FILM

Frequency improved Legendre Memory Model for Long-term Time Series Forecasting

Tian Zhou, Ziqing Ma, Xue Wang, Qingsong Wen, Liang Sun,Tao Yao,Wotao Yin,Rong Jin

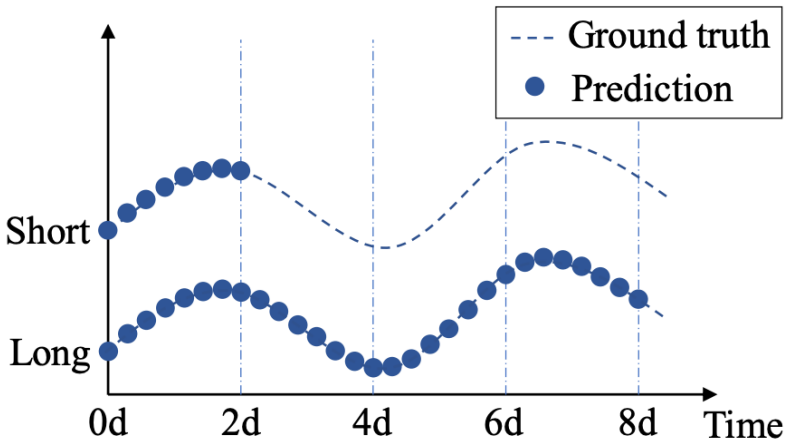
Neurips22

Contents

目录

Problem&Motivations

Problem: Long term Forecasting



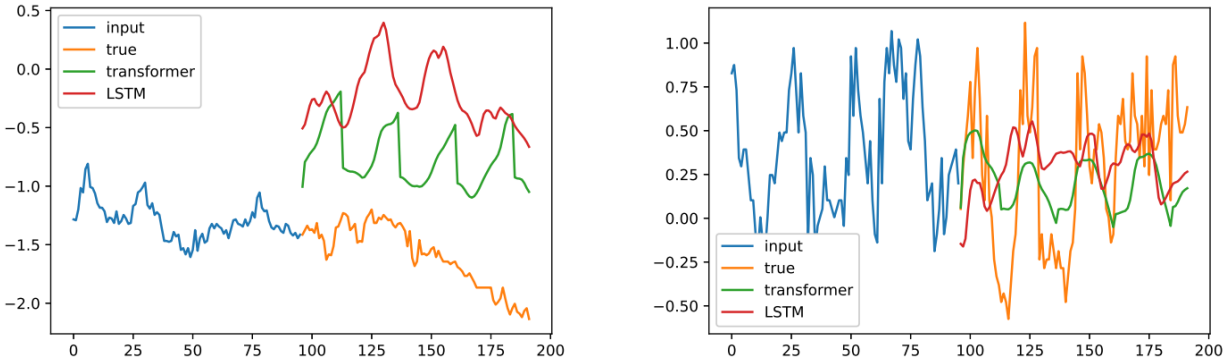


Figure 1: The discrepancy between ground truth and forecasting output from vanilla Transformer and LSTM on a real-world ETTh1 dataset Left: trend shift. Right: seasonal shift.

1) How to capture critical historical information as complete as possible
2) How to effectively remove the noise

Motivations

We can get a compact Representation of Time Series using Legendre Polynomials projection

Theorem 1 (*Similar to Proposition 6 in Gu et al. (2020)*). *If $f(x)$ is L-Lipschitz, then $\|f_{[t-\theta,t]}(x) - g^{(t)}(x)\|_{\mu^{(t)}} \leq \mathcal{O}(\theta L/\sqrt{N})$. Moreover, if $f(x)$ has k-th order bounded derivatives, we have $\|f_{[t-\theta,t]}(x) - g^{(t)}(x)\|_{\mu^{(t)}} \leq \mathcal{O}(\theta^k N^{-k+1/2})$.*

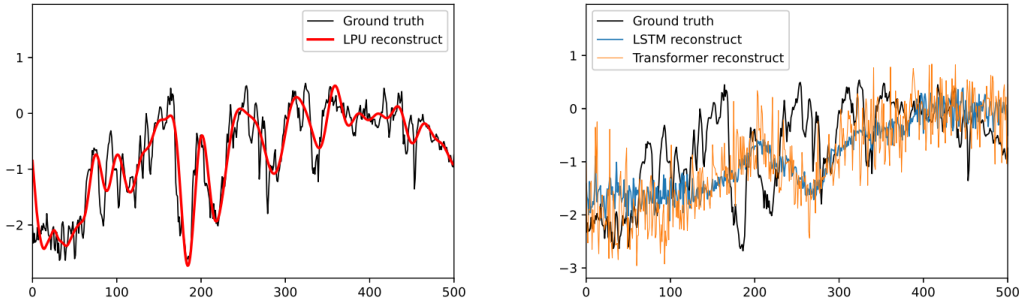
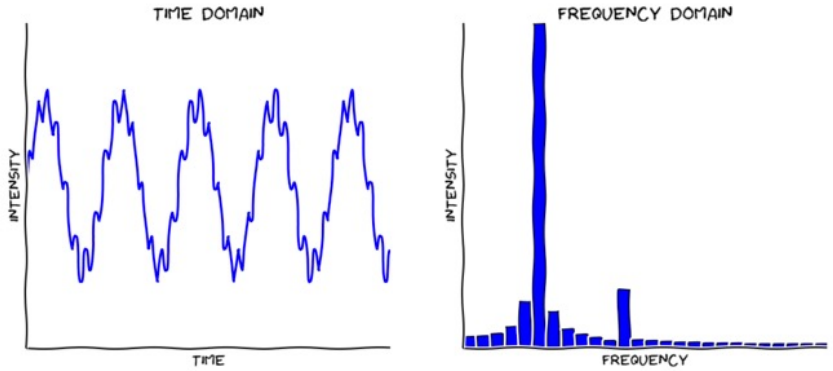


Figure 2: Data recovery with Autoencoder structure: recovery a 1024-length data with a bottleneck of 128 parameters. Left: Legendre Projection Unit. Right: LSTM and vanilla Transformer.

Theorem 3. *Let $A \in \mathbb{R}^{d\times n}$ be the Fourier coefficients matrix of an input matrix $X \in \mathbb{R}^{d\times n}$, and $\mu(A)$, the coherence measure of matrix A, is $\Omega(k/n)$. We assume there exist s and a positive $a_{\min}$ such that the elements in last $d-s$ columns of A is smaller than $a_{\min}$. If we keep first s columns selected and randomly choose $\mathcal{O}(k^2/\epsilon^2 - s)$ columns from the remaining parts, with high probability*

$$\|A - P(A)\|_F \leq \mathcal{O}\left[(1+\epsilon)a_{\min} \cdot \sqrt{(n-s)d}\right],$$

where $P(A)$ denotes the matrix projecting A onto the column selected column space.



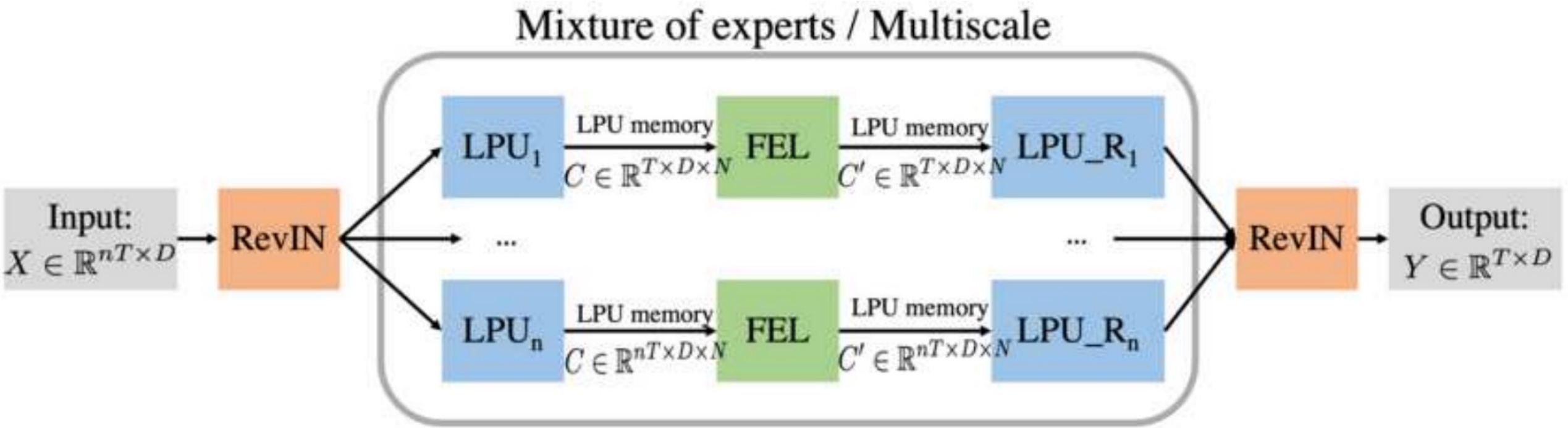
Fourier transform

$$\hat{f}(\xi) = \int_{-\infty}^{\infty} f(x)\, e^{-i2\pi\xi x}\, dx, \quad \forall\, \xi \in \mathbb{R}.$$

Fourier inverse transform

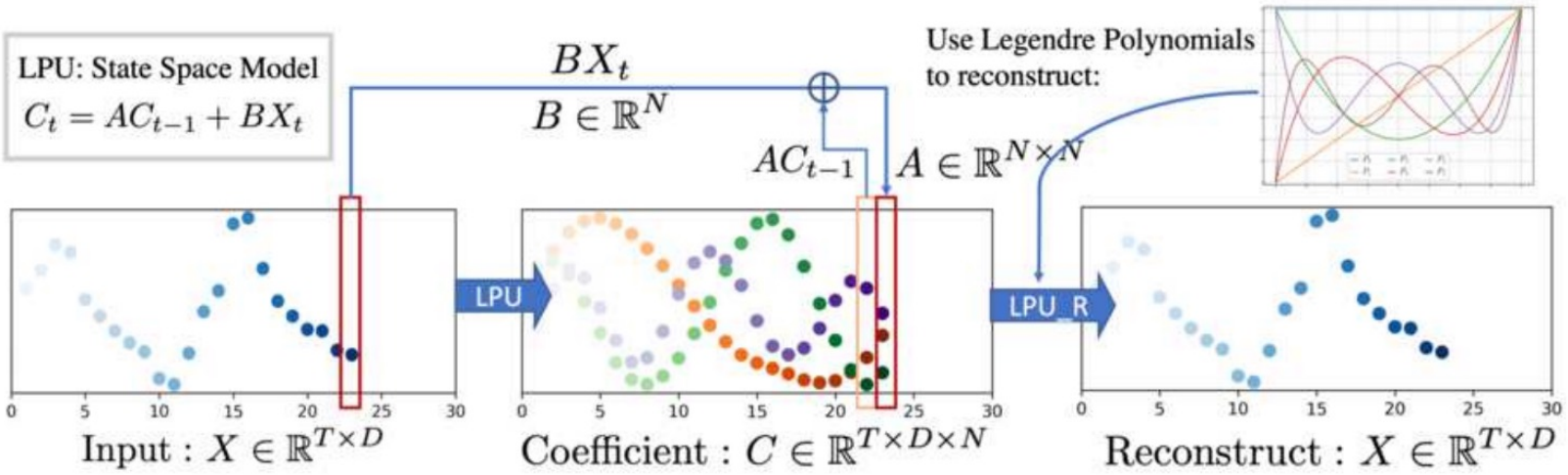
$$f(x) = \int_{-\infty}^{\infty} \hat{f}(\xi)\, e^{i2\pi\xi x}\, d\xi, \quad \forall\, x \in \mathbb{R},$$

Model Structures



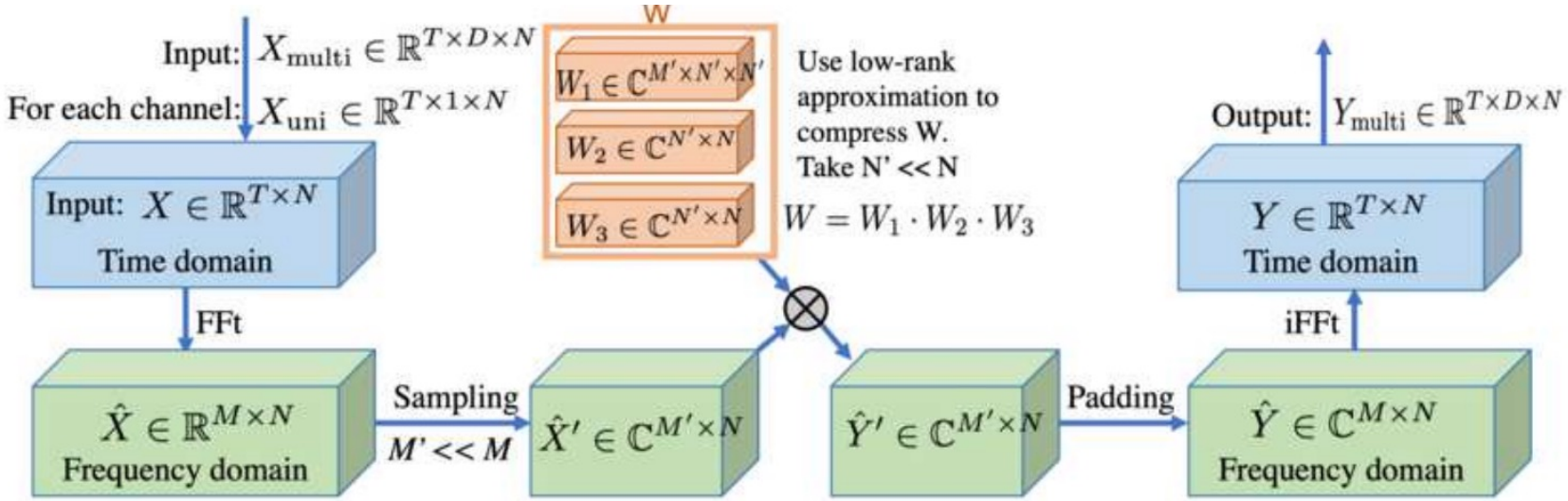
LPU: Legendre Projection Unit for **memory compression**
FEL: Frequency Enhanced Layer for frequency domain **feature transform** and **noise reduction**
LPU_R: reverse Legendre Projection Unit for **output generation**
RevIN: reversible instance **normalization**

Model Structures



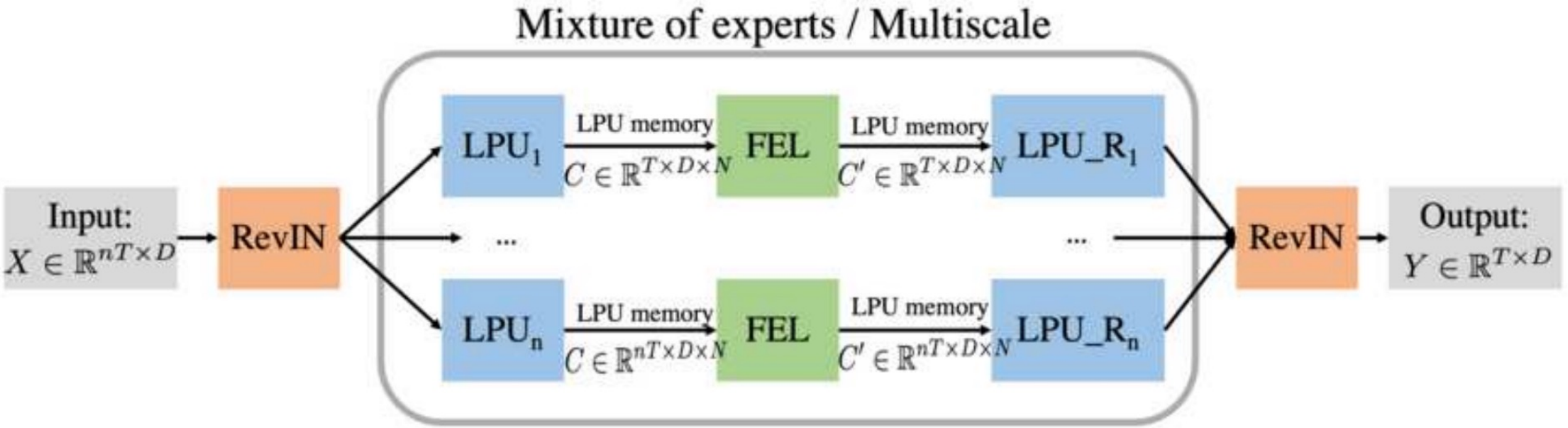
LPU contains two states: **Projection & Reconstruction.**
C(t) is the compressed memory for historical input up to time t. x(t) is the original input signal at time t.
A, B are two pre-fixed projection matrices.
C(t) is reconstructed to original input by multiplying a discrete Legendre Polynomials matrix.

Model Structures



Frequency Enhanced Layer (FEL): we use a **lowest mode sampling** mechanism to remove noise
And a **low rank compressed** weight W to process the features.

Model Structures



Mixture of experts/multiscale mechanism: introduce **multiscale phenomena** bias

Reversible instance normalization: ease the **distribution shift**, slow down the training 2-5 times.

Experiments

Table 1: multivariate long-term series forecasting results on six datasets with various input length and prediction length $O \in \{96, 192, 336, 720\}$ (For ILI dataset, we set prediction length $O \in \{24, 36, 48, 60\}$). A lower MSE indicates better performance. All experiments are repeated 5 times.

Methods		FiLM		FEDformer		Autoformer		S4		Informer		LogTrans		Reformer	
Metric		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTm2	96	**0.165**	**0.256**	0.203	0.287	0.255	0.339	0.705	0.690	0.365	0.453	0.768	0.642	0.658	0.619
	192	**0.222**	**0.296**	0.269	0.328	0.281	0.340	0.924	0.692	0.533	0.563	0.989	0.757	1.078	0.827
	336	**0.277**	**0.333**	0.325	0.366	0.339	0.372	1.364	0.877	1.363	0.887	1.334	0.872	1.549	0.972
	720	**0.371**	**0.389**	0.421	0.415	0.422	0.419	0.877	1.074	3.379	1.338	3.048	1.328	2.631	1.242
Electricity	96	**0.154**	**0.267**	0.183	0.297	0.201	0.317	0.304	0.405	0.274	0.368	0.258	0.357	0.312	0.402
	192	**0.164**	**0.258**	0.195	0.308	0.222	0.334	0.313	0.413	0.296	0.386	0.266	0.368	0.348	0.433
	336	**0.188**	**0.283**	0.212	0.313	0.231	0.338	0.290	0.381	0.300	0.394	0.280	0.380	0.350	0.433
	720	0.236	**0.332**	**0.231**	0.343	0.254	0.361	0.262	0.344	0.373	0.439	0.283	0.376	0.340	0.420
Exchange	96	**0.086**	**0.204**	0.139	0.276	0.197	0.323	1.292	0.849	0.847	0.752	0.968	0.812	1.065	0.829
	192	**0.188**	**0.292**	0.256	0.369	0.300	0.369	1.631	0.968	1.204	0.895	1.040	0.851	1.188	0.906
	336	**0.356**	**0.433**	0.426	0.464	0.509	0.524	2.225	1.145	1.672	1.036	1.659	1.081	1.357	0.976
	720	**0.727**	**0.669**	1.090	0.800	1.447	0.941	2.521	1.245	2.478	1.310	1.941	1.127	1.510	1.016
Traffic	96	**0.416**	**0.294**	0.562	0.349	0.613	0.388	0.824	0.514	0.719	0.391	0.684	0.384	0.732	0.423
	192	**0.408**	**0.288**	0.562	0.346	0.616	0.382	1.106	0.672	0.696	0.379	0.685	0.390	0.733	0.420
	336	**0.425**	**0.298**	0.570	0.323	0.622	0.337	1.084	0.627	0.777	0.420	0.733	0.408	0.742	0.420
	720	**0.520**	**0.353**	0.596	0.368	0.660	0.408	1.536	0.845	0.864	0.472	0.717	0.396	0.755	0.423
Weather	96	**0.199**	**0.262**	0.217	0.296	0.266	0.336	0.406	0.444	0.300	0.384	0.458	0.490	0.689	0.596
	192	**0.228**	**0.288**	0.276	0.336	0.307	0.367	0.525	0.527	0.598	0.544	0.658	0.589	0.752	0.638
	336	**0.267**	**0.323**	0.339	0.380	0.359	0.395	0.531	0.539	0.578	0.523	0.797	0.652	0.639	0.596
	720	**0.319**	**0.361**	0.403	0.428	0.578	0.578	0.419	0.428	1.059	0.741	0.869	0.675	1.130	0.792
ILI	24	**1.970**	**0.875**	2.203	0.963	3.483	1.287	4.631	1.484	5.764	1.677	4.480	1.444	4.400	1.382
	36	**1.982**	**0.859**	2.272	0.976	3.103	1.148	4.123	1.348	4.755	1.467	4.799	1.467	4.783	1.448
	48	**1.868**	**0.896**	2.209	0.981	2.669	1.085	4.066	1.36	4.763	1.469	4.800	1.468	4.832	1.465
	60	**2.057**	**0.929**	2.545	1.061	2.770	1.125	4.278	1.41	5.264	1.564	5.278	1.560	4.882	1.483

Improvement over sota

Table 3: Low-rank Approximation (LRA) study for Frequency Enhanced Layer (FEL): Comp. K=0 means default version without LRA, 1 means the largest compression using K=1.

Comp. K		0		16		4		1	
Metric		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	96	**0.371**	**0.394**	0.371	0.396	0.371	0.398	0.400	0.421
	192	0.414	**0.423**	0.411	0.423	0.414	0.426	0.435	0.444
	336	**0.442**	0.445	0.443	0.446	0.443	0.444	0.492	0.478
	720	**0.454**	**0.451**	0.464	0.474	0.468	0.478	0.501	0.499
Weather	96	**0.199**	**0.262**	0.199	0.263	0.197	0.262	0.198	0.263
	192	0.228	0.288	0.225	0.285	0.226	0.285	0.225	0.286
	336	0.267	0.323	0.266	0.321	0.263	0.314	0.264	0.316
	720	0.319	0.361	0.314	0.355	0.315	0.354	0.318	0.357
Parameter size		100%		1.95%		0.41%		0.10%	

Parameter saving using LRA

Table 4: Mode selection policy study for frequency enhanced layer. Lowest: select the lowest m frequency mode; Random: select m random frequency mode; Low random: select the $0.8 * m$ lowest frequency mode and $0.2 * m$ random high frequency mode.

Policy		Lowest		Random		Low random	
Metric		MSE	MAE	MSE	MAE	MSE	MAE
Exchange	96	**0.086**	**0.204**	0.086	0.208	0.087	0.210
	192	0.188	**0.292**	**0.187**	0.318	0.207	0.340
	336	0.356	**0.433**	0.358	0.437	**0.353**	0.461
	720	**0.727**	**0.669**	0.788	0.680	0.748	0.674
Weather	96	0.199	0.262	0.197	0.256	**0.196**	**0.254**
	192	**0.228**	**0.288**	0.234	0.300	0.234	0.301
	336	0.267	0.323	0.266	0.319	**0.263**	**0.316**
	720	0.319	0.361	0.317	0.356	**0.316**	**0.354**

Lowest mode policy is the most stable one

Experiments

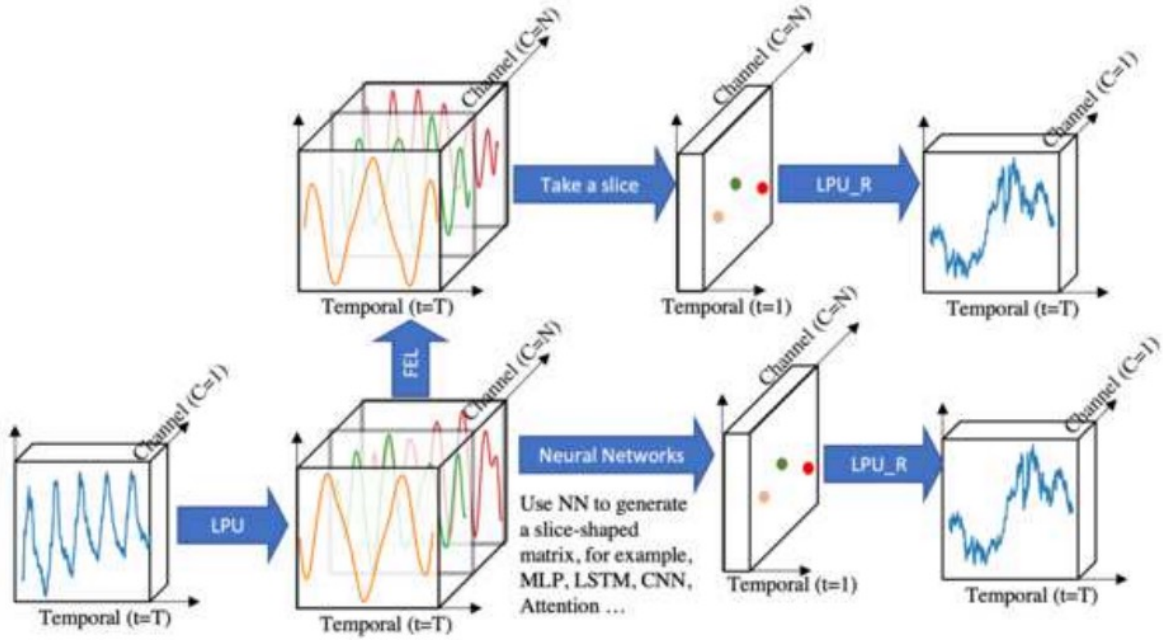


Figure 6: LPU boosting effect. LPU can serve as a plug-in block in various backbones, e.g., FEL, MLP, LSTM, CNN, and Attention. Replacing LPU with a comparable-sized linear layer will always lead to degraded performance.

LPU serve as a general **boosting plugin** Block for neural layers.

Table 2: Boosting effect of LPU layer for common deep learning backbones: MLP, LSTM, CNN and Attention.'+' indicates degraded performance.

Methods		FEL		MLP		LSTM		lagged-LSTM		CNN		Attention	
Compare		LPU	Linear	LPU	Linear	LPU	Linear	LPU	Linear	LPU	Linear	LPU	Linear
ETTm1	96	**0.030**	+38%	0.034	+8.0%	0.049	+73%	0.093	-21%	0.116	-50%	0.243	-81%
	192	**0.047**	+9.5%	0.049	+30%	0.174	+32%	0.331	-48%	0.101	+20%	0.387	-86%
	336	**0.063**	+5.8%	0.061	+64%	0.119	+84%	0.214	-19%	0.122	+25%	1.652	+12%
	720	**0.081**	+1.4%	0.082	+62%	0.184	+32%	0.303	-6.5%	0.108	+13%	4.782	-61%
Electricity	96	**0.213**	+136%	0.431	+121%	0.291	+55.6%	0.739	-33%	0.310	+43%	0.805	+23%
	192	**0.268**	+32%	0.291	+239%	0.353	+17%	0.535	+15%	0.380	+12%	0.938	+14%
	336	**0.307**	+0.1%	0.296	+235%	0.436	-6.7%	0.517	+23%	0.359	+29%	2.043	-54%
	720	**0.321**	+37%	0.339	+196%	0.636	-11%	0.492	+28%	0.424	+18%	9.115	+298%

Conclusion

- We propose a *Frequency improved Legendre Memory model (**FiLM**)* architecture with a mixture of experts for robust multiscale time series feature extraction.

- We redesign the *Legendre Projection Unit (**LPU**)* and make it a general tool for data representation that any time series forecasting model can exploit to solve the historical information preserving problem.

- We propose *Frequency Enhanced Layers (**FEL**)* that reduce dimensionality by a combination of Fourier analysis and low-rank matrix approximation to minimize the impact of noisy signals from time series and ease the overfitting problem. The effectiveness of this method is verified both theoretically and empirically.

- We conduct extensive experiments on six benchmark datasets across multiple domains (energy, traffic, economics, weather, and disease). Our empirical studies show that the proposed model improves the performance of state-of-the-art methods by **19.2%** and **26.1%** in multivariate and univariate forecasting, respectively. In addition, our empirical studies also reveal a dramatic improvement in computational efficiency through dimensionality reduction.

Thank You

https://arxiv.org/pdf/2205.08897.pdf

https://github.com/DAMO-DI-ML/NeurIPS2022-FiLM