Revisiting Attention for Multivariate Time Series Forecasting

Haixiang Wu

School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang, China
2232208015@stmail.ujs.edu.cn

Abstract

Current Transformer methods for Multivariate Time-Series Forecasting (MTSF) are all based on the conventional attention mechanism. They involve sequence embedding and performing a linear projection for Q, K, and V, and then computing attention within this latent space. We have not yet delved into the attention mechanism to explore whether such a mapping space is optimal for MTSF. To investigate this issue, we first propose Frequency Spectrum attention (FSatten), a novel attention mechanism based on the frequency domain space. It employs the Fourier transform for embedding and introduces Multi-head Spectrum Scaling (MSS) to replace the conventional linear mapping for Q and K. FSatten can accurately capture the periodic dependencies between sequences and outperform the conventional attention, without necessitating changes to mainstream architectures. We further design a more general method dubbed Scaled Orthogonal attention (SOatten). We propose an orthogonal embedding and a Head-Coupling Convolution (HCC) based on the neighboring similarity bias to guide the model in learning comprehensive dependency patterns. Experiments show that FSatten and SOatten surpass the SOTA which uses conventional attention, making it a good alternative as a basic attention mechanism for MTSF.

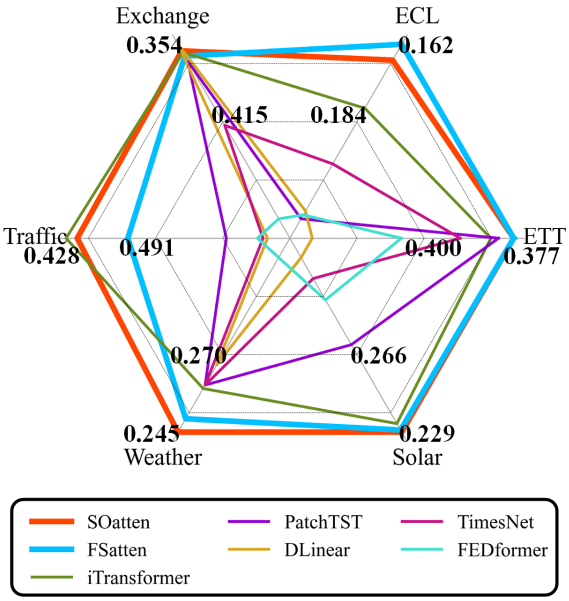

Figure 1: Performance of FSatten and SOatten.

Introduction

Multivariate Time Series Forecasting (MTSF) is extensively applied in real-world scenarios such as finance, electricity, and transportation. Benefiting from the attention mechanism's (Vaswani et al. 2017) ability to effectively capture both long- and short-term dependencies, many Transformer-based methods have demonstrated remarkable performance. These methods mainly include the Temporal Transformer, which evolves from applying attention between time steps (Zhou et al. 2021) (Li et al. 2019) (Liu et al. 2021) to applying attention between subseries (Wu et al. 2021) (Zhou et al. 2022) (Nie et al. 2022), and the Variate Transformer, which explicitly models the correlations between variates through attention (Zhang and Yan 2022) (Liu et al. 2023).

Temporal and Variate Transformers, as shown in Figure 2, primarily apply attention mechanisms to time series sequences and have become mainstream, with many subsequent studies (Zhou et al. 2023) (Jin et al. 2023) building upon these architectures. We aim to understand where these correlations between sequences manifest. In the attention mechanism, sequences are mapped to a learnable space by embedding and linear projection, and then the correlations are calculated within this latent space. Although we cannot provide physical interpretations for the learned characteristics of black-box neural networks, it is worth considering whether the dependency capturing within this latent space is optimal.

To explore the interpretability of time-series attention and make further improvements for MTSF, we propose Frequency Spectrum attention (FSatten), which is based on the frequency domain space. The consideration is that dependencies between non-stationary sequences are complex and can be synchronous or asynchronous at different frequencies. It is appropriate to consider these dependencies from the frequency domain perspective, as previous works have made improvements (Zhou et al. 2022) (Xu, Zeng, and Xu 2023). In FSatten, Fourier transform is utilized for the embedding, and Query and Key are projected by a proposed

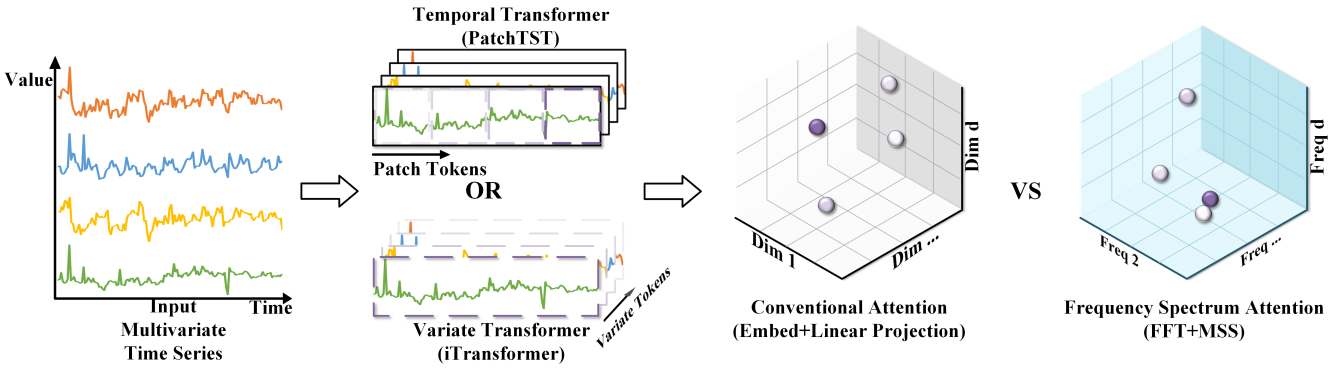

Figure 2: (Left) Temporal Transformer and Variate Transformer. (Right) Comparison of mapping space from conventional attention and FSatten

Multi-head Spectrum Scaling (MSS) instead of the conventional linear projection. MSS scales amplitude for different frequency components under each of the multiple heads, identifying clear frequency spectral relationships between sequences.

Experimental results in Figure 1 and Table 1 show that without modifying the architecture, simply replacing the conventional attention with FSatten yields significant improvement over the state-of-the-art (SOTA). This suggests that the conventional attention mechanism is not optimal for MTSF. However, the frequency domain space cannot meet all the characteristics of different scenes. Also, FSatten is good for capturing same-frequency correlations between variates but may not be highly appropriate for Temporal Transformers, as sequences split from one variate naturally tend to exhibit the same periodic frequency.

To find a more general method for various scenarios, we propose Scaled Orthogonal attention (SOatten), which creates a learnable orthogonal transformation beyond the Fourier transform. In SOatten, we propose a Head Coupling Convolution (HCC) to guide the updating of learnable orthogonal spaces by leveraging the similarity between adjacent sequences. Experiments show that SOatten enhances overall performance compared to FSatten when applied to Variate Transformer, iTransformer (Liu et al. 2023) and makes significant improvements when applied to a general Temporal Transformer, PatchTST (Nie et al. 2022), showcasing stronger adaptability. We hope the proposed methods may inspire future work in time series analysis and offer contributions to other deep learning fields.

The main contributions of this work are as follows:

- We propose FSatten, a more interpretable and effective model than conventional attention for MTSF, which replaces the learnable latent space by Frequency domain.
- Through the proposed MSS mapping for Query and Key, FSatten accurately identifies the frequency correlations between sequences. This specific dependency is more effective than what is provided by conventional linear projections.
- We propose SOatten, a more general attention than FSatten which provides a learnable orthogonal latent space facilitated by a designed HCC module for capturing comprehensive dependencies.
- On six real-world long-term forecasting benchmarks, our FSatten and SOatten outperform the SOTA method which utilizes conventional attention by an overall of 8.1% and 21.8% on MSE, demonstrating their superior effectiveness for MTSF.

Preliminaries

A Multivariate Time Series (MTS) sampling with look back window L is denoted by $X = \{x_1, ..., x_L\} \epsilon R^{C \times L}$, where each x_l at time step l is a vector of dimension C. The task is to forecast T future values $\{x_{L+1}, ..., x_{L+T}\}$. In this work, the proposed FSatten and SOatten are applied to two SOTA Transformers to compare with conventional attention, as shown in Figure 2 Left: (1) Variate Transformer, iTransformer (Liu et al. 2023), and (2) Temporal Transformer, PatchTST (Nie et al. 2022). Many subsequent approaches (Zhou et al. 2023) (Jin et al. 2023) are based primarily on these two mainstream architectures. Detailed illustrations are as follows:

Temporal Transformer

The initial Transformer-based MTS models take time steps as tokens and apply temporal attention between them. Subsequent works demonstrated that temporal attention at a sub-series level with fewer tokens is more effective and can greatly reduce the complexity. PatchTST (Nie et al. 2022) provides a general paradigm of the Temporal Transformer at the sub-series level. In PatchTST, each of the C variates $X^{(i)}_{1:L} = \left\{x^{(i)}_1, \dots x^{(i)}_L\right\} \epsilon R^{1 \times L}$ is converted to sub-series Patches $X^{(i)}_P = \left\{x^{(i)}_1, \dots x^{(i)}_N\right\} \epsilon R^{P \times N}$, where $N = \left\lfloor \frac{(L-P)}{S} \right\rfloor + 2$, P is length of patches and S is the stride - the non-overlapping region between two consecutive patches. Temporal attention is applied to capture the dependencies between patches of each variate. The $X^{(i)}_P$ is first embedded to tokens $Z^{(i)}_P = W^P_H X^{(i)}_P$, Where $W^P_H \epsilon R^{D \times P}$ and D

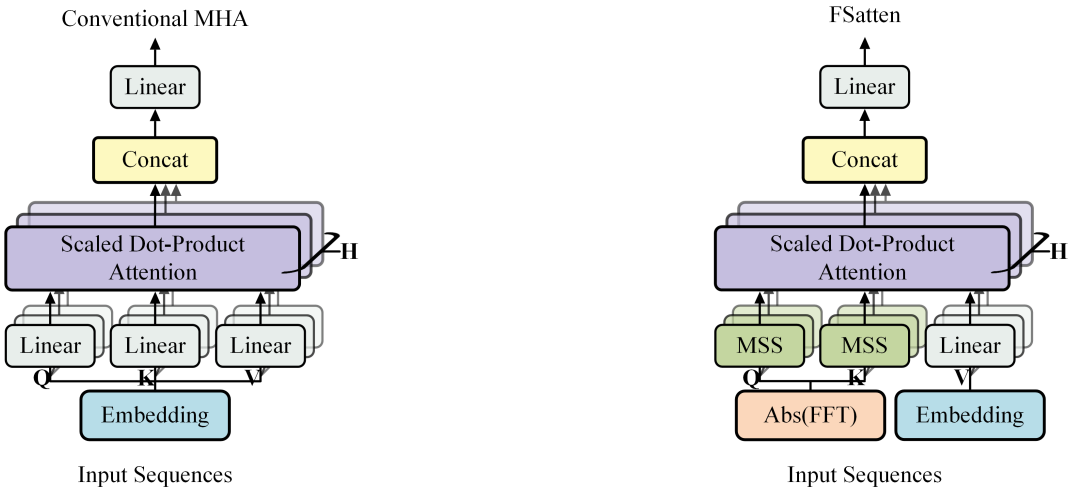

Figure 3: (left) Multi-Head Attention. (right) FSatten. On the left side of the figures is the shape of the data at each stage, and adding batch size to the front is the shape in training.

is the number of dimensions. Then the attention weight is calculated:

$$A_h^{(i)} = Sofmax(\frac{((Z_P^{(i)})^T W_h^Q)((Z_P^{(i)})^T W_h^K)^T}{\sqrt{d_K}})((Z_P^{(i)})^T W_h^V) \tag{1}$$

Where $W_h^{\{Q,K,V\}} \epsilon\ R^{D\times\frac{D}{H}}$, and H is number of attention heads.

Variate Transformer

The Temporal Transformer directly follows the paradigm in NLP. But unlike natural language, MTS has multiple parallel sequence inputs. Variate Transformer explicitly models the complex correlations between variable sequences. Typically, iTrasformer (Liu et al. 2023) embed the whole time series of each variate $X^{(i)}$ independently into a (variate) token as $Z_V = XW_H^V$, where $W_H^V \epsilon\ R^{L\times D}$. Then it adopts attention to multivariate correlations as follows:

$$A_h = Sofmax(\frac{(Z_V W_h^Q)(Z_V W_h^K)^T}{\sqrt{d_K}})(Z_V W_h^V) \tag{2}$$

Whether temporal or variate as mentioned above, both transform sequences of MTS to a latent space to provide the dependency pattern between sequences. The point of our research is to demonstrate whether the mapping to latent space under conventional attention is optimal or if we can find a better one for MTSF.

FSatten

FSatten is an innovative attention mechanism that we propose to explore the effectiveness of conventional attention. The intuitive difference from the conventional attention, as depicted in Figure 3, is that FSatten replaces the embedding by a Fourier transform and the linear projection for the query and key by a proposed MSS.

Method

We apply the FSatten to Variate Transformer for MTSF. As shown in Figure 3 right, each discrete variate sequence of the input X is first transformed by the Fast Fourier Transform (FFT) (Brigham and Morrow 1967), which efficiently computes the Discrete Fourier Transform (DFT) from the time domain to the complex frequency domain as:

$$X_k^F = \sum_{t=0}^{L} Xe^{-i(2\pi/l)kt}, 0 \le k \le F \tag{3}$$

Here, i is the imaginary unit, and the exponential term represents the Fourier basis associated with the different k frequencies. The value of F is typically half the number of data points L in FFT. According to our consideration, correlations can be made up of associated frequency components with different phases. Thus, we extracted the amplitudes of different frequencies from the complex domain as follows:

$$A_k = |X_k^F| = \sqrt{Re(X_k^F)^2 + Im(X_k^F)^2} \tag{4}$$

Where Re represents the real part of X_k^F and Im represents the imaginary part. We then apply the MSS module for the projection of queries and keys, replacing the conventional linear projection. The aim is to compare the learnable latent space for generating attention weights with the fixed frequency domain space. Since predictions are made in the time domain, the embedding and linear projection for the value remain unchanged.

$$Q = MSS_Q(A_k), K = MSS_K(A_k), V = Linear_V(Emb(X)) \tag{5}$$

After the multi-head dot product, the subsequent Feed-Forward Network (FFN) provides complicated representations by adding random noise for each variate token (Hornik 1991).

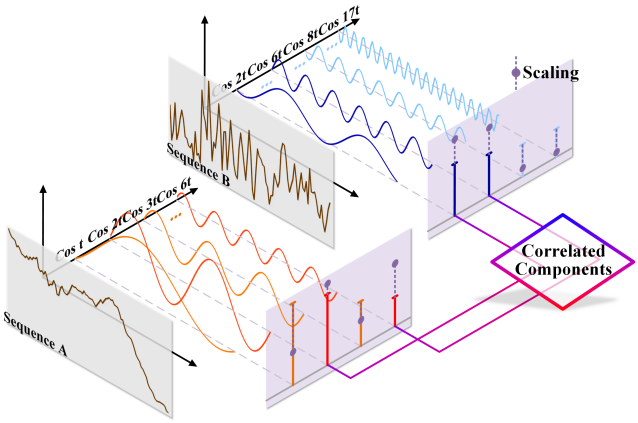

Figure 4: Multi-head Spectrum Scaling. After the Fast Fourier Transform (FFT), the correlated frequency components within the frequency domain between A and B are determined by scaled amplitude values as indicated by the purple points.

MSS

Although we remove the phase interference, it enables us to capture both synchronous and asynchronous associations. A more critical problem, however, is identifying the accurate associated frequency components. This is due to significant differences in amplitude values and information intensity for the same frequency across different sequences. As illustrated in Figure 4, the frequency from sequence B that is potentially associated with sequence A may not represent the most significant periodic characteristics of sequence A. To efficiently obtain potentially correlated frequency components, we designed a Multi-head Spectrum Scaling (MSS) projection for queries and keys. For each attention head h, we scale the amplitude value A_k of each frequency dimension using the Hadamard Product:

$$MSS(A_k) = A_k \circ W_h \tag{6}$$

where $W_h \epsilon\ R^{C\times F}$. Therefore, MSS uses H different W_h matrices to map the A_k. After learnable scaling, some frequency components can adaptively align with potentially correlated components from other sequences so that more accurate dependency patterns between sequences can be found in the subsequent dot product attention. From another consideration of maintaining the orthogonality of the frequency bias, we replace the fully connected projection, which might alter the angles between vectors and disrupt the orthogonality. Ablation experiments as well as those applied to the subsequent SOatten, demonstrate the effectiveness of MSS.

SOatten

Indeed, experiments show that FSatten outperforms conventional attention, but its performance across six real-world datasets exhibits significant variance, as shown in Figure 1 and Table 1. This suggests that a fixed frequency domain mapping may not be universally applicable. Determining the optimal configuration manually for each scenario is challenging, given the numerous unexplained physical transformations. Therefore, we aim to further develop FSatten to design a method with better generalization capabilities. Furthermore, FSatten may not be fully compatible with Temporal Transformers, considering that sequences derived from a single variate naturally exhibit identical periodic frequencies.

Method

This is a research direction with many possible approaches, but the most straightforward idea is to extend by leveraging the orthogonality of the Fourier transform. Therefore, we propose SOatten, as depicted in Figure 5, which improves the frequency domain to a more general orthogonal domain through a learnable orthogonal transformation, described as:

$$Orth(X) = W^T X,\ where\ W^T W = I \tag{7}$$

In fact, the orthogonality of the learnable space is not guaranteed during backpropagation for parameter updates. We could apply a measure such as QR decomposition, but this could result in the loss of some gradient information. We made a different trade-off for MTSF, considering that the scale of models that match the size of the dataset is usually not very large, with relatively few layers. Therefore, we only perform orthogonal initialization for the embedding, rather than enforcing a completely orthogonal space during backpropagation. Subsequently, SOatten also applies the MSS projection for query and key and applies linear projection for the value. Additionally, in dot product attention, we propose a Head Coupling Convolution (HCC) module operating on the heads of attention weights, which serves as an important guidance for the mapping space learning of SOatten.

HCC

FSatten provides sequence dependencies based on explicit spectral information, but extending this to a learnable orthogonal space makes it difficult to effectively determine valid characteristics as a defined periodicity in FSatten. In other words, data-driven approaches that learn an effective orthogonal space without any restrictions have requirements for the size and distribution of the dataset.

We propose a general enhancement method called Head Coupling Convolution (HCC), which leverages the constraint of similarity between neighboring sequences to guide the model in exploring feature spaces. Specifically, HCC involves performing convolution operations on the attention weights within the dot product attention mechanism as:

$$Atten = Softmax(\frac{QK^T}{\sqrt{d_K}}) \tag{8}$$

$$HCC(Atten) = ReLU\left\{Conv_{H\to H}(Atten, S, K)\right\} \tag{9}$$

Where S is stride, K is the kernel size, $Conv_{H\to H}$ is channel fusion convolution that maps from H heads to H heads and padding is necessary for keeping the size of the weight matrix. For most time-series data, contrastive learning methods (Yue et al. 2022) (Kiyasseh, Zhu, and Clifton 2021) (Yèche et al. 2021) (Tonekaboni, Eytan, and Goldenberg

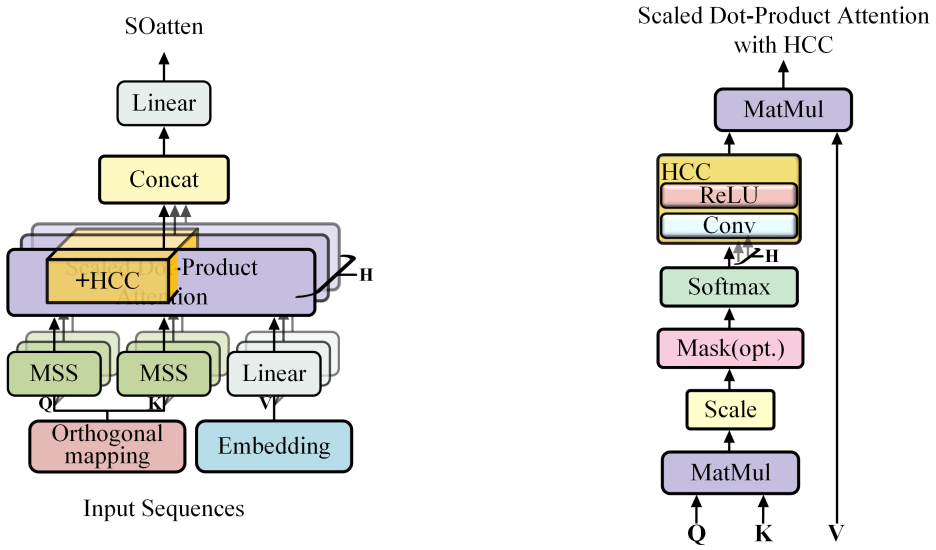

Figure 5: (left) SOatten. (right) Scaled Dot-Product Attention with HCC. On the left side of the Soatten is the shape of the data at each stage, and adding batch size to the front is the shape in training.

Models	**SOatten(V) (Ours)**		**FSatten (Ours)**		iTransformer (2024)		PatchTST (2023)		Crossformer (2023)		TiDE (2023)		TimesNet (2023)		DLinear (2023)		SCINet (2022)		FEDformer (2022)	
Metric	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTm1	0.394	0.402	0.394	0.405	0.407	0.410	**0.387**	**0.400**	0.513	0.496	0.419	0.419	0.400	0.406	0.403	0.407	0.485	0.481	0.448	0.452
ETTm1	0.287	0.331	0.286	0.331	0.288	0.332	**0.281**	**0.326**	0.757	0.610	0.358	0.404	0.291	0.333	0.350	0.401	0.571	0.537	0.305	0.349
ETTh1	0.447	0.440	0.446	**0.439**	0.454	0.447	0.469	0.454	0.529	0.522	0.541	0.507	0.458	0.450	0.456	0.452	0.747	0.647	**0.440**	0.460
ETTh2	**0.379**	**0.405**	0.381	0.407	0.383	0.407	0.387	0.407	0.942	0.684	0.611	0.550	0.414	0.427	0.559	0.515	0.954	0.723	0.437	0.449
ECL	0.166	0.259	**0.162**	**0.257**	0.178	0.270	0.216	0.304	0.244	0.334	0.251	0.344	0.192	0.295	0.212	0.300	0.268	0.365	0.214	0.327
Exchange	0.359	0.404	0.363	0.406	0.360	**0.403**	0.367	0.404	0.940	0.707	0.370	0.413	0.416	0.443	**0.354**	0.414	0.750	0.626	0.519	0.429
Traffic	0.437	0.286	0.477	0.291	**0.428**	**0.282**	0.555	0.362	0.550	0.304	0.760	0.473	0.620	0.336	0.625	0.383	0.804	0.509	0.610	0.376
Weather	**0.245**	**0.273**	0.249	0.275	0.258	0.279	0.259	0.281	0.259	0.315	0.271	0.320	0.259	0.287	0.265	0.317	0.292	0.363	0.309	0.360
Solar-Energy	**0.229**	0.261	0.230	**0.259**	0.233	0.262	0.270	0.307	0.641	0.639	0.347	0.417	0.301	0.319	0.330	0.401	0.282	0.375	0.291	0.381

Table 1: Long-term MTSF. Results are averaged from all prediction lengths. The input sequence length $L = 96$ and the prediction lengths $T = \{96, 192, 336, 720\}$. The **red** is the best and blue is the second. Full results are listed in Appendix C.1.

2021) have demonstrated the effectiveness of assumption: neighboring similarity, the similarity between sequences of the same time series decreases as the time lag increases. In fact, similar variates are arranged together in most datasets (detailed presentation shown in Appendix A.2). By applying a convolution operation to the attention weights, more critical correlated patterns between local neighboring sequences are extracted, guiding the parameter updates in the feature space during backpropagation. The diverse features extracted by H heads are all predicated on neighboring similarity, multi-head coupling helps to obtain more precise associative features than single-channel convolution.

Experiments

We extensively evaluate the proposed FSatten and SOatten on six real-world datasets, including ECL, ETT (4 subsets), Exchange, Traffic, Weather (Wu et al. 2021), and Solar-Energy (Lai et al. 2018). Detailed dataset descriptions are provided in Appendix B.1. We choose 9 well-known forecasting models as our baselines. The experimental setting is the same as in iTransformer (Liu et al. 2023).

Long-term MTSF

Compared to the baselines presented in Table 1, FSatten, based on the Variate Transformer, shows overall better forecasting performance than the SOTA which uses conventional attention mechanisms. Particularly for datasets with more pronounced periodicity, such as on ECL, FSatten significantly improves performance by an overall 8.1% compared to SOTA and exhibits greater stability for longer prediction sequences. These improvements demonstrate that FSatten effectively captures the accurate correlation at the same frequency, which is more suitable for application in Transformers for MTSF.

Periodicity is one of the most fundamental characteristics of time series, but not all datasets exhibit strong periodicity. Thus, as a more general approach that can be adapted to both Temporal and Variate transformers, SOatten achieves more

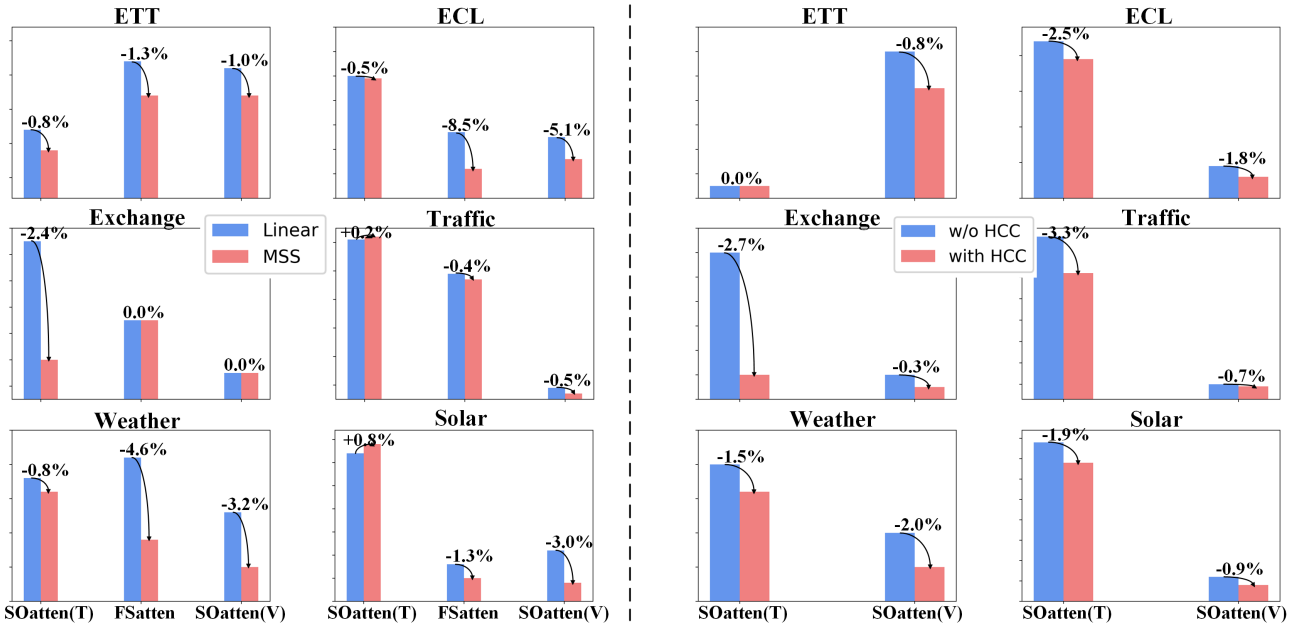

Figure 6: Ablations. (Left) for the MSS. (Right) for the HCC module. w/o represents SOatten without HCC module. MSE scores are averaged from all prediction lengths.

comprehensive improvements relative to FSatten across different scenarios. We can observe in Table 2 that, although each Transformer excels on certain datasets, SOatten consistently outperforms conventional attention mechanisms, regardless of the architecture. Of course, FSatten can provide superior performance for datasets that are known to exhibit stronger periodicity.

Ablation Studies

The effectiveness of the MSS mapping module, used in both methods, is compared with that of applying a linear mapping to FSatten and SOatten, as shown in Figure 6 left. MSS significantly outperforms the fully connected layer, corroborating its ability to identify more accurate associated components. We validate the effectiveness of the HCC module in SOatten in Figure 6 right. The HCC is an important design for SOatten, significantly enhancing forecasting performance. Especially under the Temporal Transformer, the HCC demonstrates better generalizability. These results prove that neighboring similarity is crucial for the formation of an effective orthogonal mapping space and the generation of accurate attention weights.

Visualized Analysis

First, we make visualizations of generated attention matrices and analyze the advantages of the proposed two attention mechanisms. In the upper part of Figure 7, under the Variate Transformer, SOatten and FSatten generate smaller ranges but more refined weight values than the conventional attention applied by the SOTA, iTransformer (compare the value range on the right side of heatmaps). There are three main points of analysis:

Models	Temporal Transformer				Variate Transformer			
	SOatten(T)		PatchTST		**SOatten(V)**		iTransformer	
Metric	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTm1	**0.380**	**0.395**	0.387	0.400	**0.394**	**0.402**	0.407	0.410
ETTm2	**0.280**	0.326	0.281	0.326	**0.287**	**0.331**	0.288	0.332
ETTh1	**0.451**	**0.441**	0.469	0.454	**0.447**	**0.440**	0.454	0.447
ETTh2	**0.366**	**0.395**	0.387	0.407	**0.379**	**0.405**	0.383	0.407
ECL	**0.199**	**0.282**	0.216	0.304	**0.166**	**0.259**	0.178	0.270
Exchange	**0.360**	**0.398**	0.367	0.404	**0.359**	0.404	0.360	0.403
Traffic	**0.492**	**0.310**	0.555	0.362	0.437	0.286	0.428	0.282
Weather	**0.256**	**0.280**	0.259	0.281	**0.245**	**0.273**	0.258	0.279
Solar	**0.259**	**0.284**	0.270	0.307	**0.229**	**0.261**	0.233	0.262

Table 2: Forecasting results of SOatten under Temporal and Variate Transformers. **Bolded** results are superior to conventional attention. Full results are listed in Appendix C.1.

(1) In the generated attention weight maps, the patterns of the conventional attention and FSatten show similarities, presenting dependencies that are based on the sequence periodicity. However, FSatten significantly reflects complex associations from more variable sequences, which is a benefit from the designed spectrum correlating in frequency domain space.

(2) The weights map generated by SOatten is significantly different, seemingly finding accurate dependencies based on other associated physical characteristics in addition to periodicity (as seen in the upper left part of SOatten's attention map). Furthermore, if the HCC module is not used (shown in the Appendix D), SOatten finds new physical quantities but fails to produce a comprehensive dependency pattern, proving the effectiveness of the neighboring similarity design.

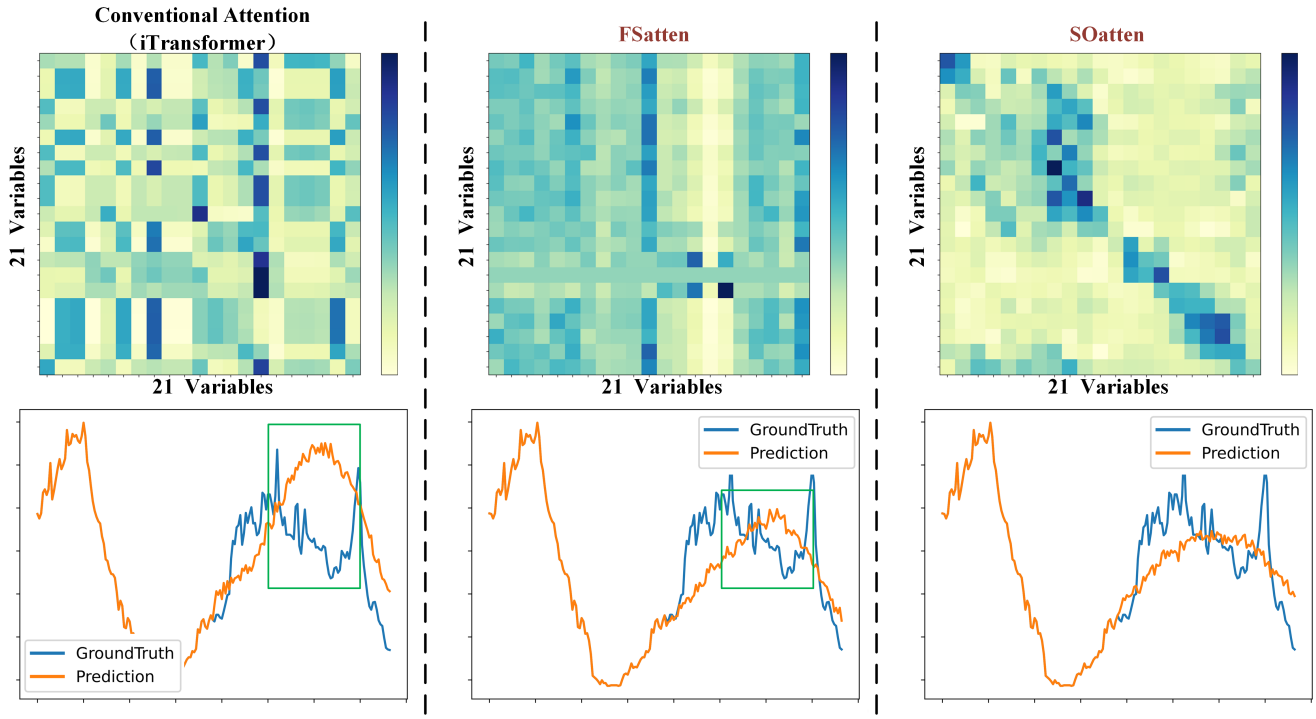

Figure 7: Attention maps and the forecasting of a few time series from Weather dataset under Variate Transformer. The attention map is calculated by averaging the attention matrices over all the heads and across all the layers.

(3) Numerical analysis of the weight matrices (in Appendix D) shows that the proposed FSatten and SOatten are both full rank (21), while the conventional attention weight matrix is not full rank (19). The condition numbers of the weight matrices generated by FSatten $(1, 519)$ and SOatten $(1, 480)$ are much smaller than that of the conventional attention $(78, 596, 560)$. These indicate that the orthogonal spaces of FSatten and SOatten are more informative and have better stability against noise than the latent space of conventional attention.

Secondly, in the lower part of Figure 7, the predictions indicate that the conventional attention mechanism's fits are poor, which appears to have learned inaccurate periodic patterns. By leveraging the frequency domain and the MSS module, FSatten finds a more accurate pattern that combines periodic dependencies based on frequency spectrum. SOatten finds an even better pattern by combining periodicity and other key physical characteristics thereby avoiding prediction errors caused by an exclusive reliance on periodicity, as shown in the green box in Figure 7.

Hyperparameter Sensitivity

Compared to conventional attention, the new hyperparameters are the dimension size of the orthogonal mapping space F in MSS, and the stride S, kernel size K in HCC. In FSatten, F is typically set to $F = (\frac{L}{2} + 1)$. Experiments on various F show small variance, demonstrating that the performance of the proposed attention mechanisms is not coincidental. Secondly, we compared different HCC stride S and kernel size K under both the Temporal and Variate Transformers. The experiment shows that the performance is optimal when the stride value is set to 1 and the model performance using HCC with different kernel sizes is consistently better than SOTA, demonstrating the effectiveness of local neighboring similarity. A 3×3 convolutional kernel is found to be the most appropriate setting. (detailed results are presented in Appendix C.3).

Efficiency

It can be observed in Appendix C.4 that FSatten has slightly improved efficiency by replacing the original linear mapping of Query and Key with an FFT. Secondly, because MSS is a Hadamard product, it can enhance the efficiency relative to fully connected layers to some extent.

Limitation and Future Work

For the MTSF problem, we propose two innovative attentions that are superior to conventional attention. We started from the frequency domain and made preliminary explorations based on the mainstream Temporal and Variate Transformers. Limitation is the performance of scenarios with a large number of variables, like Traffic. In the future, based on the periodicity and learnable characteristics, we will take advantage of modern state-space models such as Mamba (Gu and Dao 2023) (Patro and Agneeswaran 2024) that can compress the larger variable background and selectively retain the most important information.

References

Brigham, E. O.; and Morrow, R. 1967. The fast Fourier transform. *IEEE spectrum*, 4(12): 63–70.

Gu, A.; and Dao, T. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

Hornik, K. 1991. Approximation capabilities of multilayer feedforward networks. *Neural networks*, 4(2): 251–257.

Jin, M.; Wang, S.; Ma, L.; Chu, Z.; Zhang, J. Y.; Shi, X.; Chen, P.-Y.; Liang, Y.; Li, Y.-F.; Pan, S.; et al. 2023. Time-llm: Time series forecasting by reprogramming large language models. *arXiv preprint arXiv:2310.01728*.

Kiyasseh, D.; Zhu, T.; and Clifton, D. A. 2021. Clocs: Contrastive learning of cardiac signals across space, time, and patients. In *International Conference on Machine Learning*, 5606–5615. PMLR.

Lai, G.; Chang, W.-C.; Yang, Y.; and Liu, H. 2018. Modeling long-and short-term temporal patterns with deep neural networks. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, 95–104.

Li, S.; Jin, X.; Xuan, Y.; Zhou, X.; Chen, W.; Wang, Y.-X.; and Yan, X. 2019. Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting. *Advances in neural information processing systems*, 32.

Liu, S.; Yu, H.; Liao, C.; Li, J.; Lin, W.; Liu, A. X.; and Dustdar, S. 2021. Pyraformer: Low-complexity pyramidal attention for long-range time series modeling and forecasting. In *International conference on learning representations*.

Liu, Y.; Hu, T.; Zhang, H.; Wu, H.; Wang, S.; Ma, L.; and Long, M. 2023. itransformer: Inverted transformers are effective for time series forecasting. *arXiv preprint arXiv:2310.06625*.

Nie, Y.; Nguyen, N. H.; Sinthong, P.; and Kalagnanam, J. 2022. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*.

Patro, B. N.; and Agneeswaran, V. S. 2024. Simba: Simplified mamba-based architecture for vision and multivariate time series. *arXiv preprint arXiv:2403.15360*.

Tonekaboni, S.; Eytan, D.; and Goldenberg, A. 2021. Unsupervised representation learning for time series with temporal neighborhood coding. *arXiv preprint arXiv:2106.00750*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wu, H.; Xu, J.; Wang, J.; and Long, M. 2021. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in Neural Information Processing Systems*, 34: 22419–22430.

Xu, Z.; Zeng, A.; and Xu, Q. 2023. FITS: Modeling time series with $10k$ parameters. *arXiv preprint arXiv:2307.03756*.

Yèche, H.; Dresdner, G.; Locatello, F.; Hüser, M.; and Rätsch, G. 2021. Neighborhood contrastive learning applied to online patient monitoring. In *International Conference on Machine Learning*, 11964–11974. PMLR.

Yue, Z.; Wang, Y.; Duan, J.; Yang, T.; Huang, C.; Tong, Y.; and Xu, B. 2022. Ts2vec: Towards universal representation of time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 8980–8987.

Zhang, Y.; and Yan, J. 2022. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *The Eleventh International Conference on Learning Representations*.

Zhou, H.; Zhang, S.; Peng, J.; Zhang, S.; Li, J.; Xiong, H.; and Zhang, W. 2021. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 11106–11115.

Zhou, T.; Ma, Z.; Wen, Q.; Wang, X.; Sun, L.; and Jin, R. 2022. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *International Conference on Machine Learning*, 27268–27286. PMLR.

Zhou, T.; Niu, P.; Sun, L.; Jin, R.; et al. 2023. One fits all: Power general time series analysis by pretrained lm. *Advances in neural information processing systems*, 36: 43322–43355.