国际人工智能会议

AAAI 2021 论文北京预讲会

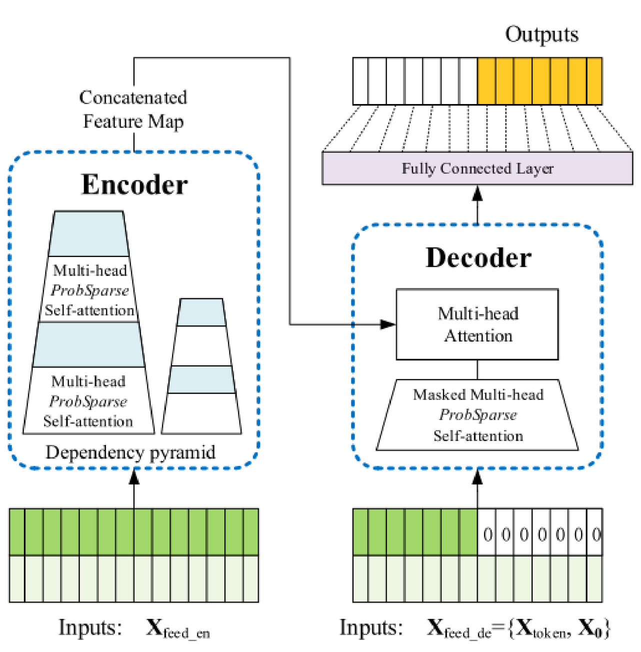
Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting

Haoyi Zhou[1], Shanghang Zhang[2], Jieqi Peng[1], Shuai Zhang[1], Jianxin Li[1], Xiong Hui[3], Wancai Zhang[4]

[1] Beihang University, [2] UC Berkeley, [3] Rutgers University,
[4] Beijing Guowang Fuda Science & Technology Development Company

Highlights

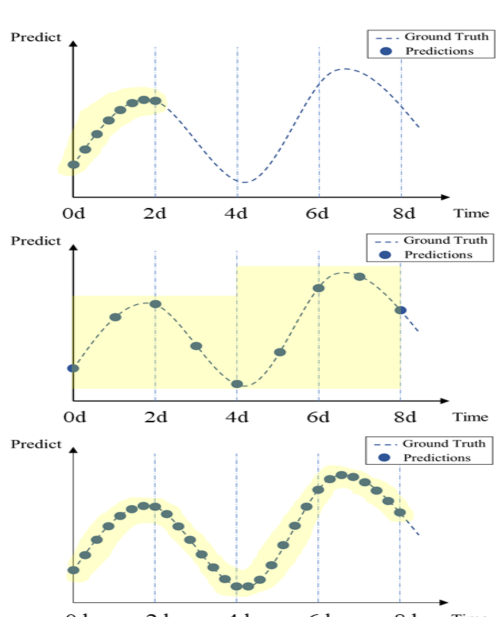
- Propose Informer to successfully enhance the prediction capacity in the Long Sequence Time-series Forecasting (LSTF) problem.



- Propose ***ProbSparse Self-Attention*** mechanism that achieves $\mathcal{O}(L\log L)$ in time complexity and memory usage.
- Propose ***Self-attention Distilling*** to sharply reduce space complexity to $\mathcal{O}((2-\epsilon)L\log L)$.
- Propose ***Generative Style Decoder*** to acquire long sequence output with only one forward step needed.

Introduction

Long Sequence Time-series Forecasting (LSTF)



- Near future predictions
- **Limited** adjustment

- Coarse predictions
- **Inefficient** adjustment

- Long sequence predictions
- **Proper** adjustment

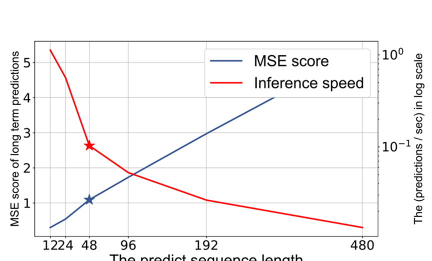
The major challenge for LSTF in **enhancing the prediction capacity to meet the increasingly long sequences demand**

- Extraordinarily ***long-range alignment ability***
- ***Efficient operations*** on long sequence inputs and outputs

Applying Transformer models in LSTF problem

Limitations of LSTM in LSTF

- The inference speed of LSTM decrease rapidly.
- *Continuous accumulation of errors causes* the MSE score *increase rapidly.*



Advances of Transformer

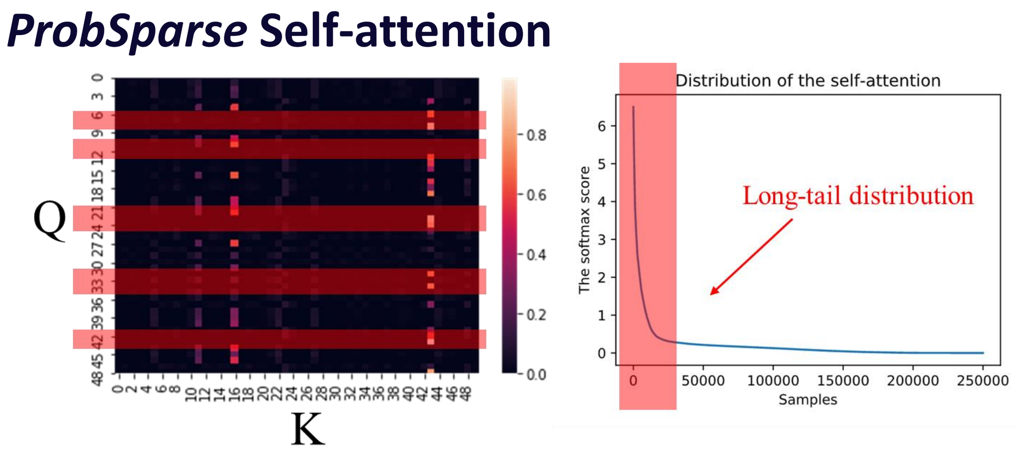
- The self-attention mechanism **reduce max path of network signals to $\mathcal{O}(1)$** and avoids recurrent structure.

Limitations of Transformer

- ***The quadratic computation of self-attention.*** The atom operation of self-attention mechanism, namely canonical dot-product, causes the time complexity and memory usage per layer to be $\mathcal{O}(L^2)$.
- ***The memory bottleneck in stacking layers for long inputs.*** The stack of J encoder/decoder layer makes total memory usage to be $\mathcal{O}(J\cdot L^2)$ which limits the model scalability on receiving long sequence inputs.
- ***The speed plunge in predicting long outputs.*** The dynamic decoding of vanilla Transformer makes the step-by-step inference as slow as RNN-based model.

Methodology

ProbSparse Self-attention



Rewrite the self-attention into the probability formulation

$$\mathcal{A}(\mathbf{q}_i,\mathbf{K},\mathbf{V})=\sum_j \frac{k(\mathbf{q}_i,\mathbf{k}_j)}{\sum_l k(\mathbf{q}_i,\mathbf{k}_l)}\mathbf{v}_j=\mathbb{E}_{p(\mathbf{k}_j|\mathbf{q}_i)}[\mathbf{v}_j]$$

Establish the measurement

Attention probability $p(\mathbf{k}_j|\mathbf{q}_i)$? Uniform probability $q(\mathbf{k}_j|\mathbf{q}_i)=\frac{1}{L_K}$

$$KL(q||p)=\sum_{j=1}^{L_K}\frac{1}{L_K}\ln\frac{1/L_K}{k(\mathbf{q}_i,\mathbf{k}_j)/\sum_l k(\mathbf{q}_i,\mathbf{k}_l)}$$

$$=\ln\sum_{l=1}^{L_K}e^{\frac{\mathbf{q}_i\mathbf{k}_l^\top}{\sqrt{d}}}-\frac{1}{L_K}\sum_{j=1}^{L_K}\frac{\mathbf{q}_i\mathbf{k}_j^\top}{\sqrt{d}}-\ln L_K$$

Dropping the constant, and **use an approximation** to the measurement for computation simplicity:

$$M(\mathbf{q}_i,\mathbf{K})=\boxed{\ln\sum_{j=1}^{L_K}e^{\frac{\mathbf{q}_i\mathbf{k}_j^\top}{\sqrt{d}}}}-\frac{1}{L_K}\sum_{j=1}^{L_K}\frac{\mathbf{q}_i\mathbf{k}_j^\top}{\sqrt{d}} \quad \textbf{Replace} \quad \boxed{\max_j\{\frac{\mathbf{q}_i\mathbf{k}_j^\top}{\sqrt{d}}\}}$$

Define *ProbSparse* self-attention

$$\mathcal{A}(\mathbf{Q},\mathbf{K},\mathbf{V})=\mathrm{Softmax}(\frac{\overline{\mathbf{Q}}\mathbf{K}^\top}{\sqrt{d}})\mathbf{V}$$

Where $\overline{\mathrm{Q}}$ is a sparse matrix of the same size of q and it only contains the Top-u queries under the sparsity measurement $M(q,K)$.

Algorithm 1 ProbSparse self-attention

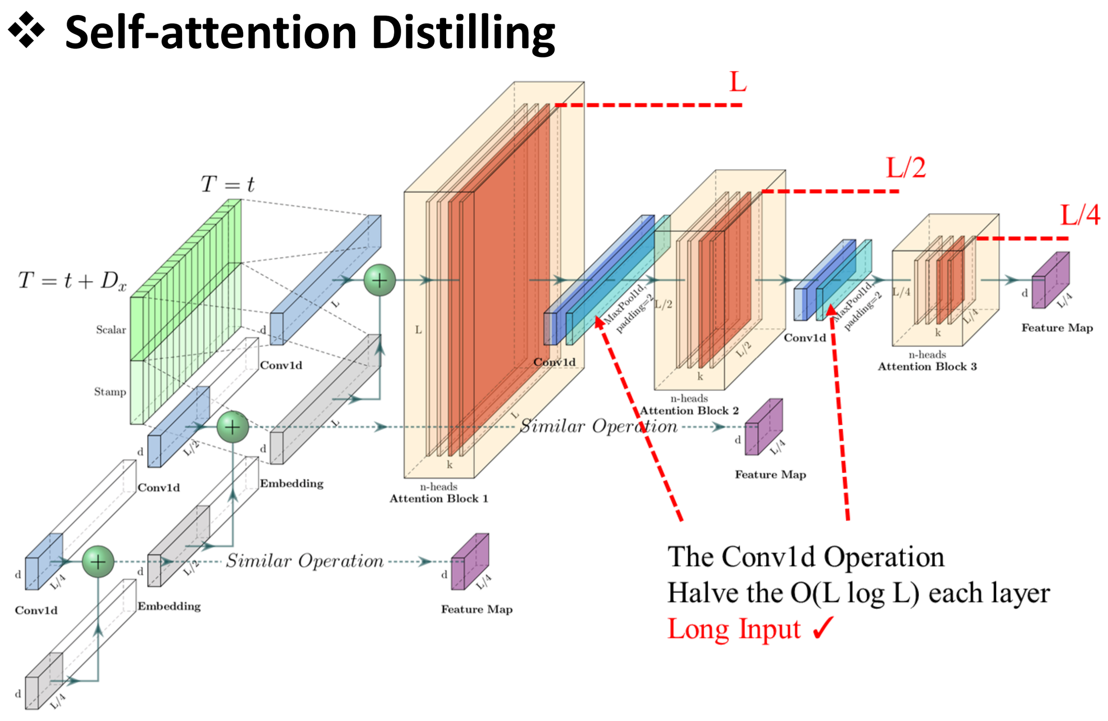
Input: Tensor $\mathbf{Q}\in\mathbb{R}^{m\times d}$, $\mathbf{K}\in\mathbb{R}^{n\times d}$, $\mathbf{V}\in\mathbb{R}^{n\times d}$

1: **initialize:** set hyperparameter c, $u=c\ln m$ and $U=m\ln n$
2: randomly select U dot-product pairs from $\mathbf{K}$ as $\bar{\mathbf{K}}$
3: set the sample score $\bar{\mathbf{S}}=\mathbf{Q}\bar{\mathbf{K}}^\top$
4: compute the measurement $M=\max(\bar{\mathbf{S}})-\mathrm{mean}(\bar{\mathbf{S}})$ by row
5: set Top-u queries under M as $\bar{\mathbf{Q}}$
6: set $\mathbf{S}_1=\mathrm{softmax}(\bar{\mathbf{Q}}\mathbf{K}^\top/\sqrt{d})\cdot\mathbf{V}$
7: set $\mathbf{S}_0=\mathrm{mean}(\mathbf{V})$
8: set $\mathbf{S}=\{\mathbf{S}_1,\mathbf{S}_0\}$ by their original rows accordingly

Output: self-attention feature map $\mathbf{S}$.

Sample $\boldsymbol{L\log L}$ dot-product pairs $\rightarrow\boldsymbol{\mathcal{O}(L\log L)}$

Self-attention Distilling



Generative-style Decoder

$$\mathbf{X}_{\mathrm{feed\_de}}^t=\mathrm{Concat}(\mathbf{X}_{\mathrm{token}}^t,\mathbf{X}_{\mathbf{0}}^t)\in\mathbb{R}^{(L_{\mathrm{token}}+L_y)\times d_{\mathrm{model}}}$$

Predicts all the outputs by one forward procedure.

- $\mathrm{X}_{\mathrm{token}}^t$: **Generative start token**, instead of choosing a specific flag as the token, we sample a "shorter" long sequence in input sequence, which is an earlier slice before output sequence.

Experiment

Univariate Time-series Forecasting

Table 1: Univariate long sequence time-series forecasting results on four datasets (five cases)

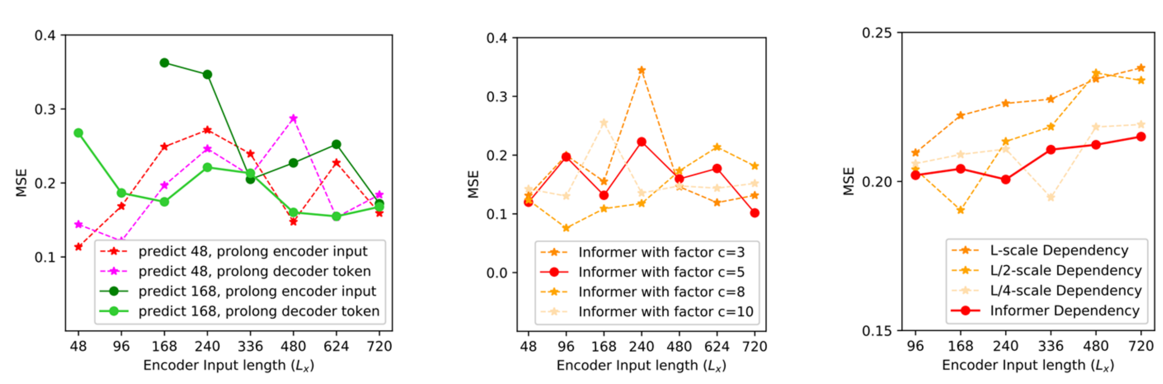
Methods	Metric	ETTh$_1$ 24	48	168	336	720	ETTh$_2$ 24	48	168	336	720	ETTm$_1$ 24	48	96	288	672	Weather 24	48	168	336	720	ECL 48	168	336	720	960	count
Informer	MSE	0.062	**0.108**	**0.146**	0.208	**0.193**	**0.079**	**0.103**	**0.143**	0.171	0.184	**0.051**	0.092	0.119	**0.181**	**0.204**	0.107	**0.164**	**0.226**	**0.241**	**0.259**	0.335	0.408	0.451	**0.466**	**0.470**	28
	MAE	0.178	0.245	**0.294**	0.363	0.365	**0.206**	**0.240**	**0.296**	0.327	0.339	**0.153**	0.217	0.249	**0.320**	**0.345**	0.223	**0.282**	**0.338**	**0.352**	**0.367**	0.423	0.466	0.488	**0.499**	**0.520**	
Informer†	MSE	**0.046**	0.129	0.183	**0.189**	0.201	0.083	0.111	0.154	**0.166**	**0.181**	0.054	**0.087**	**0.115**	0.182	0.207	0.107	0.167	0.237	0.252	0.263	0.304	0.416	0.479	0.482	0.538	14
	MAE	**0.152**	0.274	0.337	**0.346**	**0.357**	0.213	0.249	0.306	**0.323**	**0.338**	0.160	**0.210**	**0.248**	0.323	0.353	**0.220**	0.284	0.352	0.366	0.374	0.404	0.478	0.508	0.515	0.560	
LogTrans	MSE	0.059	0.111	0.155	0.196	0.217	0.080	0.107	0.176	0.175	0.185	0.061	0.156	0.229	0.362	0.450	0.120	0.182	0.267	0.299	0.274	0.360	0.410	0.482	0.522	0.546	0
	MAE	0.191	0.263	0.309	0.370	0.379	0.221	0.262	0.344	0.345	0.349	0.192	0.322	0.397	0.512	0.582	0.247	0.312	0.387	0.416	0.387	0.455	0.481	0.521	0.551	0.563	
Reformer	MSE	0.172	0.228	1.460	1.728	1.948	0.235	0.434	0.961	1.532	1.862	0.055	0.229	0.854	0.962	1.605	0.197	0.268	0.590	1.692	1.887	0.917	1.635	3.448	4.745	6.841	0
	MAE	0.319	0.395	1.089	0.978	1.226	0.369	0.505	0.797	1.060	1.543	0.170	0.340	0.675	1.107	1.312	0.329	0.381	0.552	0.945	1.352	0.840	1.515	2.088	3.913	4.913	
LSTMa	MSE	0.094	0.175	0.210	0.556	0.635	0.135	0.172	0.359	0.516	0.562	0.099	0.289	0.255	0.480	0.988	**0.107**	0.166	0.305	0.404	0.784	0.475	0.703	1.186	1.473	1.493	1
	MAE	0.232	0.322	0.352	0.644	0.704	0.275	0.318	0.470	0.548	0.613	0.201	0.371	0.370	0.528	0.805	0.222	0.298	0.404	0.476	0.709	0.509	0.617	0.854	0.910	0.9260	
DeepAR	MSE	0.089	0.126	0.213	0.403	0.614	0.080	0.125	0.179	0.568	0.367	0.075	0.197	0.336	0.908	2.371	0.108	0.177	0.259	0.535	0.407	**0.188**	**0.295**	**0.388**	0.471	0.583	6
	MAE	0.242	0.291	0.382	0.496	0.643	0.229	0.283	0.346	0.555	0.488	0.205	0.332	0.450	0.739	1.256	0.242	0.313	0.397	0.580	0.506	**0.317**	**0.398**	**0.471**	0.507	0.583	
ARIMA	MSE	0.086	0.133	0.364	0.428	0.613	3.538	3.168	2.768	2.717	2.822	0.074	0.157	0.242	0.424	0.565	0.199	0.247	0.471	0.678	0.996	0.861	1.014	1.102	1.213	1.322	1
	MAE	0.190	**0.242**	0.456	0.537	0.684	0.407	0.440	0.555	0.680	0.952	0.168	0.274	0.357	0.500	0.605	0.321	0.375	0.541	0.666	0.853	0.726	0.797	0.834	0.883	0.908	
Prophet	MSE	0.093	0.150	1.194	1.509	2.685	0.179	0.284	2.113	2.052	3.287	0.102	0.117	0.146	0.414	2.671	0.280	0.421	2.409	1.931	3.759	0.506	2.711	2.220	4.201	6.827	0
	MAE	0.241	0.300	0.721	1.766	3.155	0.345	0.428	1.018	2.487	4.592	0.256	0.273	0.304	0.482	1.112	0.403	0.492	1.092	2.406	1.030	0.557	1.239	3.029	1.363	4.184	

Multivariate Time-series Forecasting

Table 2: Multivariate long sequence time-series forecasting results on four datasets (five cases)

Methods	Metric	ETTh$_1$ 24	48	168	336	720	ETTh$_2$ 24	48	168	336	720	ETTm$_1$ 24	48	96	288	672	Weather 24	48	168	336	720	ECL 48	168	336	720	960	count
Informer	MSE	**0.509**	**0.551**	**0.878**	0.884	**0.941**	**0.446**	**0.934**	1.512	1.665	**2.340**	0.325	0.472	**0.642**	1.219	**1.651**	**0.353**	**0.464**	**0.592**	**0.623**	0.685	0.269	0.300	0.311	**0.308**	**0.328**	32
	MAE	**0.523**	**0.563**	**0.722**	0.753	**0.768**	**0.523**	**0.733**	**0.996**	**1.035**	**1.209**	0.440	0.537	0.626	**0.871**	**1.002**	**0.381**	**0.455**	**0.531**	**0.546**	0.575	0.351	0.376	0.385	**0.385**	**0.406**	
Informer†	MSE	0.550	0.602	0.893	**0.836**	0.981	0.683	0.977	1.873	1.374	2.493	**0.324**	**0.446**	0.651	1.342	1.661	0.355	0.471	0.613	0.626	**0.680**	0.269	**0.281**	0.309	0.314	0.356	12
	MAE	0.551	0.581	0.733	**0.729**	0.779	0.637	0.793	1.094	0.935	1.253	**0.440**	**0.508**	**0.616**	0.927	1.001	0.383	0.456	0.544	0.548	**0.569**	0.351	**0.366**	**0.383**	0.388	0.394	
LogTrans	MSE	0.656	0.670	0.888	0.942	1.109	0.726	1.728	3.944	3.711	2.817	0.341	0.495	0.674	1.728	1.865	0.365	0.496	0.649	0.666	0.741	**0.267**	0.290	**0.305**	0.311	0.333	2
	MAE	0.600	0.611	0.766	0.766	0.843	0.638	0.944	1.573	1.587	1.356	0.495	0.527	0.674	1.656	1.721	0.405	0.485	0.573	0.584	0.611	0.366	0.382	0.395	0.397	0.413	
Reformer	MSE	0.887	1.159	1.686	1.919	2.177	1.381	1.715	4.484	3.798	5.111	0.598	0.952	1.267	1.632	1.943	0.583	0.633	1.228	1.770	2.548	1.312	1.453	1.507	1.883	1.973	0
	MAE	0.630	0.750	0.996	1.090	1.218	1.475	1.585	1.650	1.508	1.793	0.489	0.645	0.795	0.886	1.006	0.497	0.556	0.763	0.997	1.407	0.911	0.975	0.978	1.002	1.185	
LSTMa	MSE	0.536	0.616	1.058	1.152	1.682	1.049	1.331	3.987	3.276	3.711	0.511	1.280	1.195	1.598	2.530	0.476	0.763	0.948	1.497	1.314	0.388	0.492	0.778	1.528	1.343	0
	MAE	0.528	0.577	0.725	0.794	1.018	0.689	0.805	1.560	1.375	1.520	0.517	0.819	0.785	0.952	1.259	0.464	0.589	0.713	0.889	0.875	0.444	0.498	0.629	0.945	0.886	
LSTnet	MSE	1.175	1.344	1.865	2.477	1.925	2.632	3.487	**1.442**	**1.372**	2.403	1.856	1.909	2.654	**1.009**	1.681	0.575	0.622	0.676	0.714	0.773	0.279	0.318	0.357	0.442	0.473	4
	MAE	0.793	0.864	1.092	1.193	1.084	1.337	1.577	2.389	2.429	3.403	1.058	1.085	1.378	1.902	2.701	0.507	0.553	0.585	0.607	0.643	**0.337**	0.368	0.391	0.433	0.443	

Parameter Sensitivity



(a) Input length. (b) Sampling. (c) Stacking.

Ablation study

Table 3: Ablation of *ProbSparse* mechanism

Prediction length		336			720		
Encoder's input		336	720	1440	720	1440	2880
Informer	MSE	0.243	0.225	0.212	0.258	0.238	0.224
	MAE	0.487	0.404	0.381	0.503	0.399	0.387
Informer†	MSE	0.214	0.205	-	0.235	-	-
	MAE	0.369	0.364	-	0.401	-	-
LogTrans	MSE	0.256	0.233	-	0.264	-	-
	MAE	0.496	0.412	-	0.523	-	-
Reformer	MSE	1.848	1.832	1.817	2.094	2.055	2.032
	MAE	1.054	1.027	1.010	1.363	1.306	1.334

[1] Informer† uses the canonical self-attention mechanism.
[2] The '-' indicates failure for out-of-memory.

Table 4: Ablation of Self-attention Distilling

Prediction length		336					480				
Encoder's input		336	480	720	960	1200	336	480	720	960	1200
Informer†	MSE	0.201	0.175	0.215	0.185	0.172	0.136	0.213	0.178	0.146	0.121
	MAE	0.360	0.335	0.366	0.355	0.321	0.282	0.382	0.345	0.296	0.272
Informer‡	MSE	0.187	0.182	0.177	-	-	0.208	0.182	0.168	-	-
	MAE	0.330	0.341	0.329	-	-	0.384	0.337	0.304	-	-

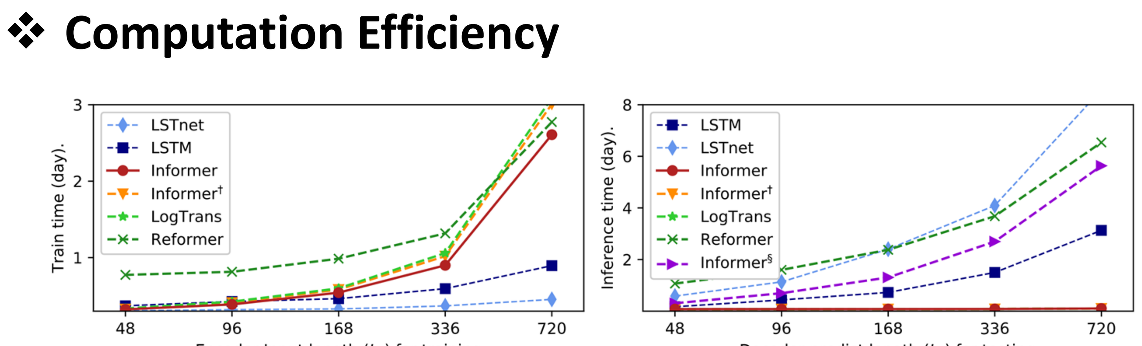
[1] Informer‡ removes the self-attention distilling from Informer†.
[2] The '-' indicates failure for out-of-memory.

Table 5: Ablation of Generative Style Decoder

Prediction length		336				480			
Prediction offset		+0	+12	+24	+48	+0	+48	+96	+168
Informer‡	MSE	0.101	0.102	0.103	0.103	0.155	0.158	0.160	0.165
	MAE	0.215	0.218	0.223	0.227	0.317	0.397	0.399	0.406
Informer§	MSE	0.152	-	-	-	0.462	-	-	-
	MAE	0.294	-	-	-	0.595	-	-	-

[1] Informer§ replaces our decoder with dynamic decoding one in Informer‡.
[2] The '-' indicates failure for the unacceptable metric results.

Computation Efficiency



论文预印版本&补充材料：https://arxiv.org/abs/2012.07436

论文项目地址 开源数据集地址 北航ACTBD公众号

主办方：中国中文信息学会青年工作委员会

承办方：智源社区