西南財經大學
SOUTHWESTERN UNIVERSITY OF FINANCE AND ECONOMICS

A TIME SERIES IS WORTH 64 WORDS: LONG-TERM FORECASTING WITH TRANSFORMERS

ICLR 2023

主讲人：陈凯伦
2024.3.27

A TIME SERIES IS WORTH 64 WORDS: LONG-TERM FORECASTING WITH TRANSFORMERS

文章背景介绍:

- LTSF-Linear论文提出对Transformer预测时间序列能力的质疑，并在文章中做出了充分的实验论证，并从多个角度说明Transformer的预测能力甚至可能不如传统的线性模型。

- 2023年初，PatchTST的提出打破了这一质疑，并迅速成为后续论文中的用于实验对比的时序预测领域最新的SOTA模型，目前从谷歌学术搜索结果来看，改论文引用次数已经达到299次。 ㄲ 引用 被引用次数：299 相关文章 所有 5 个版本

- 文章知名度高、效果显著，但文中模型实际使用的方法并不复杂，主体的Patch方法类似于VIT对图像的分块处理，文中也写到：PatchTST没有延续过去改变Transformer架构的研究思路，而保留了最纯粹的Transformer架构。

点此简单了解
LTSF-Linear

01 Introduction

Introduction

targets：

1. 多变量时间序列预测

2. 自监督表示学习

给出输入多变量时间序列:

$$L : (x_1, ..., x_L)$$

预测未来指定时间步内的序列:

$$(x_{L+1}, ..., x_{L+T}).$$

channel-independence patch time series Transformer (PatchTST)

1. **Patching***：** 输入数据不再是简单的单个时间步，而是时间序列进行分割后的子序列（patches/segments） -patch is all your need

2. **Channel-independence：** 每个通道只输入单变量时间序列

3. **self-supervised learning：** 通过掩码的方式，让模型学习输入时间序列中的特征（作为辅助任务）

Introduction

advantages：

1.对比于传统的Transformer架构模型，大幅降低了空间复杂度和时间复杂度

2.扩大了回顾窗口（输入时间序列）的长度，提高预测性能

3.采用自监督学习，提高模型整体预测性能

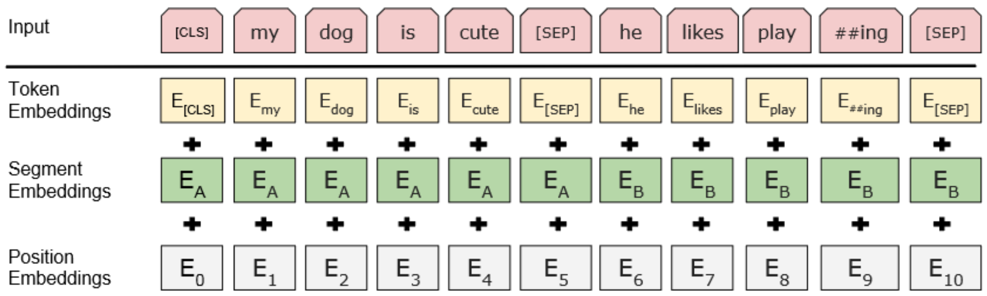
Models	L	N	patch	method	MSE
Channel-independent PatchTST	96	96			0.518
	380	96		down-sampled	0.447
	336	336			0.397
	336	42	✓		0.367
	336	42	✓	self-supervised	**0.349**
Channel-mixing FEDFormer	336	336			0.597
DLinear	336	336			0.410

Running time (s) with $L = 336$			
Dataset	w. patch	w.o. patch	Gain
Traffic	464	10040	x 22
Electricity	300	5730	x 19
Weather	156	680	x 4

Introduction

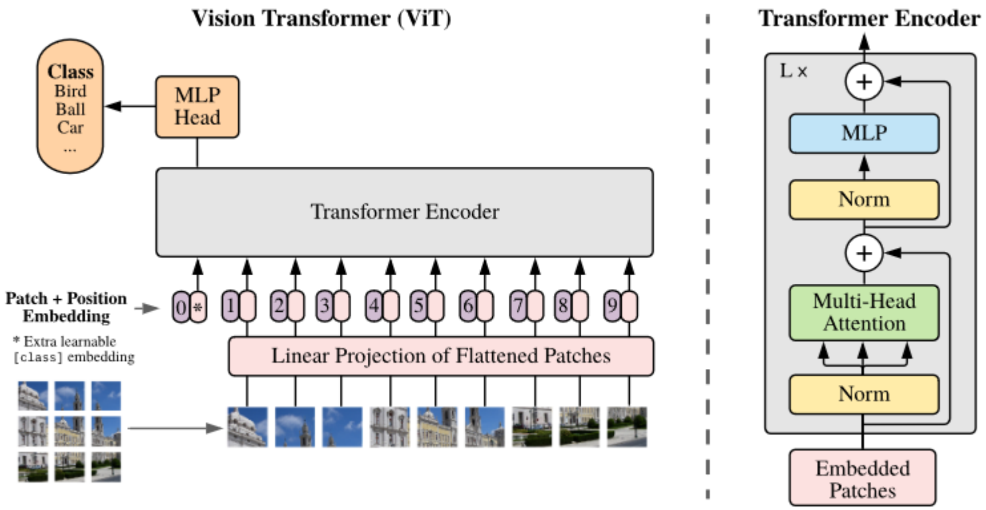
related work/相关知识补充——Patch

NLP领域Patch方法的示例:



Input/Output Representations To make BERT handle a variety of down-stream tasks, our input representation is able to unambiguously represent both a single sentence and a pair of sentences (e.g., ⟨ Question, Answer ⟩) in one token sequence. Throughout this work, a "sentence" can be an arbitrary span of contiguous text, rather than an actual linguistic sentence. A "sequence" refers to the input token sequence to BERT, which may be a single sentence or two sentences packed together.

CV领域Patch方法的示例:



An overview of the model is depicted in Figure 1. The standard Transformer receives as input a 1D sequence of token embeddings. To handle 2D images, we reshape the image $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$ into a sequence of flattened 2D patches $\mathbf{x}_p \in \mathbb{R}^{N \times (P^2 \cdot C)}$, where (H, W) is the resolution of the original image, C is the number of channels, (P, P) is the resolution of each image patch, and $N = HW/P^2$ is the resulting number of patches, which also serves as the effective input sequence length for the Transformer. The Transformer uses constant latent vector size D through all of its layers, so we flatten the patches and map to D dimensions with a trainable linear projection (Eq. 1). We refer to the output of this projection as the patch embeddings.

Introduction

related work/相关知识补充——Channel Independence

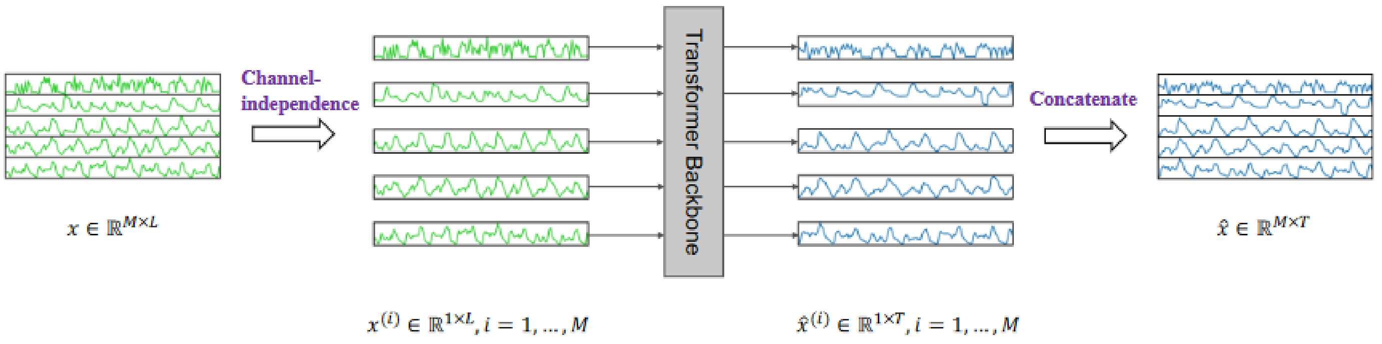
通道独立（Channel-Independence）和**通道融合**（Channel-mixing）

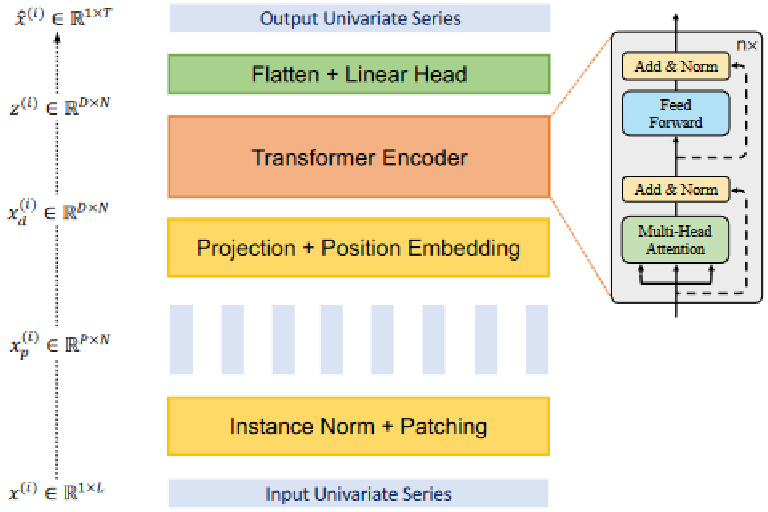
通道混合强调不同通道之间的相关性和交互性，提高模型的表达能力和泛化能力

- **通道独立：（文中举例）**
  - **CNN：**个人理解其通道独立体现在卷积核会分为三个通道对RGB分别提取特征/卷积核只会局部感受野内的信息
  - **Linear：**即全连接层，个人理解其通道独立性出现在不会出现$\alpha_{1,2}x_1x_2$这种体现不同特征交互的项出现。
- **通道融合：**
  - **Transformer：**每个位置的表示都会与其他所有位置做内积，算相似度/使用多头注意力机制，学习不同的特征，最后进行拼接。

02 Proposed Method

Channel-Independence





(b) Transformer Backbone (Supervised)

通道独立：将原来的M维特征的多变量数据转换为1维M个单变量数据，经过Transformer预测后再拼接

数学表达：L：$(x_1, x_2, \dots\dots, x_L)$

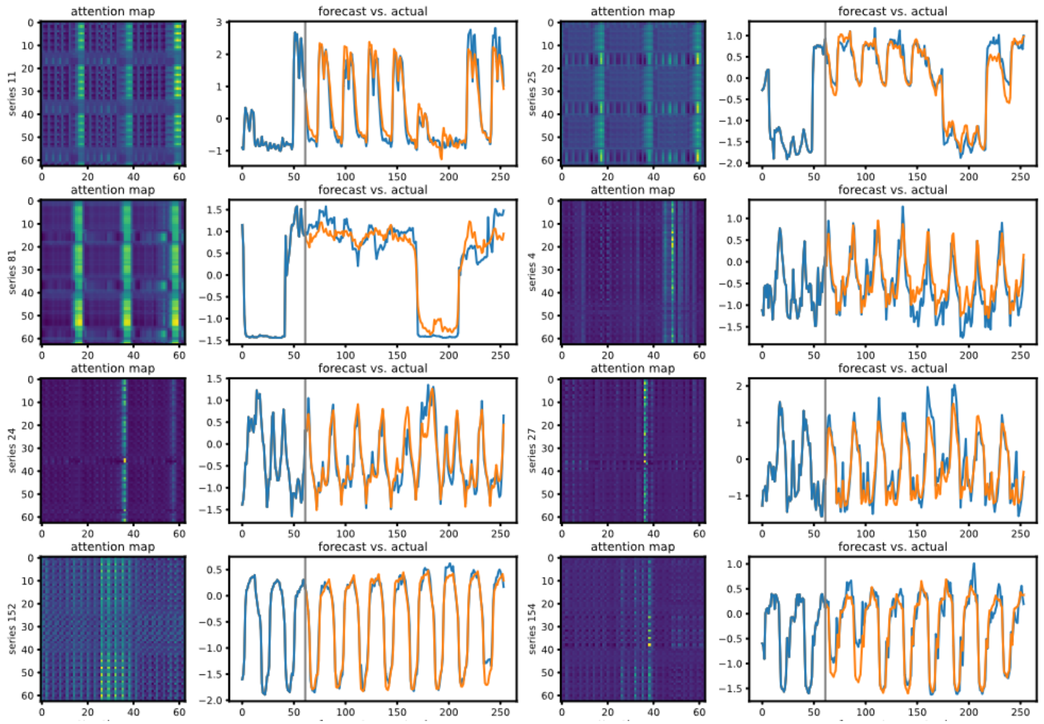
$\Downarrow$

$x_{1:L}^{(i)} = x_1^{(i)}, x_2^{(i)}, \dots\dots, x_L^{(i)}$,其中$i = 1,2, \dots\dots, M$

$\Downarrow$

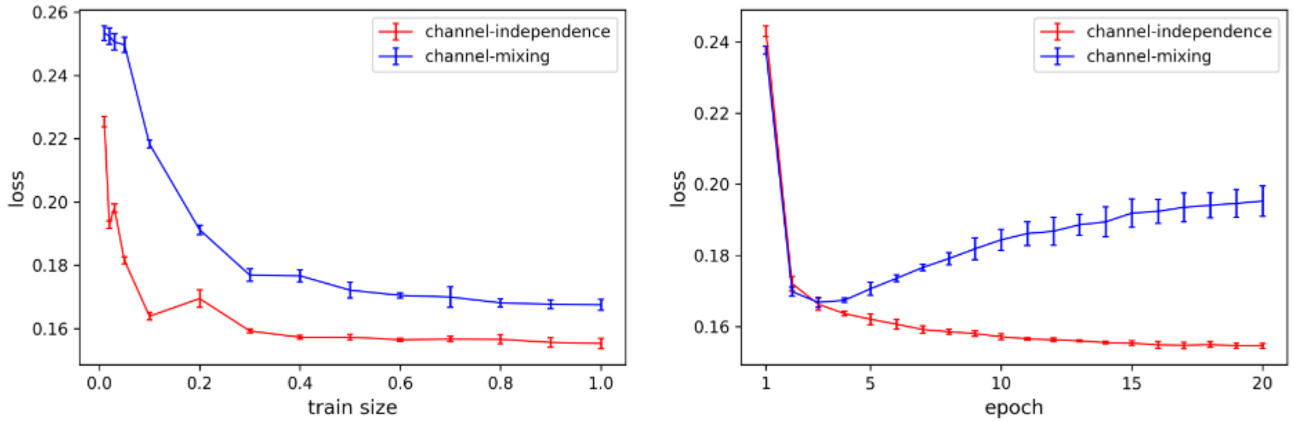
预测结果：$\hat{x}^{(i)} = (\hat{x}_{L+1}^{(i)}, \dots\dots, \hat{x}_{L+T}^{(i)})$

Channel-Independence



通道独立的优点:

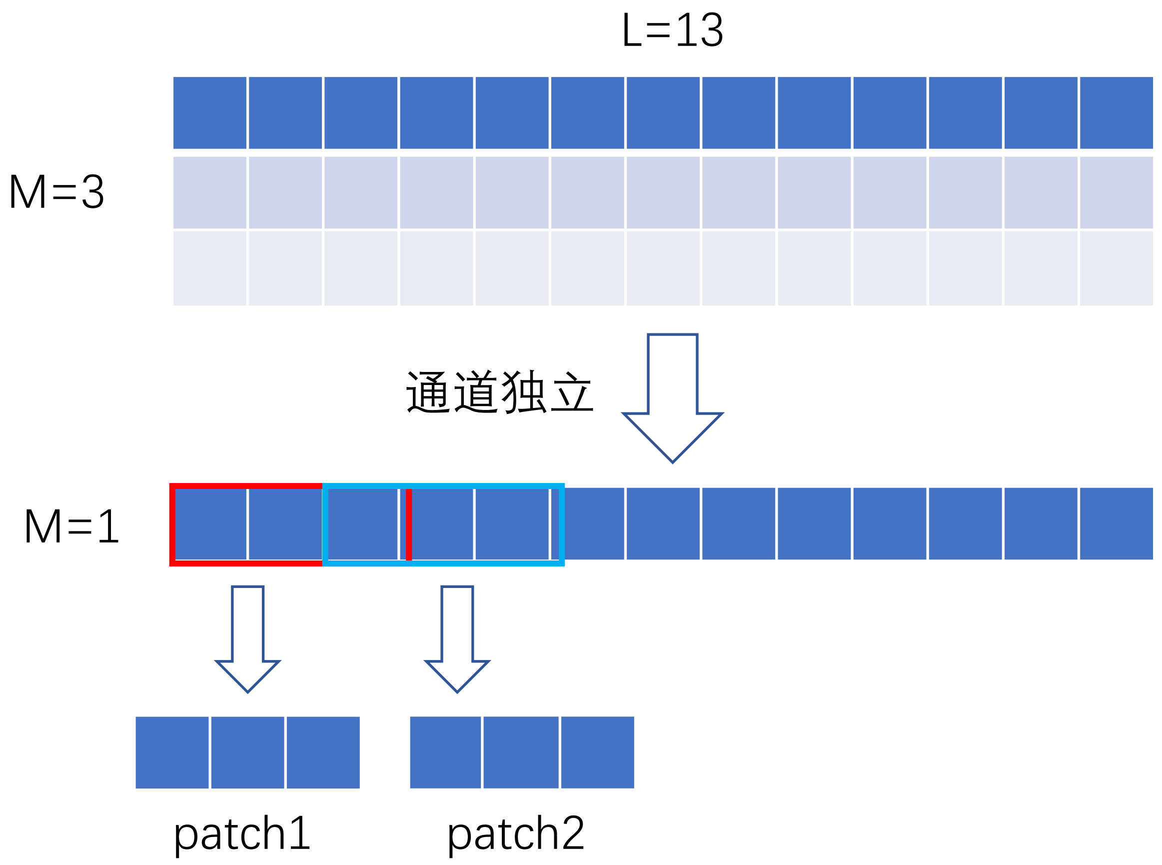
- 可以学到不同变量变化的不同模式
- 在有限的数据集内收敛的更快
- 在有限数据集内训练不容易过拟合



通道独立的缺点:

- 最直观的缺少了不同序列变量之间的交互，对于特定的下游任务很不友好

Patch



patch可分为有重叠部分和无重叠部分，这里展示的是有重叠部分

在这里演示时patch大小为3，滑动步长为2，patch数量的计算公式：

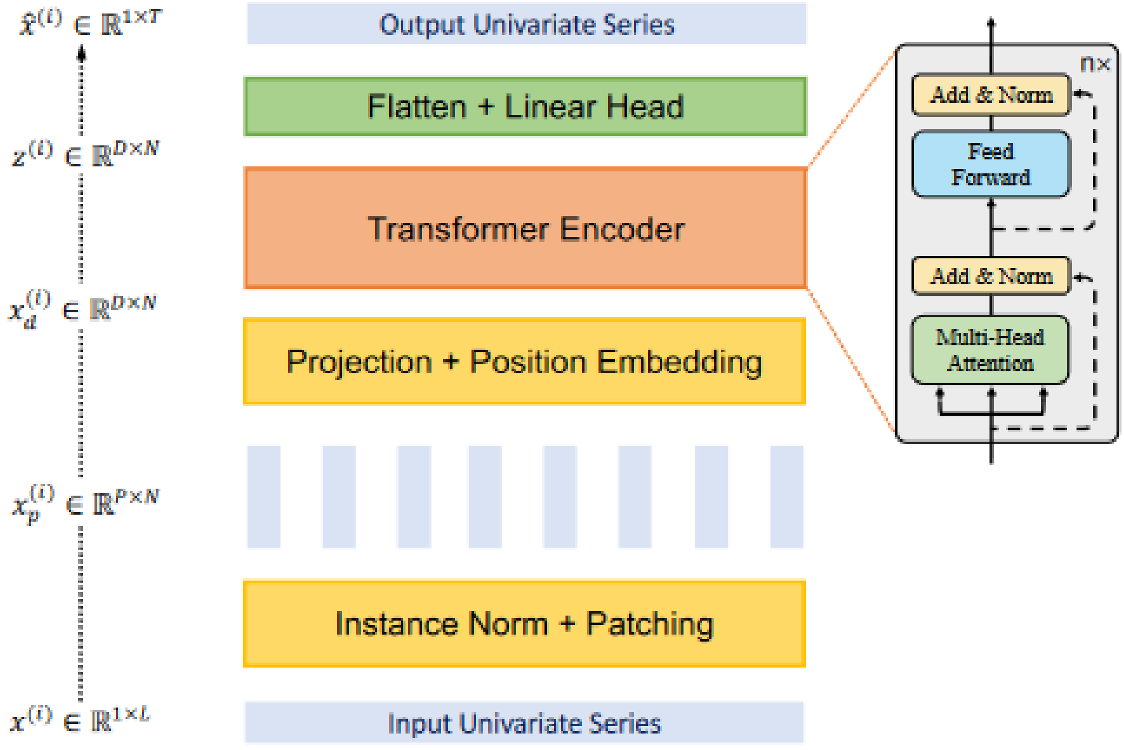
$$N = \lfloor \frac{(L-P)}{S} \rfloor + 2.$$

所以输入序列为：大小为：$x_p^{(i)} \in \mathbb{R}^{P \times N}$

其中P为一个patch中序列长度，N为patch的数量

进而完成了Token数量上的大量减少，从原来的L减少到L/S，时间复杂度和空间复杂度实现了大幅降低，让Transformer模型关注到更长的输入序列成为可能。

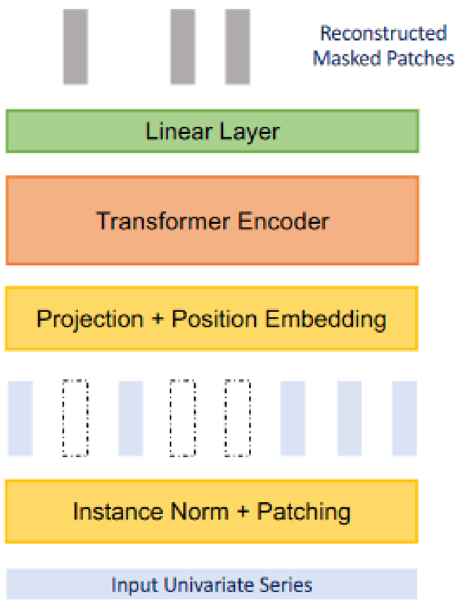
回顾transformer监督学习的整个网络结构



(b) Transformer Backbone (Supervised)

$$\mathcal{L} = \mathbb{E}_{\boldsymbol{x}} \frac{1}{M} \sum_{i=1}^{M} \|\hat{\boldsymbol{x}}_{L+1:L+T}^{(i)} - \boldsymbol{x}_{L+1:L+T}^{(i)}\|_2^2.$$

Representation Learning



(c) Transformer Backbone (Self-supervised)

Pretrain+Finetune

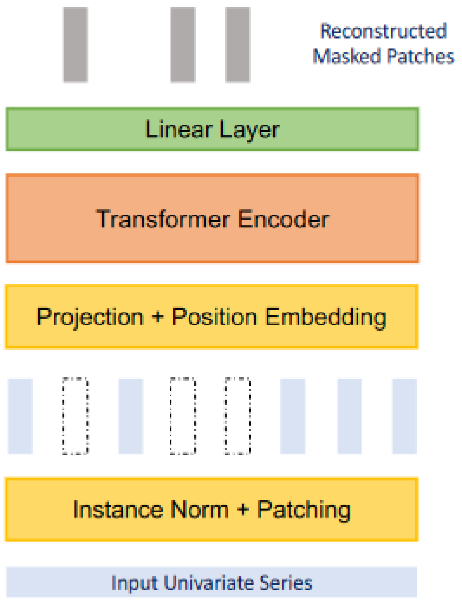
模型特点：

1. 将部分patches进行掩码（此处做patch不重复）
- **区别于之前对于时间步进行掩码：**时间步掩码没有太大意义，因为被掩码的时间步可以通过简单插值就能得出，并不包含太多实际意义，但对patch进行掩码有可能就包含局部（Local）信息

2. 原来的transformer模型预测头被改为了Linear层
- 避免了预测头矩阵规模过大，而导致下游任务样本过少而出现的对预训练样本的过拟合问题

3. 利用均方损失进行训练

Representation Learning



(c) Transformer Backbone (Self-supervised)

Pretrain+Finetune

三种训练方法：

1. 直接用目标数据集训练模型（监督学习）

2. 在目标数据集上进行表示学习训练，得到backbone，然后对在目标训练集上对预测头进行finetune-20epoch

3. 在目标数据集上进行表示学习训练，得到backbone，然后对在目标训练集上对预测头进行finetune-10epoch，最后对整个网络finetune-20epoch

03 Experiment

Experiment：数据集

Datasets	Weather	Traffic	Electricity	ILI	ETTh1	ETTh2	ETTm1	ETTm2
Features	21	862	321	7	7	7	7	7
Timesteps	52696	17544	26304	966	17420	17420	69680	69680

Table 2: Statistics of popular datasets for benchmark.

数据集：包括天气、交通、电力、流感门诊情况、变压器温度前三个为大型数据集

ETT数据集示例：

date	HUFL	HULL	MUFL	MULL	LUFL	LULL	OT
2016/7/1 0:00	5.8270001	2.0090001	1.599	0.462	4.2030001	1.34	30.531
2016/7/1 1:00	5.6929998	2.076	1.492	0.426	4.1420002	1.3710001	27.787001
2016/7/1 2:00	5.1570001	1.7410001	1.279	0.355	3.777	1.2180001	27.787001
2016/7/1 3:00	5.0900002	1.942	1.279	0.391	3.8069999	1.279	25.044001
2016/7/1 4:00	5.3579998	1.942	1.492	0.462	3.868	1.279	21.948
2016/7/1 5:00	5.6259999	2.1429999	1.528	0.533	4.0510001	1.3710001	21.174
2016/7/1 6:00	7.1669998	2.947	2.132	0.782	5.026	1.858	22.792
2016/7/1 7:00	7.4349999	3.2820001	2.3099999	1.031	5.0869999	2.224	23.143999
2016/7/1 8:00	5.559	3.0139999	2.4519999	1.173	2.9549999	1.432	21.667
2016/7/1 9:00	4.5549998	2.5450001	1.919	0.817	2.6800001	1.3710001	17.445999
2016/7/1 10:00	4.9569998	2.5450001	1.99	0.853	2.9549999	1.492	19.979
2016/7/1 11:00	5.7600002	2.5450001	2.2030001	0.853	3.4419999	1.492	20.118999
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Experiment：长时预测

Models		PatchTST/64		PatchTST/42		DLinear		FEDformer		Autoformer		Informer		Pyraformer		LogTrans	
Metric		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Weather	96	0.149	0.198	0.152	0.199	0.176	0.237	0.238	0.314	0.249	0.329	0.354	0.405	0.896	0.556	0.458	0.490
	192	0.194	0.241	0.197	0.243	0.220	0.282	0.275	0.329	0.325	0.370	0.419	0.434	0.622	0.624	0.658	0.589
	336	0.245	0.282	0.249	0.283	0.265	0.319	0.339	0.377	0.351	0.391	0.583	0.543	0.739	0.753	0.797	0.652
	720	0.314	0.334	0.320	0.335	0.323	0.362	0.389	0.409	0.415	0.426	0.916	0.705	1.004	0.934	0.869	0.675
Traffic	96	0.360	0.249	0.367	0.251	0.410	0.282	0.576	0.359	0.597	0.371	0.733	0.410	2.085	0.468	0.684	0.384
	192	0.379	0.256	0.385	0.259	0.423	0.287	0.610	0.380	0.607	0.382	0.777	0.435	0.867	0.467	0.685	0.390
	336	0.392	0.264	0.398	0.265	0.436	0.296	0.608	0.375	0.623	0.387	0.776	0.434	0.869	0.469	0.734	0.408
	720	0.432	0.286	0.434	0.287	0.466	0.315	0.621	0.375	0.639	0.395	0.827	0.466	0.881	0.473	0.717	0.396
Electricity	96	0.129	0.222	0.130	0.222	0.140	0.237	0.186	0.302	0.196	0.313	0.304	0.393	0.386	0.449	0.258	0.357
	192	0.147	0.240	0.148	0.240	0.153	0.249	0.197	0.311	0.211	0.324	0.327	0.417	0.386	0.443	0.266	0.368
	336	0.163	0.259	0.167	0.261	0.169	0.267	0.213	0.328	0.214	0.327	0.333	0.422	0.378	0.443	0.280	0.380
	720	0.197	0.290	0.202	0.291	0.203	0.301	0.233	0.344	0.236	0.342	0.351	0.427	0.376	0.445	0.283	0.376
ILI	24	1.319	0.754	1.522	0.814	2.215	1.081	2.624	1.095	2.906	1.182	4.657	1.449	1.420	2.012	4.480	1.444
	36	1.579	0.870	1.430	0.834	1.963	0.963	2.516	1.021	2.585	1.038	4.650	1.463	7.394	2.031	4.799	1.467
	48	1.553	0.815	1.673	0.854	2.130	1.024	2.505	1.041	3.024	1.145	5.004	1.542	7.551	2.057	4.800	1.468
	60	1.470	0.788	1.529	0.862	2.368	1.096	2.742	1.122	2.761	1.114	5.071	1.543	7.662	2.100	5.278	1.560
ETTh1	96	0.370	0.400	0.375	0.399	0.375	0.399	0.376	0.415	0.435	0.446	0.941	0.769	0.664	0.612	0.878	0.740
	192	0.413	0.429	0.414	0.421	0.405	0.416	0.423	0.446	0.456	0.457	1.007	0.786	0.790	0.681	1.037	0.824
	336	0.422	0.440	0.431	0.436	0.439	0.443	0.444	0.462	0.486	0.487	1.038	0.784	0.891	0.738	1.238	0.932
	720	0.447	0.468	0.449	0.466	0.472	0.490	0.469	0.492	0.515	0.517	1.144	0.857	0.963	0.782	1.135	0.852
ETTh2	96	0.274	0.337	0.274	0.336	0.289	0.353	0.332	0.374	0.332	0.368	1.549	0.952	0.645	0.597	2.116	1.197
	192	0.341	0.382	0.339	0.379	0.383	0.418	0.407	0.446	0.426	0.434	3.792	1.542	0.788	0.683	4.315	1.635
	336	0.329	0.384	0.331	0.380	0.448	0.465	0.400	0.447	0.477	0.479	4.215	1.642	0.907	0.747	1.124	1.604
	720	0.379	0.422	0.379	0.422	0.605	0.551	0.412	0.469	0.453	0.490	3.656	1.619	0.963	0.783	3.188	1.540
ETTm1	96	0.293	0.346	0.290	0.342	0.299	0.343	0.326	0.390	0.510	0.492	0.626	0.560	0.543	0.510	0.600	0.546
	192	0.333	0.370	0.332	0.369	0.335	0.365	0.365	0.415	0.514	0.495	0.725	0.619	0.557	0.537	0.837	0.700
	336	0.369	0.392	0.366	0.392	0.369	0.386	0.392	0.425	0.510	0.492	1.005	0.741	0.754	0.655	1.124	0.832
	720	0.416	0.420	0.420	0.424	0.425	0.421	0.446	0.458	0.527	0.493	1.133	0.845	0.908	0.724	1.153	0.820
ETTm2	96	0.166	0.256	0.165	0.255	0.167	0.260	0.180	0.271	0.205	0.293	0.355	0.462	0.435	0.507	0.768	0.642
	192	0.223	0.296	0.220	0.292	0.224	0.303	0.252	0.318	0.278	0.336	0.595	0.586	0.730	0.673	0.989	0.757
	336	0.274	0.329	0.278	0.329	0.281	0.342	0.324	0.364	0.343	0.379	1.270	0.871	1.201	0.845	1.334	0.872
	720	0.362	0.385	0.367	0.385	0.397	0.421	0.410	0.420	0.414	0.419	3.001	1.267	3.625	1.451	3.048	1.328

Table 3: Multivariate long-term forecasting results with supervised PatchTST. We use prediction lengths $T \in \{24, 36, 48, 60\}$ for ILI dataset and $T \in \{96, 192, 336, 720\}$ for the others. The best results are in **bold** and the second best are underlined.

回望窗口长度：
为了更大限度的体现实验的合理性
FEDformer、Autoformer、Informer的回望窗口长度从$L \in \{24、48、96、192、336、720\}$中选择最好的

DLinear模型选择原文中默认的336

PatchTST参数：
P=16,S=9
PatchTST/64：L=512，N=64
PatchTST/42：L=336，N=42
（PatchTST/42回望窗口数量与DLinear一致）

Experiment：表示学习——预测任务

Models		PatchTST						DLinear		FEDformer		Autoformer		Informer	
		Fine-tuning		Lin. Prob.		Sup.									
Metric		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Weather	96	**0.144**	**0.193**	0.158	0.209	0.152	0.199	0.176	0.237	0.238	0.314	0.249	0.329	0.354	0.405
	192	**0.190**	**0.236**	0.203	0.249	0.197	0.243	0.220	0.282	0.275	0.329	0.325	0.370	0.419	0.434
	336	**0.244**	**0.280**	0.251	0.285	0.249	0.283	0.265	0.319	0.339	0.377	0.351	0.391	0.583	0.543
	720	**0.320**	**0.335**	0.321	0.336	**0.320**	**0.335**	0.323	0.362	0.389	0.409	0.415	0.426	0.916	0.705
Traffic	96	**0.352**	**0.244**	0.399	0.294	0.367	0.251	0.410	0.282	0.576	0.359	0.597	0.371	0.733	0.410
	192	**0.371**	**0.253**	0.412	0.298	0.385	0.259	0.423	0.287	0.610	0.380	0.607	0.382	0.777	0.435
	336	**0.381**	**0.257**	0.425	0.306	0.398	0.265	0.436	0.296	0.608	0.375	0.623	0.387	0.776	0.434
	720	**0.425**	**0.282**	0.460	0.323	0.434	0.287	0.466	0.315	0.621	0.375	0.639	0.395	0.827	0.466
Electricity	96	**0.126**	**0.221**	0.138	0.237	0.130	0.222	0.140	0.237	0.186	0.302	0.196	0.313	0.304	0.393
	192	**0.145**	**0.238**	0.156	0.252	0.148	0.240	0.153	0.249	0.197	0.311	0.211	0.324	0.327	0.417
	336	**0.164**	**0.256**	0.170	0.265	0.167	0.261	0.169	0.267	0.213	0.328	0.214	0.327	0.333	0.422
	720	**0.193**	**0.291**	0.208	0.297	0.202	**0.291**	0.203	0.301	0.233	0.344	0.236	0.342	0.351	0.427

Table 4: Multivariate long-term forecasting results with self-supervised PatchTST. We use prediction lengths $T \in \{96, 192, 336, 720\}$. The best results are in **bold** and the second best are underlined.

Patch参数：不重叠、L=512、P=12、掩码率40%
step1：无监督训练100epoch

step2：
- choice1：只训练预测头20epoch，同时backbone不动
- choice2：训练预测头10epoch，然后整体有监督finetune20epoch

Experiment：表示学习——迁移学习

Models		PatchTST						DLinear		FEDformer		Autoformer		Informer	
		Fine-tuning		Lin. Prob.		Sup.									
Metric		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Weather	96	**0.145**	**0.195**	0.163	0.216	0.152	0.199	0.176	0.237	0.238	0.314	0.249	0.329	0.354	0.405
	192	**0.193**	**0.243**	0.205	0.252	0.197	**0.243**	0.220	0.282	0.275	0.329	0.325	0.370	0.419	0.434
	336	**0.244**	**0.280**	0.253	0.289	0.249	0.283	0.265	0.319	0.339	0.377	0.351	0.391	0.583	0.543
	720	0.321	0.337	**0.320**	0.336	**0.320**	**0.335**	0.323	0.362	0.389	0.409	0.415	0.426	0.916	0.705
Traffic	96	0.388	0.273	0.400	0.288	**0.367**	**0.251**	0.410	0.282	0.576	0.359	0.597	0.371	0.733	0.410
	192	0.400	0.277	0.412	0.293	**0.385**	**0.259**	0.423	0.287	0.610	0.380	0.607	0.382	0.777	0.435
	336	0.408	0.280	0.425	0.307	**0.398**	**0.265**	0.436	0.296	0.608	0.375	0.623	0.387	0.776	0.434
	720	0.447	0.310	0.457	0.317	**0.434**	**0.287**	0.466	0.315	0.621	0.375	0.639	0.395	0.827	0.466

Table 5: Transfer learning task: PatchTST is pre-trained on Electricity dataset and the model is transferred to other datasets. The best results are in **bold** and the second best are underlined.

Models		Fine-tuning	
Metric		MSE	MAE
Weather	96	**0.144**	**0.193**
	192	**0.190**	**0.236**
	336	**0.244**	**0.280**
	720	**0.320**	**0.335**
Traffic	96	**0.352**	**0.244**
	192	**0.371**	**0.253**
	336	**0.381**	**0.257**
	720	**0.425**	**0.282**
Electricity	96	**0.126**	**0.221**
	192	**0.145**	**0.238**
	336	**0.164**	**0.256**
	720	**0.193**	**0.291**

在电力数据集上预训练，然后在其他模型上微调，虽然比不上直接在对应数据集上做训练，但是也好过其他基线模型

⇨ 可以与模型原来的预测能力做对比

Experiment：表示学习——与SOTA模型对比

Models		IMP.	PatchTST Transferred		PatchTST Self-supervised		BTSF		TS2Vec		TNC		TS-TCC	
Metrics		MSE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	24	42.3%	**0.312**	**0.362**	0.322	0.369	0.541	0.519	0.599	0.534	0.632	0.596	0.653	0.610
	48	44.7%	**0.339**	**0.378**	0.354	0.385	0.613	0.524	0.629	0.555	0.705	0.688	0.720	0.693
	168	34.5%	0.424	0.437	**0.419**	**0.424**	0.640	0.532	0.755	0.636	1.097	0.993	1.129	1.044
	336	48.5%	0.472	0.472	**0.445**	**0.446**	0.864	0.689	0.907	0.717	1.454	0.919	1.492	1.076
	720	48.8%	0.508	0.507	**0.487**	**0.478**	0.993	0.712	1.048	0.790	1.604	1.118	1.603	1.206

Table 6: Representation learning methods comparison. Column name *transferred* implies pre-training PatchTST on Traffic dataset and transferring the representation to ETTh1, while *self-supervised* implies both pre-training and linear probing on ETTh1. The best and second best results are in **bold** and underlined. IMP. denotes the improvement on best results of PatchTST compared to that of baselines, which is in the range of 34.5% to 48.8% on various prediction lengths.

：表示在traffic数据集上做预训练，然后在ETTh1上做微调

：表示在ETTh1数据集上做预训练，然后在ETTh1上做微调

Q：为什么在迁移学习的性能甚至好过在目标数据集上的自学习？
A：Traffic数据集比较大，可能可以学到更多的特征

Experiment：消融实验——patch与通道独立

CI：只做通道独立；P：只做Patch

Models		PatchTST								FEDformer	
		P+CI		CI		P		Original			
Metric		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Weather	96	**0.152**	**0.199**	0.164	0.213	0.168	0.223	0.177	0.236	0.238	0.314
	192	**0.197**	**0.243**	0.205	0.250	0.213	0.262	0.221	0.270	0.275	0.329
	336	**0.249**	**0.283**	0.255	0.289	0.266	0.300	0.271	0.306	0.339	0.377
	720	**0.320**	**0.335**	0.327	0.343	0.351	0.359	0.340	0.353	0.389	0.409
Traffic	96	**0.367**	**0.251**	0.397	0.271	0.595	0.376	-	-	0.576	0.359
	192	**0.385**	**0.259**	0.411	0.276	0.612	0.387	-	-	0.610	0.380
	336	**0.398**	**0.265**	0.423	0.282	0.651	0.391	-	-	0.608	0.375
	720	**0.434**	**0.287**	0.457	0.309	-	-	-	-	0.621	0.375
Electricity	96	**0.130**	**0.222**	0.136	0.231	0.196	0.307	0.205	0.318	0.186	0.302
	192	**0.148**	**0.240**	0.164	0.263	0.215	0.323	-	-	0.197	0.311
	336	**0.167**	**0.261**	0.168	0.262	0.228	0.338	-	-	0.213	0.328
	720	**0.202**	**0.291**	0.219	0.312	0.244	0.345	-	-	0.233	0.344

Table 7: Ablation study of patching and channel-independence in PatchTST. 4 cases are included: (a) both patching and channel-independence are included in model (P+CI); (b) only channel-independence (CI); (c) only patching (P); (d) neither of them is included (Original TST model). PatchTST means supervised PatchTST/42. '-' in table means the model runs out of GPU memory (NVIDIA A40 48GB) even with batch size 1. The best results are in **bold**.

通道独立的效果比Patch好一些

Experiment：消融实验——回顾窗口长度

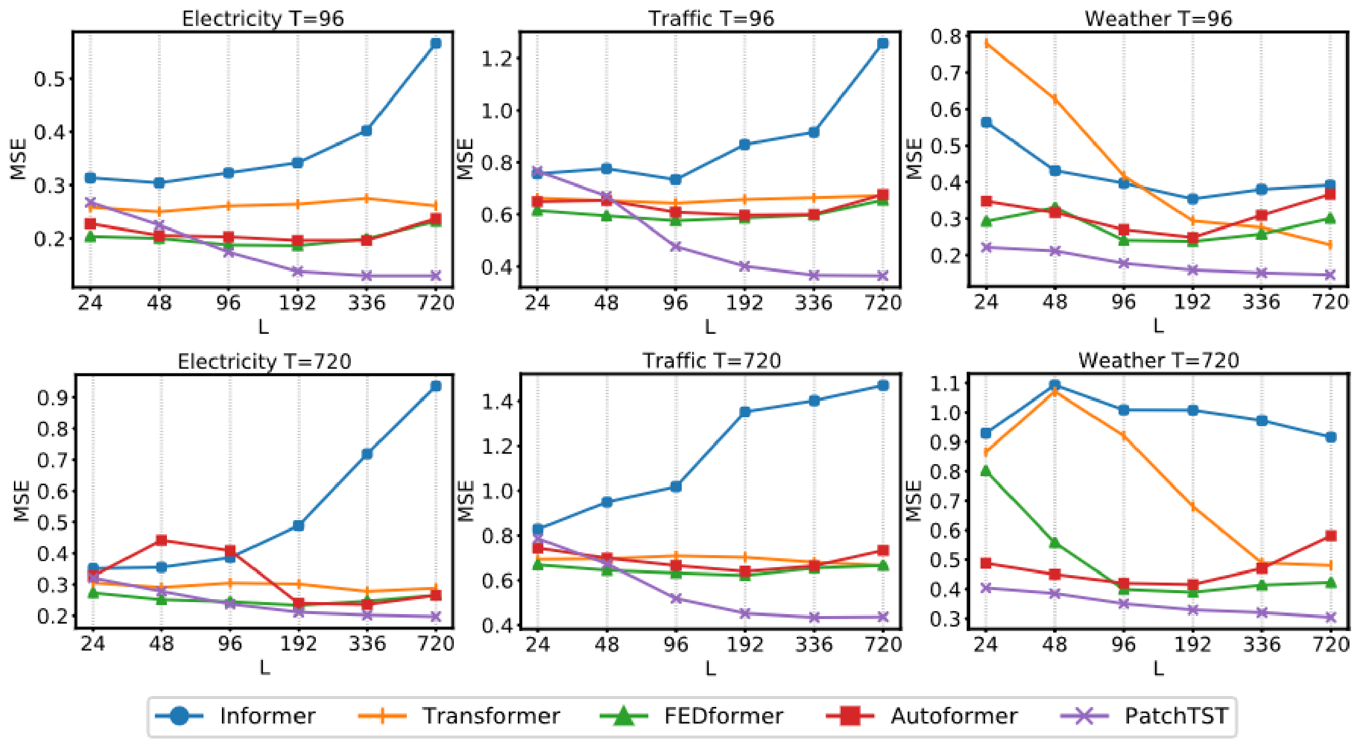


Figure 2: Forecasting performance (MSE) with varying look-back windows on 3 large datasets: Electricity, Traffic, and Weather. The look-back windows are selected to be $L = 24, 48, 96, 192, 336, 720$, and the prediction horizons are $T = 96, 720$. We use supervised PatchTST/42 and other open-source Transformer-based baselines for this experiment.

Informer很直观的随着回顾窗口的变长，效果下降，而PatchTST能够随着回顾窗口的变长提升自己的效果

西南財經大學
SOUTHWESTERN UNIVERSITY OF FINANCE AND ECONOMICS

Questions and Discussions

主讲人：陈凯伦
2024.3.27

文章背景介绍

- LTSF-Linear在2022年5月发布在arxiv上，被2023AAAI收录，与本次组会要将的论文（2022年11月发布在arxiv上）颇有针锋相对的意思。

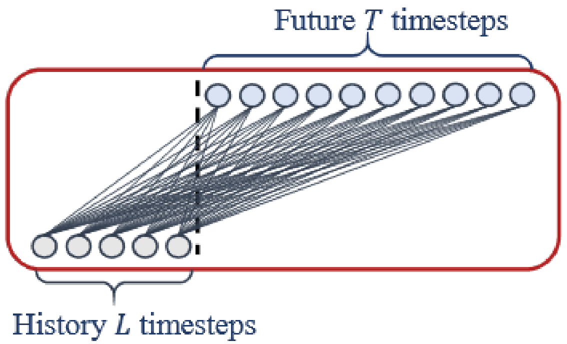


Figure 2. Illustration of the basic linear model.

- 主要对Transformer架构的模型提出了以下几个质疑：
  - 基于Transformer的模型就算用了位置编码等方法，也难以保留时间序列中最重要的时间信息。
  - 基于Transformer的模型并不能捕捉长时间序列中存在的特征，即输入序列长度变长，模型按理来说有更多的学习样本，预测效果应该更好，但是Transformer架构的模型效果提升不明显，甚至会增加误差。
  - 使用的参数量过大，是否有价值？LTSF-Linear没有像Transformer捕获变量间相关性的机制但表现出了很好的预测性能，是否说明Transfromer处理变量的方法会导致过拟合或特征冗余的问题？

点击此处回去

https://arxiv.org/abs/2205.13504

篇幅有限，这里并没有详细列出所有的研究内容，如果有兴趣，可以自行阅读论文！