First De-Trend then Attend: Rethinking Attention for Time-Series Forecasting

Xiyuan Zhang[1], Xiaoyong Jin[2], Karthick Gopalswamy[2], Gaurav Gupta[2], Youngsuk Park[2], Xingjian Shi[3], Hao Wang[2], Danielle C. Maddix[2], Yuyang Wang[2]

[1]UC San Diego [2]AWS AI Labs [3]AWS

Motivation

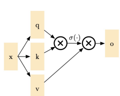
Attentions models [1] achieve promising performance for time-series forecasting. Recent works [2] explore learning attention in different domains (time, Fourier, wavelet domain).

We hope to investigate: Does learning attention in one domain offer better representation ability or empirical advantages than the other?

- *Theoretically understand their relationships: **Linear Equivalence***
- *Empirically analyze their separate advantages: **Investigation on the Role of Softmax***
- *Combine empirical advantages for a better forecasting model: **Our Method: TDformer***

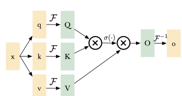
Attention Formulation

Time Attention



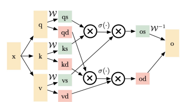
$$\mathbf{o}(\mathbf{q},\mathbf{k},\mathbf{v}) = \sigma\left(\frac{\mathbf{q}\mathbf{k}^T}{\sqrt{d_q}}\right)\mathbf{v}$$

Fourier Attention



$$\mathbf{o}(\mathbf{q},\mathbf{k},\mathbf{v}) = \mathcal{F}^{-1}\left(\sigma\left(\mathcal{F}(\mathbf{q})\overline{\mathcal{F}(\mathbf{k})}^T/\sqrt{d_q}\right)\mathcal{F}(\mathbf{v})\right)$$

Wavelet Attention



$$\mathbf{o}(\mathbf{q},\mathbf{k},\mathbf{v}) = \mathcal{W}^{-1}\left(\sigma\left(\mathcal{W}(\mathbf{q})\mathcal{W}(\mathbf{k}^T)/\sqrt{d_q}\right)\mathcal{W}(\mathbf{v})\right)$$

Linear Equivalence

Simplified assumptions without considering softmax.

Time Attention:

$$\mathbf{o}(\mathbf{q},\mathbf{k},\mathbf{v}) = \mathbf{q}\mathbf{k}^T\mathbf{v}$$

Fourier Attention:

Fourier matrix has property $\mathbf{W}^{-1} = \mathbf{W}^H, \mathbf{W}^T = \mathbf{W}$

$$\mathbf{o}(\mathbf{q},\mathbf{k},\mathbf{v}) = \mathbf{W}^H[(\mathbf{W}\mathbf{q})\overline{(\mathbf{W}\mathbf{k})}^T(\mathbf{W}\mathbf{v})] = \mathbf{q}\mathbf{k}^T\mathbf{v}$$

Wavelet Attention:

Wavelet matrix has property $\mathbf{W}^T = \mathbf{W}^{-1}$

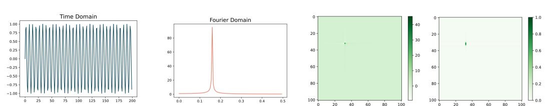
$$\mathbf{o}(\mathbf{q},\mathbf{k},\mathbf{v}) = \mathbf{W}^{-1}[(\mathbf{W}\mathbf{q})(\mathbf{W}\mathbf{k})^T(\mathbf{W}\mathbf{v})] = \mathbf{q}\mathbf{k}^T\mathbf{v}$$

Time, Fourier and wavelet attention are equivalent under linear assumptions.

Investigation on the Role of Softmax

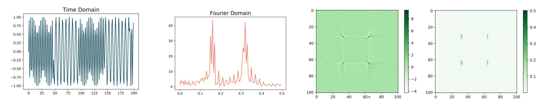
Softmax with exponential terms has the “polarization” effect: increasing the gap between large and small values

Data with fixed seasonality: Fourier attention is the most sample-efficient, as Fourier softmax attention amplifies the correct frequency modes.



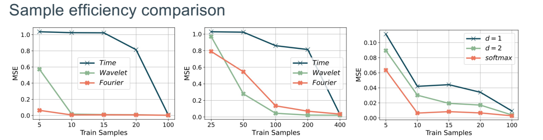
sin (x) Time | sin (x) Freq | Fourier Linear Attention | Fourier Softmax Attention

Data with varying seasonality: wavelet attention is the most effective, as wavelet softmax attention amplifies dominant frequencies, as well as keep the small-value modes that convey the information of varying seasonality.



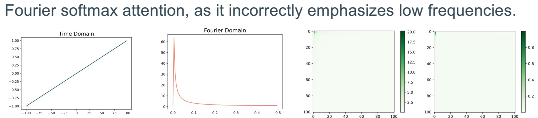
Vary Time | Vary Freq | Fourier Linear Attention | Fourier Softmax Attention

Sample efficiency comparison



Attention Models on sin (x) | Attention Models on Varying Seasonality | Softmax vs Polynomial Kernels on sin (x)

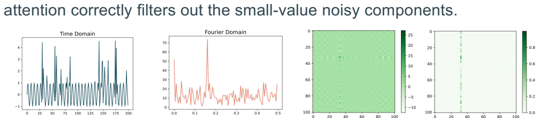
Data with trend: all attention models show inferior generalizability, especially Fourier softmax attention, as it incorrectly emphasizes low frequencies.



Trend Time | Trend Freq | Fourier Linear Attention | Fourier Softmax Attention

Metric	Time	Fourier	Wavelet	MLP
MSE	3.157 ± 0.435	8.567 ± 0.487	2.327 ± 0.689	$\mathbf{0 \pm 0}$
MAE	1.741 ± 0.121	2.880 ± 0.073	1.477 ± 0.239	$\mathbf{0.006 \pm 0.003}$

Data carrying noise: Fourier attention is the most robust, as Fourier softmax attention correctly filters out the small-value noisy components.



Noise Time | Noise Freq | Fourier Linear Attention | Fourier Softmax Attention

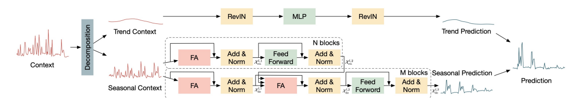
Metric	Time	Fourier	Wavelet
MSE	0.303 ± 0.002	$\mathbf{0.019 \pm 0.003}$	0.030 ± 0.008
MAE	0.495 ± 0.001	$\mathbf{0.111 \pm 0.010}$	0.137 ± 0.021

Consistent results on real-world seasonal and trend data

Method	Metric	Traffic 96	Traffic 192	Traffic 336	Traffic 720	Weather 96	Weather 192	Weather 336	Weather 720
Time	MSE	0.659	0.671	0.691	0.691	0.332	0.556	0.743	0.888
	MAE	0.358	0.358	0.368	0.363	0.395	0.533	0.622	0.702
Fourier	MSE	0.631	0.629	0.655	0.667	0.774	0.743	0.833	1.106
	MAE	0.338	0.336	0.345	0.350	0.648	0.632	0.659	0.769
Wavelet	MSE	0.622	0.629	0.640	0.655	0.358	0.564	0.815	1.312
	MAE	0.337	0.334	0.338	0.346	0.413	0.535	0.664	0.841

Our Method: TDformer

Our model design: TDformer



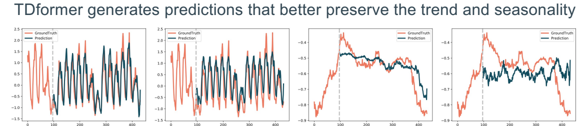
Forecasting results on benchmark multivariate time-series data

Methods		TDformer MSE	TDformer MAE	Non-stat TF MSE	Non-stat TF MAE	FEDformer MSE	FEDformer MAE	Autoformer MSE	Autoformer MAE	Informer MSE	Informer MAE	LogTrans MSE	LogTrans MAE	Reformer MSE	Reformer MAE
Electricity	96	**0.160**	**0.263**	0.169	0.273	0.193	0.308	0.201	0.317	0.274	0.368	0.258	0.357	0.312	0.402
	192	**0.172**	**0.275**	0.182	0.286	0.201	0.315	0.222	0.334	0.296	0.386	0.266	0.368	0.348	0.433
	336	**0.186**	**0.290**	0.200	0.304	0.214	0.329	0.231	0.338	0.300	0.394	0.280	0.380	0.350	0.433
	720	**0.215**	**0.313**	0.222	0.32	0.246	0.355	0.254	0.361	0.373	0.439	0.283	0.376	0.340	0.420
Exchange	96	**0.089**	**0.208**	0.111	0.237	0.148	0.278	0.197	0.323	0.847	0.752	0.968	0.812	1.065	0.829
	192	**0.183**	**0.305**	0.219	0.335	0.271	0.380	0.300	0.369	1.204	0.895	1.040	0.851	1.188	0.906
	336	**0.353**	**0.429**	0.421	0.476	0.460	0.500	0.509	0.524	1.672	1.036	1.659	1.081	1.357	0.976
	720	**0.932**	**0.725**	1.092	0.769	1.195	0.841	1.447	0.941	2.478	1.310	1.941	1.127	1.510	1.016
Traffic	96	**0.545**	**0.320**	0.612	0.338	0.587	0.366	0.613	0.388	0.719	0.391	0.684	0.384	0.732	0.423
	192	**0.571**	**0.329**	0.613	0.340	0.604	0.373	0.616	0.382	0.696	0.379	0.685	0.390	0.733	0.420
	336	**0.589**	**0.331**	0.618	0.328	0.621	0.383	0.622	0.337	0.777	0.420	0.733	0.408	0.742	0.420
	720	**0.606**	**0.337**	0.653	0.355	0.626	0.382	0.660	0.408	0.864	0.472	0.717	0.396	0.755	0.423
Weather	96	0.177	**0.215**	**0.173**	0.223	0.217	0.296	0.266	0.336	0.300	0.384	0.458	0.490	0.689	0.596
	192	**0.224**	**0.257**	0.245	0.285	0.276	0.336	0.307	0.367	0.598	0.544	0.658	0.589	0.752	0.638
	336	**0.278**	**0.290**	0.321	0.338	0.339	0.359	0.380	0.395	0.578	0.523	0.797	0.652	0.639	0.596
	720	**0.368**	**0.351**	0.414	0.410	0.403	0.428	0.419	0.428	1.059	0.741	0.869	0.675	1.130	0.792
ETTm2	96	**0.174**	**0.256**	0.192	0.274	0.203	0.287	0.255	0.339	0.365	0.453	0.768	0.642	0.658	0.619
	192	**0.243**	**0.302**	0.280	0.339	0.269	0.328	0.281	0.340	0.533	0.563	0.989	0.757	1.078	0.827
	336	**0.308**	**0.344**	0.334	0.361	0.325	0.366	0.339	0.372	1.363	0.887	1.334	0.872	1.549	0.972
	720	**0.400**	**0.400**	0.417	0.413	0.421	0.415	0.422	0.419	3.379	1.338	3.048	1.328	2.631	1.242

Ablation study by changing the trend and seasonal modules

Method	Metric	Traffic 96	Traffic 192	Traffic 336	Traffic 720	Exchange 96	Exchange 192	Exchange 336	Exchange 720
TDformer	MSE	0.545	0.571	0.589	0.606	0.089	0.183	0.353	0.932
	MAE	0.320	0.329	0.331	0.337	0.208	0.305	0.429	0.725
TDformer-MLP-TA	MSE	0.573	0.592	0.605	0.630	0.086	0.181	0.340	0.923
	MAE	0.334	0.336	0.340	0.351	0.205	0.303	0.422	0.721
TDformer-MLP-WA	MSE	0.552	0.583	0.599	0.629	0.088	0.185	0.348	0.925
	MAE	0.322	0.330	0.337	0.347	0.208	0.307	0.426	0.721
TDformer-TA-FA	MSE	0.590	0.590	0.617	0.642	0.242	0.349	0.629	0.908
	MAE	0.338	0.336	0.349	0.357	0.327	0.419	0.558	0.720
TDformer w/o RevIN	MSE	0.577	0.595	0.607	0.636	0.093	0.201	0.392	1.042
	MAE	0.320	0.325	0.328	0.339	0.222	0.330	0.474	0.763

TDformer generates predictions that better preserve the trend and seasonality



Electricity, TDformer | Electricity, FEDformer | Weather, TDformer | Weather, FEDformer

References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, £ukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in neural information processing systems, 30, 2017.

[2] Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. arXiv preprint arXiv:2201.12740, 2022.