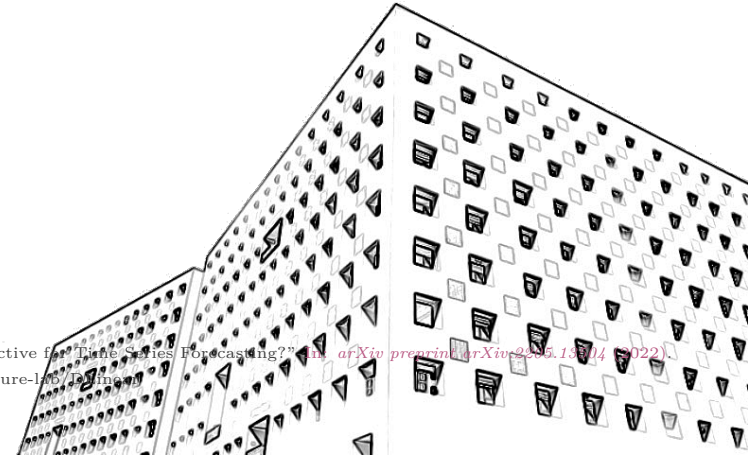
Are Transformers Effective for Time Series Forecasting?[1][2]

Paper by: Ailing Zeng, Muxi Chen, Lei Zhang, Qiang Xu

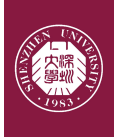
Presentation by: Tobramycin

制作于 2022 年 6 月 27 日



[1] Ailing Zeng et al. "Are Transformers Effective for Time Series Forecasting?" In: *arXiv preprint arXiv:2205.13504* (2022).

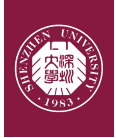
[2] Code is available at https://github.com/cure-lab/LTSF-Linear



本模版基于 Federico Zenith 发布的 SINTEF Presentation 模版二次修改制作而成

后文为 Federico Zenith 为模版提供的简明教程，版权归其所有

目录

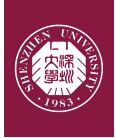


Frame Title

1 Introduction

- ”Transformer architecture relies on self-attention mechanisms to effectively extract the semantic correlations between paired elements in a long sequence, which is permutation-invariant and ’anti-ordering’ to some extent.”
- ”an embarrassingly simple architecture named DLinear ... outperforms existing complex Transformer-based models in most cases by a large margin.”

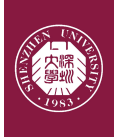
目录



Problem Formulation

2 Preliminaries

Given historical data, $\mathcal{X} = \{X_1^t, \ldots, X_C^t\}_{t=1}^{L}$, C is variates size, wherein L is the look-back window size and X_i^t is the value of the i_{th} variate at the t_{th} time step. The time series forecasting task is to predict the values $\hat{\mathcal{X}} = \left\{\hat{X}_1^t, \ldots, \hat{X}_C^t\right\}_{t=L+1}^{L+T}$ at the T future time steps.



Concepts

- Time Series Forecasting (TSF)
- Iterated Multi-Step (IMS) forecasting
  - Smaller variance
  - Error accumulation effects
  - Suitable when forecasting time step T is relatively small
- Direct Multi-Step (DMS) forecasting
  - Suitable when T is large

Time Series Decomposition

DLinear and Autoformer apply a same decomposition scheme call STL (A seasonal-trend decomposition based on Loess)[3]

- Data = trend component + seasonal component + remainder component. $v = 1$ to N

$$Y_v = T_v + S_v + R_v \tag{1}$$

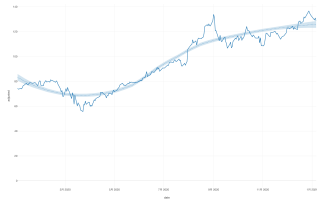


Figure: Loess regression

[3] Robert B Cleveland et al. "STL: A seasonal-trend decomposition". In: *J. Off. Stat* 6.1 (1990), pp. 3–73.

Transfomer-Like Model

2 Preliminaries

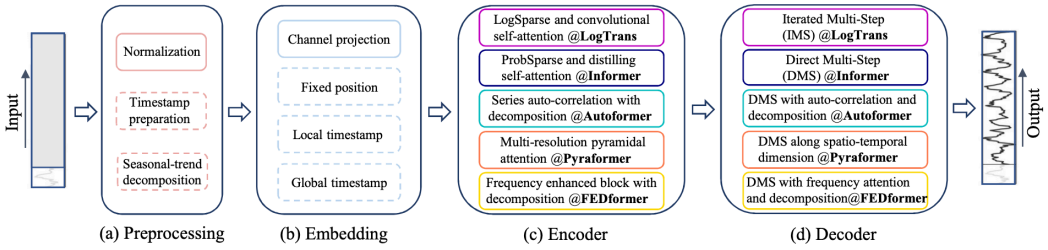


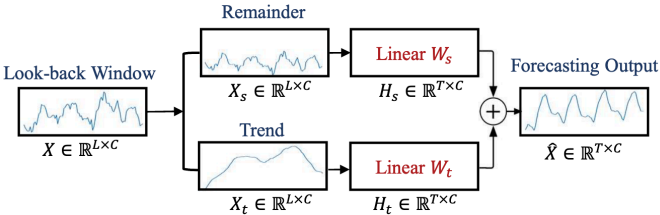
Figure 1: The pipeline of existing Transformer-based TSF solutions. In (a) and (b), the solid boxes are essential operations and the dotted boxes are applied optionally. (c) and (d) are distinct for different methods [16, 28, 27, 18, 29].

目录

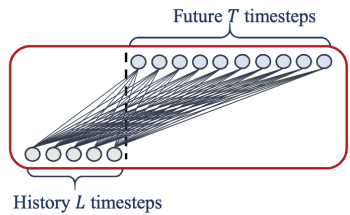
3 DLinear

Framework

3 DLinear



(a) The whole structure of *DLinear*



(b) One Linear Layer

Figure 2: Illustration of the Decomposition Linear Model.

目录

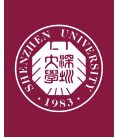
4 Experiments

Datasets

4 Experiments

- ETT (Electricity Transformer Temperature) [28][2] consists of two hourly-level datasets (ETTh) and two 15-minute-level datasets (ETTm). Each of them contains seven oil and load features of electricity transformers from July 2016 to July 2018.
- Traffic[3] describes the road occupancy rates. It contains the hourly data recorded by the sensors of San Francisco freeways from 2015 to 2016.
- Electricity[4] collects the hourly electricity consumption of 321 clients from 2012 to 2014.
- Exchange-Rate [15][5] collects the daily exchange rates of 8 countries from 1990 to 2016.
- Weather[6] includes 21 indicators of weather, such as air temperature, and humidity. Its data is recorded every 10 min for 2020 in Germany.
- ILI[7] describes the ratio of patients seen with influenza-like illness and the total number of the patients. It includes the weekly data from the Centers for Disease Control and Prevention of the United States from 2002 to 2021.

[2] https://github.com/zhouhaoyi/ETDataset
[3] http://pems.dot.ca.gov
[4] https://archive.ics.uci.edu/ml/datasets/ElectricityLoadDiagrams20112014
[5] https://github.com/laiguokun/multivariate-time-series-data
[6] https://www.bgc-jena.mpg.de/wetter/
[7] https://gis.cdc.gov/grasp/fluview/fluportaldashboard.html



Datasets

4 Experiments

Table 1: The statistics of the nine benchmark datasets.

Datasets	ETTh1&ETTh2	ETTm1 &ETTm2	Traffic	Electricity	Exchange-Rate	Weather	ILI
Variates	7	7	862	321	8	21	7
Timesteps	17,420	69,680	17,544	26,304	7,588	52,696	966
Granularity	1hour	5min	1hour	1hour	1day	10min	1week

Long-term forecasting errors

4 Experiments

Table 2: Long-term forecasting errors in terms of MSE and MAE, the lower the better. Among them, four datasets are with look-back window size $L = 96$ and forecasting horizon $T \in \{96, 192, 336, 720\}$. For the ILI dataset, $L = 36$ and $T \in \{24, 36, 48, 60\}$. Repeat-C repeats the last value of the look-back window. The best results are highlighted in **red bold** and the second best results are highlighted with a blue underline.

Methods	Metric	Electricity				Exchange-Rate				Traffic				Weather				ILI			
		96	192	336	720	96	192	336	720	96	192	336	720	96	192	336	720	24	36	48	60
DLinear-S*	MSE	0.194	0.193	0.206	0.242	**0.078**	0.159	0.274	**0.558**	0.650	**0.598**	**0.605**	0.645	0.196	0.237	0.283	0.345	**2.398**	**2.646**	2.614	2.804
	MAE	0.276	0.280	0.296	0.329	0.197	0.292	0.391	**0.574**	0.396	**0.370**	0.373	0.394	0.255	0.296	0.335	0.381	**1.040**	1.088	1.086	1.146
DLinear-I*	MSE	**0.184**	**0.184**	**0.197**	**0.234**	0.084	**0.157**	**0.236**	0.626	0.647	0.602	0.607	0.646	**0.164**	**0.209**	**0.263**	**0.338**	3.015	2.737	**2.577**	2.821
	MAE	**0.270**	**0.273**	**0.289**	**0.323**	0.216	0.298	**0.379**	0.634	0.403	0.375	0.377	0.398	**0.237**	**0.282**	**0.327**	**0.380**	1.192	**1.036**	1.043	**1.091**
FEDformer	MSE	0.193	0.201	0.214	0.246	0.148	0.271	0.460	1.195	**0.587**	0.604	0.621	**0.626**	0.217	0.276	0.339	0.403	3.228	2.679	2.622	2.857
	MAE	0.308	0.315	0.329	0.355	0.278	0.380	0.500	0.841	**0.366**	0.373	0.383	**0.382**	0.296	0.336	0.380	0.428	1.260	1.080	1.078	1.157
Autoformer	MSE	0.201	0.222	0.231	0.254	0.197	0.300	0.509	1.447	0.613	0.616	0.622	0.660	0.266	0.307	0.359	0.419	3.483	3.103	2.669	**2.770**
	MAE	0.317	0.334	0.338	0.361	0.323	0.369	0.524	0.941	0.388	0.382	**0.337**	0.408	0.336	0.367	0.395	0.428	1.287	1.148	1.085	1.125
Informer	MSE	0.274	0.296	0.300	0.373	0.847	1.204	1.672	2.478	0.719	0.696	0.777	0.864	0.300	0.598	0.578	1.059	5.764	4.755	4.763	5.264
	MAE	0.368	0.386	0.394	0.439	0.752	0.895	1.036	1.310	0.391	0.379	0.420	0.472	0.384	0.544	0.523	0.741	1.677	1.467	1.469	1.564
Pyraformer*	MSE	0.386	0.378	0.376	0.376	1.748	1.874	1.943	2.085	0.867	0.869	0.881	0.896	0.622	0.739	1.004	1.420	7.394	7.551	7.662	7.931
	MAE	0.449	0.443	0.443	0.445	1.105	1.151	1.172	1.206	0.468	0.467	0.469	0.473	0.556	0.624	0.753	0.934	2.012	2.031	2.057	2.100
LogTrans	MSE	0.258	0.266	0.280	0.283	0.968	1.040	1.659	1.941	0.684	0.685	0.734	0.717	0.458	0.658	0.797	0.869	4.480	4.799	4.800	5.278
	MAE	0.357	0.368	0.380	0.376	0.812	0.851	1.081	1.127	0.384	0.390	0.408	0.396	0.490	0.589	0.652	0.675	1.444	1.467	1.468	1.560
Reformer	MSE	0.312	0.348	0.350	0.340	1.065	1.188	1.357	1.510	0.732	0.733	0.742	0.755	0.689	0.752	0.639	1.130	4.400	4.783	4.832	4.882
	MAE	0.402	0.433	0.433	0.420	0.829	0.906	0.976	1.016	0.423	0.420	0.420	0.423	0.596	0.638	0.596	0.792	1.382	1.448	1.465	1.483
Repeat-C*	MSE	1.588	1.595	1.617	1.647	0.081	0.167	0.305	0.823	2.723	2.756	2.791	2.811	0.259	0.309	0.377	0.465	6.587	7.130	6.575	5.893
	MAE	0.946	0.950	0.961	0.975	**0.196**	**0.289**	0.396	0.681	1.079	1.087	1.095	1.097	0.254	0.292	0.338	0.394	1.701	1.884	1.798	1.677

- Methods∗ are implemented by us; Other results are from FEDformer [29].

Long-term forecasting errors

4 Experiments

Methods		DLinear-S		DLinear-I		FEDformer-f		FEDformer-w		Autoformer	
Metric		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	96	**0.375**	0.399	0.377	**0.397**	0.376	0.419	0.395	0.424	0.449	0.459
	192	**0.405**	**0.416**	0.413	0.421	0.420	0.448	0.469	0.470	0.500	0.482
	336	**0.439**	0.443	0.440	**0.439**	0.459	0.465	0.530	0.499	0.521	0.496
	720	**0.472**	0.490	0.476	**0.481**	0.506	0.507	0.598	0.544	0.514	0.512
ETTh2	96	**0.289**	**0.353**	0.438	0.451	0.346	0.388	0.394	0.414	0.358	0.397
	192	**0.383**	**0.418**	0.615	0.517	0.429	0.439	0.439	0.445	0.456	0.452
	336	**0.448**	**0.465**	0.603	0.525	0.496	0.487	0.482	0.480	0.482	0.486
	720	0.605	0.551	1.082	0.723	**0.463**	**0.474**	0.500	0.509	0.515	0.511
ETTm1	96	0.299	0.343	**0.286**	**0.334**	0.379	0.419	0.378	0.418	0.505	0.475
	192	0.335	0.365	**0.327**	**0.358**	0.426	0.441	0.464	0.463	0.553	0.496
	336	0.369	0.386	**0.367**	**0.383**	0.445	0.459	0.508	0.487	0.621	0.537
	720	**0.425**	0.421	0.429	**0.418**	0.543	0.490	0.561	0.515	0.671	0.561
ETTm2	96	**0.167**	**0.260**	0.195	0.288	0.203	0.287	0.204	0.288	0.255	0.339
	192	**0.224**	**0.303**	0.332	0.367	0.269	0.328	0.316	0.363	0.281	0.340
	336	**0.281**	**0.342**	0.545	0.476	0.325	0.366	0.359	0.387	0.339	0.372
	720	**0.397**	0.421	0.697	0.546	0.421	**0.415**	0.433	0.432	0.422	0.419
Eletricity	96	0.140	0.237	**0.133**	**0.230**	0.193	0.308	0.183	0.297	0.201	0.317
	192	0.153	0.249	**0.148**	**0.245**	0.201	0.315	0.195	0.308	0.222	0.334
	336	0.169	0.267	**0.164**	**0.263**	0.214	0.329	0.212	0.313	0.231	0.338
	720	0.203	0.301	**0.201**	**0.297**	0.246	0.355	0.231	0.343	0.254	0.361

Qualitative results

4 Experiments

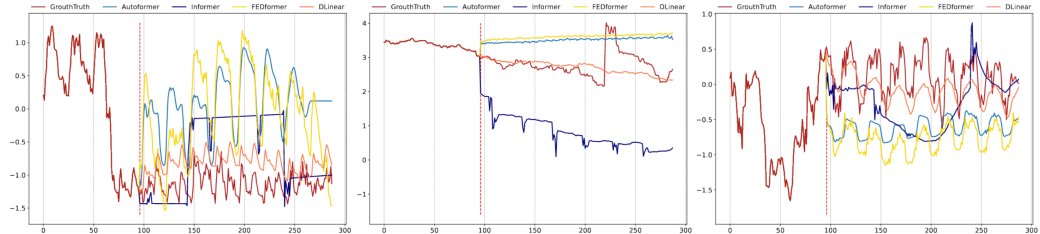
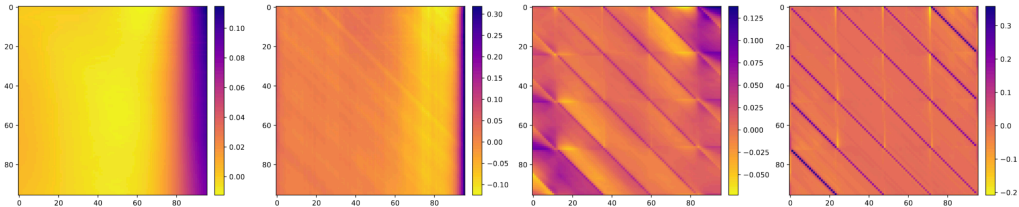


Figure 3: Illustration of the long-term forecasting output (Y-axis) of five models with an input length L=96 and output length T=192 (X-axis) on Electricity, Exchange-Rate, and ETTh2, respectively.

The interpretability of DLinear

4 Experiments



(a) Trend-Exchange (b) Remainder-Exchange (c) Trend-Electricty (d) Remainder-Electricty

Figure 4: Visualization of the weights(T*L) of *DLinear*. (a) and (c) are weights in the remainder layer. (b) and (d) are weights in the trend layer. Models are trained with a look-back window L=96 (X-axis) and a forecasting length T=96 (Y-axis) on the Exchange-Rate and Electricity datasets, respectively.

The impact of different embedding strategies

4 Experiments

Table 6: The impact of different embedding strategies on Transformer-based methods with look-back window size $L = 96$ and forecasting horizon $T \in \{96, 192, 336, 720\}$. The metric used is MSE.

Methods	Embedding	Electricity				Traffic			
		96	192	336	720	96	192	336	720
FEDformer	All	**0.189**	**0.198**	**0.210**	0.248	0.597	0.606	0.627	0.649
	wo/Pos.	0.193	0.201	0.214	**0.246**	**0.587**	**0.604**	**0.621**	**0.626**
	wo/Temp.	0.209	0.213	0.222	0.265	0.613	0.623	0.650	0.677
	wo/Pos.-Temp.	0.203	0.211	0.222	0.250	0.613	0.622	0.648	0.663
Autoformer	All	**0.193**	0.227	0.252	0.252	0.629	0.647	0.676	**0.638**
	wo/Pos.	**0.193**	**0.201**	**0.214**	**0.246**	**0.613**	**0.616**	**0.622**	0.660
	wo/Temp.	0.206	0.243	0.303	0.302	0.681	0.665	0.908	0.769
	wo/Pos.-Temp.	0.215	0.308	0.506	0.729	0.672	0.811	1.133	1.300
Informer	All	**0.274**	**0.296**	**0.300**	**0.373**	**0.719**	**0.696**	**0.777**	**0.864**
	wo/Pos.	0.436	0.599	0.599	0.670	1.035	1.186	1.307	1.472
	wo/Temp.	0.332	0.357	0.363	0.407	0.754	0.780	0.903	1.259
	wo/Pos.-Temp.	0.524	0.651	0.801	0.954	1.038	1.351	1.491	1.512

Impact of training data size

4 Experiments

"we conduct an experiment on the Traffic dataset, comparing the performance of the model trained with a full dataset (17,544*0.7hours), named Ori., and that trained with a shortened dataset (8,760 hours, i.e., 1 year), called Short."

Table 7: Comparison of forecasting errors (MSE) with different training sizes.

Methods	*DLinear*		FEDformer		Autoformer	
Dataset	*Ori.*	*Short*	*Ori.*	*Short*	*Ori.*	*Short*
96	0.650	**0.631**	0.587	**0.568**	0.613	**0.594**
192	0.598	**0.582**	0.604	**0.584**	**0.616**	0.621
336	0.605	**0.589**	0.621	**0.601**	0.622	**0.621**
720	0.645	**0.628**	0.626	**0.608**	0.660	**0.650**

Weight visualization

4 Experiments

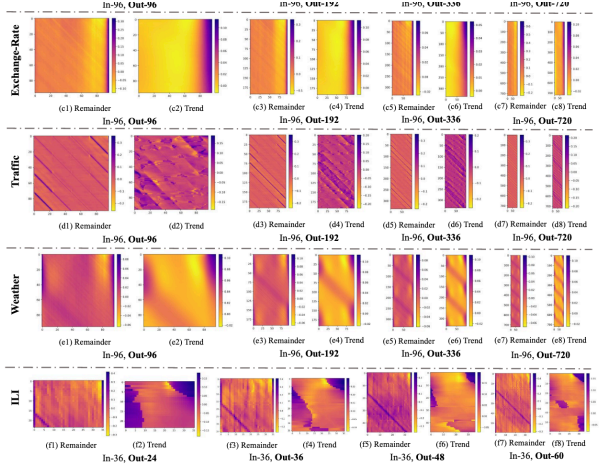


Figure 6: Visualization of the weights(T*L) of *DLinear* on several benchmarks. Models are trained with a look-back window L (X-axis) and different forecasting time steps T (Y-axis). We show weights

Q&A

感谢您的聆听和反馈