Are Transformers Effective for Time Series Forecasting?

Воробьев Дмитрий

paper

О чем статья

1) В задаче долгосрочного прогнозирования временных рядов трансформеры иногда показывают себя сильно хуже линейных моделей.

2) Ставится под сомнение способность трансформеров учитывать позиционную информацию между токенами. Позиционных эмбеддингов бывает недостаточно, когда порядок токенов очень важен.

LTSF: long-term time series forecasting

$$\text{input}: \ \{X_1^t, \ldots, X_C^t\}_{t=1}^{L}$$

$$\text{output}: \ \{X_1^t, \ldots, X_C^t\}_{t=L+1}^{L+T}$$

Рассматривается задача долгосрочного прогнозирования, то есть T >> 1.

Используется DMS (direct multi step): сразу предсказываем T точек.

Не используется (почти) IMS (iterative multi step): итеративно предсказываем по одной точке.

Как применяют трансформеры к TS.

Зачем модифицировать трансформеры:

- Квадратичная сложность классического self-attention
- Тяжелый вход, не обязательно обрабатывать весь.
- Последовательное генерирование токенов способствует накоплению ошибки
- Переменная длина выхода

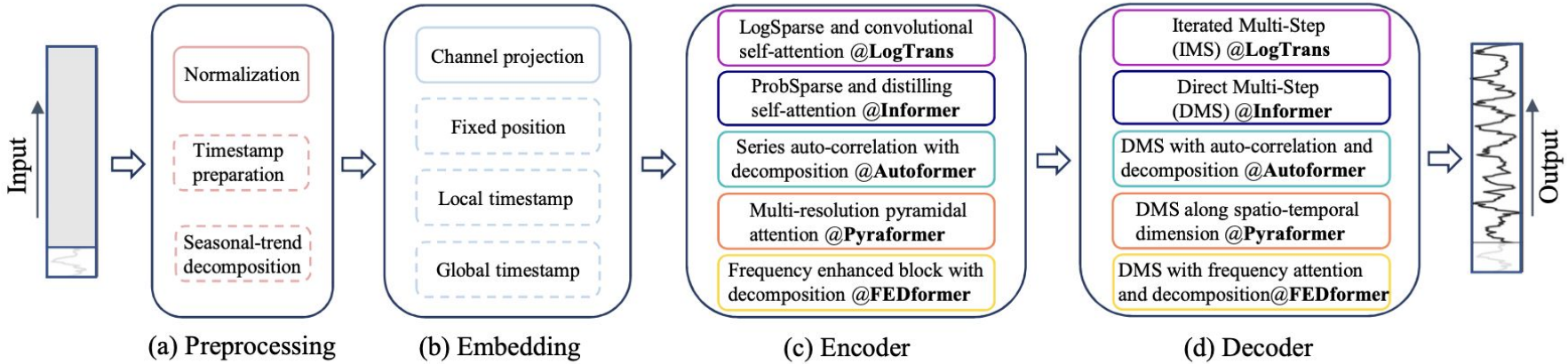


Figure 1. The pipeline of existing Transformer-based TSF solutions. In (a) and (b), the solid boxes are essential operations, and the dotted boxes are applied optionally. (c) and (d) are distinct for different methods [16, 18, 28, 30, 31].

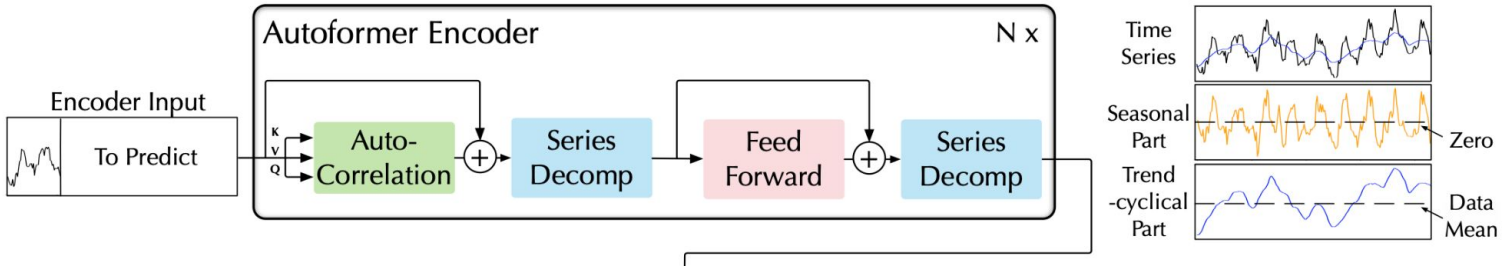
Ключевые особенности трансформеров

Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting

- **ProbSparse Self-attention Mechanism**: эффективная обработка длинных последовательностей данных, снижает вычислительную нагрузку и увеличивая диапазон внимания модели. Делает возможным более широкий обзор временных данных и прогнозирование на несколько вперед.
- **Distilling Operation**: Позволяет избирательно переносить информацию между слоями и сокращать размерность данных без значительной потери информативности.

Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting

- **Auto-correlation with decomposition**: замена полноценного self-attention блока на блок авто-корреляции, который извлекает информацию опираясь на предположении о том, что временной ряд есть смесь переодических и тренда.



Ключевые особенности трансформеров

Pyraformer: Low-complexity pyramidal attention for long-range time series modeling…

- **Pyramidal attention**: несколько уровней “внимания”, позволяет модели обращать внимание на короткие, средние и длинные зависимости в данных. Экономит вычисления за счет применения self-attention сразу к отрезкам ряда, а не к каждой точек по отдельности.

FEDformer: Frequency Enhanced Decomposed Transformer for Long-term Series Forecasting

- **Преобразование входных данных**: Временной ряд трансформируется из временной в частотную область с помощью FFT.

- **Frequency attention**: помогает модели акцентировать внимание на определенных частотах, которые могут быть более значимыми для выполнения прогнозирования.

Лиейные бейзлайны

Оказывается, что трансформеры довольно часто проигрывают простым линейным моделям.

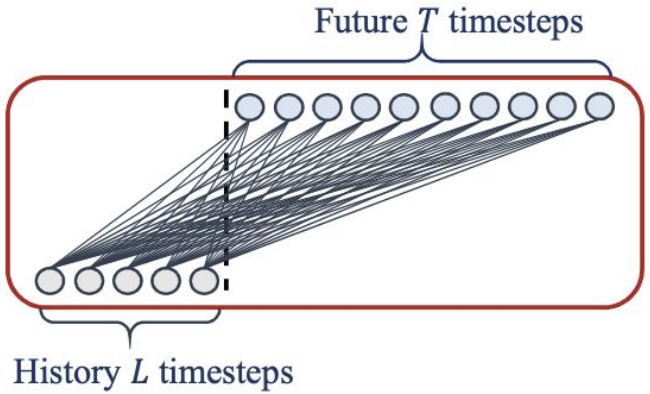


Figure 2. Illustration of the basic linear model.

DLinear:

1) Удаление тренда
2) Линейный слой
3) Добавление тренда

NLinear:

1) Вычитаем предыдущее значение
2) Линейный слой
3) Кумулятивно прибавляем предыдущее значение

Данные

1) Electricity Transformer Temperature (ETTh1, ETTh2, ETTm1, ETTm2)
2) Traffic
3) Electricity
4) Weather
5) ILI (грипп)
6) Exchange-Rate

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

Table 1. The statistics of the nine popular datasets for the LTSF problem.

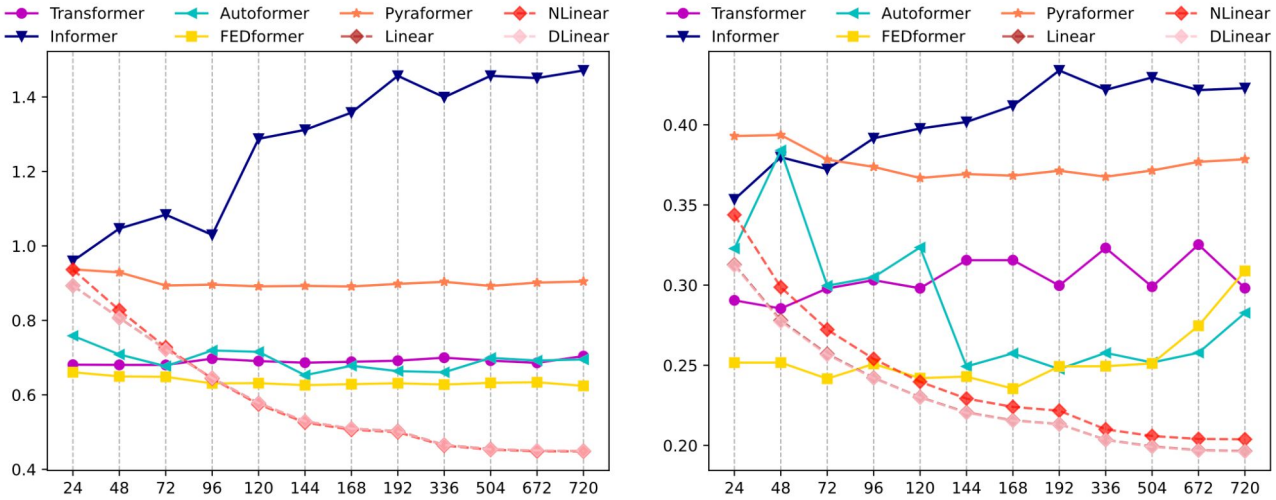
Methods		IMP.	Linear*		NLinear*		DLinear*		FEDformer		Autoformer		Informer		Pyraformer*		LogTrans		*Repeat**	
Metric		MSE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Electricity	96	27.40%	**0.140**	**0.237**	0.141	**0.237**	**0.140**	**0.237**	0.193	0.308	0.201	0.317	0.274	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

Долгосрочное прогнозирование

Methods	FEDformer		Autoformer	
Input	*Close*	*Far*	*Close*	*Far*
Electricity	0.251	0.265	0.255	0.287
Traffic	0.631	0.645	0.677	0.675

Table 3. Comparison of different input sequences under the MSE metric to explore what LTSF-Transformers depend on. If the input is *Close*, we use the $96_{th}, ..., 191_{th}$ time steps as the input sequence. If the input is *Far*, we use the $0_{th}, ..., 95_{th}$ time steps. Both of them forecast the $192_{th}, ..., (192 + 720)_{th}$ time steps.

Долгосрочное прогнозирование



(a) **720** steps-**Traffic** (b) **720** steps-**Electricity**

Figure 4. The MSE results (Y-axis) of models with different look-back window sizes (X-axis) of long-term forecasting (T=**720**) on the Traffic and Electricity datasets.

Трансформеры лишь переусложняют

Methods		Informer	*Att.-Linear*	*Embed + Linear*	Linear
Exchange	96	0.847	1.003	0.173	0.084
	192	1.204	0.979	0.443	0.155
	336	1.672	1.498	1.288	0.301
	720	2.478	2.102	2.026	0.763
ETTh1	96	0.865	0.613	0.454	0.400
	192	1.008	0.759	0.686	0.438
	336	1.107	0.921	0.821	0.479
	720	1.181	0.902	1.051	0.515

Table 4. The MSE comparisons of gradually transforming Informer to a Linear from the left to right columns. *Att.-Linear* is a structure that replaces each attention layer with a linear layer. *Embed + Linear* is to drop other designs and only keeps embedding layers and a linear layer. The look-back window size is 96.

Насколько эффективно учитывается порядок.

Methods		Linear			FEDformer			Autoformer			Informer		
Predict Length		*Ori.*	*Shuf.*	*Half-Ex.*	*Ori.*	*Shuf.*	*Half-Ex.*	*Ori.*	*Shuf.*	*Half-Ex.*	*Ori.*	*Shuf.*	*Half-Ex.*
Exchange	96	0.080	0.133	0.169	0.161	0.160	0.162	0.152	0.158	0.160	0.952	1.004	0.959
	192	0.162	0.208	0.243	0.274	0.275	0.275	0.278	0.271	0.277	1.012	1.023	1.014
	336	0.286	0.320	0.345	0.439	0.439	0.439	0.435	0.430	0.435	1.177	1.181	1.177
	720	0.806	0.819	0.836	1.122	1.122	1.122	1.113	1.113	1.113	1.198	1.210	1.196
	Average Drop	N/A	27.26%	46.81%	N/A	-0.09%	0.20%	N/A	0.09%	1.12%	N/A	-0.12%	-0.18%
ETTh1	96	0.395	0.824	0.431	0.376	0.753	0.405	0.455	0.838	0.458	0.974	0.971	0.971
	192	0.447	0.824	0.471	0.419	0.730	0.436	0.486	0.774	0.491	1.233	1.232	1.231
	336	0.490	0.825	0.505	0.447	0.736	0.453	0.496	0.752	0.497	1.693	1.693	1.691
	720	0.520	0.846	0.528	0.468	0.720	0.470	0.525	0.696	0.524	2.720	2.716	2.715
	Average Drop	N/A	81.06%	4.78%	N/A	73.28%	3.44%	N/A	56.91%	0.46%	N/A	1.98%	0.18%

Table 5. The MSE comparisons of models when shuffling the raw input sequence. *Shuf.* randomly shuffles the input sequence. *Half-EX.* randomly exchanges the first half of the input sequences with the second half. Average Drop is the average performance drop under all forecasting lengths after shuffling. All results are the average test MSE of five runs.

Различные эмбеддинги

wo/Pos: without positional embeddings

wo/Temp: without timestamp embeddings

Methods	Embedding	Traffic			
		96	192	336	720
FEDformer	All	0.597	0.606	0.627	0.649
	wo/Pos.	**0.587**	**0.604**	**0.621**	**0.626**
	wo/Temp.	0.613	0.623	0.650	0.677
	wo/Pos.-Temp.	0.613	0.622	0.648	0.663
Autoformer	All	0.629	0.647	0.676	**0.638**
	wo/Pos.	**0.613**	**0.616**	**0.622**	0.660
	wo/Temp.	0.681	0.665	0.908	0.769
	wo/Pos.-Temp.	0.672	0.811	1.133	1.300
Informer	All	**0.719**	**0.696**	**0.777**	**0.864**
	wo/Pos.	1.035	1.186	1.307	1.472
	wo/Temp.	0.754	0.780	0.903	1.259
	wo/Pos.-Temp.	1.038	1.351	1.491	1.512

Table 6. The MSE comparisons of different embedding strategies on Transformer-based methods with look-back window size 96 and forecasting lengths $\{96, 192, 336, 720\}$.

Размер обучающей выборки

Methods	FEDformer		Autoformer	
Dataset	*Ori.*	*Short*	*Ori.*	*Short*
96	0.587	**0.568**	0.613	**0.594**
192	0.604	**0.584**	**0.616**	0.621
336	0.621	**0.601**	0.622	**0.621**
720	0.626	**0.608**	0.660	**0.650**

Table 7. The MSE comparison of two training data sizes.

Ori: 100% of Traffic dataset
Short: 50% of Traffic dataset

Время работы

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

- × is modified into the same one-step decoder, which is implemented in the source code from Autoformer.
- ∗ 236.7M parameters of Pyraformer come from its linear decoder.

Table 8. Comparison of practical efficiency of LTSF-Transformers under L=96 and T=720 on the Electricity. MACs are the number of multiply-accumulate operations. We use Dlinear for comparison since it has the double cost in *LTSF-Linear*. The inference time averages 5 runs.

Заключение

1) Трансформеры кажутся очень уместной архитектурой для временных рядов, однако существующие методы не позволяют извлечь из ряда достаточно много информации для эффективного прогнозирования.
2) Показанные результаты не говорят о том, что для временных рядов надо использовать линейные модели, они показывают, что трансформеры часто можно побить простыми архитектурами.
3) В анализе временных рядов позиционные эмбеддинги не работают или работают плохо.