Applied Intelligence (2023) 53:5060–5071
https://doi.org/10.1007/s10489-022-03825-5

Hierarchical attention network for multivariate time series long-term forecasting

Hongjing Bi[1] · Lilei Lu[1] · Yizhen Meng[1]

Accepted: 27 May 2022 / Published online: 17 June 2022

Abstract

Multivariate time series long-term forecasting has always been the subject of research in various fields such as economics, finance, and traffic. In recent years, attention-based recurrent neural networks (RNNs) have received attention due to their ability of reducing error accumulation. However, the existing attention-based RNNs fail to eliminate the negative influence of irrelevant factors on prediction, and ignore the conflict between exogenous factors and target factor. To tackle these problems, we propose a novel Hierarchical Attention Network (HANet) for multivariate time series long-term forecasting. At first, HANet designs a factor-aware attention network (FAN) and uses it as the first component of the encoder. FAN weakens the negative impact of irrelevant exogenous factors on predictions by assigning small weights to them. Then HANet proposes a multi-modal fusion network (MFN) as the second component of the encoder. MFN employs a specially designed multi-modal fusion gate to adaptively select how much information about the expression of current time come from target and exogenous factors. Experiments on two real-world datasets reveal that HANet not only outperforms state-of-the-art methods, but also provides interpretability for prediction.

Keywords Multivariate time series · Hierarchical attention · Deep neural network · Long-term forecasting · Multi-modal fusion

1 Introduction

Multivariate time series (MTS) forecasting aims to study how to predict the future target based on historical target and exogenous series [1]. In recent years, MTS forecasting has been widely used in many fields, such as traffic flow forecasting [2], pedestrian behavior prediction [3], time series anomaly detection [4], and natural disaster forecasting [5]. Undoubtedly, accurate prediction can help society to operate effectively in many aspects. Many scholars devoted to developing time series prediction methods, especially based on classical statistical methods such as autoregressive comprehensive moving average (ARIMA) [6] and Holt-winter [7], as well as typical machine learning models like support vector regression (SVR) [8], deep belief network (DBN) [9], and random forests(RF) [10]. However, the above methods lack effective processing of the temporal dependence among the input series, and are difficult to cope with the evolution of long-horizon time series [11]. In addition, most of these methods solve the one-step ahead forecasting problem, which has limitations in practical applications. On the contrary, the long-term forecasting is more meaningful in actual applications, because one-step forecasting is difficult to provide a decision basis for the situation after multiple steps [12]. Recently, recurrent neural networks (RNNs) and its variants, including long short-term memory network (LSTM) [13] and gated recurrent unit (GRU) [14], have been widely used for modeling complex time series data, such as stock price prediction [15] and cyber-physical systems [16]. Among these applications, attention-based RNNs are particularly attractive to time series forecasting [17]. The attention-based RNNs utilize two independent RNNs to encode sequential inputs into context vectors and to decode these contexts into desired interpretations [18]. Attention is the information selection strategy proposed by Bahdanau et al. [19], which allows items from encoder hidden states can be selected by their importance for decoder. Qin et al. [20] added an input attention to the encoder for capturing variables interactions. Liu et al. [21] developed a dual-stage two-phase attention-based RNN model for multivariate time series long-term forecasting.

Although the attention-based RNNs have achieved encouraging performance in MTS long-term prediction, they have

✉ Hongjing Bi
oec2007@hotmail.com

[1] Department of Computer Science, Tangshan Normal University, Tangshan, Hebei 063000, People's Republic of China

several general drawbacks. **Firstly**, the multivariate time series data provide target factor and a variety of exogenous factors, but some exogenous factors present weak correlation with prediction task. For instance, crude oil prices have strong correlation with gasoline prices, but have a weak contribution to wood prices [22]. However, the RNN-based encoder blends the information of all factors into a hidden state for prediction. **Secondly**, these methods ignore the information conflict between target factor and exogenous factors, which will damage the accuracy of the model. Li et al. [15] found that linking some of the factors together did not help the stock price prediction, and even reduced the model accuracy.

To tackle these challenges, we propose a novel Hierarchical Attention Network (HANet) for multivariate time series long-term forecasting. In particular, HANet is a hierarchically structured encoder-decoder architecture, which learns both the importance factors and the long distance temporal dependence. Meanwhile, HANet employs a specially designed multi-modal fusion network (MFN) to eliminate the information conflict between target factor and exogenous factors. Specifically, we design a factor-aware attention network (FAN) as the first component of the encoder to eliminate the negative influence of irrelevant exogenous factors. FAN assigns appropriate weights to each factor in the exogenous series and converts the sequence into high-level semantics. Thus, FAN limits the contribution of irrelevant factors to high-level semantics by assigning small weights to them. To address the second challenge, we carefully designed a multi-modal fusion network (MFN) as the second component of the encoder. In the encoding stage, MFN trades off how much new information the network is considered from target and exogenous factors through specially designed multi-modal fusion gate. In addition, we introduce a temporal attention between the encoder and decoder network, which can adaptively select relevant encoder input items across time steps for improving forecasting accuracy. The main contributions of this study are as follows:

1. We propose a novel Hierarchical Attention Network for multivariate time series long-term forecasting. As a hierarchically structured neural network, HANet learns both the importance factors and the long distance temporal dependence. Meanwhile, HANet can alleviate the information conflict between target and exogenous factors.
2. We design a factor-aware attention network (FAN) to eliminate the negative influence of irrelevant exogenous factors. FAN limits the contribution of irrelevant factors to prediction by assigning small weights to them.
3. We introduce a multi-modal fusion network (MFN). MFN can alleviate the information conflict between target and exogenous factors through specially designed multi-modal fusion gate.
4. Experiments on two fields of air quality and ecological datasets show that HANet is very effective in time series long-term prediction, and outperforms the state-of-the-art methods.

2 Related work

Time series forecasting is an important field of academic research and forms part of applications such as natural disaster forecasting [5], medical diagnosis [23], traffic flow prediction [24], and stock market analysis [25]. The statistical methods, such as autoregressive integrated moving average ARIMA [6] and Holt-Winters [7] are two widely used models for time series forecasting. However, these methods only focus on univariate time series and assume the series is stationary, while practical data normally do not meet this constraint [26]. The machine learning methods, such as support vector regression (SVR) [8] and random forest [10], are also important components of time series prediction models. Although these methods can capture the interaction among features better, they are difficult to cope with the evolution of long-horizon time series due to ignore time dependence [11]. In recent years, deep neural networks have been successfully applied to time series forecasting. For example, Qin et al. [27] combined ARIMA and deep belief network (DBN) to predict the occurrence of red tides. Shin et al. [28] proved that deep neural networks had better generalization ability and higher prediction accuracy than traditional shallow neural networks. Moreover, recurrent neural networks (RNNs), especially its variants LSTM [29] and gated recurrent unit (GRU) [30], are widely used in various time series forecasting tasks, such as traffic flow forecasting [2], stock price forecasting [15], and COVID-19 prediction [23].

Taieb et al. [31] reviewed the existing long-term forecasting strategies, namely recursive, direct, and multilevel strategies. Encoder-decoder is a seq2seq model proposed by Sutskever et al. [32], which utilizes two independent RNNs to encode sequential inputs into context vectors and decode these contexts into desired sequence. Many scholars devoted to developing multivariate time series long-term forecasting methods with various encoder-decoder architectures [33]. For instance, Kao et al. [5] used the encoder-decoder model for multi-step flood forecasting and achieved good results. However, the performance of the encoder-decoder decreases with the increasing of the input sequence because the encoder compresses the input into a fixed vector. Fortunately, the attention mechanism can solve this challenge. Attention is a soft selection strategy of information, which allows items from encoder hidden states can be selected by their importance to decoder [34]. Therefore, attention-based RNNs further stimulated the related research on time series long-term prediction. Liu et al. [21] developed a dual-stage two-phase attention-based RNN (DSTP-RNN) for multivariate time series long-term forecasting. The model not only took into account the factors' spatial correlation, but also the time dependencies among the series.

3 Problem definition and notations

Multivariate time series (MTS) long-term forecasting aims to study how to predict the future multiple time steps' target series based on historical target and exogenous series. Given n (n ≥ 1) exogenous series and one target series, we utilize symbol $\boldsymbol{x}^k = \left(x_1^k, x_2^k, \ldots, x_T^k\right) \in \mathbb{R}^T$ to represent k-th exogenous series within the length of window size T, and we use symbols $\{\boldsymbol{x}_t\}_{t=1}^{T} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T\}$ and $\{y_t\}_{t=1}^{T} = \{y_1, y_2, \ldots, y_T\}$ to denote historical all exogenous and target series in T time slice. The symbol $\boldsymbol{x}_t$ $(1 \le t \le T)$ is a vector, where$\boldsymbol{x}_t = \{x_t^1, x_t^2, \ldots, x_t^n\} \in \mathbb{R}^n$ and n is the number of exogenous factors at time t. The symbol $y_t \in \mathbb{R}$ is the target factor at time t. Obviously, the output is an estimate of the target factor for Δ time steps after T, denotes as $\{\widehat{y}_t\}_{t=T+1}^{T+\Delta} = \{\widehat{y}_{T+1}, \widehat{y}_{T+2}, \ldots, \widehat{y}_{T+\Delta}\}$, where Δ $(\Delta \ge 1)$ is a variable according to the demand of the task. To sum up, HANet is to learn a nonlinear mapping from exogenous series $\{\boldsymbol{x}_t\}_{t=1}^{T}$ and target series $\{y_t\}_{t=1}^{T}$ in the history to the estimation of the future value $\{\widehat{y}_t\}_{t=T+1}^{T+\Delta}$:

$$\left[\{\boldsymbol{x}_t\}_{t=1}^{T}, \{y_t\}_{t=1}^{T}\right] \xrightarrow{f(\cdot)} \left\{\widehat{y}_t\right\}_{t=T+1}^{T+\Delta} \tag{1}$$

Where $f(\cdot)$ is the non-linear mapping function.

4 Proposed HANet model

In this section, we introduce the proposed HANet model for multivariate time series long-term forecasting problem. We first present an overview of the model. Subsequently, we detail the model with all components.

4.1 An overview of HANet

As mentioned in section 1, there are some problems to be solved in the long-term forecasting of multivariate time series: 1) eliminate the negative influence of irrelevant exogenous factors and 2) balance the information conflict between target and exogenous factors. Moreover, a successful time series long-term forecasting method should be able to capture the long-term dependence between sequences. Therefore, we propose a Hierarchical Attention Network (HANet) for multivariate time series long-term forecasting. The architecture of HANet is shown in Fig. 1.

As shown in Fig. 1, HANet is a hierarchically structured encoder-decoder model. Similarly, the encoder is also a hierarchical structure that consists of factor-aware attention network (FAN), multi-modal fusion network (MFN), and LSTM. FAN is executed on the exogenous series with the purpose of eliminating the negative effects of irrelevant factors. Specifically, FAN assigns different weights to $x_t^k \in \boldsymbol{x}^k$ $(1 \le t \le T)$ according to its importance to the prediction, and summarizes exogenous series $\boldsymbol{x}_t = \{x_t^1, x_t^2, \ldots, x_t^n\}$ into a high-level semantic $\widetilde{\boldsymbol{x}}_t$. Thus, FAN limits the contribution of irrelevant factors to high-level semantics by assigning small weights to them. Subsequently, the high-level semantics $\widetilde{\boldsymbol{x}}_t$ are fed into LSTM, and generate a hidden state $\mathbf{h}_t^x$. The hidden state $\mathbf{h}_t^x$ is entered into the MFN along with the target factory_t. MFN trades off how much new information the network is considered from target factor y_t and hidden state $\mathbf{h}_t^x$ through specially designed multi-modal fusion gate. The decoder is composed of LSTM and a single layer multilayer perceptron, which is located on the top of the encoder. Furthermore, we design a temporal attention (TA) and place it between the encoder and the decoder. TA acts as a bridge between the decoder unit i and the encoder. The function of this bridge is to select the most relevant information in the encoder for prediction. Finally, HANet leverages a single-layer multilayer to convert the state $\mathbf{d}_i$ into the predicted value $\widehat{y}_i$ $(1 \le i \le \Delta)$.

4.2 Encoding process

As shown in Fig. 1, the encoder is a hierarchically structured network which consists of a factor-aware attention network (FAN), a multi-modal fusion network (MFN), and a long short-term memory network (LSTM). In the coding phase, FAN and MFN play different roles. The former eliminates the negative effects of irrelevant factors, while the latter balances the information conflict between the target and exogenous factors. Next, we describe the encoding process in detail.

Standard LSTM LSTM has already become an extensible and useful model to address the problem of learning sequential data [13]. LSTM contains a memory cell $\mathbf{c}_t$ and three gates, i.e., forget gate $\mathbf{f}_t$, input gate $\mathbf{i}_t$, and output gate $\mathbf{o}_t$. The three gates adjust the information flow into and out of the cell. In this work, we use LSTM as the first layer of FAN. Given the input series $\{\boldsymbol{x}_t\}_{t=1}^{T} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T\}$, where $\boldsymbol{x}_t = \{x_t^1, x_t^2, \ldots, x_t^n\} \in \mathbb{R}^n$ is a vector with n exogenous factors at time t. LSTM is applied to learn a mapping from $\boldsymbol{x}_t$ to hidden state $\mathbf{h}_t$. The calculation is defined in eq. (2).

$$\begin{gathered}
\mathbf{i}_t = \sigma\left(\mathbf{W}_i \boldsymbol{x}_t + \mathbf{U}_i \mathbf{h}_{t-1} + \mathbf{b}_i\right) \\
\mathbf{f}_t = \sigma\left(\mathbf{W}_f \boldsymbol{x}_t + \mathbf{U}_f \mathbf{h}_{t-1} + \mathbf{b}_f\right) \\
\mathbf{o}_t = \sigma\left(\mathbf{W}_o \boldsymbol{x}_t + \mathbf{U}_o \mathbf{h}_{t-1} + \mathbf{b}_o\right) \\
\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot tanh\left(\mathbf{W}_c \boldsymbol{x}_t + \mathbf{U}_c \mathbf{h}_{t-1} + \mathbf{b}_c\right) \\
\mathbf{h}_t = \mathbf{o}_t \odot tanh\left(\mathbf{c}_t\right)
\end{gathered} \tag{2}$$

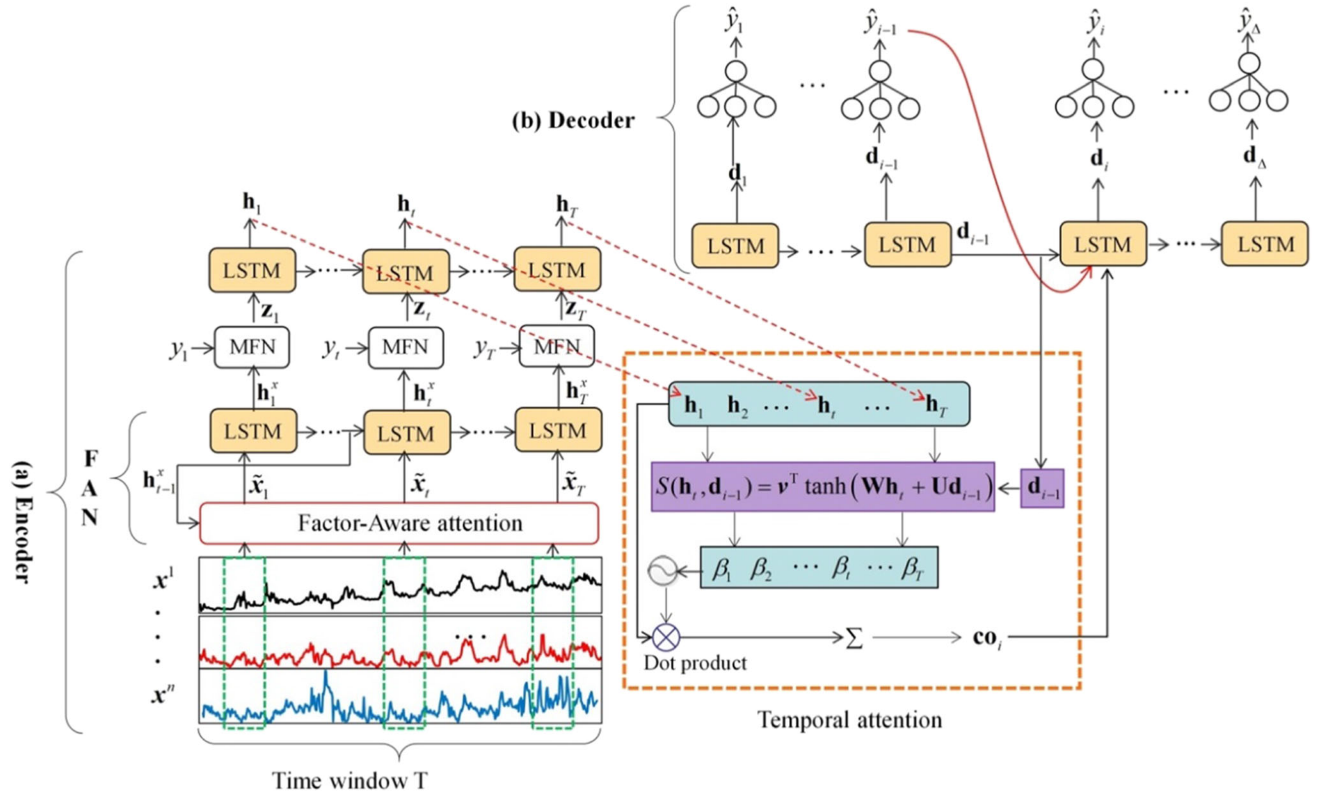


Fig. 1 A graphical illustration of Hierarchical Attention Network (HANet). The encoder of HANet consists of three components, i.e., factor-aware attention network (FAN), multi-modal fusion network (MFN), and LSTM. Here, $\boldsymbol{x}^k = \left(x_1^k, x_2^k, \ldots, x_T^k\right) \in \mathbb{R}^T$ is k-th exogenous series within the length of window size T. y_t is the target factor at time step t. $\widetilde{\boldsymbol{x}}_t$ is the high-level semantic representation of $\boldsymbol{x}_t$, where $\boldsymbol{x}_t$= $\{x_t^1, x_t^2, \ldots, x_t^n\}\in\mathbb{R}^n$ is a vector with n exogenous factors at time t. MFN is the multi-modal fusion network, which generates a hidden representation $\mathbf{z}_t$ by fusing target factor y_t and hidden state $\mathbf{h}_t^x$. $\mathbf{d}_i$ is the hidden state of decoder unit i, $\mathbf{co}_i$ is the context vector generated by temporal attention. $\widehat{y}_i$ is the predicted value

Where the symbol ⊙ is the element-wise product, and $\sigma = 1/(1 + e^{-x})$ is the sigmoid function. $\mathbf{h}_t \in \mathbb{R}^m$ is the hidden state of LSTM at time t. m is the size of LSTM hidden unit. The symbols $\mathbf{W}_* \in \mathbb{R}^{m \times m}$, $\mathbf{U}_* \in \mathbb{R}^{m \times m}$, and $\mathbf{b}_* \in \mathbb{R}^m$are learned during the training process. Obviously, the output of all three gates is between 0 and 1 after sigmoid function. Hence, if $\mathbf{f}_t$ is approximately 1 and $\mathbf{i}_t$ approaches 0. The previous memory cell $\mathbf{c}_{t-1}$ can be saved and passed to the current time step. Similarly, if $\mathbf{o}_t$ is approximately 1, we pass the information of $\mathbf{c}_t$ to $\mathbf{h}_t$. It means the hidden state $\mathbf{h}_t$ captures and retains the input sequences' historical information to the current time step.

Factor-Aware Attention As shown in eq. (2), LSTM blindly blend the information of all factors into a hidden state for prediction. Therefore, the hidden state incorporates negative information about irrelevant factors. However, real life experience shows that the contribution of each factor is different for prediction. Hence, we proposed a factor-aware attention network. The factor-aware attention network is composed of two layers feed forward neural network. In first layer, we assign appropriate weights to each factor in the exogenous series at time t. In second layer, we aggregate these hidden features to generate a high-level semantic $\widetilde{\boldsymbol{x}}_t \in \mathbb{R}^m$ corresponding to $\boldsymbol{x}_t$. Typically, given the k-th attribute vector of any exogenous series at time t (i.e., $\boldsymbol{x}^k$), we can employ the following attention mechanism:

$$e_t^k = \mathbf{v}^{\mathrm{T}} tanh\left(\mathbf{W}_e \mathbf{h}_{t-1}^x + \mathbf{U}_e \boldsymbol{x}^k\right)$$
$$\alpha_t^k = \frac{exp\left(e_t^k\right)}{\sum\limits_{m=1}^{n} exp\left(e_t^m\right)} \tag{3}$$

Where $\mathbf{v} \in \mathbb{R}^T$,$\mathbf{W}_e \in \mathbb{R}^{T \times 2m}$, and $\mathbf{U}_e \in \mathbb{R}^{T \times T}$are learnable parameters, and $(\cdot)^{\mathrm{T}}$stands for matrix transpose. Here, $\mathbf{h}_{t-1}^x \in \mathbb{R}^m$ is the hidden state of LSTM at time step t, and m is the size of LSTM unit. The attention weights are determined by historical hidden state and current input, which represent the impact of each exogenous factor on forecasting. With these attention weights, we can adaptively extract the important exogenous series:

$$\widetilde{\boldsymbol{x}}_t = \left(\alpha_t^1 x_t^1, \alpha_t^2 x_t^2, \ldots, \alpha_t^n x_t^n\right) \tag{4}$$

Then the hidden state at time t can be updated as:

$$\mathbf{h}_t^x = \boldsymbol{LSTM}\left(\mathbf{h}_{t-1}^x, \widetilde{\boldsymbol{x}}_t\right) \tag{5}$$

Where $\boldsymbol{LSTM}(\cdot)$ is an LSTM unit that can be computed according to eq. (2), $\boldsymbol{x}_t$ is replaced by the newly computed $\widetilde{\boldsymbol{x}}_t$. The symbol $\mathbf{h}_t^x \in \mathbb{R}^m$ is the output of factor-aware attention at time step t.

Multi-modal fusion network (MFN) To alleviate the information conflict between target and exogenous factors, we design a multi-modal fusion network, as shown in Fig. 1. In our opinion, target factor information is the most important feature and cannot be ignored. The exogenous factors are auxiliary information when we are trying to understand the dynamics of the target factor. Hence, in our model, we obtain the hidden representation of the target factor y_t by fusing high-level semantics. Because the hidden state $\mathbf{h}_t^x$ more or less contains noise information, we use a multi-modal fusion gate to combine features from different signals, thus better represent the information that are needed to solve a particular problem. The multi-modal fusion gate is a scalar in the range of [0, 1]. The multi-modal fusion gate is 1 when the hidden state $\mathbf{h}_t^x$ is helpful to improve the representation of target factor y_t, otherwise, the value of the gate is 0. The fusion process can be calculated by eq. (6).

$$\begin{aligned} \mathbf{s}_t &= \sigma\left[\mathbf{W}_y y_t : \mathbf{W}_x \mathbf{h}_t^x\right] \\ \mathbf{u}_t &= \mathbf{s}_t \odot \left(tanh\left(\mathbf{U}_x \mathbf{h}_t^x\right)\right) \\ \mathbf{z}_t &= \mathbf{W}_s\left[\mathbf{W}_y y_t : \mathbf{u}_t\right] \end{aligned} \tag{6}$$

Where [:] denotes concatenation, the symbols $\mathbf{W}_x$, $\mathbf{W}_y \in \mathbb{R}^{m \times m}$ are learnable parameters. The information matrix $\mathbf{U}_x \in \mathbb{R}^{2m \times m}$ converts the vector $\mathbf{h}_t^x \in \mathbb{R}^m$ into a 2 m-dimensional vector. The value of $\mathbf{s}_t(\mathbf{s}_t \in \mathbb{R}^{2m})$ is mapped to the interval [0, 1] by *logistic sigmoid* function.$\odot$is element-wise multiplication. Obviously, the multi-modal fusion gate will ignore the i-dimensional information of $\mathbf{U}_x \mathbf{h}_t^x$ when $\mathbf{s}_t^i = 0$ $(1 \leq i \leq 2m)$. The symbol $\mathbf{W}_s \in \mathbb{R}^{m \times 3m}$ is a parameter matrix. In the end, we fuse hidden state $\mathbf{h}_t^x$ and y_t into the representation $\mathbf{z}_t \in \mathbb{R}^m$. To model the temporal dependency of fusion feature sequences $\{\mathbf{z}_t\}_{t=1}^T = (\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_T)$, we utilize LSTM via the following eq. (7).

$$\begin{aligned} \mathbf{i}_t &= \sigma(\mathbf{W}_i \mathbf{z}_t + \mathbf{U}_i \mathbf{h}_{t-1} + \mathbf{b}_i) \\ \mathbf{f}_t &= \sigma\left(\mathbf{W}_f \mathbf{z}_t + \mathbf{U}_f \mathbf{h}_{t-1} + \mathbf{b}_f\right) \\ \mathbf{o}_t &= \sigma(\mathbf{W}_o \mathbf{z}_t + \mathbf{U}_o \mathbf{h}_{t-1} + \mathbf{b}_o) \\ \mathbf{c}_t &= \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot tanh(\mathbf{W}_c \mathbf{z}_t + \mathbf{U}_c \mathbf{h}_{t-1} + \mathbf{b}_c) \\ \mathbf{h}_t &= \mathbf{o}_t \odot tanh(\mathbf{c}_t) \end{aligned} \tag{7}$$

Where $\mathbf{h}_t \in \mathbb{R}^m$ is the LSTM hidden state at time step t, which is a vector with m dimension. $\mathbf{W}_* \in \mathbb{R}^{m \times m}$,$\mathbf{U}_* \in \mathbb{R}^{m \times m}$,and $\mathbf{b}_* \in \mathbb{R}^m$ are the learnable parameters.

4.3 Decoding process

Similarly, the decoding process is divided into two stages. In the first phase, it establishes the temporal correlation among encoder, previous prediction value, and the current decoder unit through temporal attention and LSTM. In the second phase, the decoder unit maps the hidden state $\mathbf{d}_i(1 \leq i \leq \Delta)$ into the final prediction result $\widehat{y}_i$.

Temporal attention The core component of temporal attention is the attention layer. The input of the attention layer consists of the encoder hidden state sequence $\{\mathbf{h}_t\}_{t=1}^T = (\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_T)$ and the hidden state of the decoder. For the convenience of description, we use symbol $\mathbf{d}_i \in \mathbb{R}^m$ to represent the hidden state of decoder unit i. For each decoder unit i, attention returns the corresponding context vector $\mathbf{co}_i$. The context vector $\mathbf{co}_i$ is the weighted sum of all hidden states of the encoder and their corresponding attention weights, as illustrated in Fig. 2.

In particular, given encoder hidden state sequence $\{\mathbf{h}_t\}_{t=1}^T = (\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_T)$ and previous decoder hidden state $\mathbf{d}_{i-1}$, the attention returns an importance score to $\mathbf{h}_t(1 \leq t \leq T)$. Then the softmax function transforms the importance score to an attention weight. The attention weight measures the similarity between $\mathbf{h}_t$ and $\mathbf{d}_{i-1}$. It also means the importance of $\mathbf{h}_t$ to $\mathbf{d}_{i-1}$. The attention mechanism's output is the weighted sum of all hidden state in $\{\mathbf{h}_t\}_{t=1}^T$, where the weight is the attention weight. The above process can be expressed by eq. (8).

$$\begin{aligned} e_i^t &= \boldsymbol{v}^{\mathrm{T}} tanh(\mathbf{W}\mathbf{h}_t + \mathbf{U}\mathbf{d}_{i-1}) \\ \beta_i^t &= \frac{exp\left(e_i^t\right)}{\sum_{k=1}^{T} exp\left(e_i^k\right)} \\ \mathbf{co}_i &= \sum_{t=1}^{T} \beta_i^t \mathbf{h}_t \end{aligned} \tag{8}$$

Where the symbol $(*)^{\mathrm{T}}$ means the transpose of matrix. $\mathbf{d}_{i-1} \in \mathbb{R}^m$ is the hidden state of decoder unit $i - 1$. e_i^t is the importance score. β_i^t is the attention weight. $\mathbf{co}_i \in \mathbb{R}^m$ is the context vector.

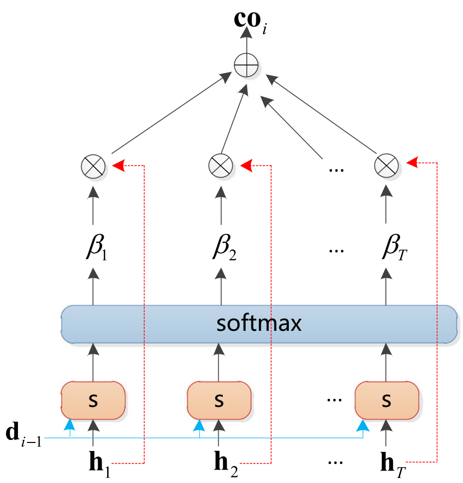


Fig. 2 Temporal attention

Subsequently, we use context vector $\mathbf{co}_i$ to update the hidden state of LSTM. Specifically, for decoder unit i, LSTM obtains its hidden state $\mathbf{d}_i$ by combining vector $\mathbf{co}_i$, previous hidden state $\mathbf{d}_{i-1}$, and predicted value $\widehat{y}_{i-1}$. The above process can be expressed by Eq. (9)

$$\begin{aligned}
\mathbf{i}_i &= \sigma\left(\mathbf{W}_*\left[\widehat{y}_{i-1} : \mathbf{co}_i\right] + \mathbf{U}_*\mathbf{d}_{i-1} + \mathbf{b}_*\right) \\
\mathbf{f}_i &= \sigma\left(\mathbf{W}_f\left[\widehat{y}_{i-1} : \mathbf{co}_i\right] + \mathbf{U}_f\mathbf{d}_{i-1} + \mathbf{b}_f\right) \\
\mathbf{o}_i &= \sigma\left(\mathbf{W}_o\left[\widehat{y}_{i-1} : \mathbf{co}_i\right] + \mathbf{U}_o\mathbf{d}_{i-1} + \mathbf{b}_o\right) \\
\widetilde{\mathbf{c}}_i &= \tanh\left(\mathbf{W}_c\left[\widehat{y}_{i-1} : \mathbf{co}_i\right] + \mathbf{U}_c\mathbf{d}_{i-1} + \mathbf{b}_c\right) \\
\mathbf{c}_i &= \mathbf{f}_t \odot \mathbf{c}_{i-1} + \mathbf{i}_t \odot \widetilde{\mathbf{c}}_i \\
\mathbf{d}_i &= \mathbf{o}_t \odot \tanh(\mathbf{c}_i)
\end{aligned} \tag{9}$$

Where [:] is the concatenation operation, and $\widehat{y}_{i-1}$ is the predicted value of the decoder unit $i-1$.

Task Learning In this paper, we use a multilayer perceptron as the task learning layer of the model with the purpose of calculating the predicted results of decoder unit i. Concretely, the predicted value $\widehat{y}_i$ is a transformation of $\mathbf{d}_i$, which can be calculated by the following equation:

$$\widehat{y}_i = \mathbf{W}_p\mathbf{d}_i \tag{10}$$

Where $\mathbf{W}_p \in \mathbb{R}^m$ is a learnable parameter.

5 Experimental results and analyses

5.1 Datasets

To compare the performance of different models on various types of MTS long-term forecasting problems, we use two available actual data sets to evaluate our proposed model and baseline method. The datasets used in our experiment are described as follows:

Beijing PM 2.5 data The dataset contains PM2.5 data and the corresponding meteorological data. We select 17,520 time series from January 1, 2013 to December 31, 2014. Each time series has 8 factors: dew point, air temperature, standard atmospheric pressure, wind direction, wind speed, hours of snow, hours of rain, PM2.5 concentration. In this paper, PM2.5 concentration is considered as the target factor. The dataset is split into training set (80%) and test set (20%) in chronological order.

Chlorophyll data The data set is taken from Tongan Bay ($118^o12^{'}$N,$24^{o}43^{'}$E). The monitoring period is from January 2009 to July 2017. There are 8733 time series. Each time series has 11 factors: Chlorophyll (Chl), Sea surface temperature (Temp), dissolved oxygen (DO), Saturated dissolved oxygen (SDO), Tide, Air temperature (Air_temp), Standard atmospheric pressure (Press), Turbidity, PH, and two meteorology wind, denoted as Wind_u and Wind_v. In this paper, chlorophyll concentration is considered as the target factor. The dataset is split into training set (90%) and test set (10%) in chronological order.

5.2 Baselines approaches

Our experiments are divided into two parts. The first part is to compare our model with the previous state-of-the-art models. The second part is ablation experiment, which compares our model with the degrade version of our model. The specific descriptions are as follows:

Seq2Seq: A model based on encoder-decoder. Kao et al. [5] applied the model to multi-step advance flood forecasting and achieved good accuracy. Therefore, the experiment uses it as one of the benchmark methods.

GED: A based-attention multivariate time series long-term forecasting model proposed by Xie et al. [17]. Experimental results on Bohai Sea and South China Sea surface temperature data sets showed that its performance was better than traditional methods, such as SVR.

STANet: A multivariate time series forecasting method proposed by He et al. [12]. They employed the model for multistep-ahead forecasting of chlorophyll. In this work, we implement it as a baseline method.

DA-RNN: DA-RNN is a one-step-ahead time-series forecasting method proposed by Qin et al. [20]. The model introduced an attention mechanism for both encoder and decoder. DA-RNN assumed the inputs must be correlated among the time. Since DA-RNN is a one-step ahead forecasting method, we implement it in the long-term prediction based on direct strategy.

DSTP-RNN: The long-term forecasting model of multivariate time series proposed by Liu et al. [21]. DSTP-RNN took into account the spatial correlation between factors and the time dependencies of the series.

HCA-GRU: A hierarchical attention-based network for sequential recommendation proposed by Cui et al. [34]. The model combined the long term dependency and user's short-term interest. In this work, we set the time frame interval to 1, and modify the output layer by mapping the learned nonlinear combination into a scalar.

MsANet: A multivariate time series forecasting model proposed by Hu et al.[37]. MsNet employed influential attention and temporal attention to extract local dependency patterns among factors and discover long-term patterns of time series trends.

Table 1 Performance of different methods compared in Chlorophyll dataset for different predictive horizons

Methods	τ= 1		τ= 6		τ= 12		τ= 24	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Seq2Seq	0.4823	0.7764	0.6439	0.9819	0.7745	1.1408	0.9086	1.2646
DA-RNN	0.4853	0.6659	0.6297	0.8388	0.6183	0.7855	0.8285	1.1767
GED	0.4267	0.6011	0.4977	0.7035	**0.5498**	0.7663	0.8845	1.1904
HCA-GRU	0.4478	0.6482	0.5082	0.7083	0.5560	0.7805	0.7038	1.1523
DSTP-RNN	0.4311	0.6121	0.5874	0.7425	0.5673	0.8113	0.7221	1.1452
STANet	0.4107	0.5823	0.5077	0.6988	0.5782	0.7694	0.7577	1.1232
MsANet	0.4415	0.6333	0.4986	0.7056	0.5604	0.7739	0.7346	1.1656
HANet	**0.3678**	**0.5689**	**0.4819**	**0.6929**	0.5565	**0.7569**	**0.6819**	**1.1064**

Besides, we implement one degraded version of our proposed model. The degrade version is used for ablation experiments:

> **HA-LSTM**: In this version, we remove the MFN component. That is to say, we blend the information of all factors (include target and exogenous factors) into a high-level semantic through factor-aware attention.

5.3 Parameters and experimental settings

In this work, we set the learning rate of all methods to 0.0001. To maintain the consistency, we use the same size for all LSTMs' hidden units. We conduct grid search for the size of LSTM's hidden state over {15, 25, 35, 45}. We set the size of time window to 24, i.e., T=24. For each sample, the first 24 time series is the input of models. We compare our model with previous state-of-the-art methods on two datasets for long-term forecasting tasks. We test all methods with horizon $\tau \in \{1, 6, 12, 24\}$ to show their effectiveness.

5.4 Experimental results and analysis

To measure the effectiveness of the proposed model in the long-term prediction of MTS, we adopt root means square error (RMSE) and mean absolute error (MAE) to assess the forecasting performance. They are calculated by the following:

$$MAE = \frac{\sum_{t=1}^{N} \left| y_t - \widehat{y}_t \right|}{N} \tag{11}$$

and

$$\boldsymbol{RMSE} = \sqrt{\frac{\sum_{t=1}^{N} \left(y_t - \widehat{y}_t \right)^{2}}{N}} \tag{12}$$

Where N is the number of samples, y_t is the real value, and $\widehat{y}_t$ is the corresponding predicted value. The closer to 0 they are, the higher the algorithm accuracy the model has.

Performance analysis In Tables 1 and 2, we show the performance of different methods in different prediction horizons.

Table 2 Performance of different methods compared in PM 2.5 dataset for different predictive horizons

Methods	τ= 1		τ= 6		τ= 12		τ= 24	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Seq2Seq	14.3048	24.7627	29.0523	48.6893	45.4879	66.6200	58.7161	81.0068
DA-RNN	12.0539	21.3099	26.3209	40.0097	37.0001	**54.4784**	51.2154	88.6673
GED	13.6901	22.6873	25.4548	40.3872	39.8709	62.1063	56.3646	78.0019
HCA-GRU	11.1151	18.9156	25.1317	39.2720	37.9419	55.9917	50.0695	**70.4730**
DSTP-RNN	11.9881	20.4029	26.7392	42.0621	38.4374	56.7373	50.3837	71.3543
STANet	12.0001	19.6192	27.2872	49.8296	38.6140	58.6099	53.2927	74.0792
MsANet	13.0729	20.2910	25.1451	39.2832	37.7276	56.5410	50.3315	73.7386
HANet	**11.0054**	**18.1005**	**24.6212**	**39.0167**	**36.4204**	54.8148	**48.3800**	71.7957

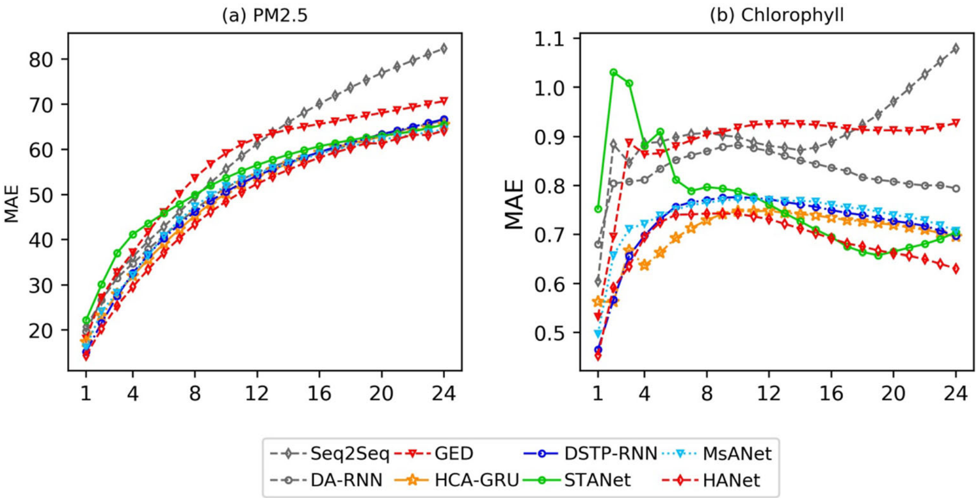


Fig. 3 Performance comparisons among different methods and different datasets when horizon is 24

The best performance is highlighted in bold. According to Tables 1 and 2, HANet model is most suitable for long-term forecasting task of multivariate time series in the mentioned models (13 out of 16 cases). However, the performance of all methods decreases with the increasing of horizon. That is to say, different model has different negative effects with the increasing of horizon. DA-RNN is a one-step ahead forecasting model, which needs to train a model for each time step in the future. In fact, the time series is a time-varying process, and its uncertainty increases as the time interval continues to increase. Since the negative effect of uncertainty, the performance of DA-RNN decreases with the increase in horizon. On the contrary, the other models only train one model to implement multi-step ahead forecasting. However, these models take the previous predicted value as the current input, so the error will gradually accumulate as the horizon continues to increase. The performance of HANet is better than baseline approaches according to Tables 1 and 2. Obviously, the increase of horizon has less negative impact on HANet than baseline methods. The experimental results show that the performance of GED is lower than other attention-based models. This is because other models analyze time series from different perspectives. Specifically, HA-LSTM and HCA-GRU attenuate the negative effects of irrelevant factors by distinguishing the contribution of each factor. STANet, DA-RNN, and MsANet consider the correlation among exogenous factors. DSTP-RNN not only considers the correlation among exogenous factors, but also focuses on the correlation between exogenous and target factors. For the same purpose, HANet adaptively filters out untrustworthy exogenous factors through factor-aware attention. According to Tables 1 and 2, we find that the attention-based model outperforms Seq2Seq in different prediction horizons. Therefore, we conclude the attention mechanism can effectively establish long-range temporal dependencies, thereby improving the performance of the model's long-term prediction.

Visualization of analysis To further illustrate the performance of HANet, we visualize the experimental results. For space reason, we only show the situation of horizon = 24 in Fig. 3. We show the MAE of HANet and other state-of-the-art models at different time steps. Here, the x-axis indicates different time steps, and the y-axis is the corresponding MAE value. Intuitively, compared with other baseline approaches, the visualization results also support the superiority of HANet. Furthermore, we see that the performance of Seq2Seq degrades faster than other attention-based methods with the increasing of predicted horizon. This result demonstrates the effectiveness of introducing an attention mechanism.

Table 3 Ablation experiments in Chlorophyll dataset for different predictive horizons

Methods	τ= 1		τ= 6		τ= 12		τ= 24	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Seq2Seq	0.4823	0.7764	0.6439	0.9819	0.7745	1.1408	0.9086	1.2646
GED	0.4267	0.6011	0.4977	0.7035	**0.5498**	0.7663	0.8845	1.1904
HA-LSTM	0.4112	0.6065	0.4961	0.7440	0.5684	0.7622	0.7070	1.1314
HANet	**0.3678**	**0.5689**	**0.4819**	**0.6929**	0.5565	**0.7569**	**0.6819**	**1.1064**

Table 4 Ablation experiments in PM 2.5 dataset for different predictive horizons

Methods	τ= 1		τ= 6		τ= 12		τ= 24	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Seq2Seq	14.3048	24.7627	29.0523	48.6893	45.4879	66.6200	58.7161	81.0068
GED	13.6901	22.6873	25.4548	40.3872	39.8709	62.1063	56.3646	78.0019
HA-LSTM	12.4656	21.8553	25.2919	40.7350	37.0144	55.4472	51.5809	72.2052
HANet	**11.0054**	**18.1005**	**24.6212**	**39.0167**	**36.4204**	54.8148	**48.3800**	**71.7957**

Ablation experiments analysis To study the performance gain of HANet's components, we conduct ablation study by implementing one degraded version of HANet, i.e., HA-LSTM. In HA-LSTM, we remove the MFN component. Therefore, HA-LSTM can be seen as GED adds a FAN module before its encoder. In Tables 3 and 4, we show the results of ablation experiments, and the best performance is highlighted in bold. According to Tables 3 and 4, the evaluation results on two real datasets show that the prediction performance of HA-LSTM is better than GED, which prove the effectiveness of FAN. Besides, the performance of HANet is better than HA-LSTM. Obviously, the experimental results show the effectiveness of MFN.

Attention weight analysis To further analyze HANet, we visualize the weight distribution of factor-aware attention, as shown in Fig. 4. For reason of space, we only show the situation of $\{1,6,12,24\}^{th}$ time step when the horizon is 24 in chlorophyll dataset. In Fig. 4, the x-axis indicates different time steps, and the y-axis is corresponding attention weight of each exogenous factor. The experimental results show that each exogenous factor has a different contribution to a given time step (i.e., the x-axis value). In other words, HANet model can not only distinguishes the importance of each factor, but also captures its dynamic changes. Moreover, the experimental results show that Sea surface temperature (Temp), air temperature (Air_temp), and PH are the most important factors, which are consistent with the studies in [35]. Meanwhile, we also found that the irrelevant factors' weight of factor-aware attention is approximately 0, such as standard atmospheric pressure (Press). Obviously, our model is credible and can provide interpretability for the research object.

Statistical analysis Statistical tests are used to examine the forecasting performance between HANet and the Baseline methods. In Table 5, we show the paired two-tailed t-tests results of all methods. In addition, we calculate and compare the average RMSE for different prediction horizon because the t-test results are easily affected by the sample size. The results prove that HANet is superior to other state-of-the-art

Fig. 4 Weight distribution of factor-aware attention

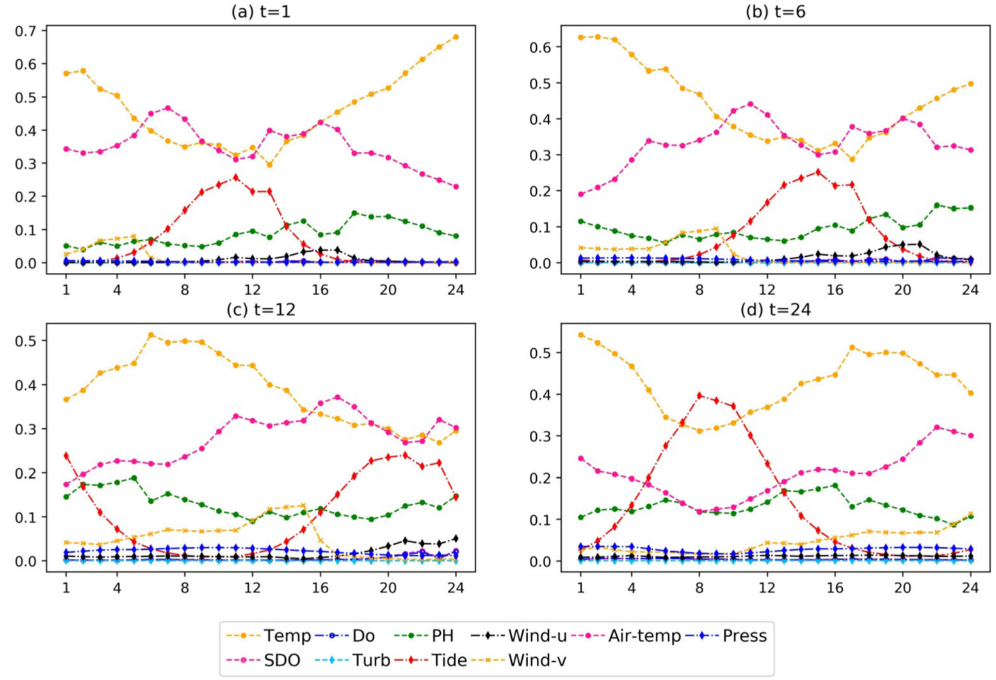


Table 5 Paired 2-tailed t-tests with HANet. Confidence level = 0.05

Methods	PM2.5			Chlorophyll		
	p value	t-statistic	avg.RMSE	p value	t-statistic	avg.RMSE
Seq2Seq	0.0000	−4.5627	55.2697	0.0002	−5.8847	1.0409
DA-RNN	0.0053	−2.7922	51.1163	0.0005	−3.0309	0.8667
GED	0.0004	−3.5140	50.7957	0.0000	−4.6342	0.8153
HCA-GRU	0.0498	−1.9619	46.1631	0.0128	−2.4871	0.8223
DSTP-RNN	0.0432	−2.0220	47.6367	0.1268	−1.5271	0.8277
STANet	0.0030	−3.1048	50.5345	0.0010	−3.2949	0.7934
MsAnet	0.0024	−3.0429	47.4635	0.0272	−2.5780	0.8196
HA-LSTM	0.0053	−2.7877	47.5607	0.1064	−1.6150	0.8410
HANet	–	–	45.9319	–	–	0.7813

methods at the 5% statistical significance level on PM2.5 data. The paired two-tailed t-test show that certain models (i.e., DSTP-RNN and HA-LSTM) are as accurate as HANet on chlorophyll data set. However, the proposed HANet has a smaller average RMSE value according to Table 5. Therefore, HANet has better predictive performance. In summary, HANet provides better predictive performance than other state-of-the-art forecasting methods.

5.5 Limitations of HANet

Although HANet can handle multivariate time series long-term forecasting problem and has a certain interpretability, which is proved by the 'Performance analysis', 'Attention weight analysis', and 'Statistical analysis' parts, it still has some limitations:

1) To effectively improve multivariate time series long-term prediction performance, HANet sacrifices more computing resources compared with other baseline approaches.
2) Inspired by human attention mechanism including the dual-stage two-phase (DSTP) model and the impact mechanism of exogenous and target factors [36], we design the FAN as a component of HANet. However, DSTP assumes that the input of the model is time-dependent, but this assumption may not always be true in practical applications.

6 Conclusion

Multivariate time series long-term forecasting has always been the subject of research in various fields such as economics, finance, and traffic. In recent years, attention-based recurrent neural networks (RNNs) have received attention due to their ability of reducing error accumulation. Since RNNs blindly blend the information of the target and non-predictive variables into a hidden state for prediction, the existing attention-based RNNs cannot eliminate the negative influence of irrelevant factors. Meanwhile these models ignore the conflict between target and exogenous factors. In this work, we propose a Hierarchical Attention Network (HANet) for multivariate time series long-term forecasting. HANet is a hierarchically structured encoder-decoder architecture, which learns both the importance factors and the long-distance temporal dependence. Specifically, we design a factor-aware attention (FA) as the first component of the encoder to eliminate the negative influence of irrelevant exogenous factors. To address the second challenge, we carefully develop a multi-modal fusion network (MFN) as the second component of the encoder. In the encoding stage, MFN trades off how much new information the network is considering from target and exogenous factors through specially designed multi-modal fusion gate. Besides, we introduce a temporal attention between the encoder and decoder network, which can adaptively select relevant encoder input items across time steps for accurate forecasting. Experimental results show that HANet is very effective and outperforms the state-of-the-art methods. We also visualize the weight distribution of factor-aware attention. The visualization of the attention weight shows that our model has great interpretability.

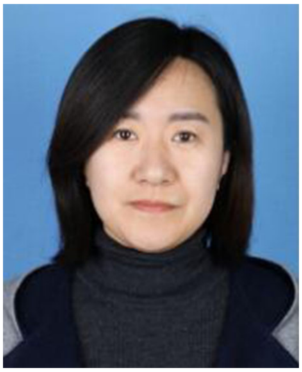
Another challenge for multivariate time series long-term forecasting is how to maintain the trend consistency between the prediction series and the original real series. In future work, we will further focus on maintaining the trend consistency at less computational costs.

Acknowledgments This work is supported by Science and Technology Project of Hebei Education Department (Grant No.ZC2021030) and Doctoral Fundation Project of Tangshan Normal University (Grant No.2022B02).

References

1. Chen T, Yin H, Chen H, Wu L, Wang H, Zhou X, Li X (2018) TADA: trend alignment with dual-attention multi-task recurrent neural networks for sales prediction. 2018 IEEE international conference on data mining (ICDM), 49–58
2. Qu L, Li W, Li W, Ma D, Wang Y (2019) Daily long-term traffic flow forecasting based on a deep neural network[J]. Expert Syst Appl 121:304–312
3. Chen K, Song X, Han D, Sun J, Cui Y, Ren X (2020) Pedestrian behavior prediction model with a convolutional LSTM encoder–decoder[J]. Phys A: Stat Mech Appl 560:125132
4. Shen L, Li Z, Kwok J (2020) Time series anomaly detection using temporal hierarchical one-class network[J]. Adv Neural Inf Proces Syst 33:13016–13026
5. Kao IF, Zhou Y, Chang LC, Chang FJ (2020) Exploring a long short-term memory based encoder-decoder framework for multi-step-ahead flood forecasting[J]. J Hydrol 583:124631
6. Hernandez-Matamoros A, Fujita H, Hayashi T, Perez-Meana H (2020) Forecasting of COVID19 per regions using ARIMA models and polynomial functions[J]. Appl Soft Comput 96:106610
7. Syafei AD, Ramadhan N, Hermana J et al (2018) Application of Exponential Smoothing Holt Winter and ARIMA Models for Predicting Air Pollutant Concentrations[J]. EnvironmentAsia 11(3)
8. Chen Y, Xu P, Chu Y, Li W, Wu Y, Ni L, Bao Y, Wang K (2017) Short-term electrical load forecasting using the support vector regression (SVR) model to calculate the demand response baseline for office buildings[J]. Appl Energy 195:659–670
9. Kuremoto T, Kimura S, Kobayashi K, Obayashi M (2014) Time series forecasting using a deep belief network with restricted Boltzmann machines[J]. Neurocomputing 137:47–56
10. Lahouar A, Slama JBH (2017) Hour-ahead wind power forecast based on random forests[J]. Renew Energy 109:529–541
11. Yin C, Dai Q (2021) A deep multivariate time series multistep forecasting network[J]. Appl Intell 52:1–19
12. He X, Shi S, Geng X, Xu L, Zhang X (2021) Spatial-temporal attention network for multistep-ahead forecasting of chlorophyll[J]. Appl Intell 51:1–13
13. Yu Y, Si X, Hu C, Zhang J (2019) A review of recurrent neural networks: LSTM cells and network architectures[J]. Neural Comput 31(7):1235–1270
14. Shen G, Tan Q, Zhang H, Zeng P, Xu J (2018) Deep learning with gated recurrent unit networks for financial sequence predictions[J]. Procedia Computer Science 131:895–903
15. Li H, Shen Y, Zhu Y (2018) Stock price prediction using attention-based multi-input LSTM[C]. Asian conference on machine learning. PMLR. 454–469
16. Muralidhar N, Muthiah S, Ramakrishnan N (2019) DyAt Nets: Dynamic Attention Networks for State Forecasting in Cyber-Physical Systems[C]. IJCAI. 3180–3186
17. Xie J, Zhang J, Yu J, Xu L (2019) An adaptive scale sea surface temperature predicting method based on deep learning with attention mechanism[J]. IEEE Geosci Remote Sens Lett 17(5):740–744
18. Lu E, Hu X (2021) Image super-resolution via channel attention and spatial attention[J]. Appl Intell 9:1–9
19. Bahdanau D, Cho KH, Bengio Y (2015) Neural machine translation by jointly learning to align and translate, 3rd International Conference on Learning Representations, ICLR 2015
20. Qin Y, Song D, Chen H, et al (2017) A Dual-Stage Attention-Based Recurrent Neural Network for Time Series Prediction[C]. IJCAI
21. Liu Y, Gong C, Yang L, Chen Y (2020) DSTP-RNN: a dual-stage two-phase attention-based recurrent neural network for long-term and multivariate time series prediction[J]. Expert Syst Appl 143: 113082
22. Shih SY, Sun FK, Lee H (2019) Temporal pattern attention for multivariate time series forecasting[J]. Mach Learn 108(8):1421–1441
23. Marques G, Agarwal D, de la Torre Díez I (2020) Automated medical diagnosis of COVID-19 through EfficientNet convolutional neural network[J]. Appl Soft Comput 96:106691
24. Huang X, Ye Y, Wang C, Yang X, Xiong L (2021) A multi-mode traffic flow prediction method with clustering based attention convolution LSTM[J]. Appl Intell:1–14
25. Chatzis SP, Siakoulis V, Petropoulos A, Stavroulakis E, Vlachogiannakis N (2018) Forecasting stock market crisis events using deep and statistical machine learning techniques[J]. Expert Syst Appl 112:353–371
26. Yin J, Rao W, Yuan M, et al (2019) Experimental study of multivariate time series forecasting models[C]. Proceedings of the 28th ACM International Conference on Information and Knowledge Management. 2833–2839
27. Qin M, Li Z, Du Z (2017) Red tide time series forecasting by combining ARIMA and deep belief network[J]. Knowl-Based Syst 125:39–52
28. Shin Y, Kim T, Hong S, Lee S, Lee EJ, Hong SW, Lee CS, Kim TY, Park MS, Park J, Heo TY (2020) Prediction of chlorophyll-a concentrations in the Nakdong River using machine learning methods[J]. Water 12(6):1822
29. Sagheer A, Kotb M (2019) Time series forecasting of petroleum production using deep LSTM recurrent networks[J]. Neurocomputing 323:203–213
30. Xue X, Gao Y, Liu M, Sun X, Zhang W, Feng J (2021) GRU-based capsule network with an improved loss for personnel performance prediction [J]. Appl Intell 51(7):4730–4743
31. Taieb SB, Atiya AF (2015) A bias and variance analysis for multistep-ahead time series forecasting[J]. IEEE Trans Neural Netw Learn Syst 27(1):62–76
32. Sutskever I, Vinyals O, Le QV (2014) Sequence to sequence learning with neural networks[C]. Advances in neural information processing systems. 3104–3112
33. Ma X, He K, Zhang D, Li D (2021) PIEED: position information enhanced encoder-decoder framework for scene text recognition[J]. Appl Intell 51(10):6698–6707
34. Cui Q, Wu S, Huang Y, Wang L (2019) A hierarchical contextual attention-based GRU network for sequential recommendation[J]. Neurocomputing 358:141–149
35. Liu X, Feng J, Wang Y (2019) Chlorophyll a predictability and relative importance of factors governing lake phytoplankton at different timescales[J]. Sci Total Environ 648:472–480
36. Hübner R, Steinhauser M, Lehle C (2010) A dual-stage two-phase model of selective attention[J]. Psychol Rev 117(3):759–784

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Hongjing Bi She received the Master of Engineering degree in Computer Applications Technology from Inner Mongolia University of Science and Technology, China, in 2010. She is currently a lecturer in the Department of Computer Science, Tangshan Normal University. Her research interests include privacy protection, data mining and time series forecasting.

Lilei Lu She received the Ph.D. in Software Engineering from Beijing University of Posts and Telecommunications, Beijing, China and M.S. in Software Engineering from Peking University, Beijing, China. She is currently an Associate Professor at the Department of Computer Science, Tangshan Normal University, Hebei, China. Her research interests include trustworthy service, cloud computing and information security.

Yizhen Meng She received a B.E. degree from North China University of Science and Technology, Tangshan, China, in 2003, and the M.S. degree in school of software engineering from Tongji University, Shanghai, China, in 2009. Now she is a Lecturer in the Computer Science Department of TangShan Normal University, TangShan, China. Her research interests include data mining and image processing.