SDformer: Transformer with Spectral Filter and Dynamic Attention for Multivariate Time Series Long-term Forecasting

Ziyu Zhou\textsuperscript{1}, Gengyu Lyu\textsuperscript{1*}, Yiming Huang\textsuperscript{1}, Zihao Wang\textsuperscript{1}, Ziyu Jia\textsuperscript{2} and Zhen Yang\textsuperscript{1}

\textsuperscript{1}Faculty of Information Technology, Beijing University of Technology, Beijing, China
\textsuperscript{2}Institute of Automation, Chinese Academy of Sciences, Beijing, China
{ziyuzhou30, rex.wangzhao} @gmail.com, huangyiming2002@126.com, jia.ziyu@outlook.com, {lyugengyu, yangzhen} @bjut.edu.cn

Abstract

Transformer has gained widespread adoption in modeling time series due to the exceptional ability of its self-attention mechanism in capturing long-range dependencies. However, when processing time series data with numerous variates, the vanilla self-attention mechanism tends to distribute attention weights evenly and smoothly, causing row-homogenization in attention maps and further hampering time series forecasting. To tackle this issue, we propose an advanced Transformer architecture entitled SDformer, which designs two novel modules, Spectral-Filter-Transform (SFT) and Dynamic-Directional-Attention (DDA), and integrates them into the encoder of Transformer to achieve more intensive attention allocation. Specifically, the SFT module utilizes the Fast Fourier Transform to select the most prominent frequencies, along with a Hamming Window to smooth and denoise the filtered series data; The DDA module applies a specialized kernel function to the query and key vectors projected from the denoised data, concentrating this innovative attention mechanism more effectively on the most informative variates to obtain a sharper attention distribution. These two modules jointly enable attention weights to be more salient among numerous variates, which in turn enhances the attention’s ability to capture multivariate correlations, improving the performance in forecasting. Extensive experiments on public datasets demonstrate its superior performance over other state-of-the-art models. Code is available at https://github.com/zhouziyu02/SDformer.

1 Introduction

Time series analysis holds significant value in a wide array of practical applications, including weather forecasting [Huang et al., 2023], energy management [Dong et al., 2023] and public opinion analysis [O’Connor et al., 2010]. Recently, Transformer [Vaswani et al., 2017] has fostered the modeling of long-range dependencies in sequential data, making them particularly suited for time series analysis [Wen et al., 2023]. For instance, Pyraformer [Liu et al., 2021] proposes a hierarchical pyramidal attention mechanism for Transformer architecture, which captures temporal dependencies at different ranges while ensuring its linear complexity in both time and memory. Autoformer [Wu et al., 2021] introduces a seasonal-trend decomposition operation into Transformer, which incorporates an auto-correlation mechanism to achieve more precise series-wise correlations discovery.

Although the above methods have achieved superior performance, they still suffer from some critical shortcomings. Especially, when modeling some datasets with a large number of variates (e.g., Traffic dataset with 862 variates), the vanilla self-attention mechanism often fails to effectively allocate attention weights among multiple variates, which is intuitively reflected in the homogenization between rows of the attention map. As shown in Figure 1(a)-(b), the color distribution of the attention map in iTransformer [Liu et al., 2023] and vanilla Transformer [Vaswani et al., 2017] is quite sparse with fewer intense areas, indicating a dispersed focus and an inability to prioritize critical inter-variate correlations.

To tackle this issue, we propose an innovative Transformer-based model named SDformer, which integrates two strategic modular designations, Spectral-Filter-Transform (SFT) and
Dynamic-Directional-Attention (DDA), to reallocate attention weights for increasing the heterogeneity of the attention map, further boosting the forecasting performance. Technically, in the SFT module, we utilize the Fast Fourier Transform to filter out insignificant frequencies including meaningless noise and fluctuations, accordingly preserving the essential features of the time series, such as continuity, periodicity and trend. Afterward, a bell-shaped Hamming Window is applied to the filtered data, which utilizes its spectral properties to minimize edge effects and enhance the smoothness of the series [Mottaghi-Kashiban and Shayesteh, 2011]. The above two operations significantly improve data quality by effectively reducing noise, further facilitating better representation learning in subsequent modules. In the DDA module, we introduce a novel kernel function, equipped with dynamic parameters and directional weights, which is applied to Query (Q) and Key (K) simultaneously (where Q and K are linear projected from the smoothed series in the SFT) to bring similar Q-K pairs closer to their nearest axis, yielding higher attention scores, while distancing dissimilar Q-K pairs to their opposite axis, resulting in relatively lower scores. Such operation sharpens the distribution of attention weights across numerous variates, and consequently makes the self-attention mechanism more capable of identifying and prioritizing key inter-variate patterns, thereby achieving higher expressiveness and effectively mitigating the ‘smooth’ problem.

As a result, Figure 1(c) visually shows more intense colors in the attention map compared with the other two maps on certain regions, indicating a stronger focus capacity on important variates. In general, these two novel components collectively overcome the limitation of the vanilla attention mechanism in Transformer when modeling time series with a large number of variates, enabling a stronger capacity for forecasting multivariate time series data. The contributions of our paper are summarized in three folds:

1. We propose a novel Transformer architecture (named SDformer) for long-term time series forecasting. To the best of our knowledge, it is the first time to address the problem of smooth attention distribution when modeling time series data with a large number of variates.

2. The Spectral-Filter-Transform and Dynamic-Directional Attention modules are designed for filtering pivotal frequency features and sharpening attention distribution respectively, which jointly enable the attention weights to be more salient among variates, further fostering the attention’s ability to capture multivariate correlations and improve the final forecasting performance.

3. Extensive experiments on various datasets demonstrate the effectiveness of SDformer against other state-of-the-art methods. Especially, it achieves superior performance on some datasets with numerous variates, such as 11.6% forecasting error reduction on Traffic dataset.

2 Related Work

2.1 Forecasting with Special Attention Mechanism

In multivariate time series forecasting, several models focus on innovating attention mechanisms to enhance the forecasting performance. For example, Informer [Zhou et al., 2021] and Reformer [Kitaev et al., 2020] leverage ProbSparse and locality-sensitive hashing mechanisms in modeling long sequences data. Pyraformer [Liu et al., 2021] employs a pyramidal attention module to capture multi-resolution temporal dependency, while ContiFormer [Chen et al., 2023] merges attention with Neural ODEs for irregular time series. Moreover, PatchTST [Nie et al., 2023] focuses on excavating patch-wise correlations using self-attention. All of these attention mechanisms are designed from the perspective of balancing between computational complexity and forecasting performance. To the best of our knowledge, none of the existing attention mechanisms effectively address the issue of overly smooth attention distribution in time series analysis.

2.2 Forecasting in the Frequency Domain

Except for the above attention innovation-based methods, various methods harness strategies in the frequency domain with different foundation models to improve forecasting results. SMF [Zhang et al., 2017] decomposes time series into distinct frequency components for varied time horizon forecasts. StemGNN [Cao et al., 2020] combines GNN and Discrete Fourier Transforms to examine both inter-series and intra-series correlations. CoST [Woo et al., 2022a] and ATFN [Yang et al., 2020] focus on seasonal-trend separation and dynamic periodic pattern capture. FrTTS [Yi et al., 2023] leverages its frequency-domain MLPs to explore global dependencies and enhance key frequency component learning. Although these models signify a trend towards exploring frequency domain strategies in time series analysis, they often directly utilize enhanced data for subsequent learning without considering potential noise. This oversight leads to the inadvertent amplification of noise components during frequency decomposition, compromising the forecasting accuracy.

2.3 Forecasting with Special Attention and Methods in the Frequency Domain

Based on the above two types of methods, some recent methods combine both innovative attention mechanisms and frequency domain analysis to achieve notable progress in time series forecasting. For example, Autoformer [Wu et al., 2021] utilizes an Auto-Correlation mechanism for improved temporal pattern recognition. FEDformer [Zhou et al., 2022] fuses Transformer architecture with seasonal-trend decomposition, exploiting frequency domain sparsity for thorough analysis. ETSformer [Woo et al., 2022b] introduces exponential smoothing attention, enhancing both efficiency and interpretability. However, existing models overlook the amplification of noise during frequency decomposition and the challenge of overly smooth attention distribution. To the best of our knowledge, SDformer is the first model to tackle these overlooked issues by integrating the SFT and DDA into the Transformer architecture, which significantly enhances the accuracy of multivariate time series forecasting.
3 SDformer

3.1 Problem Formulation

Formally speaking, we denote a multivariate time series by \([X_1, X_2, \ldots, X_T] \in \mathbb{R}^{N \times T}\), where each \(X_t \in \mathbb{R}^N\) corresponds to the observations of \(N\) variates at the \(t\)-th timestamp for \(T\) total timestamps. For any given time \(t\), the input to the model is a window of the preceding \(L\) observations, designated as \(X_t = [X_{t-L+1}, X_{t-L+2}, \ldots, X_t] \in \mathbb{R}^{N \times L}\). The forecasting objective at time \(t\) aims to predict the subsequent \(\tau\) values, represented as \(Y_t = [X_{t+1}, X_{t+2}, \ldots, X_{t+\tau}] \in \mathbb{R}^{N \times \tau}\). The forecasting model, denoted by \(f_\theta\), utilizes the historical data \(X_t\) to estimate the future values \(\hat{Y}_t\), such that the forecast is given by \(\hat{Y}_t = f_\theta(X_t)\).

3.2 Overall Architecture

Figure 2 shows the overall framework of SDformer including the Spectral-Filter-Transform (SFT) module, the Embedding operation, the stacked Dynamic-Directional-Attention (DDA) Encoder Blocks module and the Projection operation. Technically, the input time series is denoised in the SFT, and its variates are independently embedded into separate tokens [Liu et al., 2023]. These tokens are then fed into stacked DDA Encoder Blocks to extract complex representation, where the DDA achieves improved inter-variate correlations discovery through its unique kernel function, while layer normalization and feed-forward network with residual connections are utilized to learn temporal dependencies and alleviate series non-stationarization. Finally, the representations are decoded for final results via the Projection operation. Since our model focuses on addressing the insufficiency of attention mechanism in modeling time series with numerous variates, in this paper, we chiefly introduce the designed SFT and DDA, and explain how they discover effective multivariate correlations.

3.3 Spectral-Filter-Transform (SFT)

The Spectral-Filter-Transform (SFT) module plays a pivotal role in denoising and smoothing the input time series data at the beginning of our model. It achieves this data augmentation through a two-stage cross-domain process as illustrated in Algorithm 1: Frequency Domain Denoising (step 5, 6 and 7).

![Figure 2: The framework of our proposed SDformer: (a) The Spectral-Filter-Transform (SFT) is a novel preprocessing method for time series, enhancing data analysis by retaining key frequency components. (b) The Dynamic-Directional-Attention (DDA) mechanism is an improvement over the vanilla self-attention, as it dynamically modulates the closeness of Q-K pairs according to their similarity.](image-url)
7) and Time Domain Smoothing (step 8 and 9). It is noteworthy that, before conducting the frequency domain denoising, we treat multivariate time series as multiple univariate time series (inspired by the ‘channel-independence’ designation [Nie et al., 2023]), processing them individually in the SFT and eventually concatenating all the processed univariate series for subsequent operations.

Frequency Domain Denoising
Denoising is crucial in time series analysis to enhance forecasting accuracy by mitigating the impact of noise on the identification of periodic or trend patterns. In this paper, we propose a new denoising strategy that converts the time series into its corresponding frequency domain to alleviate the random environmental noise and white noise in the time series. Specifically, we first utilize the Fast Fourier Transform (FFT) to convert a univariate time series \( x \in \mathbb{R}^T \) with length \( T \) into the frequency domain \( X_f \), i.e.,

\[
X_f = \sum_{t=0}^{T-1} x[t]e^{-\frac{2\pi i}{T}kt}, \quad k = 0, \ldots, T-1. \tag{1}
\]

Then, we retain the highest \( k \) frequency components to filter out insignificant frequencies by \( X_{f_k} = \text{TopK}(X_f, k) \), where \( \text{TopK}(X_f, k) \) selects the \( k \) largest amplitudes and \( k \) is the hyper-parameter. Afterward, the retained frequency spectrum \( X_{f_k} \) is transformed back into its corresponding time domain \( x_{i_f} \) by adopting the Inverse Fast Fourier Transform (IFFT) for further analysis in subsequent operations, i.e.,

\[
x_{i_f} = \frac{1}{T} \sum_{k=0}^{T-1} X_{f_k} e^{\frac{2\pi i}{T}kt}, \quad t = 0, \ldots, T-1. \tag{2}
\]

Time Domain Smoothing
After obtaining the denoised time series \( x_{i_f} \), we further process the time series by implementing smoothing techniques to mitigate spectral leakage effects [Mottaghi-Kashitiban and Shayesteh, 2011], where a convolution operation by windowing function is useful to smooth transitions at the boundaries and reduce discontinuities. Specifically, we first define a bell-shaped Hamming Window with the size of \( w \) (\( w \) is an even number) and a window function

\[
w[n] = 0.54 - 0.46 \cos \left( \frac{2\pi n}{w} \right), \quad n = 1, \ldots, w, \tag{3}
\]

where \( n \) indexes the sample points within the window. Subsequently, the denoised series \( x_{i_f} \) from Eq.2 is padded reflectively to ensure its length matches the window size \( w \) by

\[
x_p[n] = \begin{cases} x_{i_f}[\frac{w}{2} - n], & 1 \leq n \leq \frac{w}{2} \\ x_{i_f}[n - \frac{w}{2}], & \frac{w}{2} < n \leq N + \frac{w}{2} \\ x_{i_f}[N + w - n], & N + \frac{w}{2} < n \leq N + w \end{cases} \tag{4}
\]

According to Eq.4, the first and last \( \frac{w}{2} \) elements of the filtered series \( x_{i_f} \) are mirrored and appended at the beginning and end of the series respectively. Such operation effectively extends the series \( x[n] \) to a new length of \( N + w \), where \( N \) is the original length of \( x_{i_f} \), thereby adapting to the subsequent convolution smoothing operation, i.e.,

\[
x_h[t] = \frac{\sum_{n=1}^{w} x_p[t + n] \cdot w[n]}{\sum_{n=1}^{w} w[n]}, \quad t = 1, \ldots, T, \tag{5}
\]

where the smoothed series \( x_h[t] \) is calculated by taking a weighted average of the series at each time point \( t \) and \( T \) denotes the total length.

After the above operations, we concatenate all \( x_h \) corresponding to \( N \) variates to obtain a multivariate time series \( X \in \mathbb{R}^{T \times N} \) with length \( T \) and the number of variates \( N \), where the \( X \) maintains the same shape as the input sequence of the SFT, simultaneously achieving both denoising and smoothing effectively.

3.4 Dynamic-Directional-Attention (DDA)
To capture more reliable multivariate correlations, we introduce a new Dynamic-Directional-Attention (DDA) mechanism to calculate attention weights for more effective attention distribution. The DDA mechanism operates on the principle of dynamically reorienting and scaling the Query \( Q \in \mathbb{R}^{M \times T \times E} \) and Key \( K \in \mathbb{R}^{L \times T \times E} \), where \( M \) and \( L \) are sequence lengths, \( T \) is the number of heads, and \( E \) denotes dimensions per head. The attention score within each head of the DDA is formulated as:

\[
\text{Score}(Q_i, K_j) = \phi_p(Q_i) \phi_p(K_j)^T, \tag{6}
\]

where \( \phi_p(x) = f_p(\tan(x)) \) is a specially designed function that is applied to both the query and key simultaneously. Here, \( \phi_p(x) \) is defined by the composition of a non-linear mapping \( \tan(x) \) and a special kernel function \( f_p(x) = x \cdot w_{\text{dir}} \cdot (\text{std}(x))^{-p} \cdot \lambda_{\text{dyn}} \), where \( p \) is the element-wise power, \( w_{\text{dir}} \) and \( \lambda_{\text{dyn}} \) are learnable parameters representing directional weight and dynamic parameter, respectively. These two parameters contribute to the ‘dynamic’ effect of our proposed DDA. Moreover, \( \text{std}(x) \) represents the standard deviation of the input \( x \).

In addition, the DDA introduces a dynamic scaling factor \( \tau \) to calculate the attention weight \( A \), which is formulated as:

\[
A = \text{Dropout} \left( \text{Softmax} \left( \frac{\text{scale} \cdot \text{Score}}{\tau} \right) \right), \tag{7}
\]

where \( \tau = \sqrt{\text{var}(\text{Score})} \) dynamically normalizes the scores, and \( \text{var}(\text{Score}) \) calculates the variance of score, \( \text{scale} \) is a pre-softmax scaling factor. A dropout procedure follows the softmax to regularize the attention weights \( A \), reducing overfitting by randomly zeroing a portion of the weights during the training phase. Finally, the Output of this module is computed as a weighted sum of the value matrix:

\[
\text{Output} = \sum_x A \cdot V. \tag{8}
\]

Analysis on the Kernel Function
The essence of the DDA module lies in its specialized kernel function \( f_p \), which effectively sharpens the contrast in the attention distribution for all the Q-K pairs, so that the issue
of smooth attention weight distribution can be alleviated. To illustrate its effect more clearly, Figure 3 provides an example demonstrating the impact of \( f_p \). Essentially, \( f_p \) draws each vector closer to its nearest axis. The parameter \( p \) is crucial here, as it determines the extent of this vector reorientation. This process aids in categorizing the vectors into distinct groups based on their proximity to specific axes as illustrated in the two groups in Figure 3. As a result, the similarities of Q-K pairs within the same group are enhanced, while those within different groups are weakened. Here, 'similarity' represents the attention score; therefore, the amplified similarity distribution results in a sharper attention weight distribution.

4 Experiments

4.1 Experimental Setting

Datasets. We employ seven real-world time series benchmarks for comparative study, including Weather, Exchange, Traffic, Illness, Electricity and ETT (2 subsets).\(^1\)

Baselines. We compare SDformer with eleven state-of-the-art models from three categories, including (1) Transformer-based models: iTransformer [Liu et al., 2023], PatchTST [Nie et al., 2023], FEDformer [Zhou et al., 2022], Autoformer [Wu et al., 2021]. (2) TCN-based model: TimesNet [Wu et al., 2023] and (3) MLP-based models: DLinear [Zeng et al., 2022], Autoformer [Wu et al., 2021] and DLinear [Zeng et al., 2023].

Evaluation Metrics. Two widely-used evaluation metrics including Mean Squared Error (MSE) and Mean Absolute Error (MAE) are employed for quantifying the accuracy of predictions among all comparing methods.

Implementation Details. Our experiments are conducted using PyTorch on a single NVIDIA RTX3090 24GB GPU. For model optimization, we employ the ADAM optimizer [Kingma and Ba, 2014] to optimize the L2 loss, selecting an initial learning rate from \( \{ 1 \times 10^{-3}, 5 \times 10^{-4}, 1 \times 10^{-4}\} \).

4.2 Experimental Results

Table 1 summarizes the notable superiority of SDformer in long-term forecasting, particularly on some datasets with numerous variates, e.g., Traffic dataset. Specifically, SDformer has an average decrease of 15% and 8.5% in MSE and MAE over the previous SOTA iTransformer [Liu et al., 2023], highlighting its enhanced ability in modeling high-dimensional time series data as well as capturing reliable multivariate correlations. Besides, SDformer also exhibits its superior ability across other datasets. For instance, it outperforms another SOTA DLinear [Zeng et al., 2023] by 17% in MSE and MAE on ETTm2 dataset. In a nutshell, the enhanced ability to handle complex multivariate correlations makes SDformer particularly suitable for tackling the intricacies of multivariate time series forecasting challenges.

4.3 Model Analysis

The Effectiveness of the SFT Module

Figure 4 presents two comparative case analyses on the Spectral-Filter-Transform (SFT), where the solid lines represent the post-SFT series and the corresponding dashed lines depict the pre-SFT series. The solid lines exhibit fewer fluctuations, indicating a conspicuous attenuation of noise and an enhanced smoothness compared to their dashed counterparts. This denoising and smoothing effect is pivotal for the subsequent extraction of semantic temporal patterns like trends and continuity, and is also essential for discovering multivariate correlations, further benefiting precise series forecasting.

The Effectiveness of the DDA Module

Figure 5: Query and Key shows a more intensive distribution after applying the kernel function. Visualization after PCA and t-SNE.
Table 1: Long-term forecasting results with forecasting lengths $O \in \{24, 36, 48, 60\}$ for ILI, and $O \in \{96, 192, 336, 720\}$ for other datasets and fixed lookback length $I = 96$. Results are averaged from various prediction lengths. Red: best, Blue: second best.

Figure 6: The visualization of attention maps for two sets of time series from the Traffic dataset in Case 1 and Case 2. It highlights specific points within the learned score maps that represent different degrees of similarity of variates. Due to the large size of the attention maps (i.e. 866x866), a close-up section of the entire map is captured to more clearly showcase the color differences of individual ‘pixels’ in each map.
a more uneven distribution [Han et al., 2023]. As a result, SDformer consistently shows higher Gini coefficients across all layers than iTransformer, highlighting the more concentrated attention weights distribution. Furthermore, the rank in SDformer reaches 459 in the fourth layer, surpassing iTransformer’s 296, which suggests a greater diversity in feature representation. Collectively, these quantitative results corroborate the effectiveness of SFT and DDA in overcoming attention matrix homogeneity, illustrating their efficiency in identifying key variates to capture multivariate correlations.

4.4 Further Analysis

Ablation Study

We conduct ablation studies on Weather and ETTm2 datasets to validate the indispensability of our proposed SFT and DDA module from three perspectives: (1) ‘proposed’ represents the intact SDformer proposed in this paper. (2) wo/DDA: the Dynamic-Directional-Attention in SDformer is replaced with the vanilla self-attention. (3) wo/SFT: the Spectral-Filter-Transform module is removed. Specifically, Figure 7 illustrates the results of the ablation studies, where removing any modular component will lead to performance degradation (higher MSE). Such results suggest that without the SFT, the model becomes more prone to interference from environmental or white noise, making it difficult to accurately capture the temporal patterns in time series data, which negatively impacts the forecasting accuracy. Additionally, we observe that replacing our DDA with the vanilla self-attention also results in a higher forecasting error, indicating that vanilla self-attention is less effective than our Dynamic-Directional-Attention in modeling multivariate time series with numerous variates. In summary, these two components collectively improve the model’s performance in time series forecasting.

5 Conclusion

This paper proposes SDformer, a novel Transformer-based architecture that concentrates on tackling the dispersed distribution of attention weights in Transformer when modeling time series with numerous variates. Specifically, we introduce the Spectral-Filter-Transform module to denoise and enhance the smoothness of time series. Besides, we propose the Dynamic-Directional-Attention module to sharpen the distribution of attention weights on the most informative variates. These innovations jointly boost the ability of SDformer to discern and utilize multivariate correlations. Experiments across various datasets underscore its efficiency and accuracy, marking a notable advancement in long-term forecasting tasks. Future research will explore the scalability of SDformer on large-scale real-world datasets and deploy it to support public use.
Acknowledgments

This work was supported by the National Key Research and Development Program of China (No. 2023YFB3107100), the National Natural Science Foundation of China (No. 62306020), the China Postdoctoral Science Foundation (No. 2022M702302), the Beijing Postdoctoral Science Foundation (No. 2023-2z-78), the Major Research Plan of National Natural Science Foundation of China (No. 92167102).

References


[Han et al., 2023] Dongchen Han, Xuran Pan, Yizeng Han, Shiji Song, and Gao Huang. Flatten transformer: Vision transformer using focused linear attention. In IEEE/CVF International Conference on Computer Vision, pages 5961–5971, 2023.


