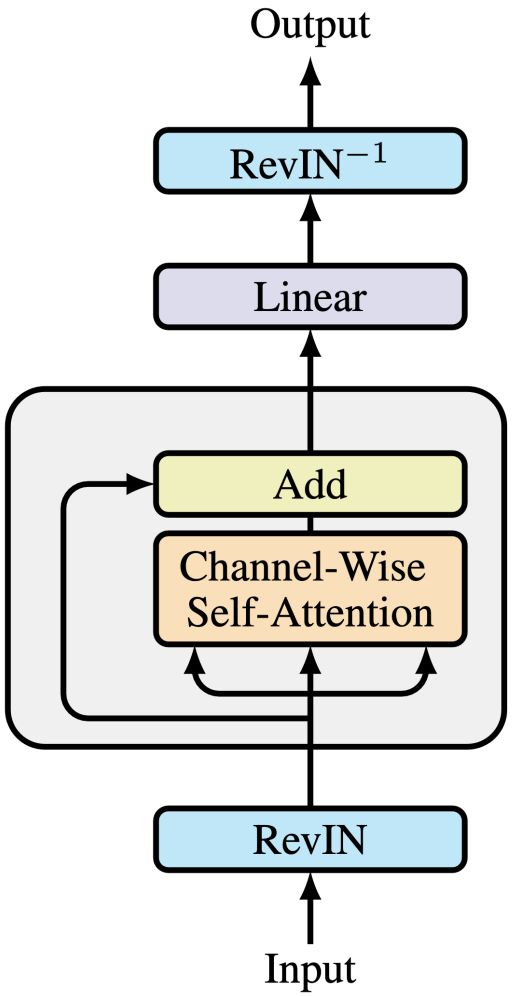
SAMformer (Ilbert et al. , ICML 2024)

A transformer-based Time Series Forecaster that actually works

SAMformer : Architecture

We can now design a simple model:

1. **RevIN layer** to be robust to train/test time shift
2. Shallow transformer with a **channel-wise attention**
3. We optimize it with **SAM**



SAMformer : Experimental results

SAMformer is better than all transformer-based models...

	iTrans*	PatchTST	FED*	TSMixer	Auto*	Pyra*	LogTrans	In*
Improvement	3.94%	11.13%	12.36%	14.33%	22.65%	61.88%	70.88%	72.20%

* formers

SAMformer : Experimental results

SAMformer is better than all transformer-based models...

	iTrans*	PatchTST	FED*	TSMixer	Auto*	Pyra*	LogTrans	In*
Improvement	3.94%	11.13%	12.36%	14.33%	22.65%	61.88%	70.88%	72.20%

* formers

... and the best mixer-based model ...

	TSMixer
Improvement	14.33%

SAMformer : Experimental results

SAMformer is better than all transformer-based models...

	iTrans*	PatchTST	FED*	TSMixer	Auto*	Pyra*	LogTrans	In*
Improvement	3.94%	11.13%	12.36%	14.33%	22.65%	61.88%	70.88%	72.20%

* formers

... and the best mixer-based model ...

	TSMixer
Improvement	14.33%

... while being much more simpler

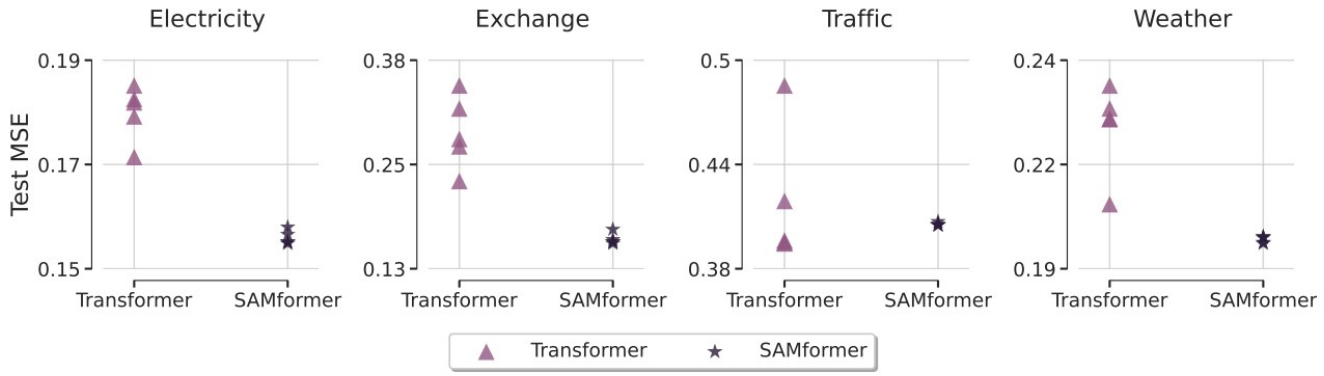
```python
import torch
import torch.nn as nn
from models.utils import RevNorm

class SAMformer(nn.Module):
    def __init__(self, pred_len, num_heads=1, d_model=16):
        super(SAMformer, self).__init__()
        self.pred_len = pred_len
        self.rev_norm = RevNorm(axis=-2)
        self.attention_layer = nn.MultiheadAttention(embed_dim=d_model, num_heads=num_heads, batch_first=True)
        self.dense = nn.Linear(d_model, pred_len)

    def forward(self, x):
        x = self.rev_norm(x, mode='norm').transpose(1, 2) #RevIN normalization
        attention_output, _ = self.attention_layer(x, x, x) #Channel-Wise Attention Computation
        x = x + attention_output #Residual Connection
        x = self.dense(x).transpose(1, 2) #Final Linear Projection
        return self.rev_norm(x, mode='denorm') #RevIn denormalization
```

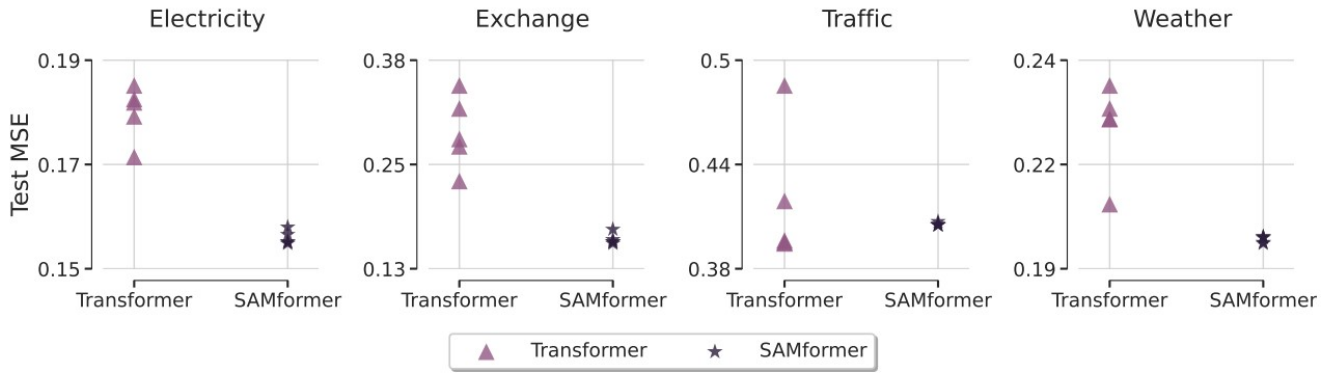
SAMformer : Experimental results

SAMformer is robust to random initialization ...



SAMformer : Experimental results

SAMformer is robust to random initialization ...

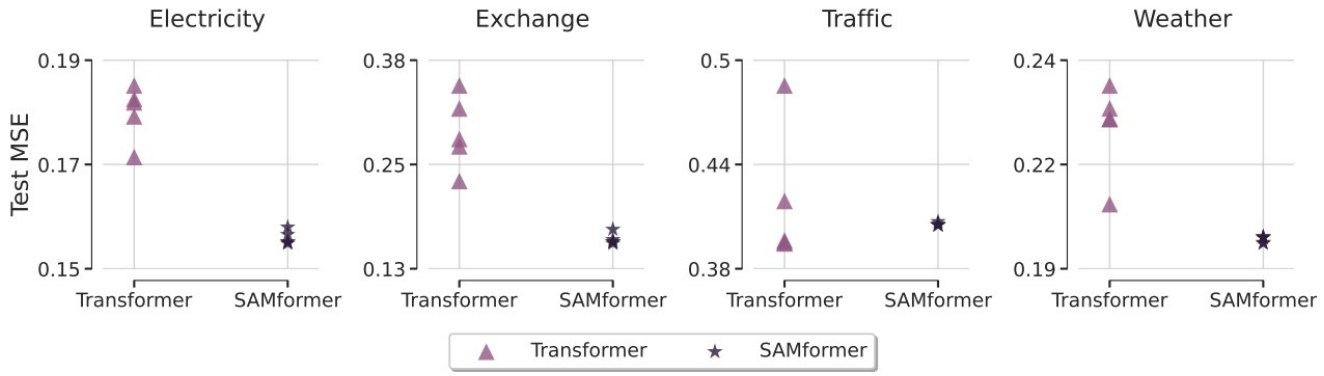


... and on par with the Foundation Model MOIRAI

Dataset	Full-shot	Zero-shot (Woo et al., 2024).		
	SAMformer	MOIRAI$_{\text{Small}}$	MOIRAI$_{\text{Base}}$	MOIRAI$_{\text{Large}}$
ETTh1	<u>0.410</u>	**0.400**	0.434	0.510
ETTh2	<u>0.344</u>	**0.341**	0.345	0.354
ETTm1	**0.373**	0.448	<u>0.381</u>	0.390
ETTm2	**0.269**	0.300	<u>0.272</u>	0.276
Electricity	**0.181**	0.233	<u>0.188</u>	<u>0.188</u>
Weather	0.260	<u>0.242</u>	**0.238**	0.259
Overall MSE improvement		**6.9%**	**1.1%**	**7.6%**

SAMformer : Experimental results

SAMformer is robust to random initialization …



… and on par with the Foundation Model MOIRAI with many fewer parameters

Dataset	Full-shot	Zero-shot (Woo et al., 2024).		
	SAMformer	MOIRAI$_{\text{Small}}$	MOIRAI$_{\text{Base}}$	MOIRAI$_{\text{Large}}$
ETTh1	0.410	**0.400**	0.434	0.510
ETTh2	0.344	**0.341**	0.345	0.354
ETTm1	**0.373**	0.448	0.381	0.390
ETTm2	**0.269**	0.300	0.272	0.276
Electricity	**0.181**	0.233	0.188	0.188
Weather	0.260	0.242	**0.238**	0.259
Overall MSE improvement		**6.9%**	**1.1%**	**7.6%**

Model	# of parameters
MOIRAI$_{small}$	14M
MOIRAI$_{base}$	91M
MOIRAI$_{large}$	311M
SAMformer	280K

SAM vs σ – Reparametrization : Comparison

σ – **Reparametrization** (Zhen et al. 2023)

- smoothes the attention matrix with attention weights reparametrization

$$\hat{\mathbf{W}} = \frac{\gamma}{\|\mathbf{W}\|_2}\mathbf{W}$$

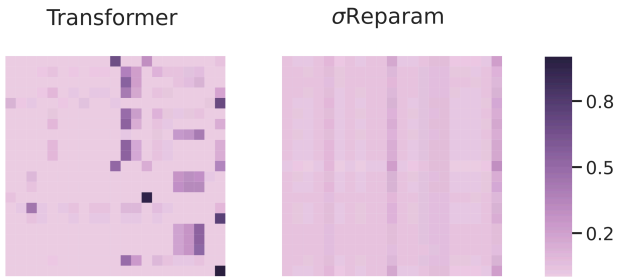
SAM vs σ – Reparametrization : Comparison

σ – **Reparametrization** (Zhen et al. 2023)

- smoothes the attention matrix with attention weights reparametrization

$$\hat{\mathbf{W}} = \frac{\gamma}{\|\mathbf{W}\|_2}\mathbf{W}$$

- proved to be efficient in NLP ... but does not work for TS Forecasting



Observations:

- Transformers ignore diagonal elements
- σ – Reparametrization over-smoothes the attention matrix

} Rank Collapse = Uninformative Channel-Wise Attention

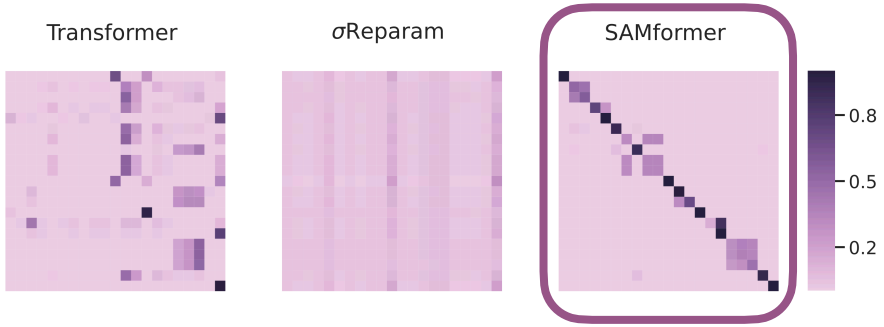
SAM vs σ – Reparametrization : Comparison

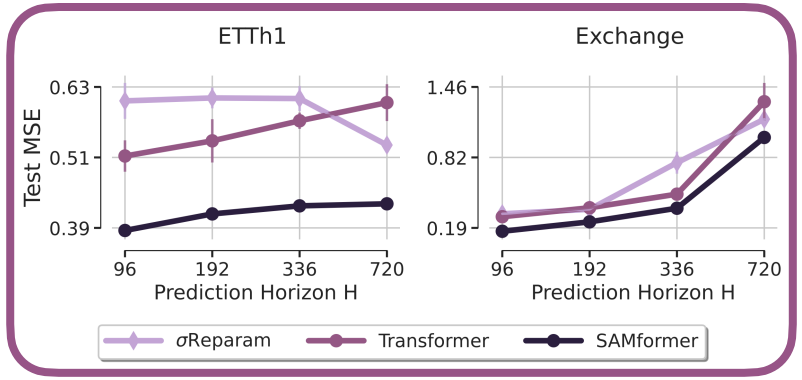
σ – **Reparametrization** (Zhen et al. 2023)

- smoothes the attention matrix with attention weights reparametrization

$$\hat{\mathbf{W}} = \frac{\gamma}{\|\mathbf{W}\|_2}\mathbf{W}$$

- proved to be efficient in NLP ... but does not work for TS Forecasting





Observations:

- Transformers ignore diagonal elements
- σ – Reparametrization over-smoothes the attention matrix
- SAMformer encourages feature self-correlation

} Rank Collapse = Uninformative Channel-Wise Attention

SAMformer : Ablation Studies

Ablation study on channel-wise attention and identity weight matrix attention ?

- Candidate 1: SAMformer with **temporal** attention (as used in all other transformers)

Model	Metrics	H	ETTh1	ETTh2	ETTm1	ETTm2	Electricity	Exchange	Traffic	Weather	**Overall Improvement**
Temporal Attention	MSE	96	$0.496_{\pm 0.009}$	$0.401_{\pm 0.011}$	$0.542_{\pm 0.063}$	$0.330_{\pm 0.034}$	$0.291_{\pm 0.025}$	$0.684_{\pm 0.218}$	$0.933_{\pm 0.188}$	$0.225_{\pm 0.005}$	12.97%
		192	$0.510_{\pm 0.014}$	$0.414_{\pm 0.020}$	$0.615_{\pm 0.056}$	$0.394_{\pm 0.033}$	$0.294_{\pm 0.024}$	$0.434_{\pm 0.063}$	$0.647_{\pm 0.131}$	$0.254_{\pm 0.001}$	
		336	$0.549_{\pm 0.017}$	$0.396_{\pm 0.014}$	$0.620_{\pm 0.046}$	$0.436_{\pm 0.081}$	$0.290_{\pm 0.016}$	$0.473_{\pm 0.014}$	$0.656_{\pm 0.113}$	$0.292_{\pm 0.000}$	
		720	$0.604_{\pm 0.017}$	$0.396_{\pm 0.010}$	$0.694_{\pm 0.055}$	$0.469_{\pm 0.005}$	$0.307_{\pm 0.014}$	$1.097_{\pm 0.084}$	-	$0.346_{\pm 0.000}$	
	MAE	96	$0.488_{\pm 0.007}$	$0.434_{\pm 0.006}$	$0.525_{\pm 0.040}$	$0.393_{\pm 0.020}$	$0.386_{\pm 0.014}$	$0.589_{\pm 0.096}$	$0.598_{\pm 0.072}$	$0.277_{\pm 0.004}$	18.09%
		192	$0.492_{\pm 0.010}$	$0.443_{\pm 0.015}$	$0.566_{\pm 0.032}$	$0.421_{\pm 0.019}$	$0.385_{\pm 0.014}$	$0.498_{\pm 0.033}$	$0.467_{\pm 0.072}$	$0.294_{\pm 0.001}$	
		336	$0.517_{\pm 0.012}$	$0.440_{\pm 0.012}$	$0.550_{\pm 0.024}$	$0.443_{\pm 0.039}$	$0.383_{\pm 0.009}$	$0.517_{\pm 0.008}$	$0.469_{\pm 0.070}$	$0.320_{\pm 0.000}$	
		720	$0.556_{\pm 0.009}$	$0.442_{\pm 0.006}$	$0.584_{\pm 0.027}$	$0.459_{\pm 0.004}$	$0.396_{\pm 0.012}$	$0.782_{\pm 0.041}$	-	$0.356_{\pm 0.000}$	

SAMformer : Ablation Studies

Ablation study on channel-wise attention and identity weight matrix attention ?

- Candidate 1: SAMformer with **temporal** attention (as used in all other transformers)

Model	Metrics	H	ETTh1	ETTh2	ETTm1	ETTm2	Electricity	Exchange	Traffic	Weather	Overall Improvement
Temporal Attention	MSE	96	$0.496_{\pm 0.009}$	$0.401_{\pm 0.011}$	$0.542_{\pm 0.063}$	$0.330_{\pm 0.034}$	$0.291_{\pm 0.025}$	$0.684_{\pm 0.218}$	$0.933_{\pm 0.188}$	$0.225_{\pm 0.005}$	12.97%
		192	$0.510_{\pm 0.014}$	$0.414_{\pm 0.020}$	$0.615_{\pm 0.056}$	$0.394_{\pm 0.033}$	$0.294_{\pm 0.024}$	$0.434_{\pm 0.063}$	$0.647_{\pm 0.131}$	$0.254_{\pm 0.001}$	
		336	$0.549_{\pm 0.017}$	$0.396_{\pm 0.014}$	$0.620_{\pm 0.046}$	$0.436_{\pm 0.081}$	$0.290_{\pm 0.016}$	$0.473_{\pm 0.014}$	$0.656_{\pm 0.113}$	$0.292_{\pm 0.000}$	
		720	$0.604_{\pm 0.017}$	$0.396_{\pm 0.010}$	$0.694_{\pm 0.055}$	$0.469_{\pm 0.005}$	$0.307_{\pm 0.014}$	$1.097_{\pm 0.084}$	-	$0.346_{\pm 0.000}$	
	MAE	96	$0.488_{\pm 0.007}$	$0.434_{\pm 0.006}$	$0.525_{\pm 0.040}$	$0.393_{\pm 0.020}$	$0.386_{\pm 0.014}$	$0.589_{\pm 0.096}$	$0.598_{\pm 0.072}$	$0.277_{\pm 0.004}$	18.09%
		192	$0.492_{\pm 0.010}$	$0.443_{\pm 0.015}$	$0.566_{\pm 0.032}$	$0.421_{\pm 0.019}$	$0.385_{\pm 0.014}$	$0.498_{\pm 0.033}$	$0.467_{\pm 0.072}$	$0.294_{\pm 0.001}$	
		336	$0.517_{\pm 0.012}$	$0.440_{\pm 0.012}$	$0.550_{\pm 0.024}$	$0.443_{\pm 0.039}$	$0.383_{\pm 0.009}$	$0.517_{\pm 0.008}$	$0.469_{\pm 0.070}$	$0.320_{\pm 0.000}$	
		720	$0.556_{\pm 0.009}$	$0.442_{\pm 0.006}$	$0.584_{\pm 0.027}$	$0.459_{\pm 0.004}$	$0.396_{\pm 0.012}$	$0.782_{\pm 0.041}$	-	$0.356_{\pm 0.000}$	

- Candidate 2: SAMformer with **identity weight matrix** attention

Model	Metrics	H	ETTh1	ETTh2	ETTm1	ETTm2	Electricity	Exchange	Traffic	Weather	Overall Improvement
Identity Attention	MSE	96	$0.477_{\pm 0.059}$	$0.346_{\pm 0.055}$	$0.345_{\pm 0.027}$	$0.201_{\pm 0.035}$	$0.175_{\pm 0.015}$	$0.179_{\pm 0.031}$	$0.416_{\pm 0.037}$	$0.206_{\pm 0.019}$	11.93%
		192	$0.467_{\pm 0.074}$	$0.374_{\pm 0.031}$	$0.384_{\pm 0.042}$	$0.248_{\pm 0.016}$	$0.189_{\pm 0.022}$	$0.320_{\pm 0.070}$	$0.437_{\pm 0.041}$	$0.236_{\pm 0.002}$	
		336	$0.512_{\pm 0.070}$	$0.372_{\pm 0.024}$	$0.408_{\pm 0.032}$	$0.303_{\pm 0.022}$	$0.211_{\pm 0.019}$	$0.443_{\pm 0.071}$	$0.500_{\pm 0.155}$	$0.277_{\pm 0.003}$	
		720	$0.505_{\pm 0.107}$	$0.405_{\pm 0.012}$	$0.466_{\pm 0.043}$	$0.397_{\pm 0.029}$	$0.233_{\pm 0.019}$	$1.123_{\pm 0.076}$	$0.468_{\pm 0.021}$	$0.338_{\pm 0.009}$	
	MAE	96	$0.473_{\pm 0.041}$	$0.395_{\pm 0.033}$	$0.376_{\pm 0.019}$	$0.294_{\pm 0.027}$	$0.283_{\pm 0.023}$	$0.320_{\pm 0.023}$	$0.301_{\pm 0.039}$	$0.259_{\pm 0.021}$	4.18%
		192	$0.463_{\pm 0.055}$	$0.413_{\pm 0.022}$	$0.399_{\pm 0.030}$	$0.321_{\pm 0.012}$	$0.291_{\pm 0.029}$	$0.418_{\pm 0.043}$	$0.314_{\pm 0.042}$	$0.278_{\pm 0.002}$	
		336	$0.490_{\pm 0.049}$	$0.413_{\pm 0.015}$	$0.411_{\pm 0.019}$	$0.354_{\pm 0.018}$	$0.309_{\pm 0.021}$	$0.498_{\pm 0.041}$	$0.350_{\pm 0.106}$	$0.305_{\pm 0.003}$	
		720	$0.496_{\pm 0.066}$	$0.438_{\pm 0.008}$	$0.444_{\pm 0.030}$	$0.406_{\pm 0.017}$	$0.322_{\pm 0.021}$	$0.788_{\pm 0.021}$	$0.325_{\pm 0.023}$	$0.347_{\pm 0.009}$	

Conclusions on SAMformer (ILBERT et al, ICML oral 2024)

1. We studied **pitfalls of transformers** in time series forecasting
    - Sharp loss landscape = lack of generalization
2. Our proposal **SAMformer**
    - **SAMformer =** RevIN + channel-wise attention + SAM optimization
    - **SOTA** in long-term multivariate time series forecasting
    - **Consistent** = same architecture of different horizons/datasets
    - **Lightweight** = the smallest SOTA model
    - **On par with** the large foundation model **MOIRAI**
3. Can inspire further work to enhance our simple architecture.

Thank you

Check my website for code, paper and slides

Let’s discuss potential post-PhD opportunities