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

Outline

- Literature and Motivation
- Linear Equivalence
- Investigation on the Role of Softmax
- Our Method: TDformer
- Conclusion

Outline

- Literature and Motivation
- Linear Equivalence
- Investigation on the Role of Softmax
- Our Method: TDformer
- Conclusion

Literature and Motivation

- Transformer [1] for time-series forecasting
- Ability to capture long-range dependencies
- Variants of time-domain and frequency-domain Transformer
  - Informer [2], Autoformer [3], FEDformer [4], ETSformer [5], ...

Literature and Motivation

- There is a lack of understanding on attention models in different domains
  - *Does learning attention in one domain offer better representation ability than learning in the other domain?*
- We seek to
  - *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***

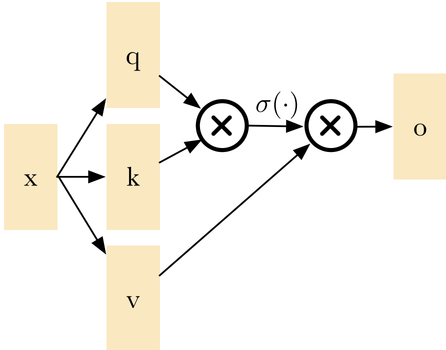
Outline

- Literature and Motivation
- Linear Equivalence
- Investigation on the Role of Softmax
- Our Method: TDformer
- Conclusion

Linear Equivalence

- Time-domain attention models

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



- Simplified assumptions without softmax

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

Linear Equivalence

- Fourier-domain attention models
- Fourier transform and inverse Fourier transform

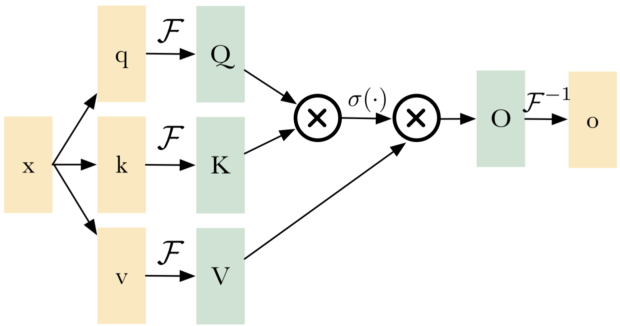
$$\mathbf{W} = (\frac{\omega^{jk}}{\sqrt{L}}) \in \mathbb{C}^{L\times L}, \omega = e^{-\frac{2\pi j}{L}} \quad \mathbf{X} = \mathbf{W}\mathbf{x} \qquad \mathbf{x} = \mathbf{W}^H\mathbf{X}$$

$$\mathbf{W}^{-1} = \mathbf{W}^H, \mathbf{W}^T = \mathbf{W}$$

- Simplified assumptions without softmax

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

- Equivalent to time-domain attention



Linear Equivalence

- Wavelet-domain attention models
- Wavelet decomposition and wavelet reconstruction

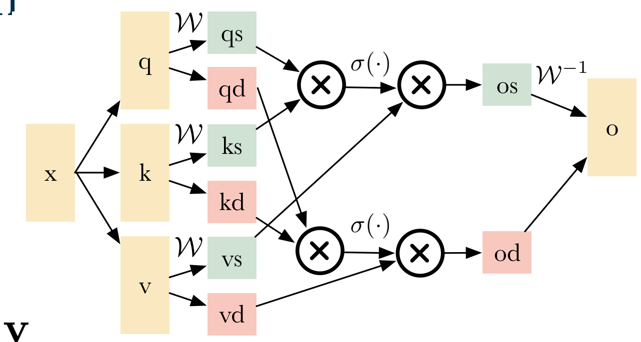
$$\mathbf{W} \in \mathbb{R}^{L \times \frac{L}{2}}, \mathbf{W}^{-1} \in \mathbb{R}^{\frac{L}{2} \times L} \qquad \mathbf{X} = \mathbf{W}\mathbf{x} \qquad \mathbf{x} = \mathbf{W}^{-1}\mathbf{X}$$

$$\mathbf{W}^T\mathbf{W} = \mathbf{I}$$

- Simplified assumptions without softmax

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

- Equivalent to time-domain attention



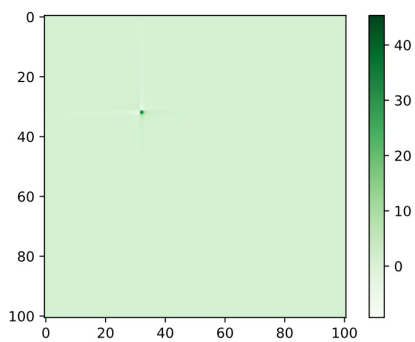
Outline

- Literature and Motivation
- Linear Equivalence
- Investigation on the Role of Softmax
- Our Method: TDformer
- Conclusion

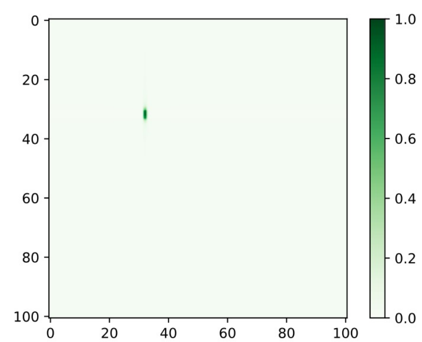
Role of Softmax: Fixed Seasonality

- sin (x): softmax in Fourier domain correctly amplifies the frequency mode

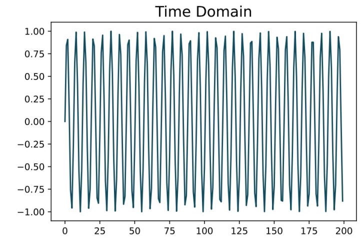
Linear Attention



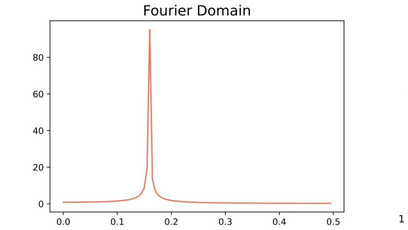
Softmax Attention

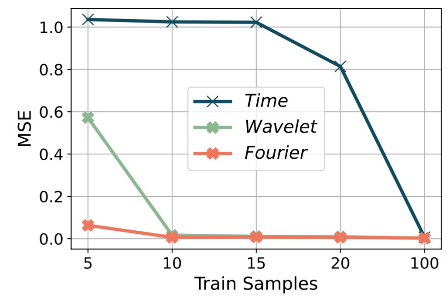


Time domain



Fourier domain

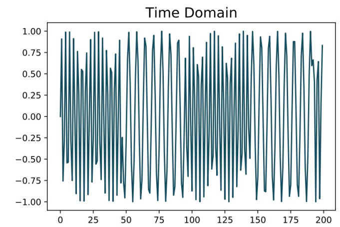




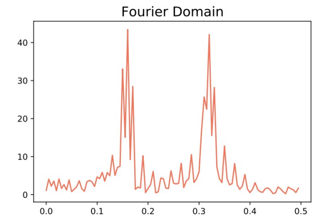
Role of Softmax: Varying Seasonality

- Alternating sin(x), sin(2x): softmax in Fourier domain loses the small-value modes that convey the information of varying frequencies

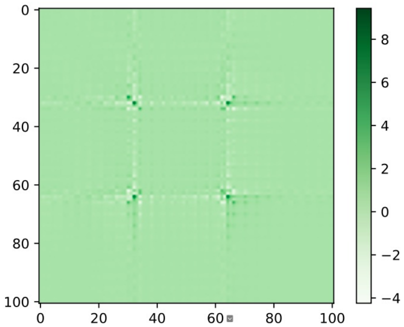
Time domain



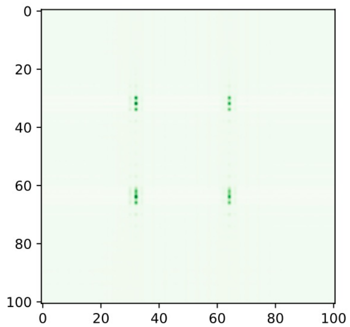
Fourier domain

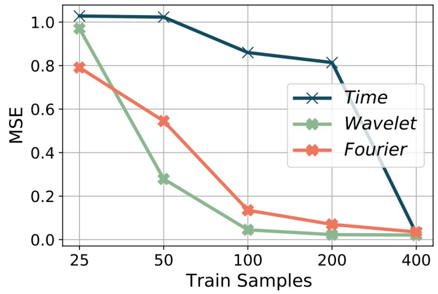


Linear Attention



Softmax Attention

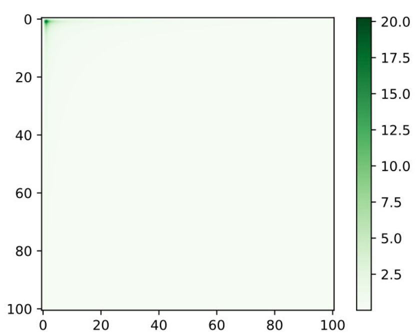




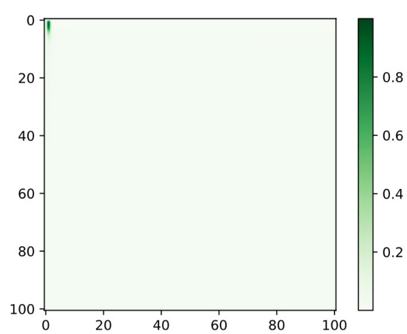
Role of Softmax: Trend Data

- Linear trend: softmax in Fourier domain incorrectly amplifies low-frequency mode

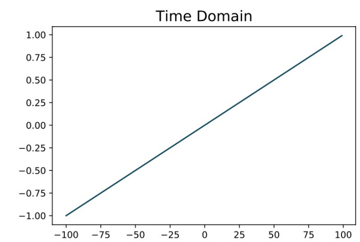
Linear Attention



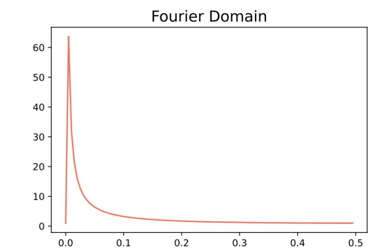
Softmax Attention



Time domain



Fourier domain



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	$\mathbf{0.006 \pm 0.003}$

Real-World Datasets

- Seasonal data (traffic): frequency attention models are better
- Trend data (weather): time attention models are better;

  all attention models do not perform well

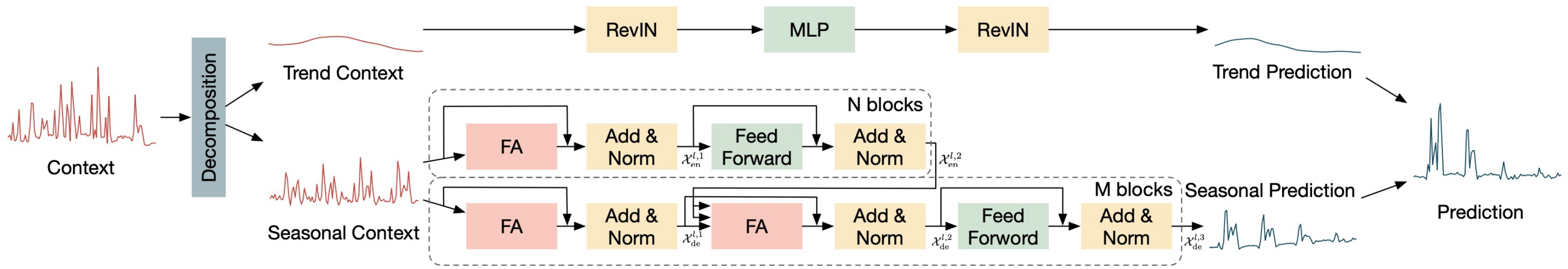
Method	Metric	Traffic				Weather			
		96	192	336	720	96	192	336	720
Time	MSE	0.659	0.671	0.691	0.691	0.332	0.556	0.743	0.888
	MAE	0.358	0.358	0.368	0.363	0.395	0.533	0.622	0.702
Fourier	MSE	0.631	0.629	0.655	0.667	0.774	0.743	0.833	1.106
	MAE	0.338	0.336	0.345	0.350	0.648	0.632	0.659	0.769
Wavelet	MSE	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

Outline

- Literature and Motivation
- Linear Equivalence
- Investigation on the Role of Softmax
- Our Method: TDformer
- Conclusion

Transformer based on Seasonal Trend Decomposition

- Seasonal trend decomposition
- MLP for trend part
- Fourier attention for seasonal part



Evaluation on Benchmarks

Methods		TDformer		Non-stat TF		FEDformer		Autoformer		Informer		LogTrans		Reformer	
Metric		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	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

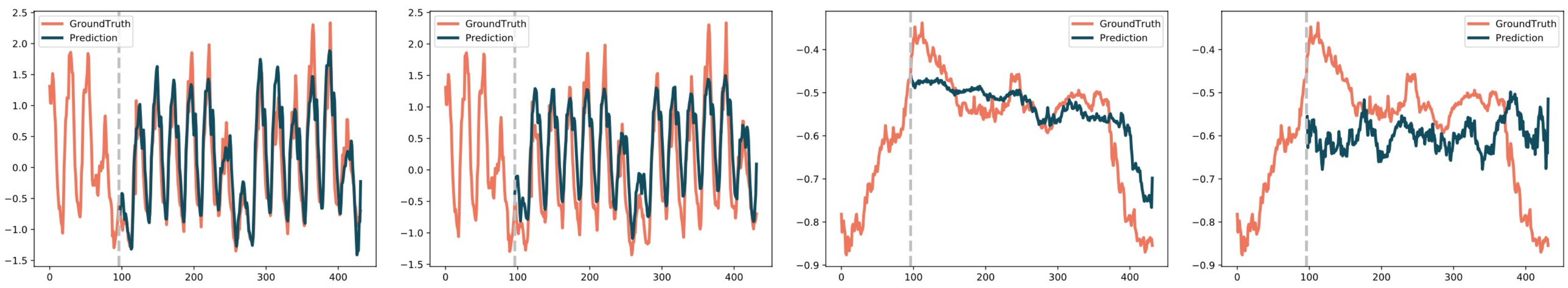
Ablation Study

- Ours-MLP-TA (WA): replace Fourier attention by time (wavelet) attention
- Ours-TA-FA: replace MLP with time attention

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

Case Study

- Compare with best performing baseline FEDformer
- Seasonal data (ECL), trend data (Weather)



(a) ECL TDformer (b) ECL FEDformer (c) Weather TDformer (d) Weather FEDformer

Outline

- Literature and Motivation
- Linear Equivalence
- Investigation on the Role of Softmax
- Our Method: TDformer
- Conclusion

Conclusion

- Attention models in different domains have linear equivalence
- Due to softmax, attention models show empirical differences
- Attention models are generally not good at generalizing trend information
- We propose TDformer that uses MLP for trend and Fourier attention for seasonality, and achieves SOTA performance against existing attention models.

Reference

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

[2] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, HuiXiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, pages 11106–11115, 2021.

[3] Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. Advances in Neural Information Processing Systems, 34:22419–22430, 2021.

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

[5] Gerald Woo, Chenghao Liu, Doyen Sahoo, Akshat Kumar, and Steven Hoi. Etsformer: Exponential smoothing transformers for time-series forecasting. arXiv preprint arXiv:2202.01381, 2022.

Thanks!

Q & A