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

Haoyi Zhou, AAAI 2021

妥布雷素

NCUT

May, 2022

1. Introduction
2. Informer
3. Experiment

研究背景与研究意义

研究背景:

- 长时间序列预测（long sequence time-series forecasting, LSTF）使得基于预测的特征变化或事件成为可能，广泛用于各行各业。
- 然而，Vanilla Transformer 有三个缺陷:
  - 时间复杂度上，自注意力需要 2 阶计算 $O(L^2)$
  - 空间复杂度上，J 个 encoder/decoder 堆叠层需要 $O(J * L^2)$
  - 训练速度与长度反比。

研究背景与研究意义

研究背景：

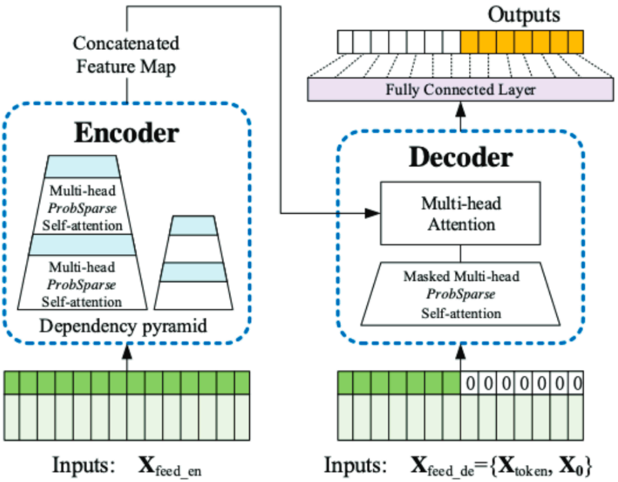
- 长时间序列预测（long sequence time-series forecasting，LSTF）使得基于预测的特征变化或事件成为可能，广泛用于各行各业。
- 然而，Vanilla Transformer 有三个缺陷：
  - 时间复杂度上，自注意力需要 2 阶计算 $O(L^2)$
  - 空间复杂度上，J 个 encoder/decoder 堆叠层需要 $O(J*L^2)$
  - 训练速度与长度反比。

研究意义：

- 提出了 Informer，提高了 Transformer 类模型的学习长时间依赖的能力
- 提出了 ProbSparse 自注意力机制，时间复杂度降低到 $O(L*logL)$，空间复杂度降低到 $O(L*logL)$
- 提出自注意力蒸馏操作，使得总空间复杂度 $O((2-\epsilon)L*logL)$
- 提出 Generative Style Decoder，降低计算量，同时避免累计

1. Introduction
2. Informer
3. Experiment

模型概览


Concatenated
Feature Map
Outputs
Fully Connected Layer
Encoder
Multi-head
ProbSparse
Self-attention
Multi-head
ProbSparse
Self-attention
Dependency pyramid
Decoder
Multi-head
Attention
Masked Multi-head
ProbSparse
Self-attention
0 0 0 0 0 0 0
Inputs: $\mathbf{X}_{\text{feed\_en}}$
Inputs: $\mathbf{X}_{\text{feed\_de}}=\{\mathbf{X}_{\text{token}}, \mathbf{X}_{\mathbf{0}}\}$

问题阐述

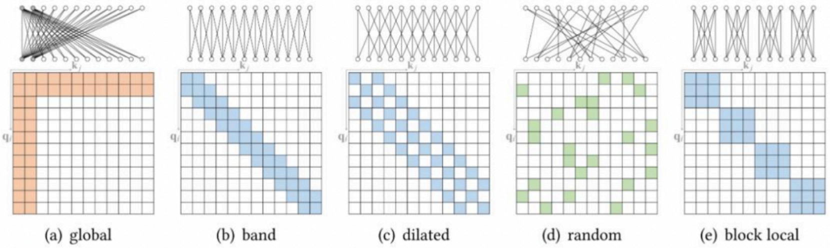
在 t 时刻，已知序列，$\mathcal{X}^t = \left\{\mathbf{x}_1^t, \ldots, \mathbf{x}_{L_x}^t \mid \mathbf{x}_i^t \in \mathbb{R}^{d_x}\right\}$
求 $\mathcal{Y}^t = \left\{\mathbf{y}_1^t, \ldots, \mathbf{y}_{L_u}^t \mid \mathbf{y}_i^t \in \mathbb{R}^{d_y}\right\}$

Encoder-decoder architecture

1. “encode” the input representations $\mathcal{X}^t$ into a hidden state representations $\mathcal{H}^t$
2. “decode” an output representations $\mathcal{Y}^t$ from $\mathcal{H}^t = \left\{\mathbf{h}_1^t, \ldots, \mathbf{h}_{L_h}^t\right\}$

注意力机制

基于位置的稀疏 attention 之一是原子稀疏 attention，如下图所示主要有五种模式。彩色方块表示计算的 attention 分数，空白方块表示放弃的 attention 分数。



ProbSparse 自注意力机制

普通的自注意力机制: $\mathcal{A}\left(\mathbf{q}_i, \mathbf{K}, \mathbf{V}\right)=\sum_j \frac{k\left(\mathbf{q}_i, \mathbf{k}_j\right)}{\sum_l k\left(\mathbf{q}_i, \mathbf{k}_l\right)} \mathbf{v}_j=\mathbb{E}_{p\left(\mathbf{k}_j \mid \mathbf{q}_i\right)}\left[\mathbf{v}_j\right]$

其中： $p\left(\mathbf{k}_j \mid \mathbf{q}_i\right)=\frac{k\left(\mathbf{q}_i, \mathbf{k}_j\right)}{\sum_l k\left(\mathbf{q}_i, \mathbf{k}_l\right)}$。 $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{L * d}$

ProbSparse 自注意力机制

普通的自注意力机制: $\mathcal{A}\left(\mathbf{q}_i,\mathbf{K},\mathbf{V}\right)=\sum_j \frac{k(\mathbf{q}_i,\mathbf{k}_j)}{\sum_l k(\mathbf{q}_i,\mathbf{k}_l)}\mathbf{v}_j=\mathbb{E}_{p(\mathbf{k}_j|\mathbf{q}_i)}\left[\mathbf{v}_j\right]$

其中: $p\left(\mathbf{k}_j \mid \mathbf{q}_i\right)=\frac{k(\mathbf{q}_i,\mathbf{k}_j)}{\sum_l k(\mathbf{q}_i,\mathbf{k}_l)}$。 $\mathbf{Q},\mathbf{K},\mathbf{V}\in\mathbb{R}^{L*d}$

ProbSparse: $\mathcal{A}(\mathbf{Q},\mathbf{K},\mathbf{V})=\operatorname{Softmax}\left(\frac{\overline{\mathbf{Q}}\mathbf{K}^{\top}}{\sqrt{d}}\right)\mathbf{V}$

其中，稀疏矩阵 $\overline{\mathbf{Q}}\in\mathbb{R}^{u*d}$，包含了 u 个最大的 $M(\mathbf{q},\mathbf{K})$。

基于 Kullback-Leibler divergence 评估概率，

$KL(q\|p)=\ln\sum_{l=1}^{L_K}e^{\mathbf{q}_i\mathbf{k}_l^{\top}/\sqrt{d}}-\frac{1}{L_K}\sum_{j=1}^{L_K}\mathbf{q}_i\mathbf{k}_j^{\top}/\sqrt{d}-\ln L_K$，作者定义稀疏性度量，$M\left(\mathbf{q}_i,\mathbf{K}\right)=\ln\sum_{j=1}^{L_K}e^{\frac{\mathbf{q}_i\mathbf{k}_j^{\top}}{\sqrt{d}}}-\frac{1}{L_K}\sum_{j=1}^{L_K}\frac{\mathbf{q}_i\mathbf{k}_j^{\top}}{\sqrt{d}}$。

ProbSparse 自注意力机制

普通的自注意力机制: $\mathcal{A}(\mathbf{q}_i, \mathbf{K}, \mathbf{V}) = \sum_j \frac{k(\mathbf{q}_i,\mathbf{k}_j)}{\sum_l k(\mathbf{q}_i,\mathbf{k}_l)} \mathbf{v}_j = \mathbb{E}_{p(\mathbf{k}_j|\mathbf{q}_i)}[\mathbf{v}_j]$

其中: $p(\mathbf{k}_j \mid \mathbf{q}_i) = \frac{k(\mathbf{q}_i,\mathbf{k}_j)}{\sum_l k(\mathbf{q}_i,\mathbf{k}_l)}$。 $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{L*d}$

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

其中，稀疏矩阵 $\overline{\mathbf{Q}} \in \mathbb{R}^{u*d}$，包含了 u 个最大的 $M(\mathbf{q}, \mathbf{K})$。

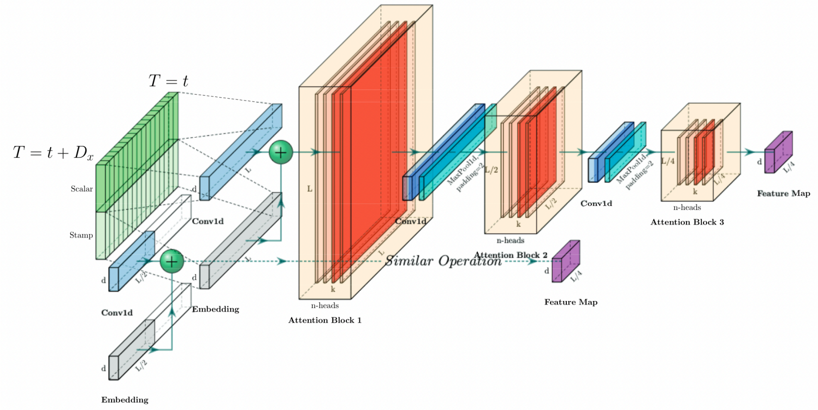
基于 Kullback-Leibler divergence 评估概率，

$KL(q\|p) = \ln \sum_{l=1}^{L_K} e^{\mathbf{q}_i\mathbf{k}_l^\top/\sqrt{d}} - \frac{1}{L_K}\sum_{j=1}^{L_K} \mathbf{q}_i\mathbf{k}_j^\top/\sqrt{d} - \ln L_K$，作者定义稀疏性度量，$M(\mathbf{q}_i, \mathbf{K}) = \ln \sum_{j=1}^{L_K} e^{\frac{\mathbf{q}_i\mathbf{k}_j^\top}{\sqrt{d}}} - \frac{1}{L_K}\sum_{j=1}^{L_K} \frac{\mathbf{q}_i\mathbf{k}_j^\top}{\sqrt{d}}$。

近似稀疏性度量，$\bar{M}(\mathbf{q}_i, \mathbf{K}) = \max_j\left\{\frac{\mathbf{q}_i\mathbf{k}_j^\top}{\sqrt{d}}\right\} - \frac{1}{L_K}\sum_{j=1}^{L_K} \frac{\mathbf{q}_i\mathbf{k}_j^\top}{\sqrt{d}}$

设 $u = c * (\mathcal{L}\backslash L)$，则时间复杂度，空间复杂度为 $\mathcal{O}(L \ln L)$

编码器

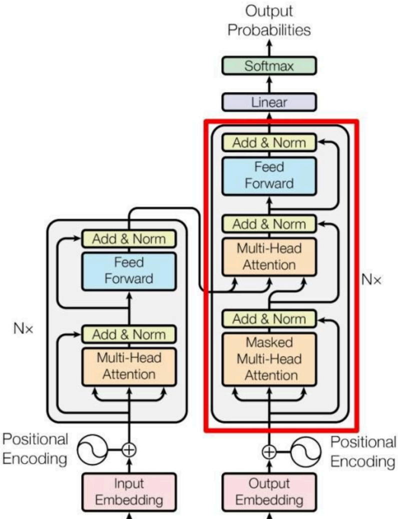


Self-attention Distilling: $\mathbf{X}_{j+1}^{t} = \mathrm{MaxPool}\left(\mathrm{ELU}\left(\mathrm{Conv\,1\,d}\left(\left[\mathbf{X}_{j}^{t}\right]_{\mathrm{AB}}\right)\right)\right)$

Self-attention Distilling reduce space complexity to be $\mathcal{O}((2-\epsilon)L\log L)$。

$\mathbf{X}_{feed}^{t} \in \mathbb{R}^{L_x \times d_{\text{model}}}$

解码器



其中:

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

数据集

- ETT (Electricity Transformer Temperature)
- ECL (Electricity Consuming Load)
- Weather (Weather)

数据集	描述	时间范围	采样频率
ETT	中国两个县的 2 年数据	2 年	1 小时/2 小时/15 分
ECL	321 个客户的电力消耗	2 年	1 小时
Weather	1,600 U.S. locations	4 年	1 小时

表 1: 数据集描述

单维时间序列预测

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

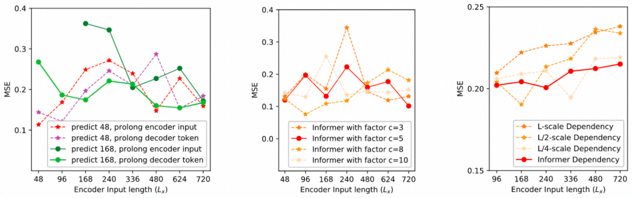
Methods	Metric	$ETTh_1$					$ETTh_2$					$ETTm_1$					Weather					ECL					count
		24	48	168	336	720	24	48	168	336	720	24	48	96	288	672	24	48	168	336	720	48	168	336	720	960	
Informer	MSE	0.062	**0.108**	**0.146**	0.208	**0.193**	**0.079**	**0.103**	**0.143**	0.171	0.184	**0.051**	0.092	0.119	**0.181**	**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	

多维时间序列预测

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

Methods	Metric	$ETTh_1$					$ETTh_2$					$ETTm_1$					Weather					ECL					count
		24	48	168	336	720	24	48	168	336	720	24	48	96	288	672	24	48	168	336	720	48	168	336	720	960	
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	

参数调优



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

Figure 4: The parameters' sensitivity of Informer.

消融实验

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.

消融实验

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.

消融实验

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.

消融实验

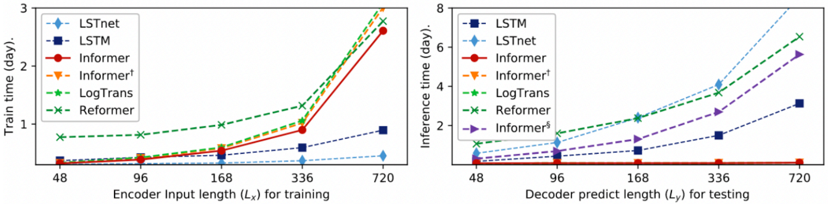


Figure 5: The total runtime of training/testing phase.

消融实验

Table 6: L-related computation statics of each layer

Methods	Training		Testing
	Time Complexity	Memory Usage	Steps
Informer	$\mathcal{O}(L \log L)$	$\mathcal{O}(L \log L)$	1
Transformer	$\mathcal{O}(L^2)$	$\mathcal{O}(L^2)$	L
LogTrans	$\mathcal{O}(L \log L)$	$\mathcal{O}(L^2)$	$1^\star$
Reformer	$\mathcal{O}(L \log L)$	$\mathcal{O}(L \log L)$	L
LSTM	$\mathcal{O}(L)$	$\mathcal{O}(L)$	L

[1] The LSTnet is hard to have a closed form.

Thank You