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

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

Check my website for code, paper and slides :

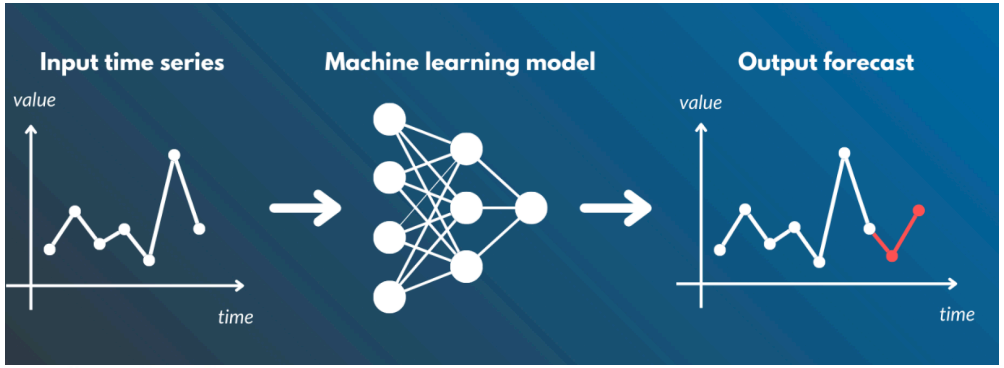
* Presenter, PhD Student at LIPADE, Paris Descartes University & Huawei Paris Research Center

Accepted as an Oral at ICML 2024

Time Series Forecasting : A Definition

Problem Setup

1. Given past observations, predict future ones



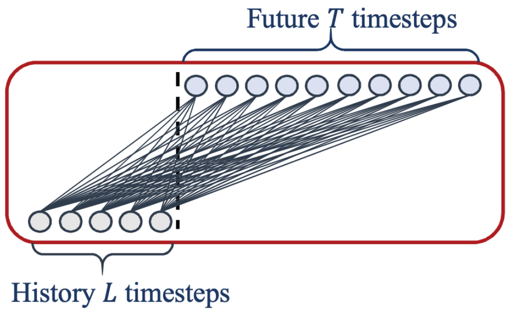
2. Univariate (single channel) vs. multivariate (multi-channels)

3. Short, medium and long-term horizon

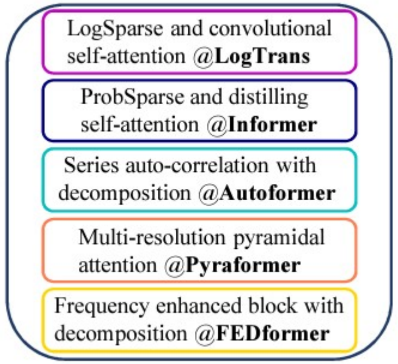
Failure of Transformers in Time Series Forecasting

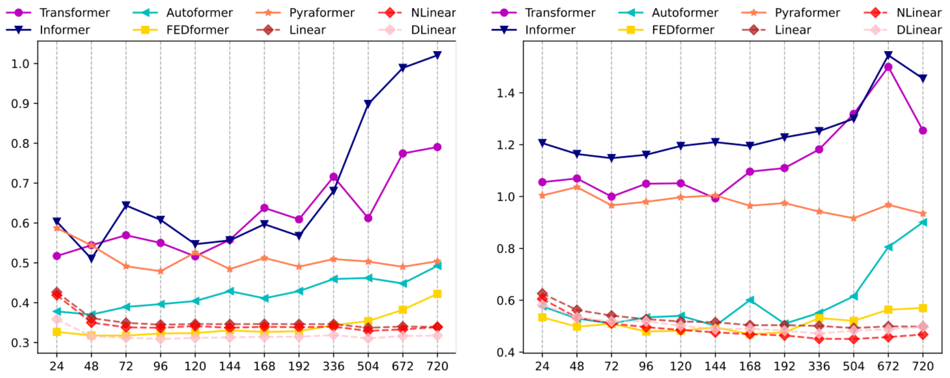
Main conclusions from *Zeng et al., AAAI’23*

1. Transformer-based methods don’t work well in forecasting
2. A Linear model surpasses the SOTA FEDformer (ICML’22) in most cases by 20%~50%



VS.





(a) **24** steps-**ETTh1** (b) **720** steps-**ETTh1**

Yet Transformers dominate in NLP and vision… Why?

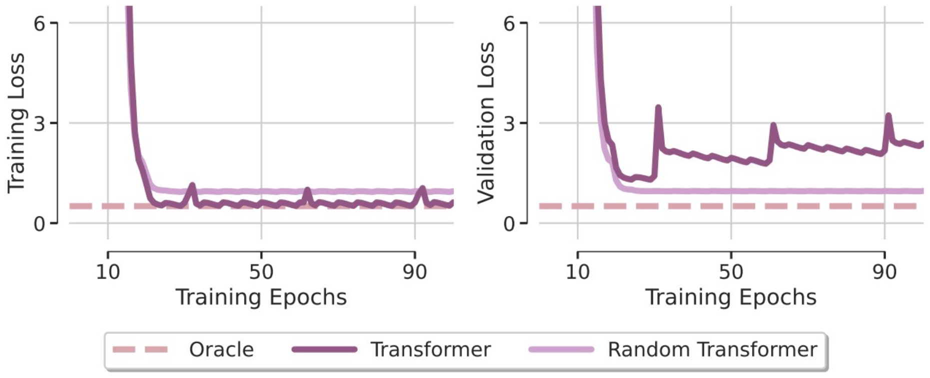
1. Are Transformers Effective for Time Series Forecasting ? Zeng et al. 2023.

SAMformer (Ilbert et al. , ICML Oral 2024)

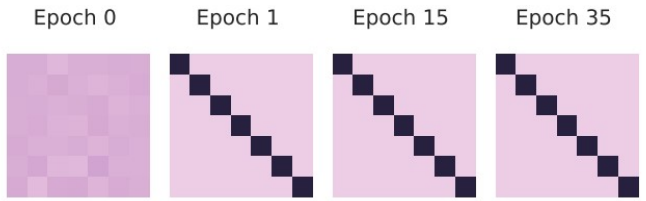
A transformer-based TS forecaster that actually works

Simple Toy Regression Example : Pitfalls

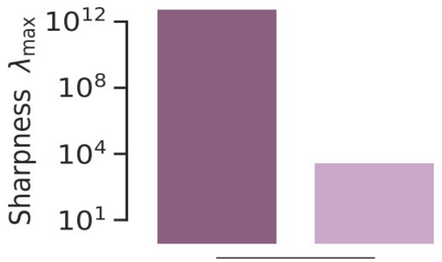
Context : Linear Regression Problem



1. Linear Transformer severely overfits…
2. … and works better if we freeze the attention…
3. … because the attention get stuck at the identity matrix and does not move afterwards



Pathological behavior suggesting sharp local minima

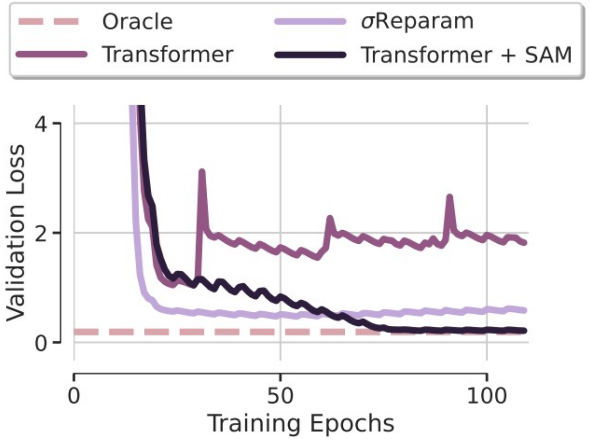


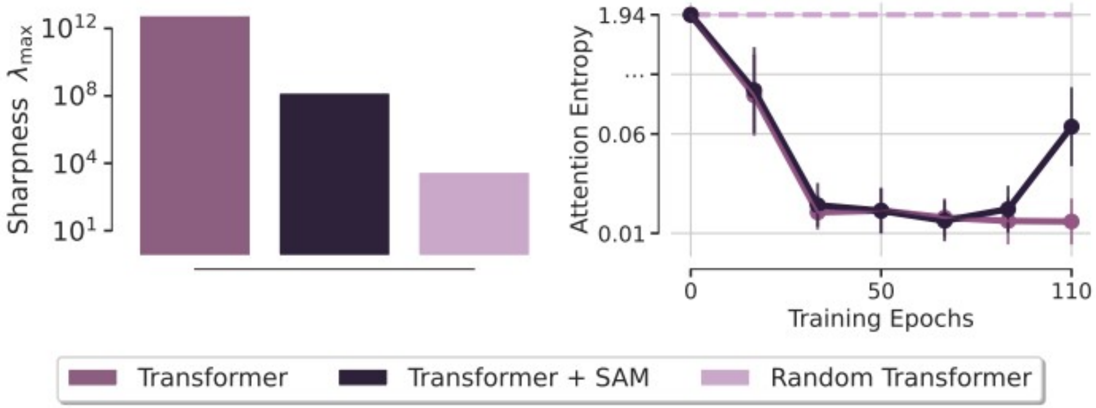
Simple Toy Regression Example : Solution

1. **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

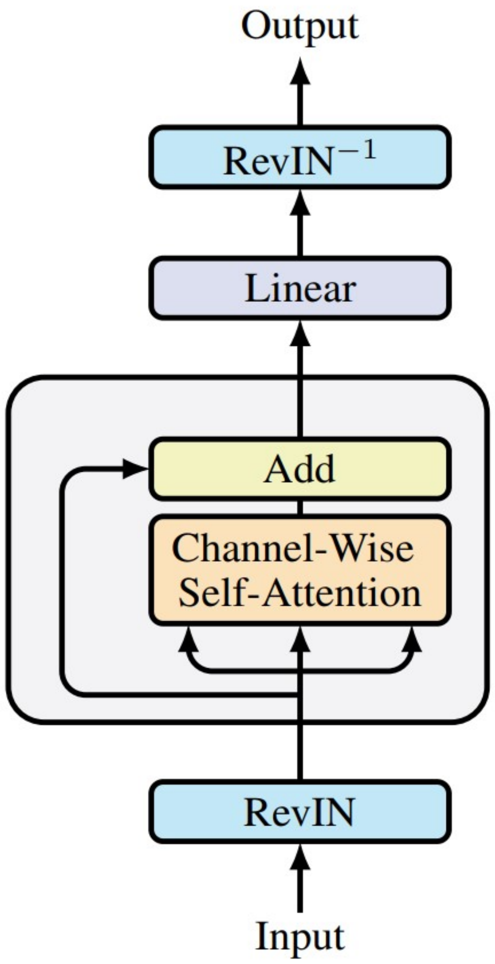
SAMformer : Architecture *(ILBERT et al, ICML 2024)*

Let's put it all together now:

1. **RevIN layer** to be robust to train/test time shift

2. Shallow transformer with a **channel-wise attention**

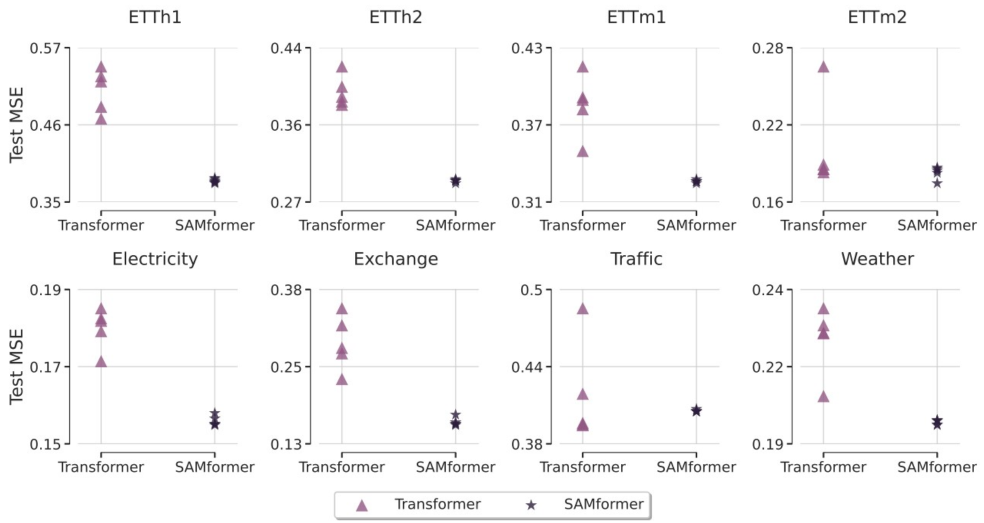
3. We optimize it with **SAM**



SAMformer : Experimental results

	PatchTST	iTransformer	TSMixer	TSMixer + SAM	Informer	FEDformer	Autoformer	Pyraformer	LogTrans
Improvement	11.13%	3.94%	14.33%	5.25%	72.20%	12.36%	22.65%	61.88%	70.88%

1. SAMformer is much better than all transformer-based models...
2. ... more robust to random initialization, smaller and more consistent ...
3. ... and on par with the Foundation Model MOIRAI



Dataset	$H=96$		$H=192$		$H=336$		$H=720$		**Total**
	SAMformer	TSMixer	SAMformer	TSMixer	SAMformer	TSMixer	SAMformer	TSMixer	
ETT	50272	124142	99520	173390	173392	247262	369904	444254	-
Exchange	50272	349344	99520	398592	173392	472464	369904	669456	-
Weather	50272	121908	99520	171156	173392	245028	369904	442020	-
Electricity	50272	280676	99520	329924	173392	403796	369904	600788	-
Traffic	50272	793424	99520	842672	173392	916544	369904	1113536	-
Avg. Ratio	6.64		3.85		2.64		1.77		**3.73**

Dataset	Full-shot	Zero-shot (Woo et al., 2024).		
	SAMformer	$\text{MOIRAI}_{\text{Small}}$	$\text{MOIRAI}_{\text{Base}}$	$\text{MOIRAI}_{\text{Large}}$
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%

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

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

Thank you.

Check code, paper and slides