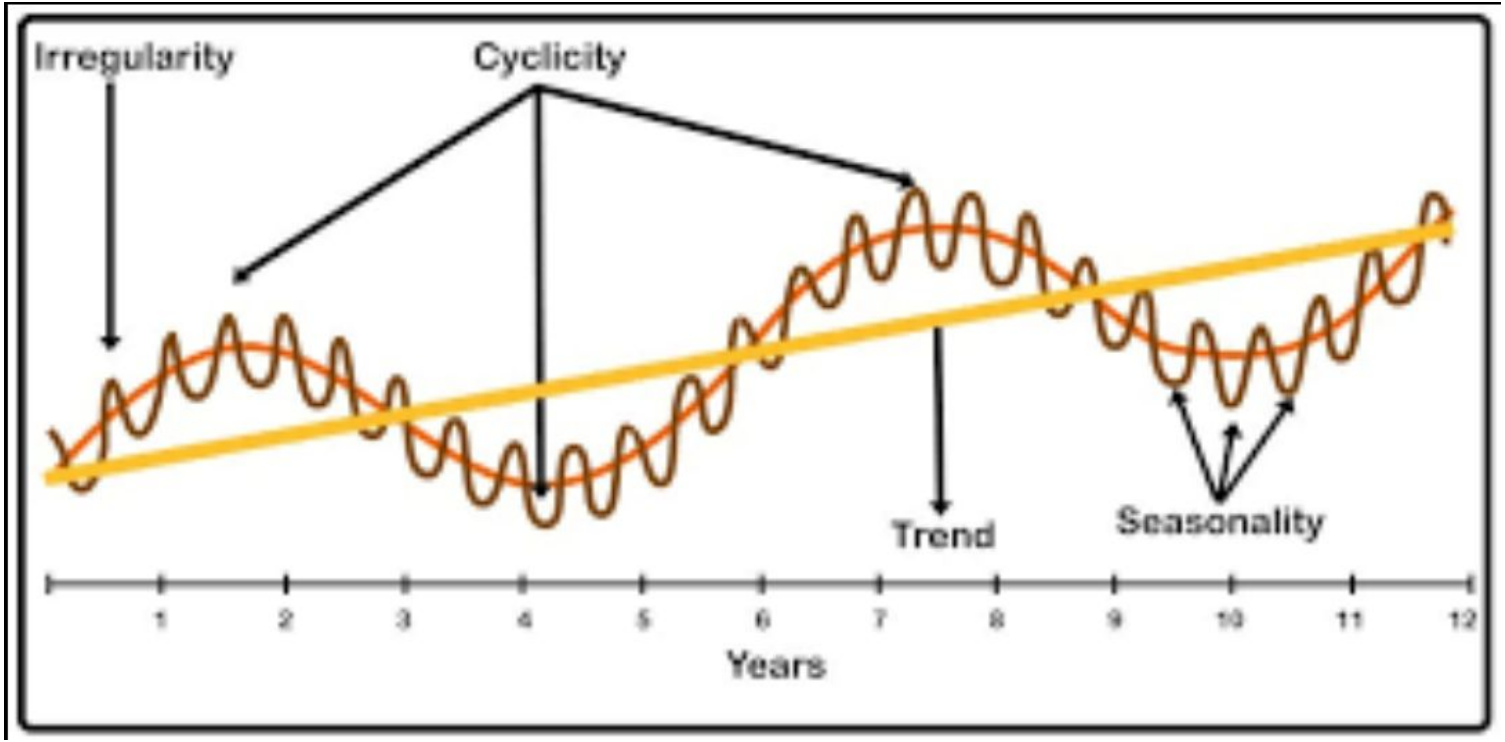
Are Transformers Effective for Time Series Forecasting?

Yusupov Viacheslav

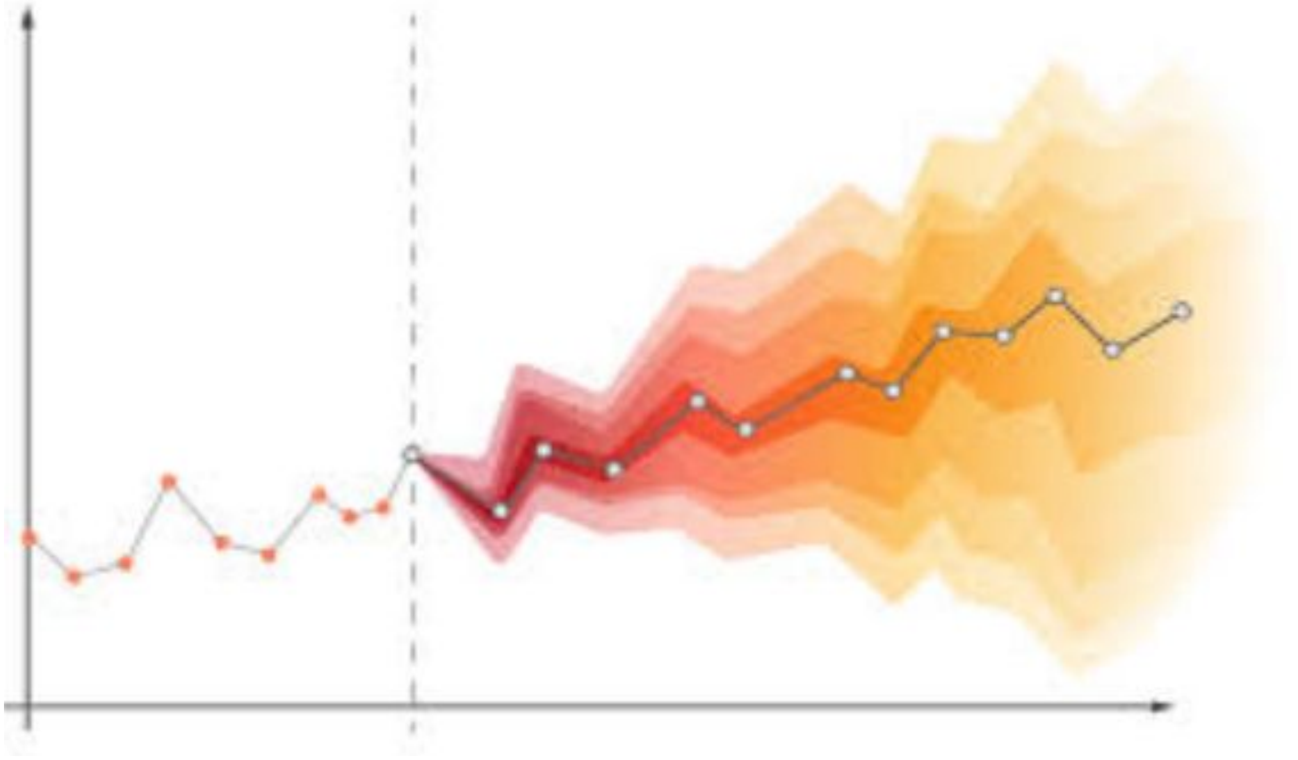
Transformers

- Natural Language Processing (GPT, Bert)
- Computer Vision (ViT, Swin)
- Video (MART)
- Graphs (Graphormer)
- Recommender Systems (SASRec)
- …

Time series


Irregularity
Cyclicity
Trend
Seasonality
1
2
3
4
5
6
7
8
9
10
11
12
Years

Long-term Time Series Forecasting



Iterated Multi-Step (IMS) forecasting

$$y_t = m(\boldsymbol{x}_{t-1};\boldsymbol{\theta}) + e_t, \quad \text{where} \quad \boldsymbol{x}_t = [y_t, \ldots, y_{t-p+1}]'$$

$$\hat{\boldsymbol{\theta}} = \underset{\theta \in \Theta}{\operatorname{argmin}} \sum_t (y_t - m(\boldsymbol{x}_{t-1};\boldsymbol{\theta}))^2,$$

Direct Multi-Step (DMS) forecasting

$$y_t = m_h(y_{t-h}, \ldots, y_{t-h-p_h}; \boldsymbol{\theta}_h) + e_{t,h},$$

$$\hat{\boldsymbol{\theta}}_h = \underset{\boldsymbol{\theta}_h \in \boldsymbol{\Theta}_h}{\operatorname{argmin}} \sum_t \left[y_t - m_h(\boldsymbol{x}_{t-h}; \boldsymbol{\theta}_h) \right]^2 .$$

Transformer models

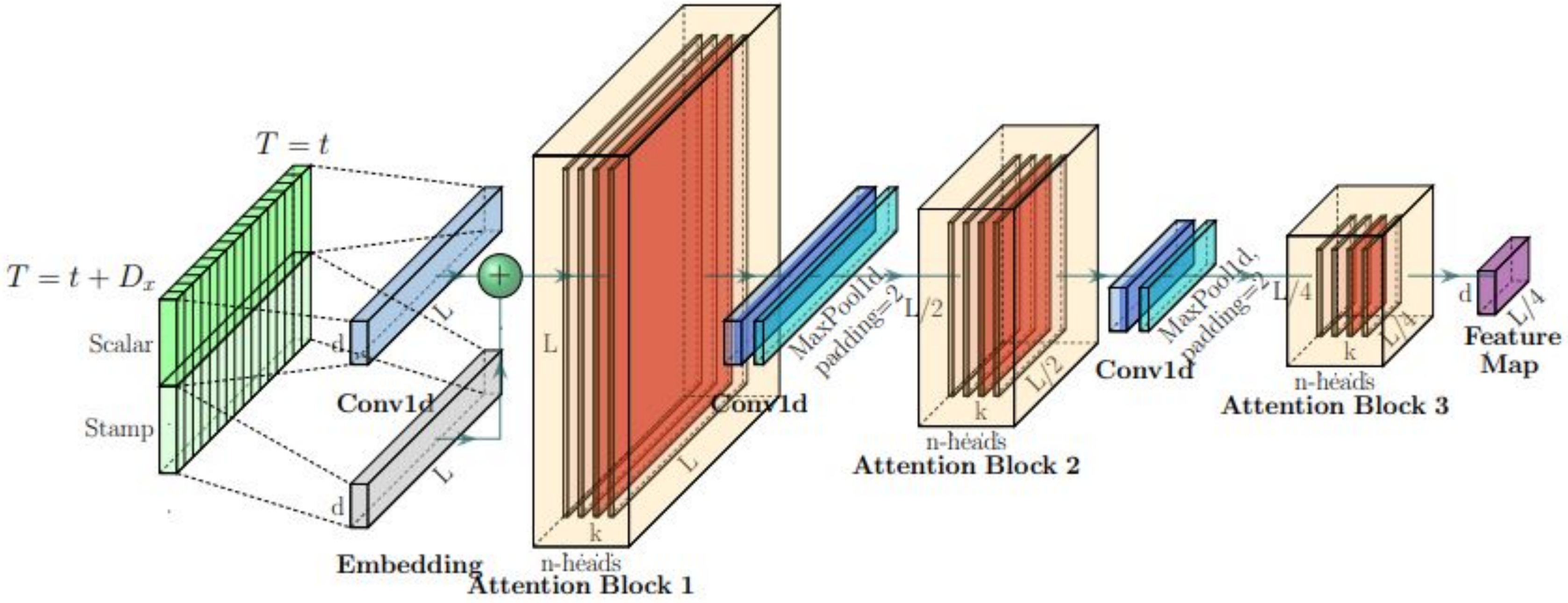
- Informer
- Pyraformer
- Autoformer
- FEDformer
- …

Informer

$$M(\mathbf{q}_i, \mathbf{K}) = \ln \sum_{j=1}^{L_K} e^{\frac{\mathbf{q}_i \mathbf{k}_j^\top}{\sqrt{d}}} - \frac{1}{L_K} \sum_{j=1}^{L_K} \frac{\mathbf{q}_i \mathbf{k}_j^\top}{\sqrt{d}} \quad ,$$

$$\mathcal{A}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{Softmax}(\frac{\overline{\mathbf{Q}}\mathbf{K}^\top}{\sqrt{d}})\mathbf{V} \quad ,$$

Informer


$T = t$
$T = t + D_x$
Scalar
Stamp
d
L
Conv1d
Embedding
L
k
n-heads
Attention Block 1
Conv1d
MaxPool1d
padding=2
L/2
k
n-heads
Attention Block 2
Conv1d
MaxPool1d,
padding=2
L/4
k
n-heads
Attention Block 3
Feature
Map

Pyraformer

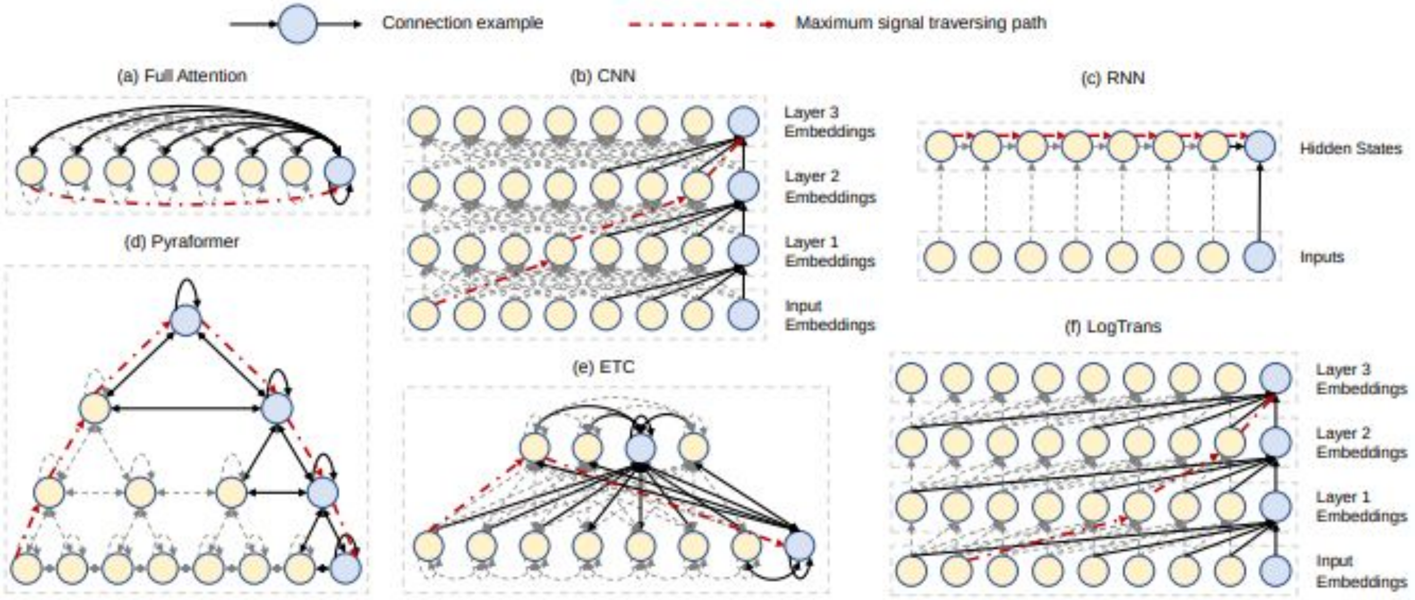


Figure 1: Graphs of commonly used neural network models for sequence data.

Pyraformer

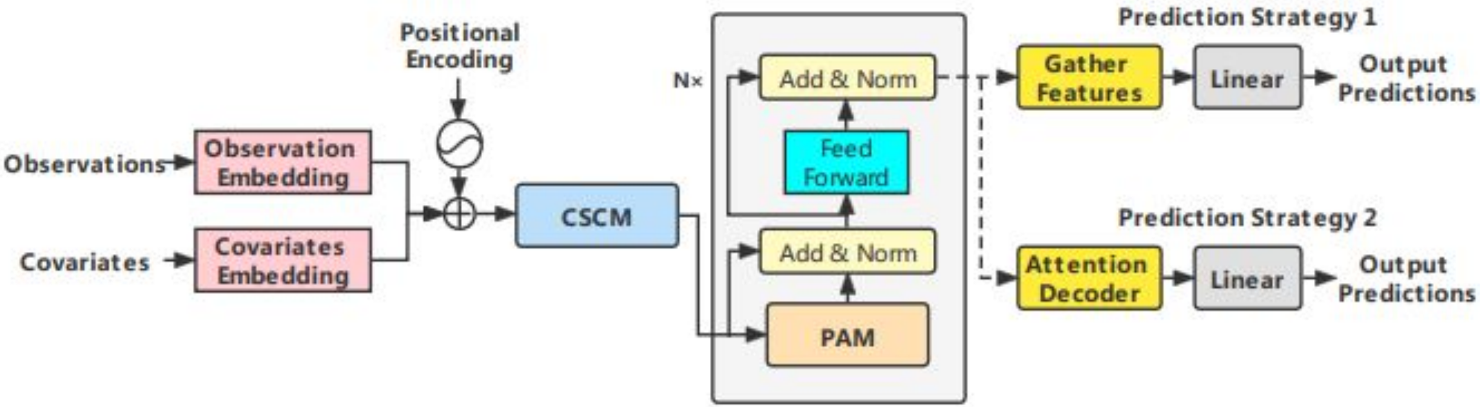
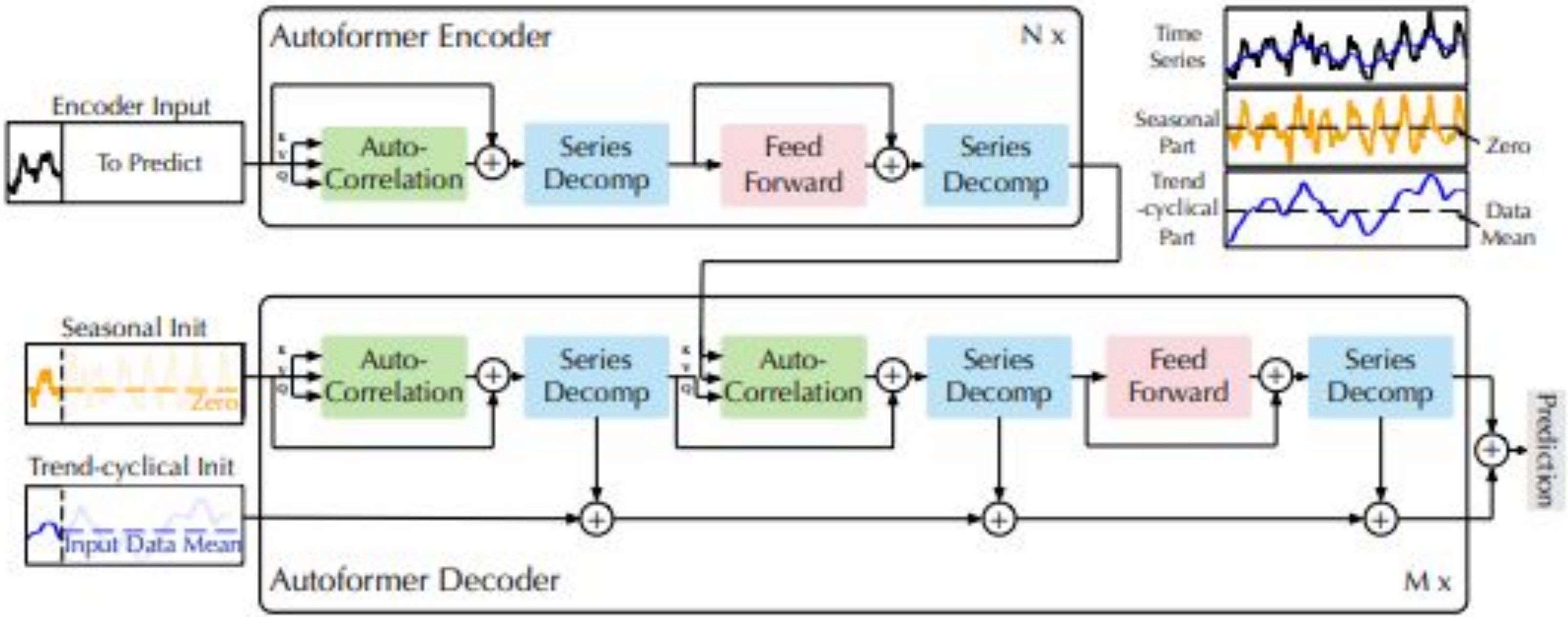


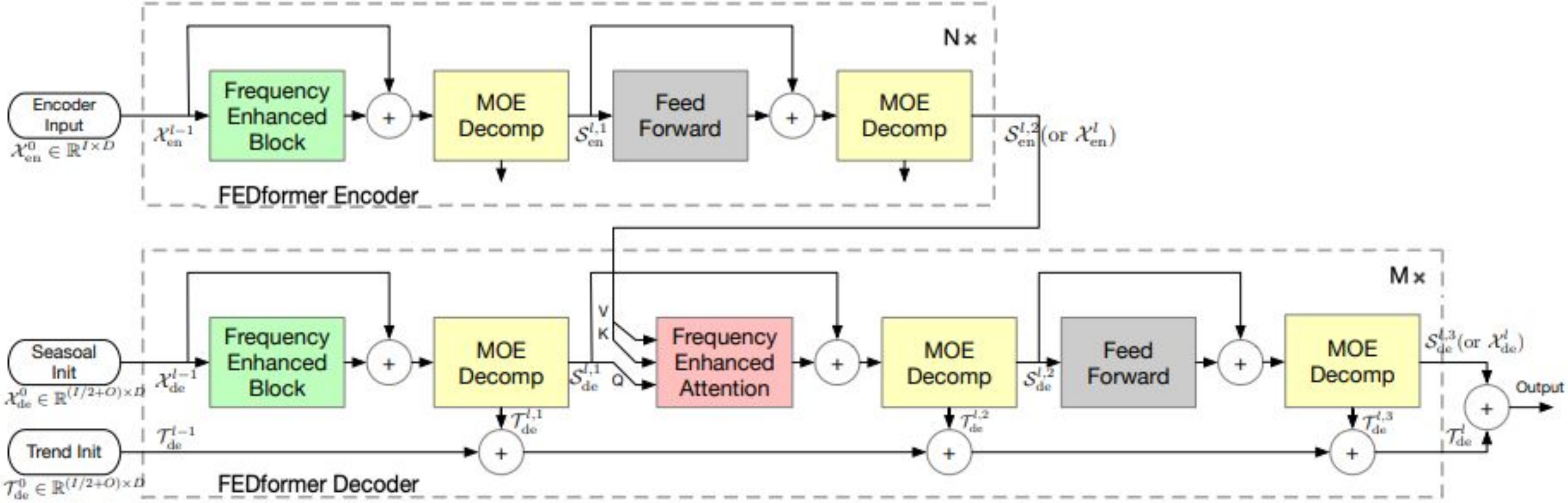
Figure 2: The architecture of Pyraformer: The CSCM summarizes the embedded sequence at different scales and builds a multi-resolution tree structure. Then the PAM is used to exchange information between nodes efficiently.

Autoformer



Autoformer Encoder
N x
Encoder Input
To Predict
Auto-Correlation
Series Decomp
Feed Forward
Series Decomp
Time Series
Seasonal Part
Zero
Trend -cyclical Part
Data Mean
Seasonal Init
Zero
Trend-cyclical Init
Input Data Mean
Auto-Correlation
Series Decomp
Auto-Correlation
Series Decomp
Feed Forward
Series Decomp
Prediction
Autoformer Decoder
M x

FEDformer



Encoder Input
$\mathcal{X}_{en}^{0} \in \mathbb{R}^{I \times D}$
$\mathcal{X}_{en}^{l-1}$
Frequency Enhanced Block
MOE Decomp
$\mathcal{S}_{en}^{l,1}$
Feed Forward
MOE Decomp
$\mathcal{S}_{en}^{l,2}$(or $\mathcal{X}_{en}^{l}$)
N ×
FEDformer Encoder
Seasoal Init
$\mathcal{X}_{de}^{0} \in \mathbb{R}^{(I/2+O) \times D}$
$\mathcal{X}_{de}^{l-1}$
Frequency Enhanced Block
MOE Decomp
$\mathcal{S}_{de}^{l,1}$
V
K
Q
Frequency Enhanced Attention
MOE Decomp
$\mathcal{S}_{de}^{l,2}$
Feed Forward
MOE Decomp
$\mathcal{S}_{de}^{l,3}$(or $\mathcal{X}_{de}^{l}$)
M ×
Output
$\mathcal{T}_{de}^{l,1}$
$\mathcal{T}_{de}^{l,2}$
$\mathcal{T}_{de}^{l,3}$
$\mathcal{T}_{de}^{l}$
Trend Init
$\mathcal{T}_{de}^{l-1}$
$\mathcal{T}_{de}^{0} \in \mathbb{R}^{(I/2+O) \times D}$
FEDformer Decoder

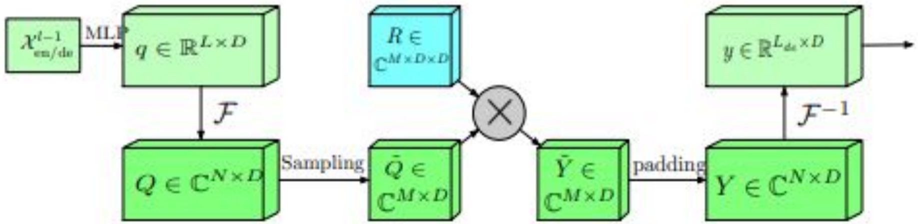


Figure 3. Frequency Enhanced Block with Fourier transform (FEB-f) structure.

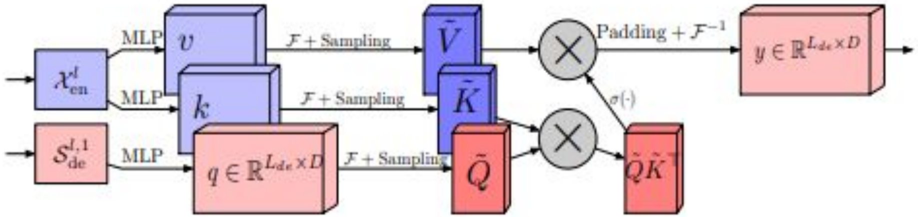
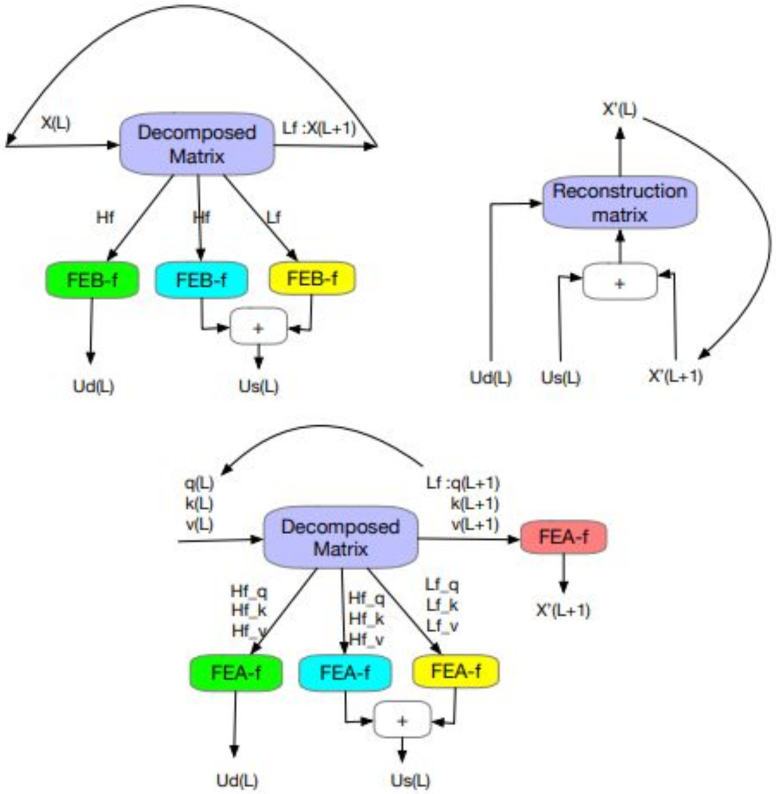


Figure 4. Frequency Enhanced Attention with Fourier transform (FEA-f) structure, $\sigma(\cdot)$ is the activation function.



Are Transformers Effective for Time Series Forecasting?

$$\hat{X}_i = W X_i,$$

$$W \in \mathbb{R}^{T \times L}$$

DLinear

Specifically, *DLinear* is a combination of a *Decomposition* scheme used in Autoformer and FEDformer with linear layers. It first decomposes a raw data input into a trend component by a moving average kernel and a remainder (seasonal) component. Then, two one-layer linear layers are applied to each component, and we sum up the two features to get the final prediction. By explicitly handling trend, *DLinear* enhances the performance of a vanilla linear when there is a clear trend in the data.

NLinear

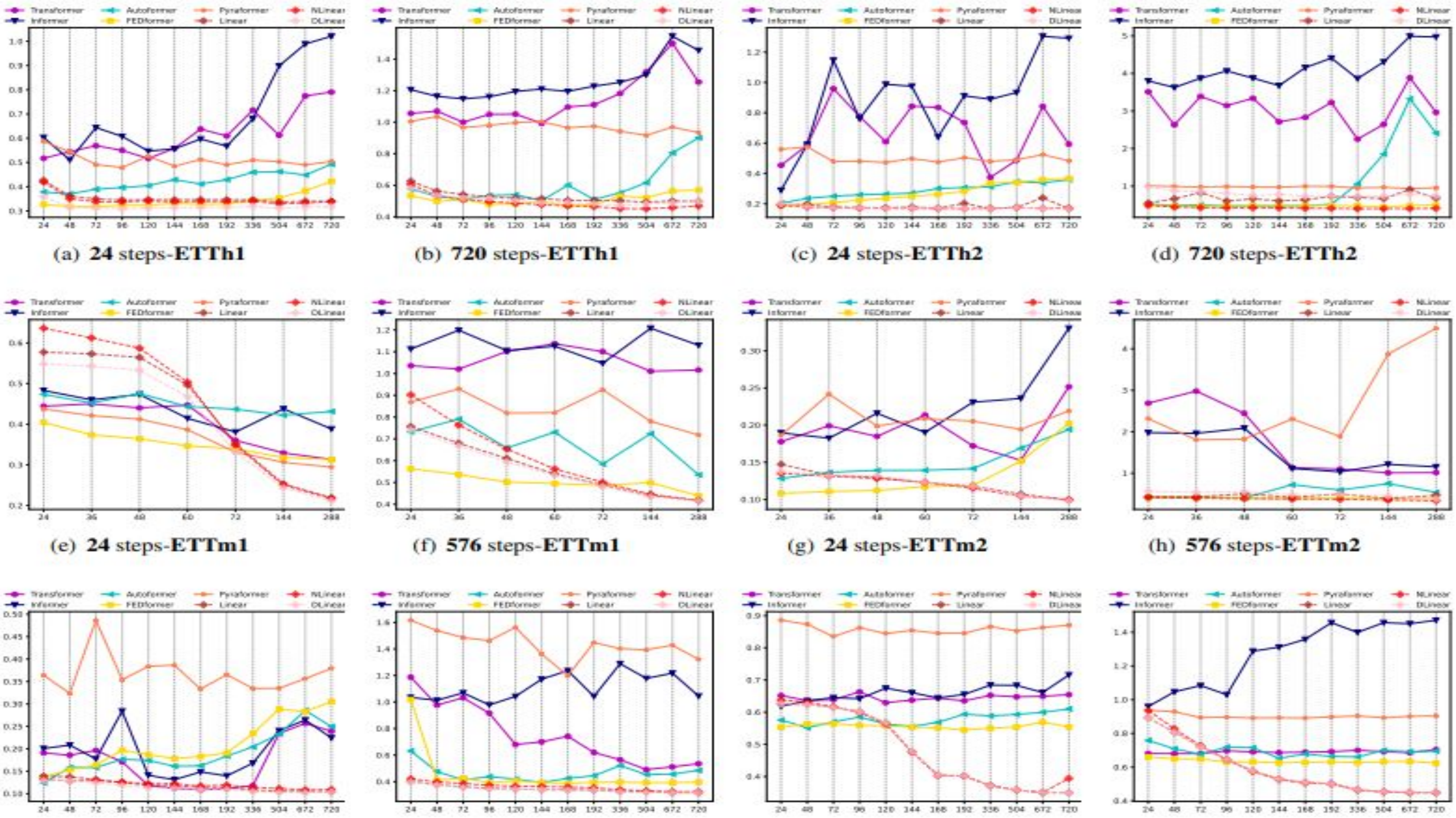
Meanwhile, to boost the performance of *LTSF-Linear* when there is a distribution shift in the dataset, *NLinear* first subtracts the input by the last value of the sequence. Then, the input goes through a linear layer, and the subtracted part is added back before making the final prediction. The subtraction and addition in *NLinear* are a simple normalization for the input sequence.

Models

Method	MACs	Parameter	Time	Memory
DLinear	**0.04G**	**139.7K**	**0.4ms**	**687MiB**
Transformer×	4.03G	13.61M	26.8ms	6091MiB
Informer	3.93G	14.39M	49.3ms	3869MiB
Autoformer	4.41G	14.91M	164.1ms	7607MiB
Pyraformer	0.80G	241.4M*	3.4ms	7017MiB
FEDformer	4.41G	20.68M	40.5ms	4143MiB

Results

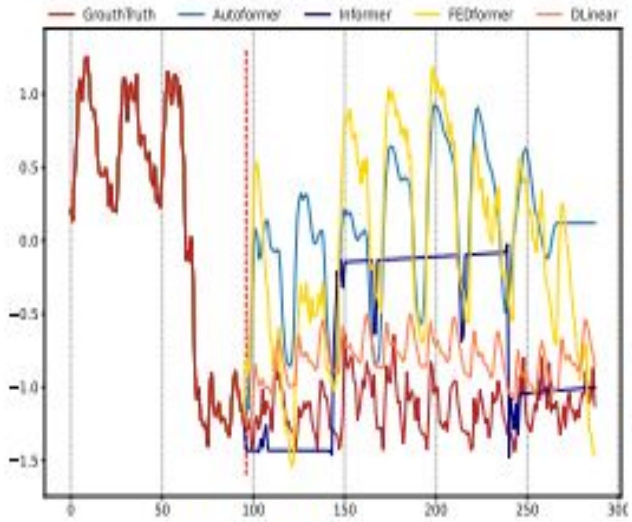
Methods		IMP.	Linear*		NLinear*		DLinear*		FEDformer		Autoformer		Informer		Pyraformer*		LogTrans		*Repeat**	
Metric		MSE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Electricity	96	27.40%	**0.140**	**0.237**	0.141	**0.237**	**0.140**	**0.237**	0.193	0.308	0.201	0.317	0.274	0.368	0.386	0.449	0.258	0.357	1.588	0.946
	192	23.88%	**0.153**	0.250	0.154	**0.248**	**0.153**	0.249	0.201	0.315	0.222	0.334	0.296	0.386	0.386	0.443	0.266	0.368	1.595	0.950
	336	21.02%	**0.169**	0.268	0.171	**0.265**	**0.169**	0.267	0.214	0.329	0.231	0.338	0.300	0.394	0.378	0.443	0.280	0.380	1.617	0.961
	720	17.47%	**0.203**	0.301	0.210	**0.297**	**0.203**	0.301	0.246	0.355	0.254	0.361	0.373	0.439	0.376	0.445	0.283	0.376	1.647	0.975
Exchange	96	45.27%	0.082	0.207	0.089	0.208	**0.081**	0.203	0.148	0.278	0.197	0.323	0.847	0.752	0.376	1.105	0.968	0.812	**0.081**	**0.196**
	192	42.06%	0.167	0.304	0.180	0.300	**0.157**	0.293	0.271	0.380	0.300	0.369	1.204	0.895	1.748	1.151	1.040	0.851	0.167	**0.289**
	336	33.69%	0.328	0.432	0.331	0.415	**0.305**	0.414	0.460	0.500	0.509	0.524	1.672	1.036	1.874	1.172	1.659	1.081	**0.305**	**0.396**
	720	46.19%	0.964	0.750	1.033	0.780	**0.643**	**0.601**	1.195	0.841	1.447	0.941	2.478	1.310	1.943	1.206	1.941	1.127	0.823	0.681
Traffic	96	30.15%	**0.410**	0.282	**0.410**	**0.279**	**0.410**	0.282	0.587	0.366	0.613	0.388	0.719	0.391	2.085	0.468	0.684	0.384	2.723	1.079
	192	29.96%	**0.423**	0.287	**0.423**	**0.284**	**0.423**	0.287	0.604	0.373	0.616	0.382	0.696	0.379	0.867	0.467	0.685	0.390	2.756	1.087
	336	29.95%	0.436	0.295	**0.435**	**0.290**	0.436	0.296	0.621	0.383	0.622	0.337	0.777	0.420	0.869	0.469	0.734	0.408	2.791	1.095
	720	25.87%	0.466	0.315	**0.464**	**0.307**	0.466	0.315	0.626	0.382	0.660	0.408	0.864	0.472	0.881	0.473	0.717	0.396	2.811	1.097
Weather	96	18.89%	**0.176**	0.236	0.182	**0.232**	**0.176**	0.237	0.217	0.296	0.266	0.336	0.300	0.384	0.896	0.556	0.458	0.490	0.259	0.254
	192	21.01%	**0.218**	0.276	0.225	**0.269**	0.220	0.282	0.276	0.336	0.307	0.367	0.598	0.544	0.622	0.624	0.658	0.589	0.309	0.292
	336	22.71%	**0.262**	0.312	0.271	**0.301**	0.265	0.319	0.339	0.380	0.359	0.395	0.578	0.523	0.739	0.753	0.797	0.652	0.377	0.338
	720	19.85%	0.326	0.365	0.338	**0.348**	**0.323**	0.362	0.403	0.428	0.419	0.428	1.059	0.741	1.004	0.934	0.869	0.675	0.465	0.394
ILI	24	47.86%	1.947	0.985	**1.683**	**0.858**	2.215	1.081	3.228	1.260	3.483	1.287	5.764	1.677	1.420	2.012	4.480	1.444	6.587	1.701
	36	36.43%	2.182	1.036	**1.703**	**0.859**	1.963	0.963	2.679	1.080	3.103	1.148	4.755	1.467	7.394	2.031	4.799	1.467	7.130	1.884
	48	34.43%	2.256	1.060	**1.719**	**0.884**	2.130	1.024	2.622	1.078	2.669	1.085	4.763	1.469	7.551	2.057	4.800	1.468	6.575	1.798
	60	34.33%	2.390	1.104	**1.819**	**0.917**	2.368	1.096	2.857	1.157	2.770	1.125	5.264	1.564	7.662	2.100	5.278	1.560	5.893	1.677
ETTh1	96	0.80%	0.375	0.397	**0.374**	**0.394**	0.375	0.399	0.376	0.419	0.449	0.459	0.865	0.713	0.664	0.612	0.878	0.740	1.295	0.713
	192	3.57%	0.418	0.429	0.408	**0.415**	**0.405**	0.416	0.420	0.448	0.500	0.482	1.008	0.792	0.790	0.681	1.037	0.824	1.325	0.733
	336	6.54%	0.479	0.476	**0.429**	**0.427**	0.439	0.443	0.459	0.465	0.521	0.496	1.107	0.809	0.891	0.738	1.238	0.932	1.323	0.744
	720	13.04%	0.624	0.592	**0.440**	**0.453**	0.472	0.490	0.506	0.507	0.514	0.512	1.181	0.865	0.963	0.782	1.135	0.852	1.339	0.756
ETTh2	96	19.94%	0.288	0.352	**0.277**	**0.338**	0.289	0.353	0.346	0.388	0.358	0.397	3.755	1.525	0.645	0.597	2.116	1.197	0.432	0.422
	192	19.81%	0.377	0.413	**0.344**	**0.381**	0.383	0.418	0.429	0.439	0.456	0.452	5.602	1.931	0.788	0.683	4.315	1.635	0.534	0.473
	336	25.93%	0.452	0.461	**0.357**	**0.400**	0.448	0.465	0.496	0.487	0.482	0.486	4.721	1.835	0.907	0.747	1.124	1.604	0.591	0.508
	720	14.25%	0.698	0.595	**0.394**	**0.436**	0.605	0.551	0.463	0.474	0.515	0.511	3.647	1.625	0.963	0.783	3.188	1.540	0.588	0.517
ETTm1	96	21.10%	0.308	0.352	0.306	0.348	**0.299**	**0.343**	0.379	0.419	0.505	0.475	0.672	0.571	0.543	0.510	0.600	0.546	1.214	0.665
	192	21.36%	0.340	0.369	0.349	0.375	**0.335**	**0.365**	0.426	0.441	0.553	0.496	0.795	0.669	0.557	0.537	0.837	0.700	1.261	0.690
	336	17.07%	0.376	0.393	0.375	0.388	**0.369**	**0.386**	0.445	0.459	0.621	0.537	1.212	0.871	0.754	0.655	1.124	0.832	1.283	0.707
	720	21.73%	0.440	0.435	0.433	0.422	**0.425**	**0.421**	0.543	0.490	0.671	0.561	1.166	0.823	0.908	0.724	1.153	0.820	1.319	0.729
ETTm2	96	17.73%	0.168	0.262	**0.167**	**0.255**	**0.167**	0.260	0.203	0.287	0.255	0.339	0.365	0.453	0.435	0.507	0.768	0.642	0.266	0.328
	192	17.84%	0.232	0.308	**0.221**	**0.293**	0.224	0.303	0.269	0.328	0.281	0.340	0.533	0.563	0.730	0.673	0.989	0.757	0.340	0.371
	336	15.69%	0.320	0.373	**0.274**	**0.327**	0.281	0.342	0.325	0.366	0.339	0.372	1.363	0.887	1.201	0.845	1.334	0.872	0.412	0.410
	720	12.58%	0.413	0.435	**0.368**	**0.384**	0.397	0.421	0.421	0.415	0.433	0.432	3.379	1.338	3.625	1.451	3.048	1.328	0.521	0.465



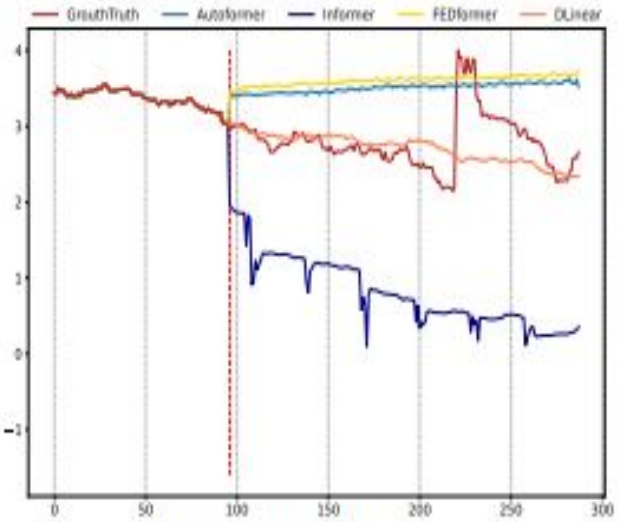
(a) **24** steps-**ETTh1** (b) **720** steps-**ETTh1** (c) **24** steps-**ETTh2** (d) **720** steps-**ETTh2**

(e) **24** steps-**ETTm1** (f) **576** steps-**ETTm1** (g) **24** steps-**ETTm2** (h) **576** steps-**ETTm2**

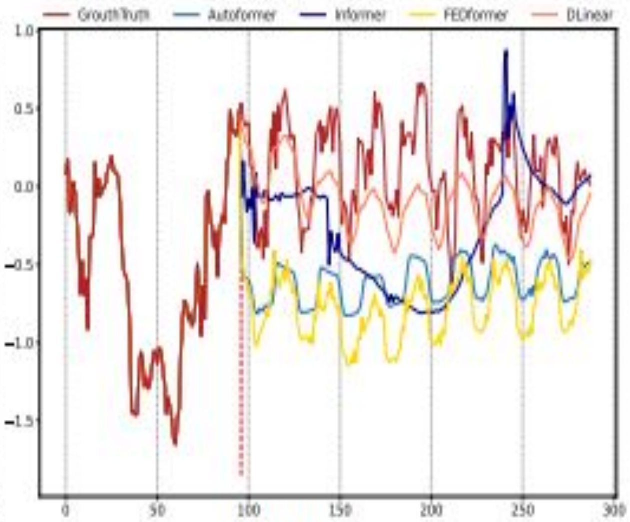
(i) **24** steps-**Weather** (j) **720** steps-**Weather** (k) **24** steps-**Traffic** (l) **720** steps-**Traffic**



(a) Electricity



(b) Exchange-Rate



(c) ETTh2

References

- Souhaib Ben Taieb, Rob J Hyndman, et al. Recursive and direct multi-step forecasting: the best of both worlds, volume 19. Citeseer, 2012.
- Shizhan Liu, Hang Yu, Cong Liao, Jianguo Li, Weiyao Lin, Alex X Liu, and Schahram Dustdar. Pyraformer: Low-complexity pyramidal attention for long-range time series modeling and forecasting. In International Conference on Learning Representations, 2021.
- Jiehui Xu, Jianmin Wang, Mingsheng Long, et al. Autoformer: Decomposition transformers with autocorrelation for long-term series forecasting. Advances in Neural Information Processing Systems, 34, 2021.

References

- Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In The Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Virtual Conference, volume 35, pages 11106– 11115. AAAI Press, 2021.
- Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. Fedformer: Frequency enhanced decomposed transformer for long-term series 14 forecasting. In International Conference on Machine Learning, 2022.