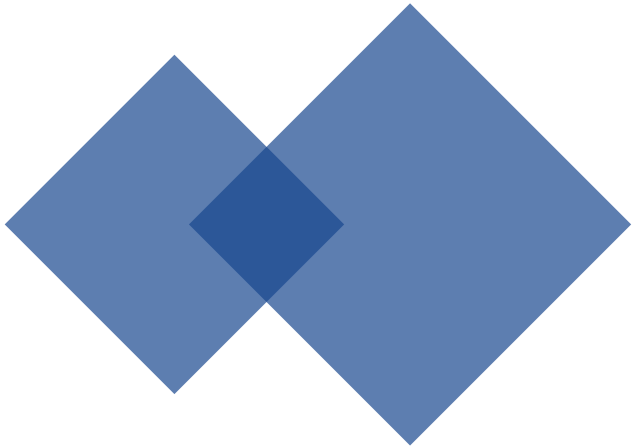
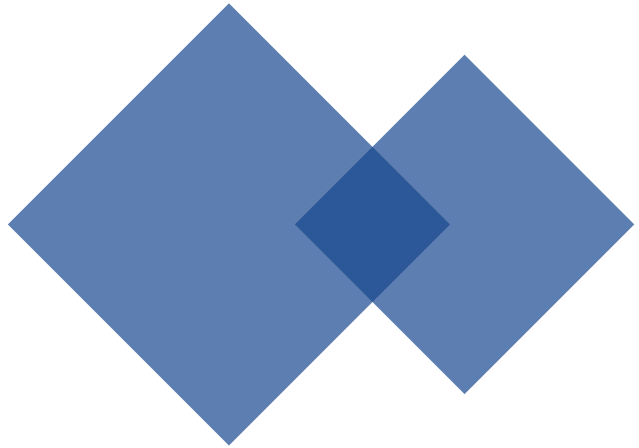


TDformer

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

Xiyuan Zhang (UC San Diego)
AWS AI Labs
AWS

23.11.28



Presented by Yyyq

01 问题引入

- 基于Attention的Transformer做时间序列预测

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

- **时域 Time domain Attention（Informer、Autoformer）**
- **傅里叶频域 Fourier domain Attention（FEDformer）**
- **小波变换频域 Wavelet domain Attention**

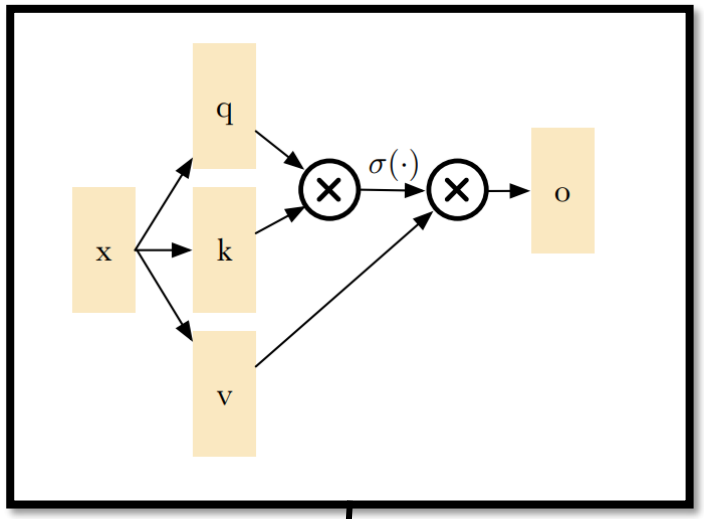
1. Zhou, Haoyi, et al. “Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting.” Proceedings of the AAAI Conference on Artificial Intelligence, Sept. 2022
2. Wu, Haixu, et al. “Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting.” Cornell University - arXiv,Cornell University - arXiv, June 2021.
3. Zhou, Tian, et al. "Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting." International Conference on Machine Learning. PMLR, 2022.

02 主要贡献

- **数学证明：**在线性条件下，时域、傅立叶域和小波域的注意模型具有相同的表征能力
- **考虑Softmax的非线性：**
  - 对于**具有较强季节性的数据：**频域注意模型比时域注意模型表现更好，更具样本效率
  - 对于**有趋势的数据：**注意模型一般表现出较差的泛化能力
  - 对于**带有噪声尖峰的数据：**频域注意力模型对这种尖峰数据更加稳健
- **提出了Tdformer模型（Trend Decomposition Transformer）**

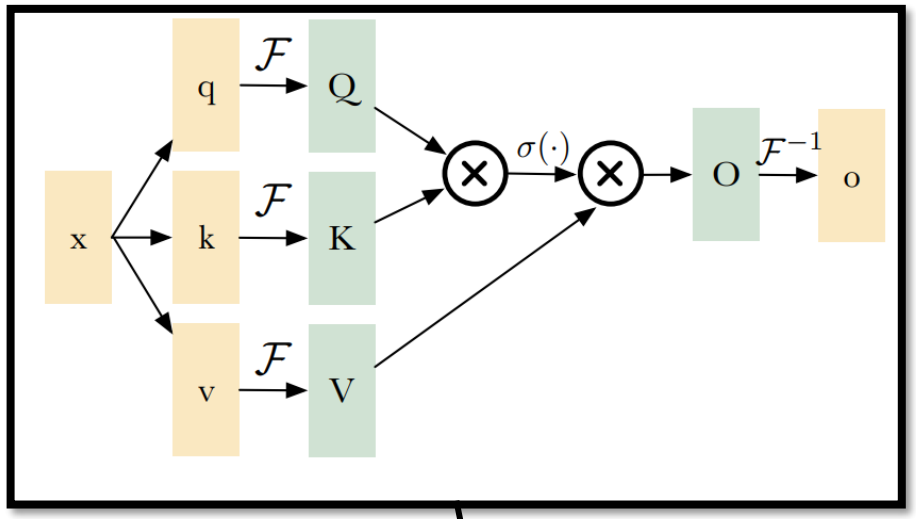
03 Linear Equivalence of Attention in Various Domains

Time domain Attention



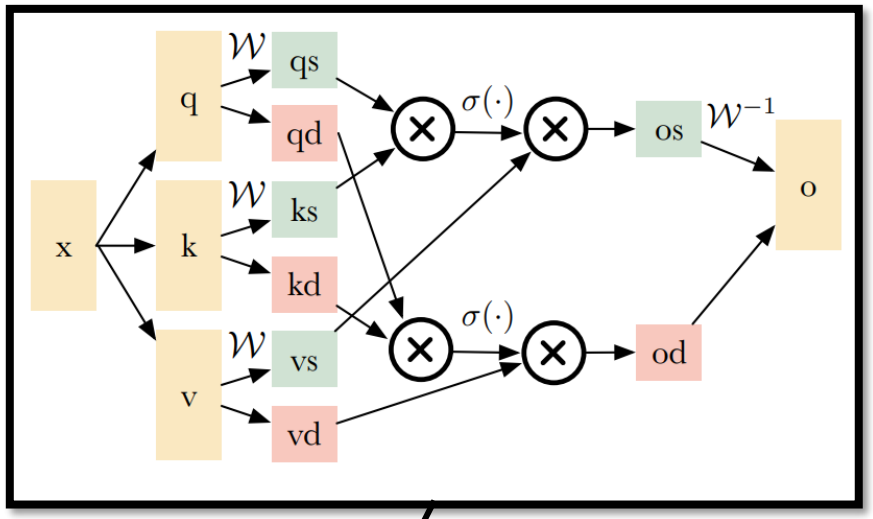
$$\mathbf{o}(\mathbf{q},\mathbf{k},\mathbf{v}) = \sigma\left(\frac{\mathbf{q}\mathbf{k}^T}{\sqrt{d_q}}\right)\mathbf{v},$$

Fourier domain 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 domain Attention



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

03 Linear Equivalence of Attention in Various Domains

Lemma 3.1. When $\sigma(\cdot) = \mathrm{Id}(\cdot)$ (linear attention).

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

$\mathbf{x} = \mathbf{W}^{H}\mathbf{X}$ $\quad$ $\mathbf{X} = \mathbf{W}\mathbf{x},$ $\quad$ W表示傅里叶矩阵：$\mathbf{W}^{-1} = \mathbf{W}^{H}, \mathbf{W}^{T} = \mathbf{W}.$

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

② $\mathbf{o}(\mathbf{q},\mathbf{k},\mathbf{v}) = \mathcal{W}^{-1}\left(\sigma\left(\mathcal{W}(\mathbf{q})\mathcal{W}(\mathbf{k}^{T})/\sqrt{d_q}\right)\mathcal{W}(\mathbf{v})\right).$ $\Leftrightarrow$ $\mathbf{o}(\mathbf{q},\mathbf{k},\mathbf{v}) = \sigma\left(\frac{\mathbf{q}\mathbf{k}^{T}}{\sqrt{d_q}}\right)\mathbf{v},$

$\mathbf{x} = \mathbf{W}^{-1}\mathbf{X}$ $\quad$ $\mathbf{X} = \mathbf{W}\mathbf{x},$ $\quad$ W表示小波分解矩阵（正定性）：$\mathbf{W}^{T}\mathbf{W} = \mathbf{I}.$

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

结论：时间、傅立叶和小波注意模型在给定线性假设下是等价的。

04 非线性的Softmax对不同注意力的影响

- **Softmax函数：归一化指数函数** $\sigma(\mathbf{z})_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}}$ for $j = 1, \ldots, K$.

含任意实数的K维向量$z \longrightarrow \sigma(z)$另一个K维实向量（将一个向量的元素转化为一个概率分布）

- 各个元素的值都在0到1之间
- 并且所有元素的和等于1

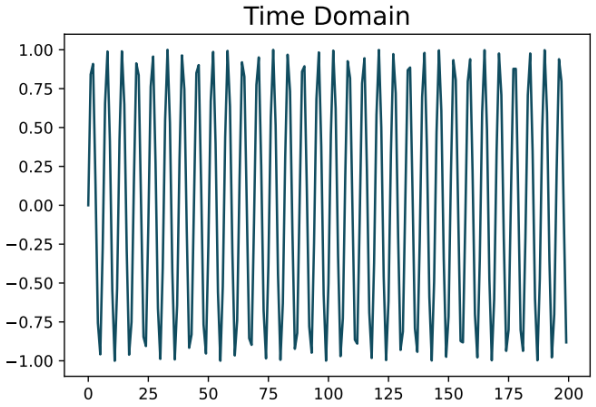
- **注意力机制中的Softmax函数：概率分布 → 权重分配（权重越高，受关注程度越高）**

1. 先通过一个线性变换将查询向量q和键k进行映射，
2. 然后计算查询向量q和每个键k的相似度得分，
3. 最后使用softmax函数将得分转化为注意力权重。
4. 注意力权重可以用于对值v进行加权求和，得到最终的注意力输出。

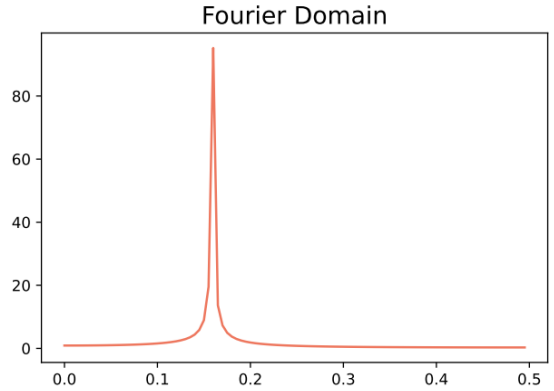
04 非线性的Softmax对不同注意力的影响

4.1 具有季节性的数据

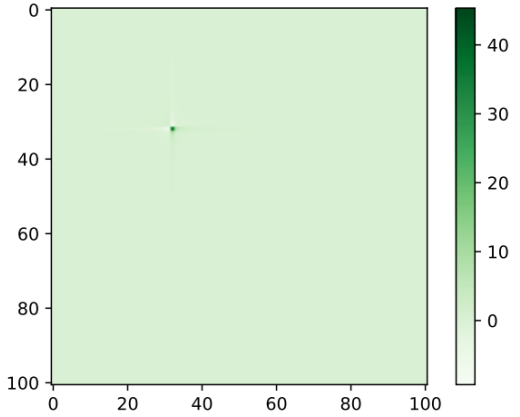
固定季节性数据： $EX.\sin(x)$



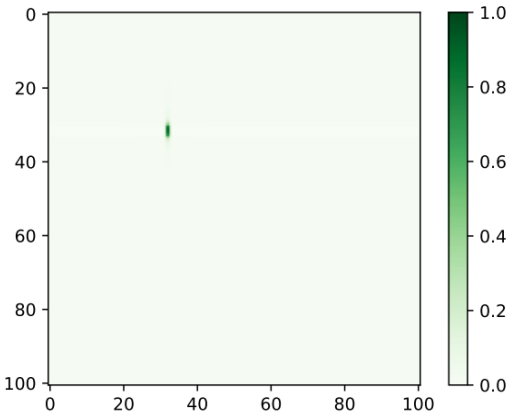
(a) sin(x) Time



(b) sin(x) Freq



(c) sin(x) Linear



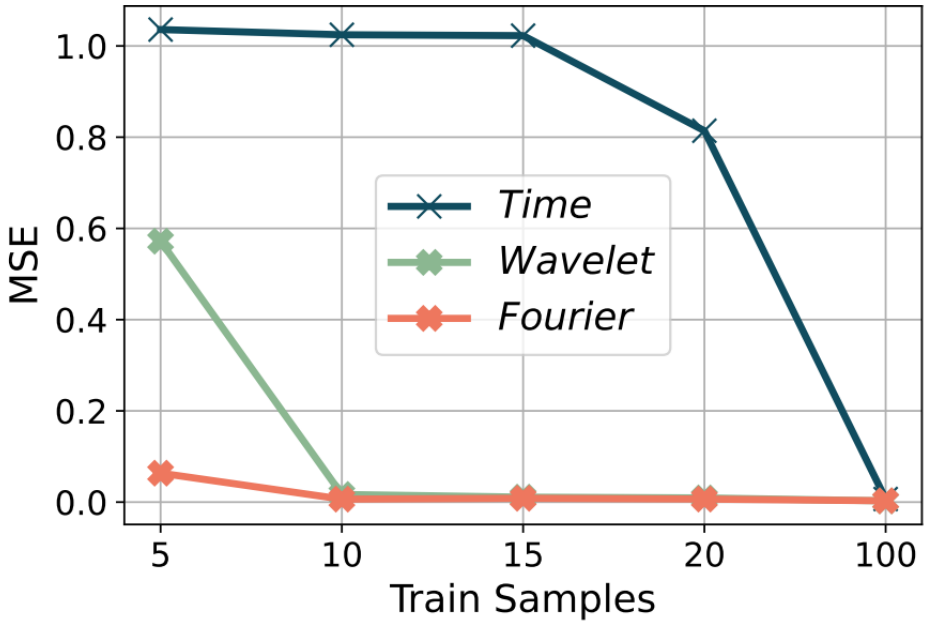
(d) sin(x) Softmax

结论：softmax的极化效应将分数集中在主导频率上，有助于模型更好地捕捉季节信息。

04 非线性的Softmax对不同注意力的影响

4.1 具有季节性的数据

固定季节性数据： $EX.\sin(x)$

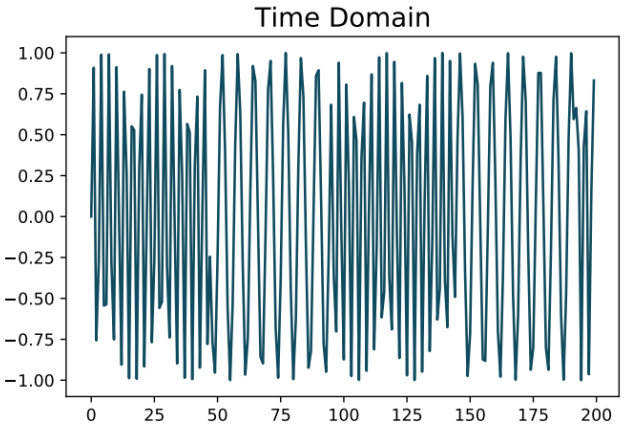


结论：频域注意力模型能够快速识别主频率模式(采样效率更高)。

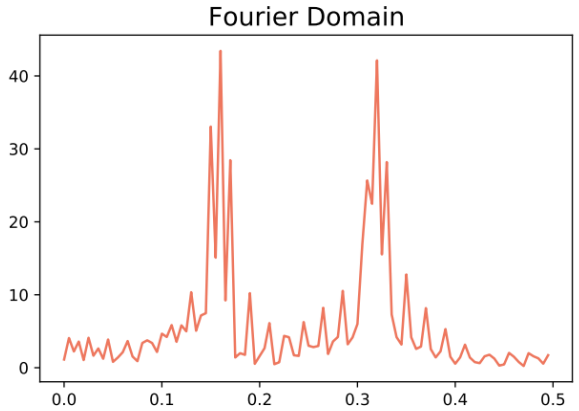
04 非线性的Softmax对不同注意力的影响

4.1 具有季节性的数据

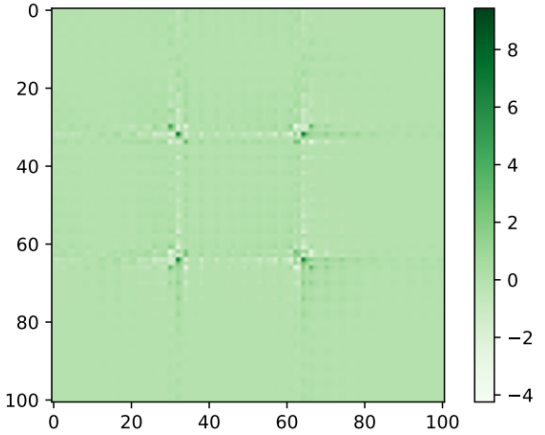
变化季节性数据： $EX.\sin(x)$ & $\sin(2x)$交替



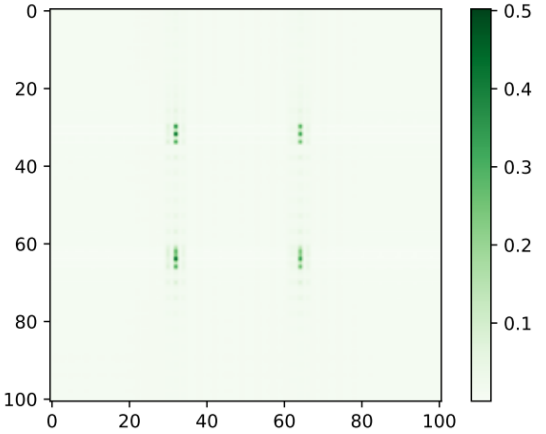
(e) Vary Time



(f) Vary Freq



(g) Vary Linear

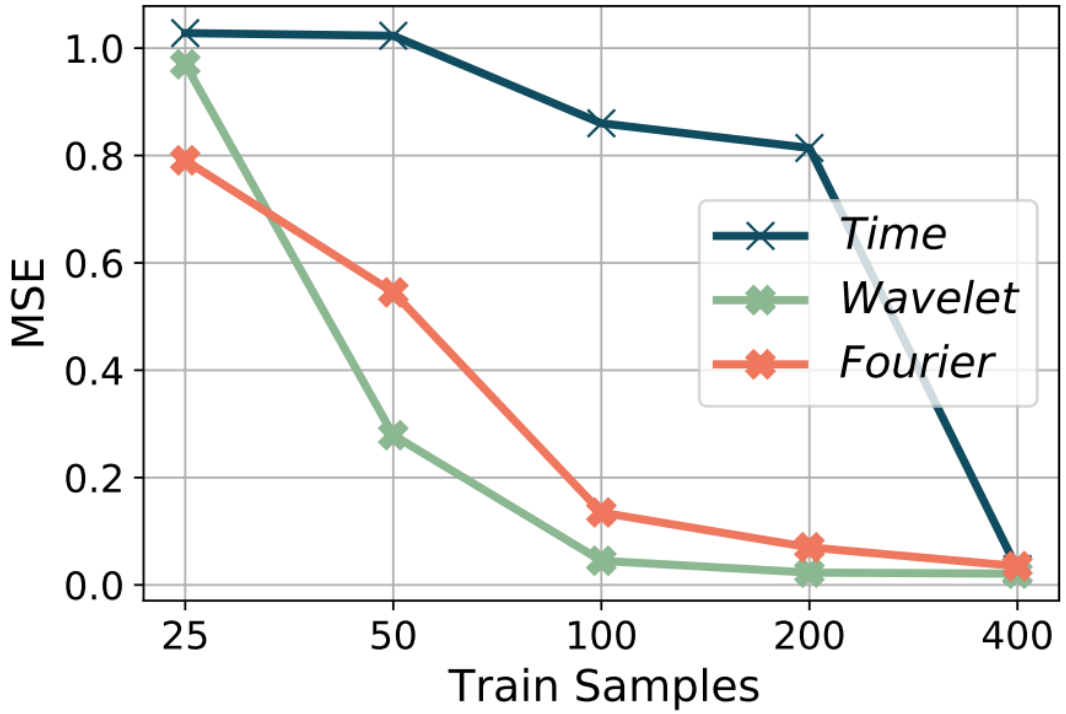


(h) Vary Softmax

04 非线性的Softmax对不同注意力的影响

4.1 具有季节性的数据

变化季节性数据： $EX.\sin(x)$ & $\sin(2x)$交替

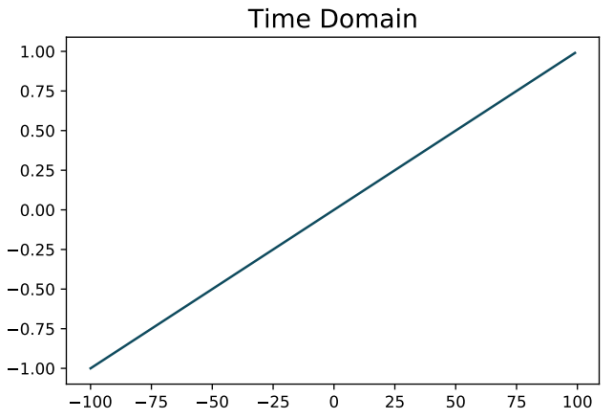


结论：小波注意结合多尺度时频表示，提供了更好的局域频率信息。

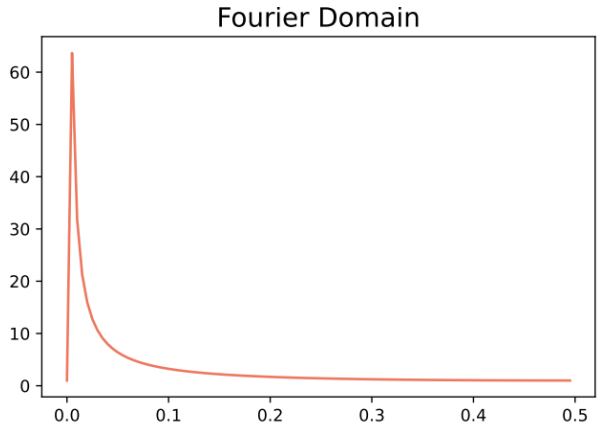
04 非线性的Softmax对不同注意力的影响

4.2 具有趋势性的数据

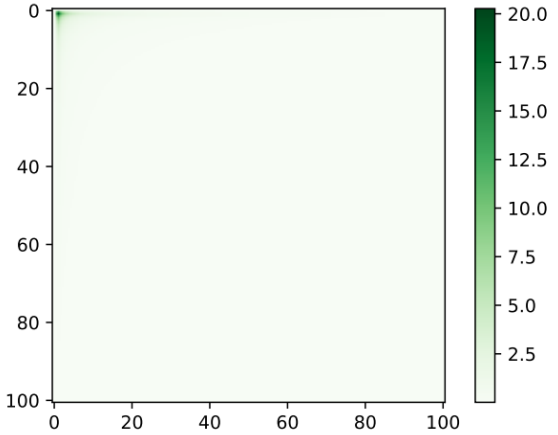
EX.线性趋势数据



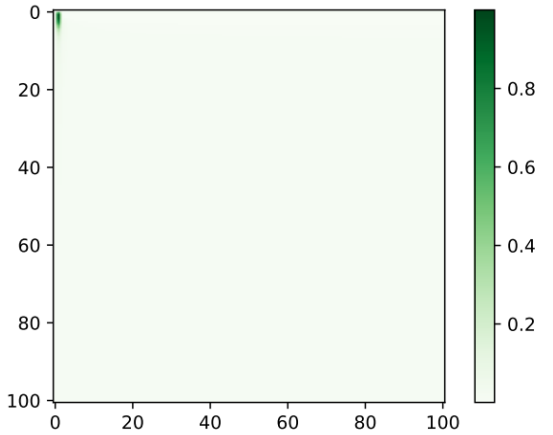
(i) Trend Time



(j) Trend Freq



(k) Trend Linear



(l) Trend Softmax

结论：在softmax的极化效应下，注意力分数更强调低频成分，导致模型在趋势性数据上泛化能力差。

04 非线性的Softmax对不同注意力的影响

4.2 具有趋势性的数据

EX.线性趋势数据

Table 1: MSE and MAE of attention models and MLP with linear-trend data.

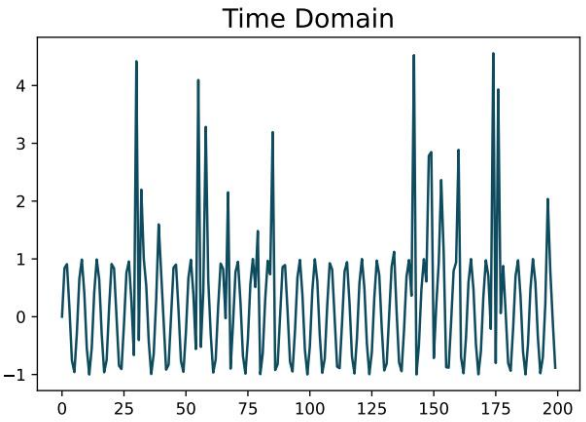
Metric	Time	Fourier	Wavelet	MLP
MSE	3.157 ± 0.435	8.567 ± 0.487	2.327 ± 0.689	$\mathbf{0} \pm \mathbf{0}$
MAE	1.741 ± 0.121	2.880 ± 0.073	1.477 ± 0.239	$\mathbf{0.006} \pm \mathbf{0.003}$

结论：MLP能够完美预测趋势信号。

04 非线性的Softmax对不同注意力的影响

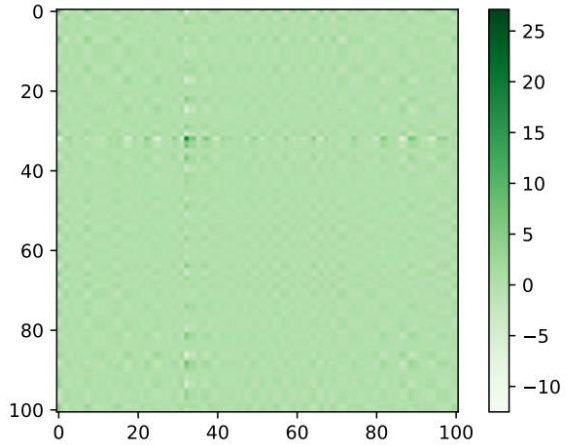
4.3 带有尖峰的数据

EX. 将极大峰值注入到 $sin(x)$ 中

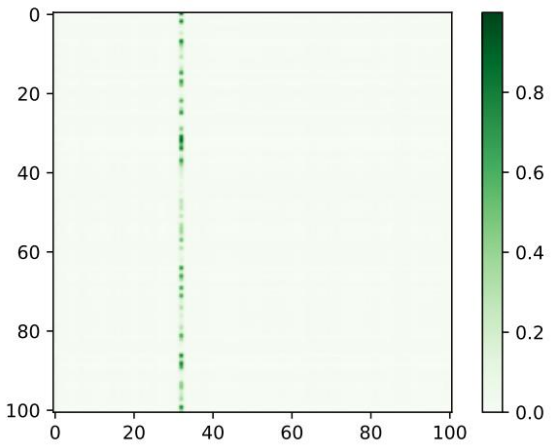


(m) Spike Time

(n) Spike Freq



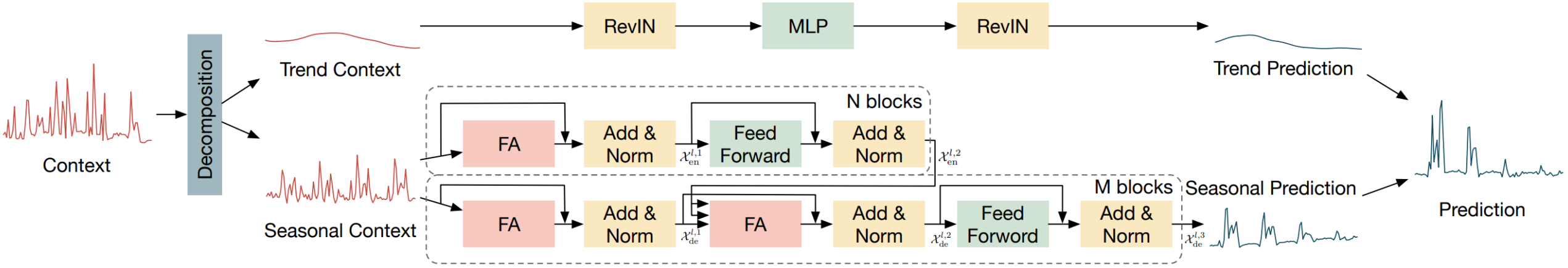
(o) Spike Linear



(p) Spike Softmax

结论：频域注意力模型对峰值的鲁棒性更强。

05 TDformer：模型结构



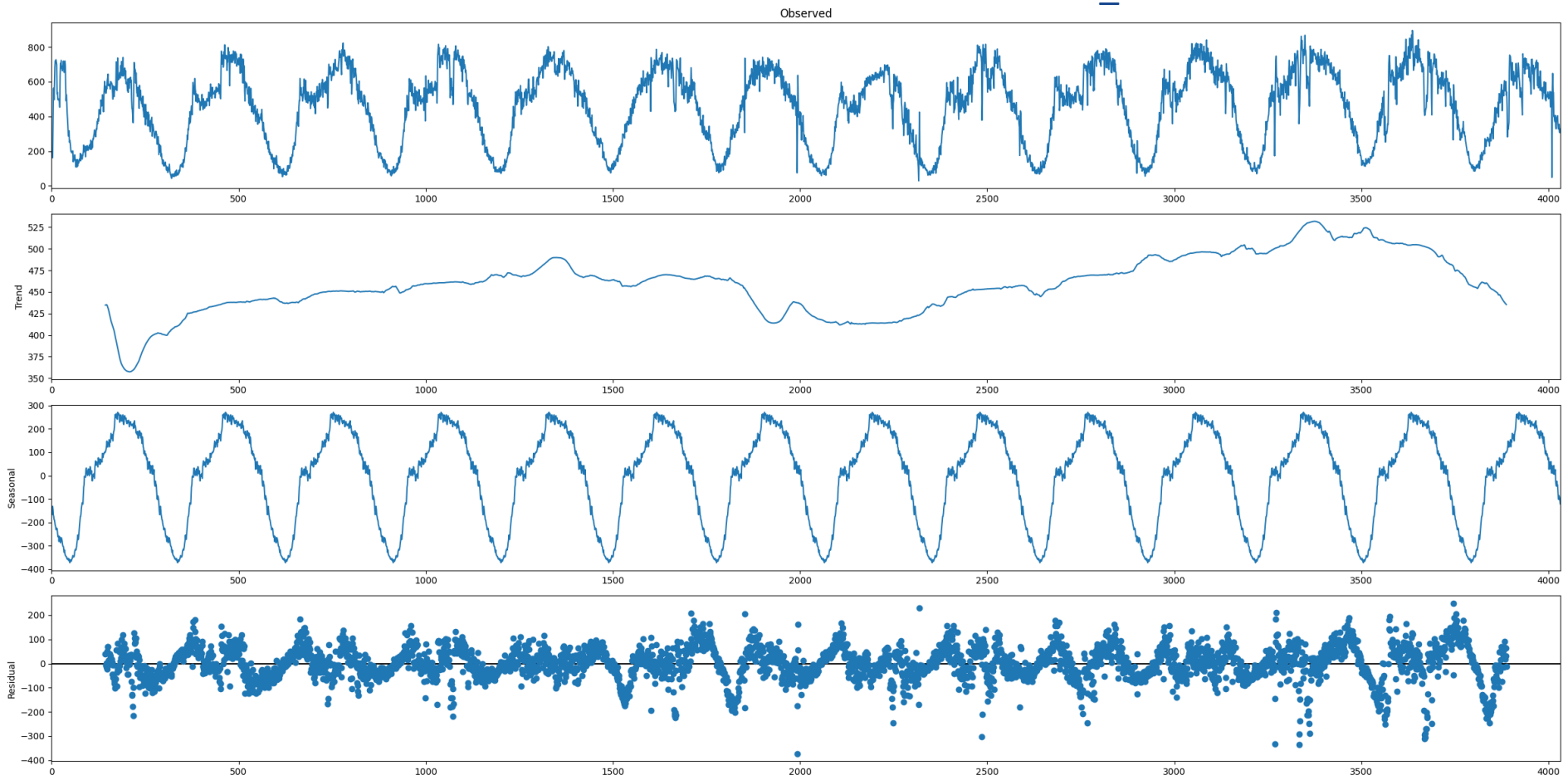
时间序列

- 趋势项　(RevIN) MLP (RevIN)
- 季节项　Transformer（使用傅里叶频域注意力）

05 TDformer：时间序列分解

STL分解： Seasonal-Trend decomposition using LOESS

Real_Data=Trend+Seasonal+Residual



05 TDformer：时间序列分解

STL分解： Seasonal-Trend decomposition using LOESS

Real_Data=Trend+Seasonal+Residual

$$X_t = T_t + S_t + R_t,$$

STL指标：量化时序数据的趋势性和季节性

$$\mathrm{S} = \max(0, 1 - \frac{\mathrm{Var}(R_t)}{\mathrm{Var}(S_t) + \mathrm{Var}(R_t)}).$$

Table 6: Seasonality strength of benchmark datasets.

Dataset	Electricity	Exchange	Traffic	Weather	ETTm2
Seasonality Strength	0.998	0.299	0.998	0.476	0.993

电力消耗　　汇率　交通道路占用量　天气温湿度　变压器油温

05 TDformer：时间序列分解

➢ 用于季节性趋势分解的专家混合（FEDformer）

文章设计了一个混合专家分解块，它包含一组不同大小的平均滤波器，从输入信号中提取多个趋势成分，以及一组数据相关的权重，将它们组合成最终趋势。形式化如下：

$$\mathbf{X}_{\mathbf{trend}} = \sigma(w(\mathbf{X})) * f(\mathbf{X}), \mathbf{X}_{\mathbf{seasonal}} = \mathbf{X} - \mathbf{X}_{\mathbf{trend}},$$

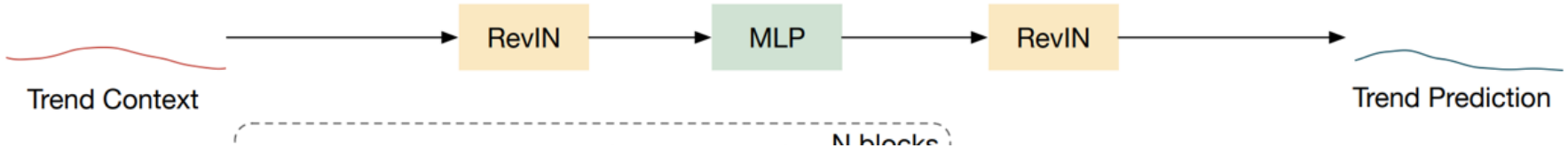
1. Zhou, Tian, et al. "Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting." International Conference on Machine Learning. PMLR, 2022.

05 TDformer：MLP

- 可逆实例归一化RevIN（Non-stationary transformers）

  有效地去除和恢复主要存在于趋势中的非平稳信息

$$\mathcal{X}_{\text{trend}} = \text{RevIN}(\text{MLP}(\text{RevIN}(\mathbf{x_{trend}}))).$$



1. Liu, Yong, et al. "Non-stationary transformers: Exploring the stationarity in time series forecasting." Advances in Neural Information Processing Systems 35 (2022): 9881-9893.

06 实验1：不同注意力机制对于季节性和趋势性数据的预测结果

Table 6: Seasonality strength of benchmark datasets.

Dataset	Electricity	Exchange	Traffic	Weather	ETTm2
Seasonality Strength	0.998	0.299	0.998	0.476	0.993

Table 3: MSE and MAE of different attention models with real-world seasonal and trend data.

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

06 实验2：对比实验

Table 4: MSE and MAE of multivariate time-series forecasting on benchmark datasets with input context length 96 and forecasting horizon $\{96, 192, 336, 720\}$. We **bold** the best performing results.

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

06

实验3：消融实验

Table 6: Seasonality strength of benchmark datasets.

Dataset	Electricity	Exchange	Traffic	Weather	ETTm2
Seasonality Strength	0.998	0.299	0.998	0.476	0.993

Method	Metric	Traffic				Exchange			
		96	192	336	720	96	192	336	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

谢谢观看

MANY THANKS !

23.11.28