DSformer: A Double Sampling Transformer for Multivariate Time Series Long-term Prediction

Chengqing Yu
Institute of Computing Technology,
Chinese Academy of Sciences,
Beijing, China
University of Chinese Academy of Sciences, Beijing, China
yuchengqing22b@ict.ac.cn

Fei Wang*
Institute of Computing Technology,
Chinese Academy of Sciences,
Beijing, China
University of Chinese Academy of Sciences, Beijing, China
wangfei@ict.ac.cn

Zezhi Shao
Institute of Computing Technology,
Chinese Academy of Sciences,
Beijing, China
University of Chinese Academy of Sciences, Beijing, China
Shaozezhi19b@ict.ac.cn

Tao Sun
Institute of Computing Technology,
Chinese Academy of Sciences,
Beijing,China
suntao@ict.ac.cn

Lin Wu
Institute of Computing Technology,
Chinese Academy of Sciences,
Beijing,China
wulin@ict.ac.cn

Yongjun Xu
Institute of Computing Technology,
Chinese Academy of Sciences,
Beijing,China
xyj@ict.ac.cn

ABSTRACT

Multivariate time series long-term prediction, which aims to predict the change of data in a long time, can provide references for decision-making. Although transformer-based models have made progress in this field, they usually do not make full use of three features of multivariate time series: global information, local information, and variables correlation. To effectively mine the above three features and establish a high-precision prediction model, we propose a double sampling transformer (DSformer), which consists of the double sampling (DS) block and the temporal variable attention (TVA) block. Firstly, the DS block employs down sampling and piecewise sampling to transform the original series into feature vectors that focus on global information and local information respectively. Then, TVA block uses temporal attention and variable attention to mine these feature vectors from different dimensions and extract key information. Finally, based on a parallel structure, DSformer uses multiple TVA blocks to mine and integrate different features obtained from DS blocks respectively. The integrated feature information is passed to the generative decoder based on a multi-layer perceptron to realize multivariate time series long-term prediction. Experimental results on nine real-world datasets show that DSformer can outperform eight existing baselines.

CCS CONCEPTS

• **Information systems** → **Data mining**.

KEYWORDS

Multivariate time series long-term prediction, Double sampling transformer, temporal variable attention block

CIKM '23, October 21–25, 2023, Birmingham, United Kingdom
© 2023 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0124-5/23/10.
https://doi.org/10.1145/3583780.3614851

ACM Reference Format:
Chengqing Yu, Fei Wang, Zezhi Shao, Tao Sun, Lin Wu, and Yongjun Xu. 2023. DSformer: A Double Sampling Transformer for Multivariate Time Series Long-term Prediction. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23), October 21–25, 2023, Birmingham, United Kingdom.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3583780.3614851

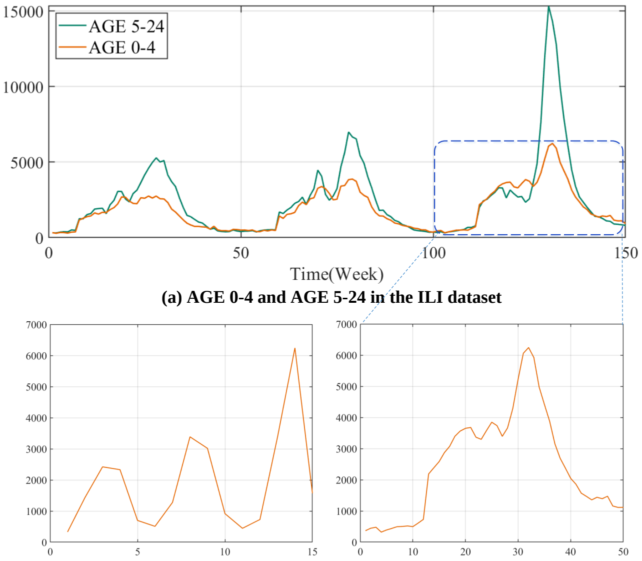
1 INTRODUCTION

Multivariate time series prediction is widely used in our life, such as weather [1], energy [31], economics [3], environment [13], traffic [33] and other fields [8] [17] [41]. Specially, multivariate time series long-term prediction can help people understand the changing trend of data for a long time in the future, which provides important references for decision-making [29] [9]. Therefore, multivariate time series long-term prediction has always been a hot topic in academia [24] and industry [4].

Multivariate time series is composed of multiple time series with correlations [23]. And these correlated time series usually fluctuate and change over time [30]. As a special sequence form different from natural language, researchers usually need to analyze the time series context relation [6] and variable correlation [39] of data to achieve long-term prediction. At present, Transformer-based models are widely studied in this field because of their powerful context relation analysis capabilities [37]. However, these models do not make full use of three features of multivariate long sequence time series. Based on Figure 1, we introduce these features next:

- **Variable correlation:** As shown in Figure 1 (a), two correlated time series show similar change patterns over time. If the model can find the relationship between these two time series, that is, variable correlation, it can mine more information and improve the modeling effect.
- **Global information:** When the sampling frequency of the AGE 0-4 data in Figure 1 (a) is increased, the raw data can be transformed into the time series shown in Figure 1 (b). By observing Figure 1 (b), we find that the processed data shows seasonality on global. In other words, the proposed data is composed of multiple similar segments. If the model

[*] Fei Wang is the corresponding author.



(a) AGE 0-4 and AGE 5-24 in the ILI dataset

(b) AGE 0-4 with larger sampling intervals

(c) A segment of AGE 0-4

Figure 1: Examples of the multivariate time series in ILI dataset. (a) Time series of variable AGE 5-24 and variable AGE 0-4 in the ILI dataset. (b) AGE 0-4 time series with larger sampling intervals. (c) A segment of AGE 0-4 time series.

finds this global information, it can predict the overall future changes of the data.

- **Local information:** As shown in Figure 1 (c), when we focus on observing one part of three similar segments of AGE 0-4 data in Figure 1 (a), we can capture more detailed local information than Figure 1 (b). Therefore, if the model can combine this information with the above global information, it won't lose local details in the process of modeling.

Based on the above analysis, if we can effectively use these three features (global information, local information and variables correlation) of multivariate long sequence time series, the model can be more suitable for long-term prediction. However, we need to address the following technical challenges: (1) How do we make our model observe these three features of the original data? (2) How to effectively integrate these features to achieve multivariate time series long-term prediction?

To mine the above three features of the multivariate long sequence time series, we propose a double sampling (DS) block and a temporal variable attention (TVA) block, which can mine these features from the following aspects: (1) The DS block uses two components (down-sampling method and piecewise sampling method) to process the raw data. The down-sampling method obtains the feature vector by extracting the original data with a larger sampling interval, as shown in Figure 1 (b). Observing the data with larger sampling intervals can reduce the influence of local noise and obtain more global information. And the piecewise sampling method obtains the local time series by splitting the original data proportionally, as shown in Figure 1 (c). Observing a continuous segment of a long sequence can enhance the utilization of local information. After processing by the DS block, we can obtain two feature vectors containing global information or local information respectively. (2) The TVA block uses a parallel modeling structure to combine temporal attention and variable attention, and mine above feature vectors. Specifically, temporal attention analyzes context relation and captures the information from temporal dimension (global information or local information). And variable attention focuses on analyzing the variable correlation. Besides, different from the traditional idea of stacking multiple layers, we use temporal attention and variable attention to mine feature vectors respectively, and then integrate the extracted information. Based on the above ideas, the TVA block can mine and integrate temporal information (global information or local information) and variable correlation. Then, we need to further integrate above three key features.

To further mine and integrate above three key features (local information, global information and variable correlation), we still use the idea of parallel modeling to mine and integrate the two feature vectors obtained by DS block. Specifically, multiple TVA blocks are used to model feature vectors obtained by DS block separately and integrate the processed features. Firstly, two TVA blocks are used to separately mine two different feature vectors obtained by DS block. And the TVA block introduce the variable correlation while mining the global information or local information owned by above two feature vectors respectively. Then, we use a TVA block to combine the above feature vectors and obtain the feature vector that integrates these key information. Finally, the integrated feature vector is transmitted to the generative decoder for prediction modeling. Based on the above blocks and modeling steps, we finally proposed the double sampling transformer (DSformer). **In general, the main contributions of this paper are shown as follows**:

- We propose a novel model for multivariate time series long-term prediction, which is called DSformer. It learns and integrates global information, local information and variables correlation of multivariate time series.
- We design a double sampling block to preprocess the original data and help the model mine the global information and local information. Besides, we propose a temporal variable attention block to mine the data from the temporal dimension and variable dimension. These two blocks are combined by a parallel structure for information integration.
- We conduct comparative experiments on nine real world data sets. The results demonstrate that DSformer can outperform eight existing SOTA models.

2 RELATED WORK

2.1 Deep learning based methods

At present, deep learning has been widely studied in the field of multivariate time series long-term prediction [32]. As one of the most classical deep learning algorithms in time series prediction, recurrent Neural Network (RNN) [22] has been widely studied. As the most classical variant of RNN, the long short-term memory network (LSTM) [19] and the gated recurrent unit (GRU) [5] have made progress in the field of time series prediction. Compared with RNN, LSTM and GRU effectively solve the gradient problem and improve their prediction accuracy [16]. In addition to RNN-based models, the convolutional neural network (CNN) [25] based models have also been proven to have effects in the field of multivariate time series long-term prediction. For example, Temporal Convolutional Network (TCN) [40] improves the ability of sequence modeling by introducing Dilated Causal Convolutions and residual connections. Besides, with the improvement of computer performance, the idea of fusing different network structures is constantly proposed [38]. LSTMa [50] improved the ability of the model to mine temporal information by combining LSTM and attention mechanism. Besides, by effectively combining LSTM, CNN and attention mechanism, LSTNet [14] achieved better results than traditional methods in multivariate time series long-term prediction. However, the above models have limitations in mining the key context information of long sequence and the correlation of different variables, which limits their performance.

2.2 Transformer based methods

At present, Transformer variants have seen rapid growth in multivariate time series long-term prediction [43]. Li et al. [15] used the convolutional self-attention mechanism to improve the sequence modeling ability of the traditional transformer and proposed the LogSparse transformer (LogTrans). Kitaev et al. [12] combined Locality sensitive hashing attention with reversible residual layers to improve the ability to analyze long-term dependencies and proposed Reformer. Zhou et al. [47] proposed Informer by introducing a ProbSparse self-attention mechanism and the generative decoder. Liu et al. [21] proposed Pyraformer, which introduces the pyramidal attention module and multi-resolution modeling approach. The above models focus on optimizing the ability of attention to analyze the long-term dependence, but they do not fully analyze the characteristics of time series. Different from the above methods, Autoformer [35] introduces autoregressive attention and deep decomposition structure to realize long-term prediction of time series. The deep learning decomposition structure improves the ability of the model to analyze trends and seasons. On this basis, FEDformer [49] and TDformer [45] introduce the deep decomposition framework and Fourier Attention to realize the long-term prediction of time series. These methods improve the ability to mine time series context relation by introducing trend and seasonal modeling. However, the decomposition method transform raw data into fixed forms based on expert experience, which limits the ability of the model to mine the data itself. At present, to strengthen the model's ability to mine global information from raw data, Patch TST [27] and Crossformer [46] adopt the idea of patch modeling. In addition, Patch TST and Crossformer respectively adopt channel independence and two-stage attention to realize multivariable modeling. Due to the mining of more data features, Patch TST and Crossformer can achieve better capabilities than the transformer variants mentioned above. However, Patch TST ignores the correlation between different variables, and Crossformer ignores the role of local information. In general, the existing models do not make full use of the key features of multivariate long sequence time series, which limits their performance.

3 METHODOLOGY

3.1 Preliminaries

In this section, we introduce the basic definition of multivariate time series and multivariate time series long-term prediction.

Multivariate time series. The multivariate time series is a data form composed of multiple sequences that change over time [48]. The representation of multivariate time series can usually be defined as a tensor $X \in R^{N*T}$ [2]. N is the number of variables. T is the length of the time step.

Multivariate time series long-term prediction. Given a historical sequence $X \in R^{N*H}$ from H time steps in history, the model can predict the value $Y \in R^{N*L}$ of the nearest L time steps in the future [28]. The main purpose of multivariate time series long-term prediction is to establish the mapping relationship between input $X \in R^{N*T}$ and label $Y \in R^{N*L}$ [36].

3.2 Overall framework of the proposed model

The overall framework of the DSformer is given in Figure 2. And it can be found that DSformer contains two important component: the double sampling block and the temporal variable attention block. And DSformer combines a double sampling block and three temporal variable attention blocks to mine three features and fully perform information integration. In this section, we intuitively discuss each block of DSformer and its parallel structure.

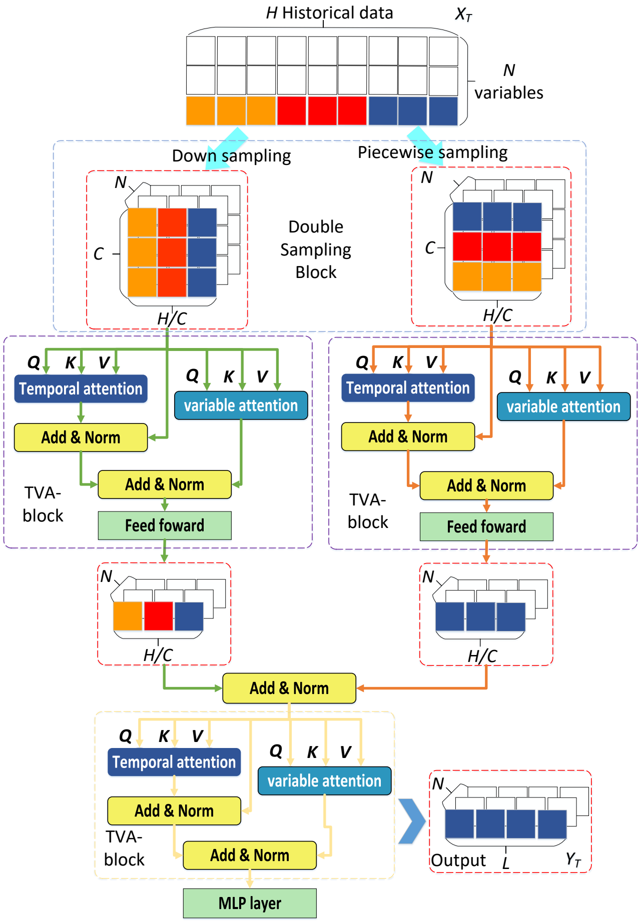


Figure 2: Overall framework of the proposed DSformer, the DS block and the TVA block.

Firstly, we discuss the DS block, which uses the downsampling and the piecewise sampling to process the original input features respectively. The downsampling converts the original sequence into multiple subsequences with simliar length by increasing the sampling interval. The global information of subsequences with larger time intervals is more significant than that of the original sequence [26]. The piecewise sampling can divide the long sequence into multiple contiguous fragments. Because the observation length of the time series is reduced, the model can mine the local information more intensively [44]. At the same time, to reduce the information loss caused by sampling, we connect the subsequences obtained from the sampling method and convert them into 3D tensors.

Second, the TVA block aims to mine the 3D tensors processed by the DS block from the temproal dimension and the variable dimension. Based on the parallel structure, the TVA block enable the temporal attention and the variable attention to mine input features respectively. Different from the traditional stacked multi-layer structure, the parallel structure enables the model to mine information more centrally [7]. Then the effective integration of temporal information and variable information is realized through addition and layer normalization.

Finally, the overall framework of DSformer also adopts parallel structure to realize feature mining and modeling. Specifically, the two different 3D tensors obtained by the DS block are mined by two TVA blocks. And then, a new TVA block is used to achieve the fusion of above two processed tensors. Therefore, DSformer can be used to mine global information, local information and variable correlation in parallel. Based on this structure, DSformer can strengthen the ability of mining features and achieving fusion.

3.3 Double sampling block

The double sampling block consists of two important steps: the down sampling and the piecewise sampling. Figure 3 presents a schematic of these two sampling methods. These two sampling methods transform the original 2D feature vectors $X \in R^{N*H}$ into two 3D features $X_{ds} \in R^{N*C*\frac{H}{C}}$ and $X_{ps} \in R^{N*C*\frac{H}{C}}$. The feature vector obtained by downsampling contain more global information. The feature vector obtained by piecewise sampling contains more local information. In the following, we briefly describe the proposed two sampling methods.

Down sampling. For a time series with length H, we obtain C subsequences of consistent length in the same way as shown in Figure 3 (a). The subsequence obtained by the downsampling method has a larger time interval. As a special form of sequence, observing time series data with larger time intervals can obtain more intuitive global information. In addition, to avoid the information loss caused by down-sampling, we put C subsequences together and obtain the feature vector $X_{ds} \in R^{N*C*\frac{H}{C}}$ for subsequent modeling. For the jth subsequence, its main constituent form is given as follows:

$$X^{j}{}_{ds} = [x_j, x_{j+\frac{H}{C}}, x_{j+2*\frac{H}{C}}, ..., x_{j+(C-1)*\frac{H}{C}}], \tag{1}$$

Piecewise sampling. For a time series with length H, we obtain C subsequences of consistent length in the same way as shown in Figure 3 (b). The piecewise sampling method can transform the original time series into continuous subsequence with the same length. Each subsequence contains local information over a historical period of time. Unlike down sampling, piecewise sampling allows the model to focus more attention on local information, which usually reflects the details of local changes over a cycle. In addition, to avoid the information loss caused by piecewise sampling, we put C subsequences together and obtain the feature vector $X_{ps} \in R^{N*C*\frac{H}{C}}$ for subsequent modeling. For the jth subsequence, its main constituent form is given as follows:

$$X^{j}{}_{ps} = [x_{1+(j-1)*C}, x_{2+(j-1)*C}, x_{3+(j-1)*C}, ..., x_{j*C}], \tag{2}$$

After obtaining two different feature vectors $X_{ds} \in R^{N*C*\frac{H}{C}}$ and $X_{ps} \in R^{N*C*\frac{H}{C}}$ by the DS block, we next introduce how to use TVA block to mine above feature vectors from temporal dimension and variable dimension.

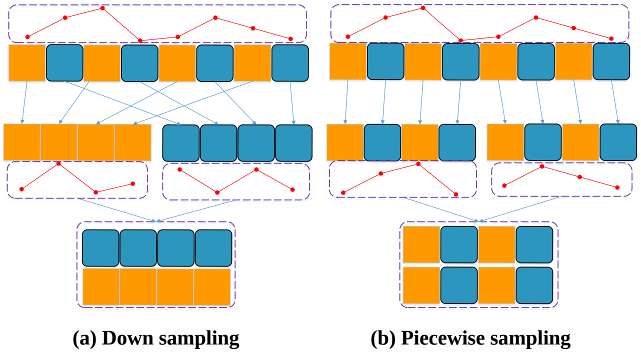

(a) Down sampling (b) Piecewise sampling

Figure 3: Schematic of the down sampling method and the piecewise sampling method.

3.4 TVA block

The proposed TVA block consists of two main components: temporal attention and variable attention. The main function of temporal attention is to mine the context information of data from the temporal dimension. The main function of variable attention is to mine the internal implicit relation between different variables. The information mined by these two components is integrated through a parallel structure. Figure 4 illustrates the detailed composition of TVA block, temporal attention and variable attention. Next, we present the modeling details of temporal attention, variable attention, and TVA blocks.

For the $X_{ds} \in R^{N*C*\frac{H}{C}}$ and $X_{ps} \in R^{N*C*\frac{H}{C}}$ obtained by the double sampling block, they are transmitted to the temporal attention and variable attention as input to the TVA block. Then, temporal attention and variable attention will process above two feature vectors in the following way:

Temporal attention. Temporal attention consists of three main components (multi-head attention, residual connection, and layer normalization).

Firstly, multi-head attention is used to mine the time dimension of the input feature vector $X_{ds} \in R^{N*C*\frac{H}{C}}$ and obtain the processed

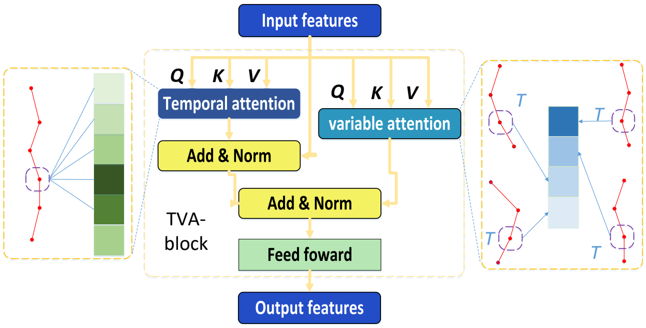


Figure 4: Schematic diagram of TVA block, temporal attention, and variable attention.

feature vector $X^{ta}{}_{ds} \in R^{N*C*\frac{H}{C}}$.

$$\begin{cases} Q = FC(X_{ds}) \\ K = FC(X_{ds}) \\ V = FC(X_{ds}), \end{cases} \tag{3}$$

$$X^{ta}{}_{ds} = softmax(Q * K^T)V, \tag{4}$$

where, $FC(.)$ stands for the fully connected layer. $softmax(.)$ stands for the normalized exponential function. K^T stands for K after converting the last two dimensions ($K^T \in R^{N*\frac{H}{C}*C}$).

Then, the output $X^{TA}{}_{ds} \in R^{N*C*\frac{H}{C}}$ of the temporal attention component is obtained by the residual connection and the layer normalization:

$$U = \frac{1}{L}\sum_{i=0}^{L}(X^{ta}{}_{dsi} + X_{dsi}), \tag{5}$$

$$\sigma = \sqrt{\frac{1}{L}\sum_{i=0}^{L}(X^{ta}{}_{dsi} + X_{dsi} - U)^2}, \tag{6}$$

$$X^{TA}{}_{ds} = \frac{g}{\sqrt{\sigma^2 + \epsilon}} \odot (X^{ta}{}_{ds} + X_{ds} - U) + b, \tag{7}$$

where, U and σ are represent the statistics of the feature vectors. g is the gain. b is the bias. σ is a small decimal number that prevents division by zero.

Variable attention. Different from temporal attention, variable attention mainly uses multi-head attention to mine data from the perspective of the number N of variables. Through the mining of variable attention, DSformer can effectively analyze the correlation between different variables and conduct information interaction. The formulas of the variable attention are given as follows:

$$\begin{cases} Q = FC(X_{ds}) \\ K = FC(X_{ds}) \\ V = FC(X_{ds}), \end{cases} \tag{8}$$

$$X^{VA}{}_{ds} = softmax(\frac{Q * K^T}{\sqrt{d_k}})V, \tag{9}$$

where, d_k can let the outcome of $Q * K^T$ satisfy the distribution with expectation 0 and variance 1. In particular, in the above formula, the corresponding matrix forms of Q and K^T are $\frac{H}{C} * C * N$ and $\frac{H}{C} * N * C$, respectively

Based on above temporal attention method and variable attention method, $X^{TA}{}_{ds} \in R^{N*C*\frac{H}{C}}$ and $X^{VA}{}_{ds} \in R^{N*C*\frac{H}{C}}$ are obtained. Then, $X^{TA}{}_{ds}$ and $X^{VA}{}_{ds}$ are integrated and the output $X'{}_{ds} \in R^{N*\frac{H}{C}}$ of TVA block is obtained by the following formula:

$$X'{}_{ds} = FC(F_{LN}(X^{TA}{}_{ds} + X^{VA}{}_{ds})), \tag{10}$$

where, $F_{LN}(.)$ stands for layer normalization. In addition, the main function of $FC(.)$ is to transform the feature vector dimension from $N * C * \frac{H}{C}$ to $N * \frac{H}{C}$.

Information integration based on TVA block. The feature vector $X_{ps} \in R^{N*C*\frac{H}{C}}$ is mined in the same way as above methods. And the output feature vector $X'{}_{ps} \in R^{N*\frac{H}{C}}$ is obtained. For $X'{}_{ps}$ and $X'{}_{ds}$, we first used the layer normalization to achieve preliminary information fusion.

$$X' = F_{LN}(X'{}_{ps} + X'{}_{ds}), \tag{11}$$

Then, the two-dimensional feature vector $X' \in R^{N*\frac{H}{C}}$ was used as the input to the TVA block and fully mined from the temporal dimension and variable dimension.

Different from the previous modeling form, the feature vectors $X'{}_{ps} \in R^{N*\frac{H}{C}}$ processed by the the information fusion method based on TVA block are 2D tensors. Therefore, the main matrix forms of the variables Q and K^T (temporal attention) modeled here are $N * \frac{H}{C}$ and $\frac{H}{C} * N$ respectively.

Finally, the feature vectors, which are further mined and integrated by TVA block, are passed to the multi-layer perceptron to effectively realize the output of the final prediction result $Y \in R^{N*L}$.

3.5 DSformer

DSformer is constructed by effectively combining the double sampling block and three TVA block. The double sampling block effectively obtains the feature vectors containing key information. TVA block mines different feature vectors and fully realizes information integration. The specific modeling steps of the proposed DSformer are given as follows:

Step I: For the original 2D input features $X \in R^{N*H}$, the data is transformed into two 3D features $X_{ds} \in R^{N*C*\frac{H}{C}}$ and $X_{ps} \in R^{N*C*\frac{H}{C}}$ by a double sampling block.

Step II: Two TVA blocks are used to model and analyze X_{ds} and X_{ps}, respectively. TVA block deeply mines feature vectors from both temporal dimension and variable dimension. In addition to this, the 3D features are transformed into 2D features $X'{}_{ds} \in R^{N*\frac{H}{C}}$ and $X'{}_{ps} \in R^{N*\frac{H}{C}}$ by the feedforward neural network.

Step III: Add $X'{}_{ds}$ and $X'{}_{ps}$. And then layer normalization is used to process the new feature vector $X' \in R^{N*\frac{H}{C}}$.

Step IV: The TVA block is used to further mine the feature vector X' from the temporal dimension and the variable dimension. At the same time, the mined feature vectors are passed to the MLP for long-term prediction.

Step V: Based on the MLP for decoding, the model finally obtains the prediction result $Y \in R^{N*L}$ with the prediction step of L. The decoding process is calculated using the following formula:

$$Y = FC(X''), \tag{12}$$

where, $X^{''}$ stands for the feature vector obtained after the TVA block processing in step IV.

Based on the above steps, DSformer can effectively analyze and mine the key features and obtain the multivariate time series long-term prediction results. In addition, to ensure the training effect of the model, we adopt the method of fusing L1 Loss and L2 Loss for error backpropagation. The formula is given as follows:

$$Loss = w_{L1} * \frac{1}{B*N*L}\sum_{k=0}^{B}\sum_{j=0}^{N}\sum_{i=0}^{L}|Y_{ijk} - {Y^{tru}}_{ijk}| + \\ (1-w_{L1}) * \frac{1}{B*N*L}\sum_{k=0}^{B}\sum_{j=0}^{N}\sum_{i=0}^{L}(Y_{ijk} - {Y^{tru}}_{ijk})^2, \tag{13}$$

where, Y represents the prediction result of the model. Y^{tru} stands for the true label. B represents the number of samples. N stands for the number of variables. L stands for prediction step size. w_{L1} represents the weight of Loss.

4 EXPERIMENT AND ANALYSIS

4.1 Experimental design

Dataset. In order to fully verify the effectiveness of the proposed DSformer in the field of multivariate time series long-term prediction, this paper selects nine classical data sets for comparative experiments. These datasets include ETT (ETTh1, ETTh2, ETTm1 and ETTm2), Exchange, ILI, Weather, Electricity and Traffic [49]. Table 1 presents the basic statistics of these datasets.

Table 1: The statistics of the nine datasets.

Datasets	Variates	Timesteps	Granularity
ETTh1	7	17420	1hour
ETTh2	7	17420	1hour
ETTm1	7	69680	15min
ETTm2	7	69680	15min
Exchange	8	7588	1day
ILI	7	966	1week
Weather	21	52696	10min
Electricity	321	26304	1hour
Traffic	862	17544	1hour

Baselines. To construct comparative experiments and prove the effectiveness of DSformer, we select eight SOTA models with excellent performance in time series long-term prediction as baselines. The main baselines include PatchTST [27], Crossformer [46], TimesNet [34], Dlinear [42], FEDformer [49], Pyraformer [21], Autoformer [35] and Informer [47].

Setting. The main hyperparameter values of the DSformer are shown in Table 2. To conduct fair comparison experiments, we designed the experiment from the following aspects: (1) These nine datasets are divided into training sets, validation sets, and test sets according to the ratio in the reference [18]. (2) These nine datasets were uniformly preprocessed by z-score normalization method. For each set of experiments, we set five different random seeds for repeated experiments. The final result of the model is obtained by averaging the repeated experiments. (3) For the ILI dataset, we set the historical looking back window H = 36 and the predicted future step size L = [24, 36, 48, 60]. For the other data sets, we set the history looking back window H = 96 and the prediction future step size L = [96, 192, 336, 720].

Table 2: Values of the corresponding hyperparameters for different prediction step sizes.

Config	Values (96,192,336,720)
optimizer	Adam [11]
learning rate	0.0001
number of multi-head attention	2/2/1/1
Dropout	0.15
sampling interval	2/2/3/3
weight of Loss	0.35/0.35/0.65/0.65
learning rate schedule	MultiStepLR
milestone	[25,50,75]
gamme	0.5
batch size	16
epoch	100

Evaluation index. The selection of appropriate evaluation indexes is the key to evaluate the prediction performance of different models. Considering the characteristics of multivariate long sequence time series prediction, we choose Mean Absolute Error (MAE) [20] and Mean Squared Error (MSE)[10] as the main evaluation indexes .

4.2 Main results

Table 3 shows the prediction results of the proposed DSformer and all baselines on nine datasets. The best results are highlighted in **bold** and the second best results are underlined. Based on Table 3, the following conclusions can be obtained: (1) Compared with other SOTA methods, Informer and Pyraformer have larger prediction errors. Although these two methods design advanced attention structures to improve the performance of the model, they fail to fully mine the core features of time series. (2) Autoformer and FEDformer improve their prediction performance by introducing trend-season decomposition. However, the decomposition method converts the original sequence into a fixed form based on expert experience, which limits the ability of the model to mine the original data. (3) Compared with transformer variants mentioned above, Patch TST, TimesNet and Crossformer focus on mining global information and variable correlation from original data, which enables them to achieve better prediction results. However, for multivariate long sequence time series, they do not make full use of the three features proposed in this paper, which makes their performance limited. (4) Compared with the existing SOTA models, the DSformer can achieve satisfactory prediction results. Firstly, two feature vectors focusing on global information and local information can be obtained through the DS block. Then, TVA block can mine and model these feature vectors from temporal dimension and variable dimension. Finally, DSformer uses a parallel structure to integrate the above feature information and realize long-term prediction. Therefore, the DSformer can achieve better performance than other SOTA methods.

Table 3: Multivariate time series prediction results on nine real-world datasets.

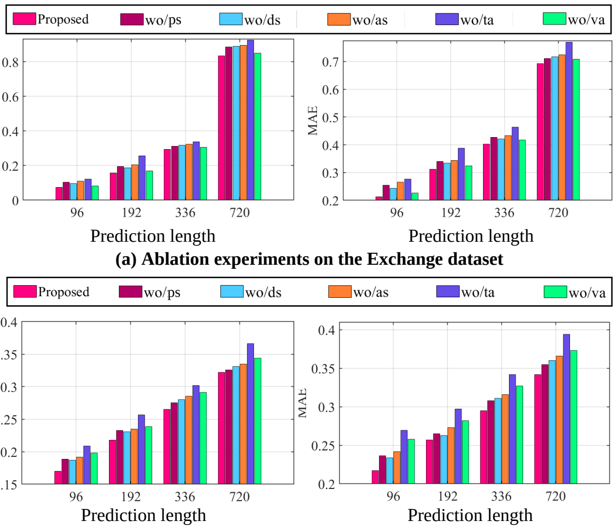
Data	L	DSformer		Patch TST *		Crossformer*		TimesNet		Dlinear		FEDformer		Autoformer		Informer		Pyraformer	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	96	**0.352**	**0.392**	0.393	0.408	0.396	0.412	0.384	0.402	0.386	0.400	0.376	0.419	0.449	0.459	0.865	0.713	0.664	0.612
	192	**0.408**	**0.425**	0.445	0.434	0.410	0.438	0.436	0.429	0.437	0.432	0.420	0.448	0.500	0.482	1.008	0.792	0.790	0.681
	336	0.448	**0.436**	0.484	0.451	**0.440**	0.461	0.491	0.469	0.481	0.459	0.459	0.465	0.521	0.496	1.107	0.809	0.891	0.738
	720	**0.469**	**0.454**	0.480	0.471	0.519	0.524	0.521	0.500	0.519	0.516	0.506	0.507	0.514	0.512	1.181	0.865	0.963	0.782
ETTh2	96	**0.268**	**0.304**	0.294	0.343	0.339	0.379	0.340	0.374	0.333	0.387	0.346	0.388	0.358	0.397	3.489	1.515	0.645	0.597
	192	**0.332**	**0.341**	0.377	0.393	0.415	0.425	0.402	0.414	0.477	0.476	0.429	0.439	0.456	0.452	3.755	1.525	0.788	0.683
	336	**0.349**	**0.387**	0.381	0.409	0.452	0.468	0.452	0.452	0.594	0.541	0.496	0.487	0.482	0.486	4.721	1.835	0.907	0.747
	720	**0.375**	**0.393**	0.412	0.433	0.455	0.471	0.462	0.468	0.831	0.657	0.463	0.474	0.515	0.511	3.647	1.625	0.963	0.783
ETTm1	96	**0.292**	0.368	0.321	**0.360**	0.320	0.373	0.338	0.375	0.345	0.372	0.379	0.419	0.505	0.475	0.672	0.571	0.543	0.510
	192	**0.351**	**0.379**	0.362	0.384	0.386	0.401	0.374	0.387	0.380	0.389	0.426	0.441	0.553	0.496	0.795	0.669	0.557	0.537
	336	**0.384**	0.408	0.392	**0.402**	0.404	0.427	0.410	0.411	0.413	0.413	0.445	0.459	0.621	0.537	1.212	0.871	0.754	0.655
	720	**0.442**	0.439	0.450	**0.435**	0.569	0.528	0.478	0.450	0.474	0.453	0.543	0.490	0.671	0.561	1.166	0.823	0.908	0.724
ETTm2	96	**0.130**	**0.231**	0.178	0.260	0.196	0.275	0.187	0.267	0.193	0.292	0.203	0.287	0.255	0.339	0.365	0.453	0.435	0.507
	192	**0.207**	**0.275**	0.249	0.307	0.248	0.317	0.249	0.309	0.284	0.362	0.269	0.328	0.281	0.340	0.533	0.563	0.730	0.673
	336	**0.262**	**0.318**	0.313	0.346	0.322	0.358	0.321	0.351	0.369	0.427	0.325	0.366	0.339	0.372	1.363	0.887	1.201	0.845
	720	**0.327**	**0.372**	0.400	0.398	0.402	0.406	0.408	0.403	0.554	0.522	0.421	0.415	0.433	0.432	3.379	1.338	3.625	1.451
Exchange	96	**0.075**	**0.213**	0.081	0.216	0.139	0.265	0.107	0.234	0.088	0.218	0.148	0.278	0.197	0.323	0.847	0.752	0.376	1.105
	192	**0.158**	**0.308**	0.169	0.317	0.241	0.375	0.226	0.344	0.176	0.315	0.271	0.380	0.300	0.369	1.204	0.895	1.748	1.151
	336	**0.294**	**0.402**	0.305	0.416	0.392	0.468	0.367	0.448	0.313	0.427	0.460	0.500	0.509	0.524	1.672	1.036	1.874	1.172
	720	**0.834**	**0.692**	0.853	0.702	1.112	0.802	0.964	0.746	0.839	0.695	1.195	0.841	1.447	0.941	2.478	1.310	1.943	1.206
ILI	24	**1.538**	0.815	1.610	**0.814**	3.041	1.186	2.317	0.934	2.398	1.040	3.228	1.260	3.483	1.287	5.764	1.677	7.042	2.012
	36	**1.546**	**0.829**	1.579	0.870	3.406	1.232	1.972	0.920	2.646	1.088	2.679	1.080	3.103	1.148	4.755	1.467	7.394	2.031
	48	**1.672**	**0.841**	1.673	0.854	3.459	1.221	2.238	0.940	2.614	1.086	2.622	1.078	2.669	1.085	4.763	1.469	7.551	2.057
	60	**1.548**	**0.803**	1.702	0.829	3.640	1.305	2.027	0.928	2.804	1.146	2.857	1.157	2.770	1.125	5.264	1.564	7.662	2.100
Weather	96	**0.170**	**0.217**	0.178	0.219	0.185	0.248	0.172	0.220	0.196	0.255	0.217	0.296	0.266	0.336	0.300	0.384	0.896	0.556
	192	**0.215**	**0.257**	0.224	0.259	0.229	0.305	0.219	0.261	0.237	0.296	0.276	0.336	0.307	0.367	0.598	0.544	0.622	0.624
	336	**0.265**	**0.295**	0.278	0.298	0.287	0.332	0.280	0.306	0.283	0.335	0.339	0.380	0.359	0.395	0.578	0.523	0.739	0.753
	720	**0.322**	**0.342**	0.350	0.346	0.356	0.398	0.365	0.359	0.345	0.381	0.403	0.428	0.419	0.428	1.059	0.741	1.004	0.934
Electricity	96	**0.163**	0.264	0.174	**0.259**	0.175	0.279	0.168	0.272	0.197	0.282	0.193	0.308	0.201	0.317	0.274	0.368	0.386	0.449
	192	**0.174**	0.272	0.178	**0.265**	0.192	0.302	0.184	0.289	0.196	0.285	0.201	0.315	0.222	0.334	0.296	0.386	0.378	0.443
	336	**0.187**	0.287	0.196	**0.282**	0.208	0.317	0.198	0.300	0.196	0.285	0.201	0.315	0.222	0.334	0.296	0.386	0.376	0.443
	720	**0.216**	**0.309**	0.237	0.316	0.225	0.337	0.220	0.320	0.245	0.333	0.246	0.355	0.254	0.361	0.373	0.439	0.376	0.445
Traffic	96	**0.458**	0.311	0.477	0.305	0.519	**0.295**	0.593	0.321	0.650	0.396	0.587	0.366	0.613	0.388	0.719	0.391	0.867	0.468
	192	**0.467**	0.323	0.471	**0.299**	0.526	0.307	0.617	0.336	0.598	0.370	0.604	0.373	0.616	0.379	0.696	0.382	0.869	0.467
	336	**0.479**	0.329	0.485	0.305	0.530	**0.300**	0.629	0.336	0.605	0.373	0.621	0.383	0.622	0.337	0.777	0.420	0.881	0.469
	720	**0.512**	0.342	0.518	0.325	0.573	**0.313**	0.640	0.350	0.645	0.394	0.626	0.382	0.660	0.408	0.864	0.472	0.896	0.473

* represents that we set uniform input and output sizes to ensure the fairness of the experiment. The results of other methods are from Timesnet [34]

4.3 Ablation experiments

The DSformer contains four key components: piecewise sampling, down sampling, temporal attention and variable attention. To prove that these components can help the DSformer to mine the key feature information, wu conducts ablation experiments from the following five perspectives: (1) wo/ ps: the piecewise sampling component is removed. (2) wo/ ds: the down-sampling component is removed. (3) wo/ as: In this part, we remove the double sampling block. (4) wo/ ta: we remove temporal attention and replace it with multi-layer perceptron. (5) wo/ va: variable attention is removed.

Figure 5 illustrates the results of the ablation experiments. Based on the experimental results, we can obtain the following conclusions: (1) When there is a correlation between different variables, deleting the variable attention will increase the prediction error of the model. According to the experimental results, the correlation between different time series in Weather data set is large, so the variable attention have a great influence on the prediction results. (2) After deleting the temporal attention, the prediction performance of DSformer decreases significantly. The main reason is that the most important step in time series modeling is mining the time series context relation. If temporal attention is lost, it is difficult for DSformer to effectively analysis the context relation of time series data. (3) When the prediction step size is long, removing down sampling significantly increases the error of the prediction. When the prediction step size is short, deleting piecewise sampling significantly increases the prediction error. The main reason is that the features obtained by down sampling contain more global information, and the features obtained by interval sampling contain more local information. Therefore, they will affect the modeling effect of different prediction steps, respectively. (4) After removing the down-sampling and piecewise sampling at the same time, the prediction error of the DSformer further increases. The main reason is that these two sampling methods can deepen the model's ability to focus on learning the global and the local respectively. When the sampling method is removed, the model may not be able to focus on the key information, which increases the prediction errors of the DSformer. (5) The results of ablation experiments can demonstrate the importance of the proposed four components, which can effectively mine the three main features of multivariate long sequence time series and reduce prediction error.



(a) Ablation experiments on the Exchange dataset

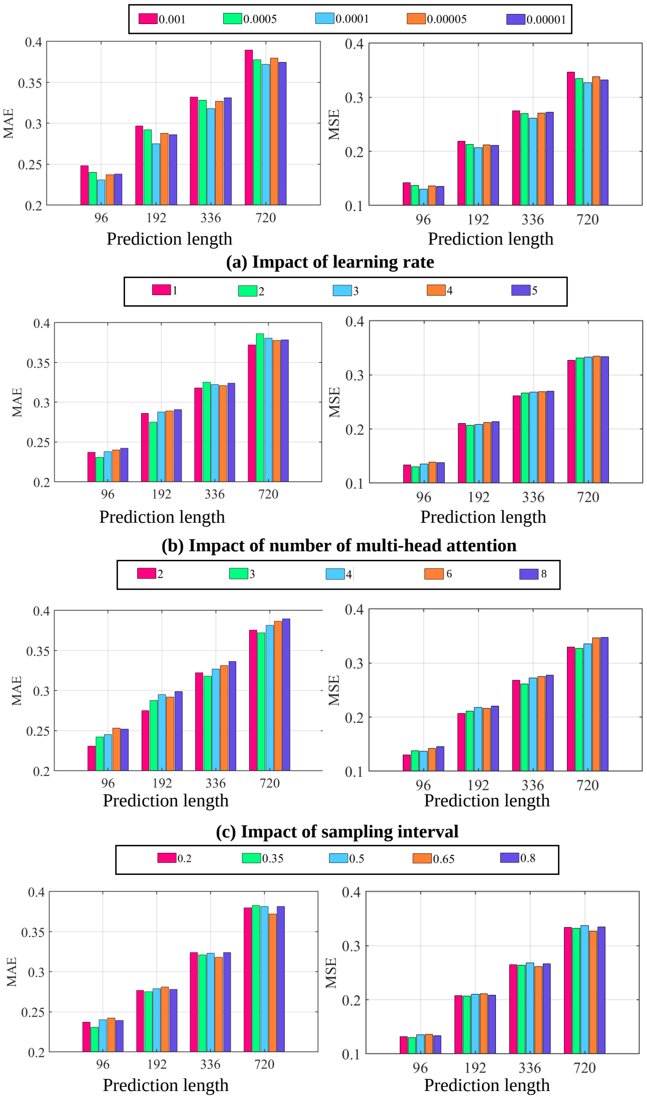
(b) Ablation experiments on the Weather dataset

Figure 5: Results of ablation experiments on Exchange and weather datasets.

4.4 Hyperparameter analysis experiments

The hyperparameters of the DSformer usually affect the final prediction results. In order to fully analyze the sensitivity of the DSformer and the role of some key hyperparameters, this section conducts an experimental analysis of the four main hyperparameters (learning rate, number of multi-head attention, sampling interval and weight of Loss) on ETTm2 dataset. Figure 6 illustrates the results of the hyperparameter analysis experiments.

Based on the experimental results, we can get the following conclusions: (1) The number of multi-head attention and the weight of Loss have relatively little impact on the prediction performance of the model. For multi-head attention, an appropriate number can prevent overfitting while ensuring modeling performance. For the weight of Loss, an appropriate value can ensure the training effect of the model and improve the overall performance. (2) The learning rate has a large impact on the model performance. The main reasons include the following two aspects: On the one hand, a large learning rate will produce phenomena such as overfitting. On the other hand, a smaller learning rate will affect the effect of training. Therefore, the setting of learning rate is very important to ensure the training effect of DSformer. (3) As one of the main hyperparameters, the sampling interval has a relatively large impact on the prediction results. When the sampling interval is small, DSformer can achieve better results with shorter prediction steps. When the sampling interval is relatively large, DSformer can achieve better results on longer prediction steps. However, when the sampling interval is too large, the prediction error of DSformer increases significantly. The main reason is that too large sampling interval makes the model lose more local information, which resulted in insufficient usage of information and reduced prediction accuracy. Therefore, setting the sampling interval appropriately can affect the effect of different prediction steps of the DSformer.(4) The sampling interval C of the double sampling block is an important parameter because it affects the input feature information of the DSformer. In the next section, we will future analyze the influence of sampling interval C and input history length H on the experimental results.



(a) Impact of learning rate

(b) Impact of number of multi-head attention

(c) Impact of sampling interval

(d) Impact of weight of Loss

Figure 6: Impact of four key hyperparameters on prediction results (ETTm2 dataset).

4.5 Effects of the sampling interval and the history length

The history length affects the information obtained by the model. And the sampling interval affects the model's utilization of information. Considering the effect of the history length and the sampling interval on the input information of DSformer, we further analyze the influence of these two hyperparameters on the prediction results in this section. To adequately evaluate these two key hyperparameters, we carried out the following comparison experiments:

(1) Based on the hyperparameter experiments, it can be found that too large sampling interval is not conducive to the experimental results. Therefore, we set the sampling interval of the double sampling block, that is, the number of subsequences $C = [2, 3, 4, 6]$. (2) Considering the characteristics of the prediction step size of the model, the history length of the model is set as $H = [96, 192, 336]$, respectively. (3) All above parameters were used to construct experiments on the ETTh2 dataset. And the future length of the DSformer is set to $L = [96, 192, 336, 720]$.

Table 4: Experimental results on ETTh2 dataset with different history length H and sampling interval C.

L	H	96		192		336	
	C	MSE	MAE	MSE	MAE	MSE	MAE
96	2	0.268	0.304	0.269	0.313	0.289	0.342
	3	0.273	0.311	0.278	0.328	0.274	0.325
	4	0.275	0.315	0.263	0.305	0.254	0.295
	6	0.278	0.327	0.265	0.307	0.262	0.302
192	2	0.332	0.341	0.340	0.351	0.356	0.364
	3	0.337	0.347	0.329	0.340	0.342	0.352
	4	0.343	0.349	0.328	0.335	0.318	0.329
	6	0.344	0.356	0.327	0.336	0.323	0.331
336	2	0.352	0.391	0.354	0.392	0.357	0.395
	3	0.349	0.387	0.352	0.393	0.353	0.391
	4	0.356	0.398	0.345	0.384	0.341	0.378
	6	0.362	0.401	0.347	0.386	0.337	0.376
720	2	0.381	0.398	0.379	0.397	0.377	0.395
	3	0.375	0.393	0.374	0.392	0.371	0.392
	4	0.389	0.403	0.367	0.385	0.366	0.389
	6	0.394	0.412	0.365	0.386	0.362	0.383

Table 4 shows the experimental results for different history lengths and sampling intervals. Based on the experimental results, the following conclusions can be drawn: (1) When the history length is short, the sampling interval C cannot be too large. If the sampling interval is large, the prediction performance of the model will degrade significantly. The main reason is that the increasing sampling interval limits the ability of DSformer to mine the local information of raw data. The loss of too much local information is not conducive to the short-term prediction effect of the DSformer. (2) When the history length is large, increasing the sampling interval C can improve the prediction performance of the model. On the one hand, increasing the sampling interval can make the feature vector obtained by down-sampling contain more global information. On the other hand, when the history length is large, the model contains more historical period information, and increasing the sampling interval can make the piecewise sampling obtain feature vectors that pay more attention to local information. (3) The history length H of DSformer can affect the amount of global and local information obtained by the model. The size of the sampling interval C can make the model focus on different informations of input features. Therefore, the appropriate balance between the history length and the sampling interval can make the model effectively use the overall information and the global information, which can improve the prediction accuracy of DSformer.

4.6 Efficiency

In this section, based on the Electricity data sets, we compare the efficiency of the DSformer and other baselines (Dlinear, Pyraformer, Crossformer, FEDformer and Autoformer). Besides, to make a fair comparison, we compare the mean training time of each epoch of several models. The experimental equipment is the Intel(R) Xeon(R) Gold 5217 CPU @ 3.00GHz, 128G RAM computing server with RTX 3090 graphics card. The batch size is set to 16.

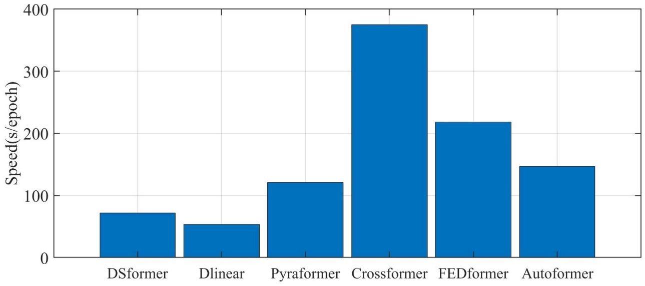


Figure 7: Training time for each epoch of different models.

Based on Figure 7, it can be found that although the computational complexity of DSformer is $O(N^2)$, the actual computational resource consumption of DSformer is not large. Specifically, most existing transformer variants use various theoretical methods to reduce computational complexity, but their actual computational resource consumption is not low due to the introduction of many tricks. Compared with above models, DSformer has two advantages: On the one hand, DSformer uses the DS block to reduce the length of the sequence that needs to be modeled. On the other hand, DSformer does not use some tricks that significantly increase computational resource consumption, such as embeding. Therefore, the results of efficiency comparison further prove the practical application value of DSformer.

5 CONCLUSION AND FUTURE WORK

In this paper, we propose DSformer, an efficient multivariate time series long-term prediction model, which contains two finely designed blocks, including the DS block and TVA block. The DS block simply and efficiently mines the global information and the local information of time series, which are significant features for long-term prediction. And the TVA block can effectively integrate the above information and variable correlation to significantly improve time series prediction accuracy. The experiments on nine real-world datasets show that DSformer achieves state-of-the-art performance for MTS long-term prediction. In the future, we will try to design a module to adaptive balance sampling interval and history length, further improving the information mining ability and long term prediction effect of the model.

ACKNOWLEDGMENTS

This work is supported by the Youth Innovation Promotion Association of the Chinese Academy of Sciences (No.2023112).

REFERENCES

[1] MA Castán-Lascorz, P Jiménez-Herrera, A Troncoso, and Gualberto Asencio-Cortés. 2022. A new hybrid method for predicting univariate and multivariate time series based on pattern forecasting. *Information Sciences* 586 (2022), 611–627.

[2] Jeongwhan Choi, Hwangyong Choi, Jeehyun Hwang, and Noseong Park. 2022. Graph neural controlled differential equations for traffic forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 6367–6374.

[3] Taesung Choi, Dongkun Lee, Yuchae Jung, and Ho-Jin Choi. 2022. Multivariate time-series anomaly detection using SeqVAE-CNN hybrid model. In *2022 International Conference on Information Networking (ICOIN)*. IEEE, 250–253.

[4] Razvan-Gabriel Cirstea, Chenjuan Guo, Bin Yang, Tung Kieu, Xuanyi Dong, and Shirui Pan. 2022. Triformer: Triangular, Variable-Specific Attentions for Long Sequence Multivariate Time Series Forecasting–Full Version. *arXiv preprint arXiv:2204.13767* (2022).

[5] Shuqin Dong, Chengqin Yu, Guangxi Yan, Jintian Zhu, and Hui Hu. 2021. A Novel ensemble reinforcement learning gated recursive network for traffic speed forecasting. In *2021 Workshop on Algorithm and Big Data*. 55–60.

[6] Yuwei Fu, Di Wu, and Benoit Boulet. 2022. Reinforcement learning based dynamic model combination for time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 6639–6647.

[7] Yuwei Fu, Di Wu, and Benoit Boulet. 2022. Reinforcement learning based dynamic model combination for time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 6639–6647.

[8] Wei Guo, Chang Meng, Enming Yuan, Zhicheng He, Huifeng Guo, Yingxue Zhang, Bo Chen, Yaochen Hu, Ruiming Tang, Xiu Li, and Rui Zhang. 2023. Compressed Interaction Graph Based Framework for Multi-Behavior Recommendation. In *Proceedings of the ACM Web Conference 2023*. Association for Computing Machinery, 960–970.

[9] Min Han, Shoubo Feng, CL Philip Chen, Meiling Xu, and Tie Qiu. 2018. Structured manifold broad learning system: A manifold perspective for large-scale chaotic time series analysis and prediction. *IEEE Transactions on Knowledge and Data Engineering* 31, 9 (2018), 1809–1821.

[10] Weiqiu Jin, Shuqing Dong, Chengqing Yu, and Qingquan Luo. 2022. A data-driven hybrid ensemble AI model for COVID-19 infection forecast using multiple neural networks and reinforced learning. *Computers in Biology and Medicine* 146 (2022), 105560.

[11] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[12] Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. 2019. Reformer: The Efficient Transformer. In *International Conference on Learning Representations*.

[13] Pu-Yun Kow, Li-Chiu Chang, Chuan-Yao Lin, Charles C-K Chou, and Fi-John Chang. 2022. Deep neural networks for spatiotemporal PM2. 5 forecasts based on atmospheric chemical transport model output and monitoring data. *Environmental Pollution* 306 (2022), 119348.

[14] Guokun Lai, Wei-Cheng Chang, Yiming Yang, and Hanxiao Liu. 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.

[15] Shiyang Li, Xiaoyong Jin, Yao Xuan, Xiyou Zhou, Wenhu Chen, Yu-Xiang Wang, and Xifeng Yan. 2019. Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting. *Advances in neural information processing systems* 32 (2019).

[16] Ke Liang, Yue Liu, Sihang Zhou, Wenxuan Tu, Yi Wen, Xihong Yang, Xiangjun Dong, and Xinwang Liu. 2023. Knowledge Graph Contrastive Learning Based on Relation-Symmetrical Structure. *IEEE Transactions on Knowledge and Data Engineering* (2023).

[17] Ke Liang, Lingyuan Meng, Sihang Zhou, Siwei Wang, Wenxuan Tu, Yue Liu, Meng Liu, and Xinwang Liu. 2023. Message Intercommunication for Inductive Relation Reasoning. *arXiv preprint arXiv:2305.14074* (2023).

[18] Yubo Liang, Zezhi Shao, Fei Wang, Zhao Zhang, Tao Sun, and Yongjun Xu. 2023. BasicTS: An Open Source Fair Multivariate Time Series Prediction Benchmark. In *Benchmarking, Measuring, and Optimizing: 14th BenchCouncil International Symposium, Bench 2022, Virtual Event, November 7-9, 2022, Revised Selected Papers*. Springer, 87–101.

[19] Hui Liu, Chengqing Yu, Haiping Wu, Zhu Duan, and Guangxi Yan. 2020. A new hybrid ensemble deep reinforcement learning model for wind speed short term forecasting. *Energy* 202 (2020), 117794.

[20] Hui Liu, Chengqing Yu, and Chengming Yu. 2021. A new hybrid model based on secondary decomposition, reinforcement learning and SRU network for wind turbine gearbox oil temperature forecasting. *Measurement* 178 (2021), 109347.

[21] Shizhan Liu, Hang Yu, Cong Liao, Jianguo Li, Weiyao Lin, Alex X Liu, and Schahram Dustdar. 2022. Pyraformer: Low-complexity pyramidal attention for long-range time series modeling and forecasting. In *International Conference on Learning Representations*.

[22] Xinwei Liu, Muchuan Qin, Yue He, Xiwei Mi, and Chengqing Yu. 2021. A new multi-data-driven spatiotemporal PM2.5 forecasting model based on an ensemble graph reinforcement learning convolutional network. *Atmospheric Pollution Research* 12, 10 (2021), 101197.

[23] Yijing Liu, Qinxian Liu, Jian-Wei Zhang, Haozhe Feng, Zhongwei Wang, Zihan Zhou, and Wei Chen. 2022. Multivariate Time-Series Forecasting with Temporal Polynomial Graph Neural Networks. In *Advances in Neural Information Processing Systems*.

[24] Chang Meng, Hengyu Zhang, Wei Guo, Huifeng Guo, Haotian Liu, Yingxue Zhang, Hongkun Zheng, Ruiming Tang, Xiu Li, and Rui Zhang. 2023. Hierarchical Projection Enhanced Multi-Behavior Recommendation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery, 4649–4660.

[25] Xiwei Mi, Chengqing Yu, Xinwei Liu, Guangxi Yan, Fuhao Yu, and Pan Shang. 2022. A dynamic ensemble deep deterministic policy gradient recursive network for spatiotemporal traffic speed forecasting in an urban road network. *Digital Signal Processing* 129 (2022), 103643.

[26] Jalal Mostafa, Sara Wehbi, Suren Chilingaryan, and Andreas Kopmann. 2022. SciTS: A Benchmark for Time-Series Databases in Scientific Experiments and Industrial Internet of Things. In *Proceedings of the 34th International Conference on Scientific and Statistical Database Management*. 1–11.

[27] Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. 2023. A Time Series is Worth 64 Words: Long-term Forecasting with Transformers. In *The Eleventh International Conference on Learning Representations*.

[28] Zezhi Shao, Fei Wang, Zhao Zhang, Yuchen Fang, Guangyin Jin, and Yongjun Xu. 2023. HUTFormer: Hierarchical U-Net Transformer for Long-Term Traffic Forecasting. *arXiv preprint arXiv:2307.14596* (2023).

[29] Zezhi Shao, Zhao Zhang, Fei Wang, Wei Wei, and Yongjun Xu. 2022. Spatial-Temporal Identity A Simple yet Effective Baseline for Multivariate Time Series Forecasting. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 4454–4458.

[30] Zezhi Shao, Zhao Zhang, Fei Wang, and Yongjun Xu. 2022. Pretraining Enhanced Spatial temporal Graph Neural Network for Multivariate Time Series Forecasting. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 1567–1577.

[31] Zezhi Shao, Zhao Zhang, Wei Wei, Fei Wang, Yongjun Xu, Xin Cao, and Christian S Jensen. 2022. Decoupled dynamic spatial-temporal graph neural network for traffic forecasting. In *Proceedings of the VLDB Endowment*, Vol. 15. 2733–2746.

[32] Jing Tan, Hui Liu, Yanfei Li, Shi Yin, and Chengqing Yu. 2022. A new ensemble spatio-temporal PM2.5 prediction method based on graph attention recursive networks and reinforcement learning. *Chaos, Solitons & Fractals* 162 (2022), 112405.

[33] Fei Wang, Di Yao, Yong Li, Tao Sun, and Zhao Zhang. 2023. AI-enhanced spatial-temporal data-mining technology: New chance for next-generation urban computing. *The Innovation* 4, 2 (2023).

[34] Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. 2023. TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis. In *The Eleventh International Conference on Learning Representations*.

[35] Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. 2021. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in Neural Information Processing Systems* 34 (2021), 22419–22430.

[36] Xinle Wu, Dalin Zhang, Chenjuan Guo, Chaoyang He, Bin Yang, and Christian S Jensen. 2021. AutoCTS: Automated correlated time series forecasting. *Proceedings of the VLDB Endowment* 15, 4 (2021), 971–983.

[37] Yongjun Xu, Xin Liu, Xin Cao, Changping Huang, Enke Liu, Sen Qian, Xingchen Liu, Yanjun Wu, Fengliang Dong, and Cheng-Wei Qiu. 2021. Artificial intelligence: A powerful paradigm for scientific research. *The Innovation* 2, 4 (2021), 100179.

[38] Guangxi Yan, Jiang Chen, Yu Bai, Chengqing Yu, and Chengming Yu. 2022. A Survey on Fault Diagnosis Approaches for Rolling Bearings of Railway Vehicles. *Processes* 10, 4 (2022), 724.

[39] Jaemin Yoo and U Kang. 2021. Attention-Based Autoregression for Accurate and Efficient Multivariate Time Series Forecasting. In *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*. SIAM, 531–539.

[40] Chengqing Yu, Guangxi Yan, Chengming Yu, Yu Zhang, and Xiwei Mi. 2023. A multi-factor driven spatiotemporal wind power prediction model based on ensemble deep graph attention reinforcement learning networks. *Energy* 263 (2023), 126034.

[41] Zhihan Yue, Yujing Wang, Juanyong Duan, Tianmeng Yang, Congrui Huang, Yunhai Tong, and Bixiong Xu. 2022. Ts2vec: Towards universal representation of time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 8980–8987.

[42] Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. 2023. Are transformers effective for time series forecasting?. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 37. 11121–11128.

[43] George Zerveas, Srideepika Jayaraman, Dhaval Patel, Anuradha Bhamidipaty, and Carsten Eickhoff. 2021. A transformer-based framework for multivariate time series representation learning. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*. 2114–2124.

[44] Tianping Zhang, Yizhuo Zhang, Wei Cao, Jiang Bian, Xiaohan Yi, Shun Zheng, and Jian Li. 2022. Less is more: Fast multivariate time series forecasting with light sampling-oriented mlp structures. *arXiv preprint arXiv:2207.01186* (2022).

[45] Xiyuan Zhang, Xiaoyong Jin, Karthick Gopalswamy, Gaurav Gupta, Youngsuk Park, Xingjian Shi, Hao Wang, Danielle C Maddix, and Yuyang Wang. 2022. First De-Trend then Attend: Rethinking Attention for Time-Series Forecasting. *arXiv preprint arXiv:2212.08151* (2022).

[46] Yunhao Zhang and Junchi Yan. 2023. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *The Eleventh International Conference on Learning Representations*.

[47] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. 2021. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 11106–11115.

[48] Tian Zhou, Ziqing Ma, Qingsong Wen, Liang Sun, Tao Yao, Wotao Yin, and Rong Jin. 2022. Film: Frequency improved legendre memory model for long-term time series forecasting. *Advances in Neural Information Processing Systems* 35 (2022), 12677–12690.

[49] Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. 2022. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *International Conference on Machine Learning*. PMLR, 27268–27286.

[50] Yang Zhou, Zhuojia Yang, Qiang Sun, Chengqing Yu, and Chengming Yu. 2023. An artificial intelligence model based on multi-step feature engineering and deep attention network for optical network performance monitoring. *Optik* 273 (2023), 170443.