IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 27, 2021, accepted December 14, 2021, date of publication December 20, 2021,
date of current version January 4, 2022.

Digital Object Identifier 10.1109/ACCESS.2021.3136712

A Dual-Staged Attention Based Conversion-Gated Long Short Term Memory for Multivariable Time Series Prediction

SHUFANG FENG AND YONG FENG
School of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237, China
Corresponding author: Shufang Feng (y30190716@mail.ecust.edu.cn)

This work was supported in part by the Key Laboratory of Smart Manufacturing in Energy Chemical Process, Ministry of Education, East China University of Science and Technology, Shanghai, China.

ABSTRACT In multivariate time series modeling, it is necessary to capture short-term mutation and long-term dependence information simultaneously. However, mechanism which can capture short-term change is difficult to be used to grasp long-term dependence information, and vice versa. In order to capture both short-term mutation and long-term dependence information in the same model, this paper proposed a dual-staged attention mechanism based on conversion-gated Long Short Term Memory network(DA-CG-LSTM). Hyperbolic tangent function is introduced into the input-gate and the forget-gate of Long Short Term Memory network(LSTM), which improves the ability of the network to extract the short-term mutation information. Further, dual-staged attention mechanism is added in the network, which includes input attention and temporal attention. Input attention adaptively extracts the feature relations of exogenous sequences, and temporal attention selects the relevant hidden layer states across all the time steps. Experiments on air quality and traffic flow time series data show that the proposed network has lower average absolute error, average absolute percentage error and root mean square error by more than 50% compared with Dual-staged Attention Recurrent Neural Network(DA-RNN) and Transformation-gated LSTM(TG-LSTM).

INDEX TERMS Conversion-gated, dual-staged attention mechanism, long short-term memory network, time series prediction.

I. INTRODUCTION

Time series prediction algorithms have been widely applied in many areas, such as environmental temperature [1], [16], financial stock [2], [17] and traffic condition [3], [27]. Although traditional models [4]–[8] have shown their effectiveness for various real world applications in multivariate time series prediction [9], they cannot model nonlinear relationships and do not differentiate among the exogenous terms. In order to address this issue, researches have turned to recurrent neural network in recent years.

Through application of gate mechanism, variants of recurrent neural network (such as LSTM [10] and GRU [11]) can not only effectively address issues of gradient disappearance and explosion, but also capture nonlinear and long-term dependence relationships. However, LSTM is unable to capture short-term information with high correlation with the target sequence effectively [22].

In order to enhance the capability of capturing long-term dependence, encoder-decoder, combined with recurrent neural network, was proposed to address variable filtering issues for multivariable time series [12]. Also, the same network has attracted wide attention in machine translation [13], where original sentences were encoded to fixed-length vectors and then decoded to generate corresponding translation. However, with increase of the length of input sequence, the performance of encoder-decoder network deteriorated rapidly. For time series forecasting problems, it is usually necessary to consider a relatively long segment of target and associated sequences. To address this issue, inspired by human attention mechanism, encoder network based on attention mechanism, which adopted appropriate mechanism to select hidden state

The associate editor coordinating the review of this manuscript and approving it for publication was Jon Atli Benediktsson.

across all time steps [14], was proposed. Attention based LSTM, was used to generate translation dictionary with relative probability distribution [15]. Double attention layers were added in LSTM to extract spatio-temporal feature from exogenous sequences, which can improve efficiency and performance in grasping long-term dependence information [16]. Dual-staged attention based recurrent neural network(DA-RNN) was proposed [17]. In the first stage, input attention and encoder network were used to select relevant driving series at each time step. In the second stage, temporal attention was combined with decoder network to select relevant hidden states across all time steps, which can effectively capture long-term dependent information of target sequence. GeoMAN [18] introduced multi-level attention mechanism to model the dynamic spatio-temporal dependencies. DSTP-RNN [19] enhanced spatial correlation between exogenous sequences through application of multi-level attention mechanism. Although DA-RNN and the similar networks [17]–[19] have advantages of extracting hidden information and obtain long-term dependence relationship, short-term mutation information can't be dig out. In these networks, the attention weights are small in this driving sequence, which means they cannot pay attention to short-term mutation information.

In order to effectively capture short-term mutation information, Parametric Sigmoid Norm Based CNN(PSNET) [20] was proposed to separate centroid of samples of different complexity by translating and scaling Sigmoid function, in order to extract relevant input sequences. However, it cannot solve the problem of Sigmoid function's oversaturation. Maxout [21], which fits convex functions as activation functions using segmented linear functions through piecewise linear functions, extracts multi-dimensional feature information, and improves the ability to capture short-term mutation information. However, it is difficult to be widely used due to the large number of parameters. TG-LSTM [22], which used hyperbolic tangent function at input-gate and cell state renewal, anti-hyperbolic tangent function at forget-gate on LSTM, can effectively capture short-term mutation information. However, important information of the data tends to be lost due to the decreasing value of the hyperbolic tangent function. In addition, hyperbolic tangent function is constantly used in the process of cell state renewal, which is not conducive to capture the law of long-term dependence, so long-term information is actively discarded.

On the whole, network based on temporal attention mechanism can capture long-term dependence information, while it is difficult to capture short-term mutation information. The improvement of neural network's activation function, which ignores long-term dependence information, can capture short-term mutation information. Therefore, how to find the balance between the two mechanisms is challenged.

Dual-Staged Attention Based Conversion-gated Long Short Term Memory (DA-CG-LSTM) was proposed in this paper to get the balance. Anti-hyperbolic tangent and hyperbolic tangent function are introduced into forget-gate and input-gate of LSTM respectively, in order to capture short-term mutation and delete redundant information. Moreover, for long-term dependence information, dual-staged attention is added. In the first stage, input attention adaptively allocates weights to feature and temporal dimension of time series. Hidden states are obtained by inputting into conversion-gated Long Short Term Memory(CG-LSTM). In the second stage, temporal attention is used to allocate weights in all time steps. With new activation functions and attention mechanism, DA-CG-LSTM can balance between long dependence and short mutation information. In order to verify the performance, DA-CG-LSTM, DA-RNN, TG-LSTM and other models were tested on Air Quality, MITV traffic flow and Appliances Energy dataset. Prediction accuracy is significantly improved. Moreover, compared with DA-RNN, the number of DA-CG-LSTM's parameters is reduced by 50%.

II. DA-CG-LSTM

A. MULTIVARIATE TIME SERIES PREDICTION PROBLEM

Given n-dimensional input series with the length of window size T, row vector $x^k = (x_1^k, x_2^k, \ldots x_T^k) \in \Re^T$ represents k-dimensional input sequence value when time window size is T; column vector $x_t = (x_t^1, \ldots x_t^n) \in \Re^n$ represents input sequence value in n-dimension at time t.

In general, given previous value $(y_{t-T+1}, \ldots, y_{t-1}), y_t \in \Re$ of target series and $(x_{t-T+1}, x_{t-T+2}, \ldots x_t), x_t \in \Re^n$ of the input sequence at current and past moments, time series prediction model aims to learn the nonlinear mapping function of current target value:

$$\hat{y}_t = F(y_{t-T+1}, y_{t-T+2}, \ldots, y_{t-1}, x_{t-T+1}, x_{t-T+2}, \ldots x_t) \tag{1}$$

where, $F(\cdot)$ is a nonlinear mapping function that needs to learn.

B. MODEL STRUCTURE

In order to effectively address issues of capturing short-term mutation information and processing long-term dependence information in the same model, this paper proposes a dual-staged attention mechanism network based on CG-LSTM. For capturing short-term mutation information, CG-LSTM adds anti-hyperbolic tangent function at forget-gate and hyperbolic tangent function at input-gate. Before CG-LSTM converges, multiple backpropagation calculations are needed to learn optimal parameters of loss function, partial derivative of conversion gate makes value range of gradient data stream constantly change. In other words, short-term mutation information will correspond to the most significant interval of range variation, thus capturing mutation pattern well.

Using temporal attention is able to enhance the capability of processing long-term dependence information [17]. Dual-staged attention mechanism, able to select hidden states that

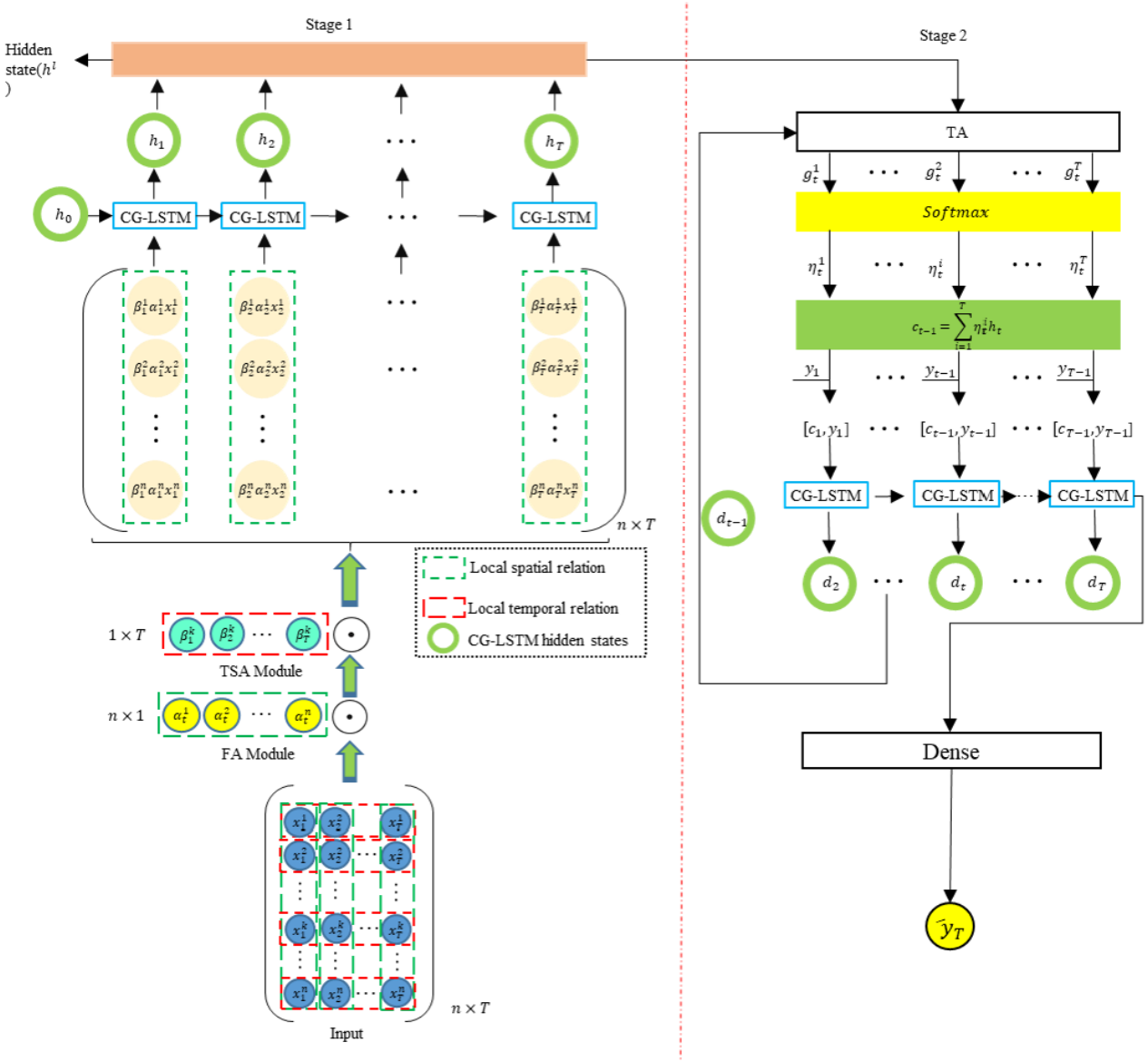


FIGURE 1. Model Structure of DA-CG-LSTM, Stage 1 represents Input Attention Mechanism, Stage 2 represents Temporal Attention Mechanism, TSA Module represents Time step dimension attention Module, FA Module represents Feature Dimension Attention Module, TA represents Temporal Attention Mechanism.

has high correlation with target sequence in the first stage, able to use some classification information to filter hidden states in the second stage, can achieve best effectiveness [25]. Therefore, this paper introduced dual-staged attention mechanism. The structural framework of DA-CG-LSTM is shown in Figure 1. In the first stage, before input sequences enter into CG-LSTM to capture short-term mutation information, input attention is used to adaptively allocate weights to feature and temporal dimension of time series, and then enters into CG-LSTM to get hidden states as local driving sequences. In the second stage, temporal attention used hidden states and cell states of CG-LSTM as classification information to adaptively assign weights to local driving sequence across all time steps, which will achieve optimal prediction effect. Finally, this paper uses DA-CG-LSTM to predict output value $\hat{y}_t$: $\hat{y}_t = F(y_{t-T+1}, \ldots, y_{t-1}, x_{t-T+1}, x_{t-T+2}, \ldots x_t) = f(h_t, c_t)$, h_t, c_t represent hidden states and cell states of CG-LSTM respectively.

The backpropagation function of DA-CG-LSTM proposed in this paper is smooth and differentiable, so root mean square error is used as error function for model training. Local driving sequence obtained by CG-LSTM can adaptively select relevant input sequence with short-term mutation information, while using temporal attention mechanism to capture long-term dependence information.

C. CONVERSION-GATED LSTM (CG-LSTM)

1. The input-gate formula of LSTM can be expressed as:

$$i_t = \sigma(W_i \cdot x_t + U_i \cdot h_{t-1} + b_i) \tag{2}$$

Considering that hyperbolic tangent function can alleviate over-saturation issue of Sigmoid function [22], the input-gate formula of CG-LSTM is shown in Equation 3:

$$i_t = \tanh(\sigma(W_i \cdot x_t + U_i \cdot h_{t-1} + b_i)) \tag{3}$$

Before convergence of conversion-gated LSTM, multiple backpropagation calculations are carried out to learn optimal parameters of loss function, and partial derivative of conversion gate makes value range of gradient data stream constantly change. At this time, short-term mutation information will correspond to the most significant interval of range change, thus capturing mutation law well. After passing through input-gate of conversion-gated LSTM, current information stream retained will be more scattered, which enables the network to consider more mutation information in the process of iteration.

In cell state renewal formula, TG-LSTM only performs hyperbolic tangent processing on the input gate, making value range of input-gate smaller, is performed on input-gate, which may cause network to focus more on capturing long-term and delete redundant long-term information. At the expense of certain current input information, long-term dependence and short-term mutation information are maximized.

2. The forget-gate output formula of LSTM is:

$$f_t = \sigma(W_f \cdot x_t + U_f \cdot h_{t-1} + b_f) \tag{4}$$

Due to saturation of Sigmoid activation function, it cannot capture short-term mutation information effectively. TG-LSTM [22] disperses data flow by introducing inverse hyperbolic tangent function according to characteristics of its partial derivative, so as to achieve the effect of capturing short-term mutation information. The forget-gate formula of TG-LSTM is:

$$f_t = 1 - \tanh(\sigma(W_f \cdot x_t + U_f \cdot h_{t-1} + b_f)) \tag{5}$$

According to the function graph of $1 - \tanh(\cdot)$, it can be obtained when independent variable is at [0, 1]. Dependent variable at this stage does not produce saturation [12], and is inferred that using arctangent operation of forget-gate can solve saturation problem of Sigmoid function. However, the slope of forget-gate's formula is formed as: $1 - (\tanh(\cdot))^2$. Since value range of its derivative is [0.25, 1], the effect of data flow dispersion is not obvious, this paper introduced an improved form of forget-gate, which is shown in Equation 6:

$$f_t = 1 - \tanh(1 - 1/(\sigma(W_f \cdot x_t + U_f \cdot h_{t-1} + b_f))^2), \sigma(\cdot) \in (0, 1) \tag{6}$$

The value range of f_t is $(0, 1)$, and it is consistent with the changing trend of $\sigma(\cdot)$, which solves this issue that TG-LSTM reduces main information, and retains the advantage of TG-LSTM capturing short-term mutation information. Forget-gate's slope of CG-LSTM is: $f_t' = (1 - (\tanh(1 - 1/(\sigma(\cdot)^2))^2) \cdot (2/\sigma(\cdot)^3), \sigma(\cdot) \in (0, 1)$. As $\sigma(\cdot)$ decreases, value first increases and then decreases.

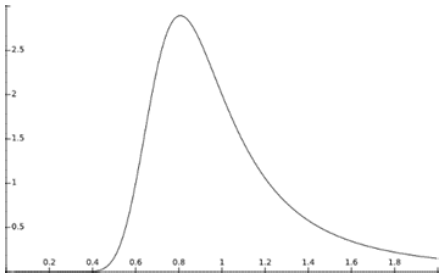


FIGURE 2. Forgetting gate of CG-LSTM's derivative graph.

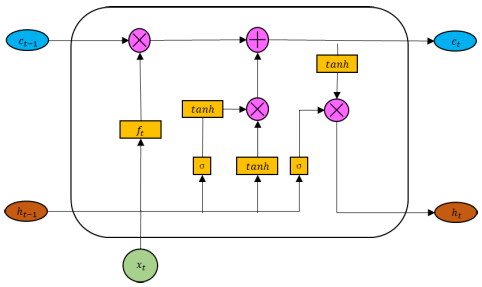


FIGURE 3. Structure of CG-LSTM.

As shown in Figure 2, when domain interval is (0, 1], value range is (0, 2.89]. Multiple backpropagation calculations are required to learn optimal parameters of loss function before CG-LSTM converges, and partial derivative of conversion gate makes value range of gradient data stream continuously change. In other words, short-term mutation information will correspond to most significant interval of range change, so that well capturing mutation law. Moreover, after forget-gate's output of CG-LSTM, retained data will be more dispersed, which causes DA-CG-LSTM considering more mutation information during iteration process, which is more effective than TG-LSTM and addresses the issue of over-saturation of Sigmoid function. In addition, this paper also experiments to prove that conversion mechanism can enhance capability of capturing short-term mutation information compared with LSTM. Moreover, when $\sigma(\cdot) = 0.5$, value of forget-gate's output is 0.005, which means that forget-gate discards more redundant information and can make better use of the short-term mutation relationship between information.

In this paper, gate mechanism of LSTM [10] is improved, and structure of CG-LSTM is shown in Figure 3.

CG-LSTM updates cell state c_t and hidden state h_t by using forget-gate f_t, output-gate o_t and input-gate i_t, as shown in Equation 7-11:

$$f_t = 1 - \tanh(1 - 1/(\sigma(W_f \cdot x_t + U_f \cdot h_{t-1} + b_f))^2) \tag{7}$$
$$i_t = \sigma(W_i \cdot x_t + U_i \cdot h_{t-1} + b_i) \tag{8}$$
$$c_t = f_t \bigodot c_{t-1} + i_t \bigodot \tanh(W_g \cdot x_t + U_g \cdot h_{t-1} + b_g) \tag{9}$$
$$o_t = \sigma(W_o \cdot x_t + U_o \cdot h_{t-1} + b_o) \tag{10}$$
$$h_t = o_t \bigodot \tanh(c_t) \tag{11}$$

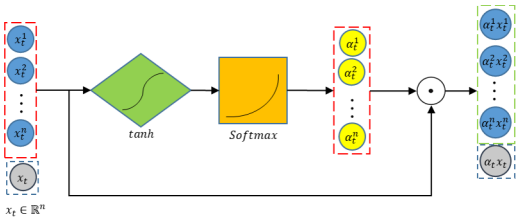


FIGURE 4. Feature scale attention weighting.

Here, $\sigma(\cdot)$ stands for Sigmoid activation function; $x_t \in \Re^n$, $h_{t-1} \in \Re^m$ represent current input sequence and previous hidden state, $W_f, W_i, W_o, W_g \in \Re^{m\times n}$,$U_f, U_i, U_o, U_g \in \Re^{m\times m}$, $b_f, b_i, b_o, b_g \in \Re^m$ represent the parameters to be learned, where $b_f, b_i, b_o, b_g \in \Re^m$ represents the deviation value.

In this paper, DA-CG-LSTM is proposed based upon CG-LSTM. The purpose is to achieve balance between capturing long-term dependence and short-term mutation information. Since the processing of anti-hyperbolic tangent of forget-gate, information is more dispersed after entering into it, which can solve problem that information captured is relatively concentrated due to saturation issue of Sigmoid function, and retain historical information. In addition, it has carried on hyperbolic tangent at input-gate, which makes short-term mutation information easier to capture. However, CG-LSTM which has improvements on forget-gate will discard some information. Therefore, attention mechanism is added to select input sequence adaptively before introducing CG-LSTM. At this time, input sequence retains more effective information, so that more invalid information can be abandoned when data passes through forget-gate. The weighted input sequence is analyzed by CG-LSTM to obtain local driving sequence, and temporal attention mechanism is used to address long-term dependence issue of CG-LSTM.

D. DUAL-STAGED ATTENTION MECHANISM

1) INPUT ATTENTION MECHANISM

In the first stage, single-layer attention can only extract information from the time axis of input sequence. However, the time axis and the feature axis of time series will have different influences on target variable, so this paper introduced double-layers attention.

Given time series X, local driving sequence x_t is mapped function to h_t by nonlinear function $f(\cdot)$:

$$h_t = f_1(h_{t-1}, o(x_t)) \tag{12}$$

Here, $h_t \in \Re^m$ is the hidden state of Conversion-gated LSTM at the moment, and $o(\cdot)$ represents weights assignment function.

Firstly, input attention is used to obtain weighted feature dimension of input sequence, as shown in Figure 4:

$$e_t^k = v_e^T \tanh(W_e \cdot x_t + B_e) \tag{13}$$

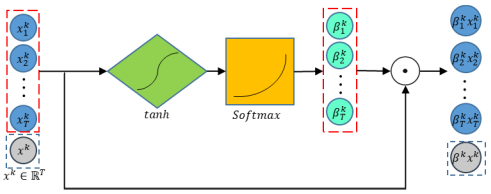


FIGURE 5. Time scale attention weighting.

$$\alpha_t^k = \exp(e_t^k) / \sum_{i=1}^{n} \exp(e_t^i) \tag{14}$$

Here, $v_e^T \in \Re^n$, $W_e \in \Re^{m\times m}$ are the parameters to learn, and $B_e \in \Re^n$ is deviation term. α_t^k represents attention weight of k-dimensional feature at time t. Softmax function is used to ensure that sum of attention weights is 1. New input sequence can be obtained according to the obtained attention weights:

$$\tilde{x}_t = (\alpha_t^1 * x_t^1, \alpha_t^2 * x_t^2, \ldots, \alpha_t^n * x_t^n)^T \tag{15}$$

Then weighted input sequence is obtained by attention of temporal dimension, as shown in Fig. 5:

$$l_t^k = v_s^T \tanh(W_s \cdot \tilde{x}_t + B_s) \tag{16}$$

$$\beta_t^k = \exp(l_t^k) / \sum_{j=1}^{T} \exp(l_t^j) \tag{17}$$

Here, $v_s^T \in \Re^n$, $W_s \in \Re^{T\times T}$ are parameters to learn, $B_s \in \Re^T$ is deviation term, β_t^k has same meaning with α_t^k, and weighted input sequence is obtained as:

$$\begin{aligned} X &= (\beta_1 \tilde{x}_1, \beta_2 \tilde{x}_2, \ldots \beta_T \tilde{x}_T)^T \\ &= \begin{bmatrix} \beta_1^1 * \alpha_1^1 * x_1^1 & \cdots & \beta_1^n * \alpha_1^n * x_1^n \\ \vdots & \ddots & \vdots \\ \beta_T^1 * \alpha_T^1 * x_T^1 & \cdots & \beta_T^n * \alpha_T^n * x_T^n \end{bmatrix} \end{aligned} \tag{18}$$

2) TEMPORAL ATTENTION MECHANISM

In order to address the issue of capturing short-term mutation information, this paper uses CG-LSTM to analyze multiple input sequences. However, as the length of input sequence increases, performance of capturing long-term dependence information will deteriorate. Therefore, in order to address this issue, temporal attention is used to adaptively select hidden states across all time steps. Specifically, according to hidden states $d_{t-1} \in \Re^l$ and cell states $c_{t-1} \in \Re^l$ of CG-LSTM, local driving sequences at time t is obtained:

$$g_t^i = z_{op}^T \tanh(W_{op} \cdot [d_{t-1}; c_{t-1}] + U_{op} \cdot h_i + b_{op}) \tag{19}$$

$$\eta_t^i = \exp(g_t^i) / \sum_{j=1}^{l} \exp(g_j^i) \tag{20}$$

Here, $[d_{t-1}; c_{t-1}] \in \Re^{2l}$ represent hidden states and cell states of conversion-gated LSTM; $z_{op}^T \in \Re^m$, $U_{op} \in \Re^{m\times m}$, $W_{op} \in \Re^{m\times 2p}$ are parameters to learn, and $b_{op} \in \Re^m$ is deviation term. η_t^i represents attention weight of the i-th

driving sequence, and temporal attention information c_t is obtained as follows:

$$c_t = \sum_{j=1}^{l} \eta_t^i * h_i \tag{21}$$

The update operations for hidden state d_{t-1} and cell state c_t of conversion-gated LSTM are as follows:

$$f_t = 1-\tanh(1-1/(\sigma(W_f \cdot [c_{t-1}; y_{t-1}]+U_f \cdot d_{t-1}+b_f))^2) \tag{22}$$

$$i_t = \sigma(W_i \cdot [c_{t-1}; y_{t-1}] + U_i \cdot d_{t-1} + b_i) \tag{23}$$

$$c_t = f_t \bigodot c_{t-1}+i_t \bigodot \tanh(W_g \cdot [x_{t-1}; y_{t-1}]+U_g \cdot d_{t-1}+b_g) \tag{24}$$

$$o_t = \sigma(W_o \cdot [c_{t-1}; y_{t-1}] + U_o \cdot d_{t-1} + b_o) \tag{25}$$

$$d_t = o_t \bigodot \tanh(c_t) \tag{26}$$

where, $W_f, W_i, W_o, W_g \in \Re^{l\times m+1}$,$U_f, U_i, U_o, U_g \in \Re^{l\times l}$, $b_f, b_i, b_o, b_g \in \Re^{l}$ represent parameters to learn, and $[c_{t-1}; y_{t-1}] \in \Re^{m+1}$ represents information selected by attention mechanism and output at the last moment.

E. TRAINING PROCESS

In this paper, Adam optimizer are used to train the model. The backpropagation function of DA-CG-LSTM is smooth and differentiable, so RMSE is chosen as objective function and parameters are learned through backpropagation:

$$RMSE(\hat{y}_t, y_t) = \sqrt{(1/N)\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{27}$$

where, N represents the number of training samples.

III. EXPERIMENT

A. DATA SETS

Air Quality dataset [23] is collected from five metal oxide chemical sensor arrays embedded in an air quality chemistry multi-sensor device, contains 9,358 time series instances. The installation, located in a heavily polluted area of an Italian city, recorded free of charge on-site air quality sensor records from February 2004 to March 2005 on a highway. Ground truth averages hourly concentrations of CO, non-paraffin, benzene, total nitrogen oxides (NOx) and nitrogen dioxide (NO2), provided by a reference certified analyzer at the same location. Missing values are marked with -200. In this experiment, the benzene content was taken as the target sequence, and 13 related exogenous sequences were selected. Considering that air pollution degree may have a certain relationship with time, time was included in the selected exogenous sequences. The first 5979 data instances were selected as the training set, the next 1495 data as the verification set, and the final 1869 data as the test set.

The Metro Interstate Traffic Volume(MITV) dataset [29] represents the hourly traffic volume of Interstate 94 westbound between Minneapolis and St. Paul at the 301 stations for 48,203 time-series instances. The hourly westbound traffic volume of this road was recorded from October 2012 to September 2018, and the climate, time, holiday or not, temperature at that time were recorded as variables. In this experiment, 8 related exogenous sequences were selected with traffic volume as the target sequence. In addition, the first 30,840 data are used as the training set, the next 7711 data are used as the verification set, and the final 9638 data are used as the test set.

The Appliances energy dataset [26] is a total of 19,735 items of electricity consumption data recorded every 10 minutes over a period of 4.5 months for a particular house. The collectors used a ZigBee wireless sensor network to collect temperature and humidity around and inside the house, uploading the data every 10 minutes for a total of 4.5 months, and used this data as features for the electricity consumption data. Finally, the dataset has 28 features. In this experiment, the initial 12611 data were used as the training set, the subsequent 3177 data were used as the validation set and the final 3947 data were used as the test set.

B. PARAMETER SETTING AND EVALUATION INDICATORS

Experiments were carried out on the platform of Python 3.7. The models are developed with Tensorflow 2.0. Each model is trained 20 times, and average values of training error and other statistical indicators were obtained. In these experiments, Check Point module was used to select the optimal model during every training procedure.

For DA-CG-LSTM, there are three parameters, including window size T, batch size of input sequence, and the number of units m in the hidden layer of CG-LSTM. Set $T \in [5, 10, 15, 20, 30, 50]$, $batch\_size \in [16, 32, 64, 128, 256, 512]$ and $m \in [10, 20, 30, 40, 50]$. Find the optimal parameters which are $T = 15$, $batch\_size = 128$ and $m = 30$.

In order to measure the effectiveness of various time series prediction methods, three different evaluation indicators were adopted in this paper. They are root mean square error (RMSE), mean absolute error (MAE) and mean absolute percentage error(MAPE). Specifically, for target value y_T and predicted value $\hat{y}_T$, the formula of mean square error is shown in Equation 27. The mean absolute error is defined as:

$$MAE(\hat{y}_t, y_t) = (1/N)\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{28}$$

When comparing prediction performance of different data sets, mean absolute percentage error is a common method because it measures proportion of prediction bias based on ground true value:

$$MAPE(\hat{y}_t, y_t) = (1/N)\sum_{i=1}^{N}|(y_i - \hat{y}_i)/y_i| \times 100\% \tag{29}$$

TABLE 1. Performance comparison of different networks in three data sets($T = 15$).

Methods	Air Quality			MITV			Appliances Energy		
	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE
LSTM	1.13±0.12	15.70±0.07	0.54±0.04	3.74±0.08	7.53±0.10	2.35±0.06	10.97±1.14	4.99±0.65	5.17±0.35
TG-LSTM	0.45±0.07	9.30±0.21	0.32±0.05	3.68±0.05	7.40±0.31	2.21±0.02	8.24±0.96	4.46±0.53	3.19±0.26
LSTM+Zonout	0.44±0.19	10.17±0.26	0.35±0.03	3.67±0.23	7.37±0.49	2.30±0.07	8.39±0.84	4.47±0.59	3.73±0.34
AT-LSTM	0.49±0.11	10.76±0.26	0.35±0.03	3.62±0.09	7.15±0.12	2.23±0.05	7.89±0.87	3.58±0.41	3.21±0.27
DA-LSTM	0.42±0.10	9.81±0.19	0.34±0.07	4.02±0.11	8.81±0.27	2.75±0.10	9.79±1.08	4.11±0.35	4.38±0.26
DA-RNN	1.05±0.11	12.15±0.13	0.44±0.06	3.15±0.09	6.15±0.08	1.92±0.07	8.07±0.27	3.71±0.12	3.54±0.16
CG-LSTM	0.36±0.06	6.39±0.16	0.22±0.04	3.51±0.05	6.71±0.12	2.09±0.06	7.78±0.68	4.03±0.56	4.02±0.54
FA+CG-LSTM	0.29±0.15	5.52±0.23	0.19±0.13	3.53±0.12	6.91±0.45	2.14±0.12	7.89±0.38	3.12±0.41	3.52±0.16
SA-CG-LSTM	0.91±0.18	11.42±0.14	0.42±0.11	3.07±0.23	5.89±0.08	1.84±0.11	7.50±0.64	3.86±0.45	3.77±0.35
DA-CG-LSTM	0.25±0.13	3.32±0.11	0.12±0.06	2.79±0.21	5.06±0.19	1.58±0.09	7.13±0.08	3.01±0.14	3.12±0.26

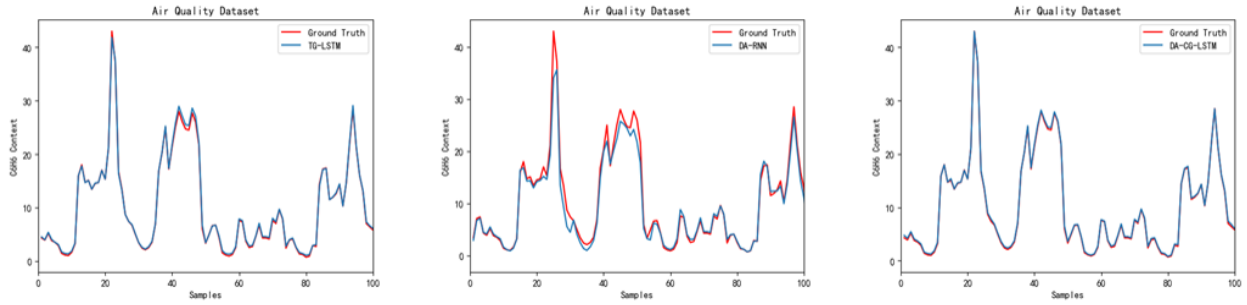
TABLE 2. Performance comparison of different networks in three data sets($T = 30$).

Methods	Air Quality			MITV			Appliances Energy		
	RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE
LSTM	1.18±0.14	18.90±0.20	0.65±0.09	4.12±0.45	7.80±0.13	2.38±0.06	12.13±1.17	6.01±0.74	5.92±0.42
TG-LSTM	0.63±0.06	9.01±0.18	0.31±0.05	3.93±0.27	7.80±0.13	2.38±0.06	8.98±1.03	4.95±0.69	3.75±0.65
LSTM+Zonout	0.52±0.21	10.17±0.31	0.35±0.12	3.67±0.41	7.52±0.18	2.34±0.08	9.24±0.97	5.01±0.75	4.22±0.43
AT-LSTM	0.42±0.12	9.59±0.14	0.33±0.07	3.61±0.19	7.05±0.11	2.20±0.10	8.56±0.91	4.23±0.51	3.75±0.41
DA-LSTM	0.57±0.11	10.67±0.16	0.47±0.06	4.03±0.09	7.85±0.07	2.46±0.06	10.86±1.32	4.68±0.3	5.23±0.39
DA-RNN	1.43±0.05	15.70±0.11	0.54±0.10	3.62±0.17	9.45±0.15	2.95±0.09	8.41±0.75	4.23±0.23	4.12±0.19
CG-LSTM	0.44±0.09	8.14±0.13	0.28±0.09	3.72±0.13	7.71±0.11	2.38±0.07	8.75±0.82	4.23±0.67	4.75±0.79
FA+CG-LSTM	0.35±0.12	8.72±0.24	0.30±0.11	3.55±0.21	6.97±0.14	2.11±0.14	8.91±0.68	3.68±0.71	4.01±0.25
SA-CG-LSTM	1.26±0.16	13.57±0.19	0.57±0.10	4.55±0.26	8.76±0.20	2.72±0.06	8.10±0.76	4.21±0.47	3.94±0.46
DA-CG-LSTM	0.28±0.11	5.52±0.20	0.19±0.10	3.12±0.23	6.65±0.17	2.09±0.09	7.75±0.39	3.47±0.23	3.96±0.33

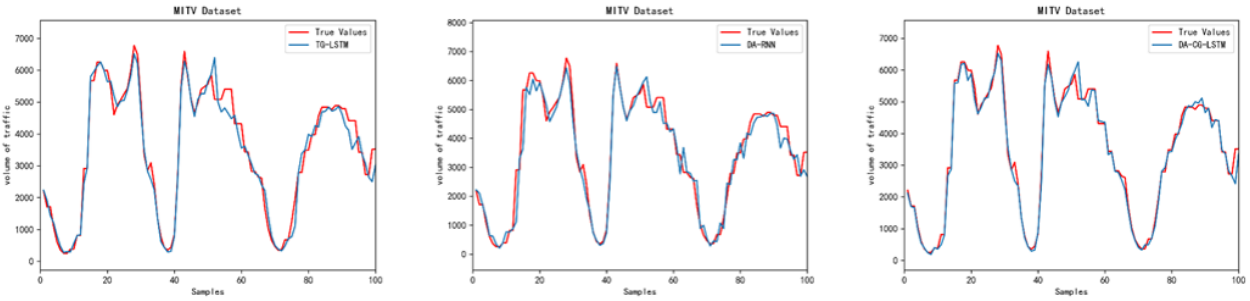
C. RESULTS

In this paper, DA-CG-LSTM, CG-LSTM and other time prediction models are compared were compared on Air Quality, MITV and Appliances Energy datasets. (1)LSTM [10] address the issue of long-term dependence through gate mechanism. (2)LSTM+Zoneout [24]: Through selectively updating neural units, weak learning ability that long-term dependence of LSTM can be greatly improved. (3)TG-LSTM [22] can effectively capture short-term mutation information. (4)ATT-LSTM [15] combines attention mechanism and LSTM to address the issue of long-term dependence. (5)DA-RNN [17] introduced dual-staged attention mechanism combined with LSTM on time series prediction, which could achieve good effect. (6)DA-LSTM [16] adds double-layer salient attention at input stage of LSTM to improve efficiency and performance, which can solve long-term dependence problem. (7)CG-LSTM is improved through gate mechanism based on LSTM. (8)The first stage of attention mechanism, FA+CG-LSTM: using two-dimensional attention in CG-LSTM. (9)The second stage attention mechanism, SA-CG-LSTM: adaptive selection of all hidden states is carried out through temporal attention.

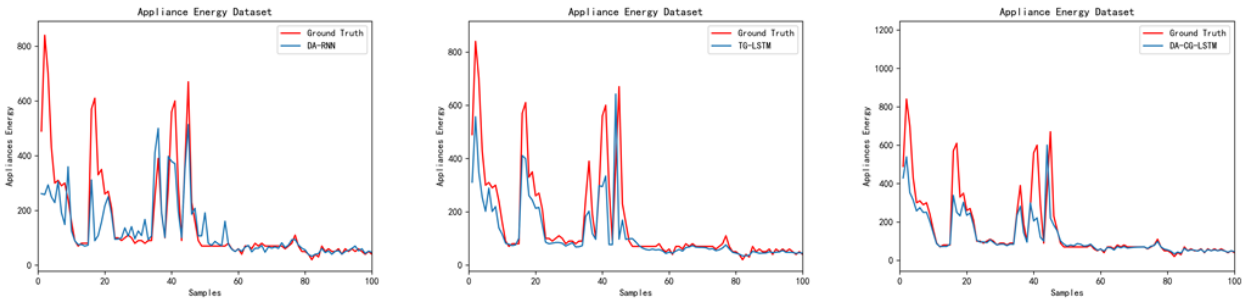
All models were compared in two cases $T = 15$ and $T = 30$. The obtained evaluations are shown in Table 1 and Table 2. It is obvious that DA-CG-LSTM performs best in both datasets. Moreover, prediction effect of CG-LSTM is better than that of TG-LSTM and LSTM, the prediction accuracy of CG-LSTM increased by 5% compared with TG-LSTM and 9% compared with LSTM. Through comparing performance of different models on these data sets, it is shown that although LSTM has certain ability of processing long-term dependence, performance of LSTM is poor when time step increases. DA-RNN has better prediction performance than other baseline models on MITV and Appliance Energy. However, prediction performance of DA-RNN is not optimal on Air Quality. It can be inferred that Air Quality has high instability and is required to capture short-term mutation information. Although the processing of Zoneout model can enhance long-term dependence learning ability of LSTM, it does not consider different effects of multivariate series on target sequence in advance, and ignores complex association of multivariate. AT-LSTM was originally applied to natural language processing, which performance is significantly weaker than that of many deep learning networks since only single-level salient attention is used to calculate scores of multivariate series. DA-LSTM, improved based on AT-LSTM, added temporal attention in output stage. Input sequence is screened through two-layer attention, and exogenous sequences, more relevant to target sequence, are obtained, which solves the problem of lack of ability to capture short-term mutation information to a certain extent. However, due to using LSTM, its performance is poor in the face of increasing time step. It performs well in Air Quality, needs to make better of short-term mutation information, but it is difficult to perform well in MITV and Appliance Energy, need to make use of long-term dependence



(a) Comparison of Air Quality Dataset Prediction



(b) Comparison of MITV Dataset Prediction



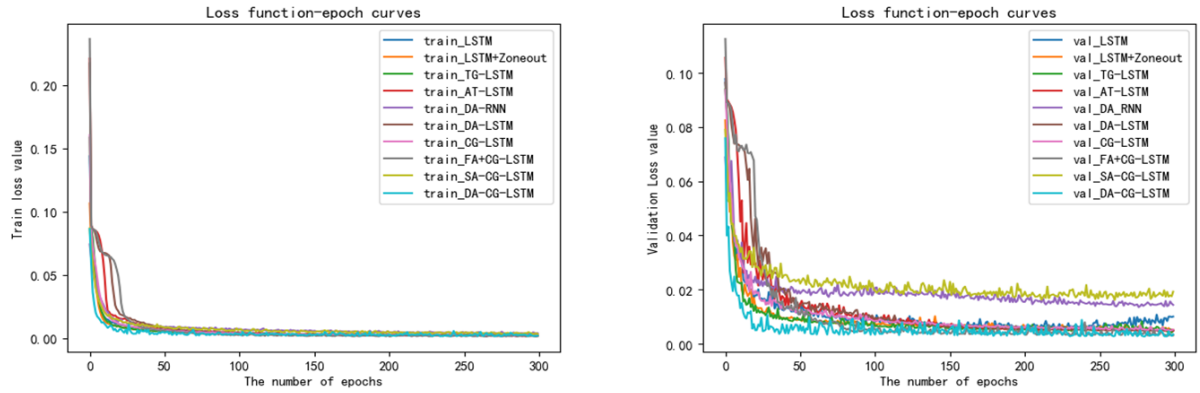
(c) Comparison of Appliances Energy Dataset Prediction

FIGURE 6. **Comparison of Data Prediction in DA-RNN, TG-LSTM and DA-CG-LSTM.**

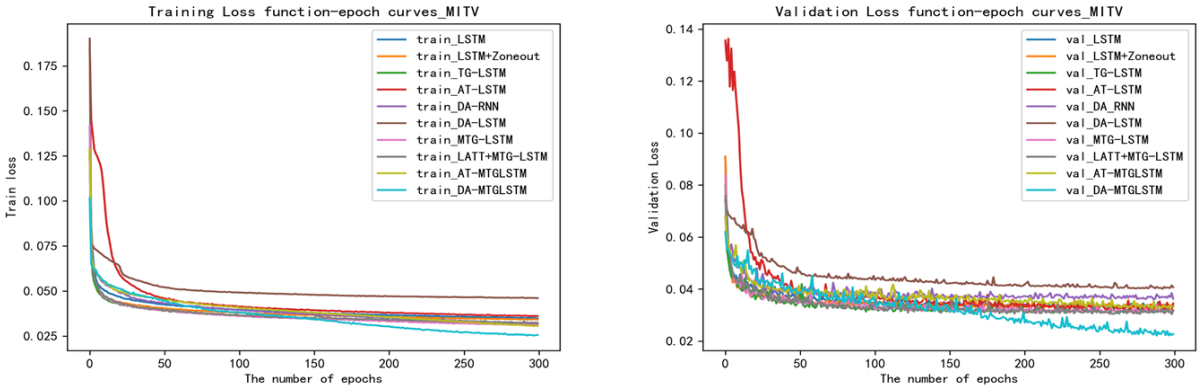
relationship. DA-CG-LSTM is built on the basis of DA-RNN and TG-LSTM, inherits their advantages. It can be seen from Figure 6(a) that the prediction of Air Quality requires more use of capturing short-term dependencies, which causes TG-LSTM performing better than DA-RNN. And long-term dependence needs to be used in forecasting MITV and Appliance Energy, as shown in Figure 6(b) and Figure 6(c), performances in DA-RNN and TG-LSTM are opposite, and DA-CG-LSTM is better than two baseline models under any circumstances.

With the increasement of time step, prediction performance becomes not so well. when the parameter time steps change from 15 to 30, the performance of DA-CG-LSTM model decreased by 15%, 17.1% and 16.4% respectively on Air Quality, MITV, Appliance Energy data set, is smaller than that of other comparison models. For example, the performance of DA-RNN decreased by 31.3%, 23.2% and 27.3% respectively. Therefore, DA-CG-LSTM has better robustness and wide applicability.

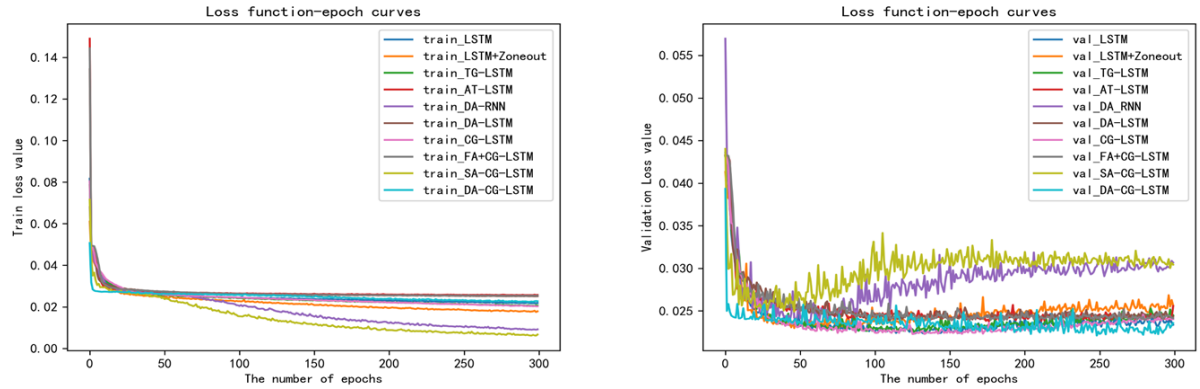
The training and validation set loss of each network in each iteration are shown in Figure 7(a), Figure 7(b) and Figure 7(c). It can be known that training and validation set loss of DA-CG-LSTM in each iteration are better than all baseline models on Air Quality. DA-CG-LSTM basically reaches stable stage after the number of iterations exceeding 25, better than other models in terms of convergence speed. According to the convergence of curve, it still has the potential to improve performance. DA-CG-LSTM does not show a large advantage in the training set loss on MITV. However, when the number of iterations exceeds 150 times, it shows an advantage compared to other. According to the convergence of curve, it still has the ability to continuously



(a) Loss of Comparison over Air Quality Dataset



(b) Loss of Comparison over MITV Dataset



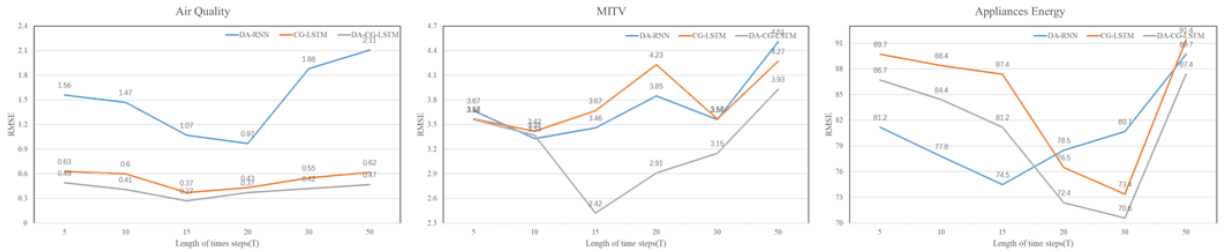
(c) Loss of Comparison over Appliances Energy Dataset

FIGURE 7. the Training-set Loss(left) and the Validation-set Loss(right).

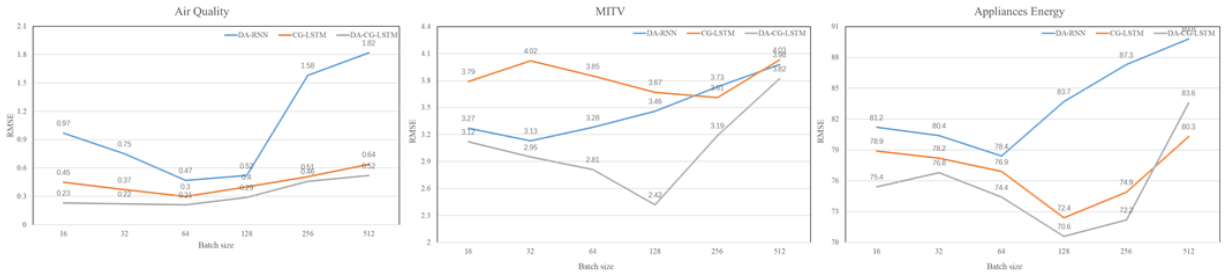
improve performance in subsequent iterations. From above analysis, it is can be known that DA-CG-LSTM has better generalization performance. It can be seen that DA-CG-LSTM does not outperform the other baseline models in terms of the training set on Appliance Energy. However, DA-CG-LSTM outperforms the other baseline models by a large margin for the loss in the test set. It is known that the model has good generalisation.

D. PARAMETER SENSITIVITY

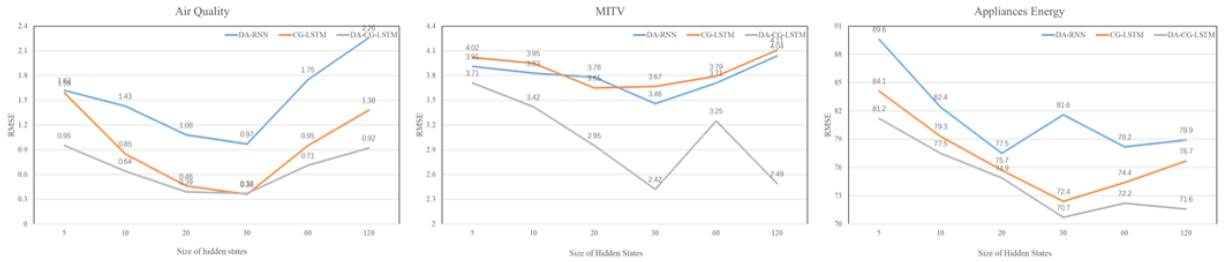
This paper explores sensitivity of networks in parameters, that is, the impact of selected parameters on the prediction performance of networks. Specifically, impacts of the changed parameters on the prediction performance can be obtained when only one parameter is changed and other parameters remain unchanged. When keeping in the hidden layer of CG-LSTM,$batch\_size = 128$, and changing size of time



(a) RMSE vs. Length of time steps T.



(b) RMSE vs. batch size.



(c) RMSE vs. Size of hidden states.

FIGURE 8. **RMSE vs. various parameters over Air Quality, MITV and Appliances Energy Dataset.**

step T, it is found that the effect of different time steps on networks is shown in Figure 8(a). By setting $T = 15$, $m = 30$, and changing *batch_size*, this paper gets the prediction effect of different batches, which is shown in Figure 8(b). After that, through changing the number m of network units, the result is shown in Figure 8(c). It can be known that with the changes of various parameters, the prediction effect of DA-CG-LSTM will produce concave changes, and optimal effect will be reached at $T = 15$, $m = 30$, *batch_size* $= 128$. The performance of DA-CG-LSTM changes little with the change of parameters, which is more robust than other models. Therefore, it can be better applied to multivariate sequence forecasting and widely applicated.

E. TIME COST ANALYSIS

When considering the performance of the model, the time cost of the model also needs to be considered [27], [28]. The model in this paper is built based on DA-RNN and CG-LSTM, optimizing dual-staged attention mechanism of DA-RNN, and the overall parameters are reduced by nearly 20% compared to DA-RNN. On Air Quality, MITV and Appliance Energy datasets, the number of parameters respectively required for DA-RNN are 39004, 37804 and 44524 and DA-CG-LSTM are 33209, 32249 and 38729 when $T = 15$, $m = 30$, *batch_size* $= 128$. Therefore, compared with DA-RNN, DA-CG-LSTM requires fewer parameters to achieve better prediction results.

To test the computational time consumption of DA-CG-LSTM. Experiments were conducted on Air Quality, MITV and Appliance Energy datasets. The experiments were run on a computer with an Intel Core i7-8700 3.20GHz CPU, 16GB of RAM and a GeForce GTX 1060-Ti 4G GPU. The numbers of training set were respectively 5979, 30840 and 12611, the numbers of validation set were respectively 1495, 7711 and 3177, and the numbers of validation set were respectively 1869, 9638 and 3947. The time cost for model training and testing are shown in Table 3. For example, on MITV data set, the training time and testing time of the DA-CG-LSTM are both less than DA-RNN, but both are more than CG-LSTM. This indicates that dual attention mechanism increases the running time and also improves

TABLE 3. Time cost comparison of different networks in three data sets($T = 15$).

Methods	Model Train(s)			Model Test(s)		
	Air Quality	MITV	Appliances Energy	Air Quality	MITV	Appliances Energy
TG-LSTM	1.47	4.13	2.04	0.05	0.12	0.07
DA-RNN	7.05	12.42	9.21	0.24	0.41	0.32
CG-LSTM	1.49	4.24	2.15	0.07	0.13	0.08
DA-CG-LSTM	7.01	11.76	8.93	0.21	0.37	0.27

the ability of the network to extract long-term dependence information.

F. ANALYSIS

DA-CG-LSTM, based on DA-RNN [17] and CG-LSTM, makes use of these advantages. CG-LSTM can better discard redundant and disperse information. Combined with attention mechanism, the ability to capture short-term mutation information can be enhanced. Temporal attention mechanism network of DA-RNN has better advantages in the extraction of hidden information, aiming to extract long-term dependence information. Through using double-layer attention to improve DA-RNN's input stage, only using single-layer attention to extract information of feature dimensions, it can effectively reduce impacts on different dimensions and time steps at target sequence. Therefore, in this section, different parameters was used to discuss and analyze DA-RNN, CG-LSTM, and DA-CG-LSTM respectively.

As time step changing, prediction performance evaluation of above three networks is shown in Figure 8(a). Also with changing of batch size and the number of hidden layer units, performances are shown in Figure 8(b) and Figure 8(c). It can be seen that DA-RNN and CG-LSTM have their own advantages. The MITV and Appliances Energy datasets require networks that are good at capturing long-term dependencies for training and prediction, while the Air Quality dataset requires networks that are good at capturing short-term mutation information. DA-RNN has better ability to deal with long-term dependence, while CG-LSTM has advantage in capturing short-term mutation information. Therefore, performance of DA-RNN is generally better than CG-LSTM in MITV and Appliance Energy, vice versa. DA-CG-LSTM uses advantages of these two models and can further improve prediction performance.

IV. CONCLUSION

Aiming at solving problems of capturing long-term dependence and short-term mutation information in same model, this paper proposes dual-staged attention mechanism based on CG-LSTM. Through introducing hyperbolic tangent function in forget-gate and input-gate of LSTM, the ability to capture short-term mutation information is improved. Combined attention mechanism with CG-LSTM, which adaptively capture short-term mutation information, using temporal attention to extract long-term dependence information. DA-CG-LSTM, based on dual-staged attention, can not only capture short-term mutation information, but also long-term dependence information. Experiments on MITV, Air Quality and Appliance Energy show that performance of DA-CG-LSTM proposed in this paper is better than some excellent time series prediction networks, and it, which can solve more multivariate time series analysis problems, has lower regression analysis errors and more accurate prediction capabilities, better robustness.

REFERENCES

[1] A. N. Sigurdarson and B. Hrafnkelsson, "Bayesian prediction of monthly precipitation on a fine grid using covariates based on a regional meteorological model," *Environmetrics*, vol. 27, no. 1, pp. 27–41, Feb. 2016.
[2] D. Durante, *Analysis of Italian Financial Market Via Bayesian Dynamic Covariance Models*. Springer, 2014, doi: 10.1007/978-3-319-02084-6_33.
[3] Y. Liu and R. Wang, "Study on network traffic forecast model of SVR optimized by GAFSA," *Chaos, Solitons Fractals*, vol. 89, pp. 153–159, Aug. 2016.
[4] P. M. Clements and E. G. Mizon, "Empirical analysis of macroeconomic time series: VAR and structural models," *Eur. Econ. Rev.*, vol. 35, no. 4, pp. 887–917, 1991.
[5] S.-J. Huang and K.-R. Shih, "Short-term load forecasting via ARMA model identification including non-Gaussian process considerations," *IEEE Trans. Power Syst.*, vol. 18, no. 2, pp. 673–679, May 2003.
[6] H. Akaike, "Autoregressive model fitting for control," *Ann. Inst. Stat. Math.*, vol. 23, no. 1, pp. 163–180, Dec. 1971.
[7] B. D. Nearing and R. L. Verrier, "Modified moving average analysis of T-wave alternans to predict ventricular fibrillation with high accuracy," *J. Appl. Physiol.*, vol. 92, no. 2, pp. 541–549, Feb. 2002.
[8] N. Gagey, E. Ravaud, and J. Lassau, "Anatomy of the acromial arch: Correlation of anatomy and magnetic resonance imaging," *Surgical Radiol. Anatomy*, vol. 15, no. 1, pp. 63–70, Mar. 1993.
[9] M. van Heel, "A new family of powerful multivariate statistical sequence analysis techniques," *J. Mol. Biol.*, vol. 220, no. 4, pp. 877–887, Aug. 1991.
[10] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[11] J. Chung, C. Gulcehre, and K. Cho, "Empirical evaluation of gated recurrent neural networks on sequence modeling," in *Proc. NIPS Workshop Deep Learn.*, 2014, vol. 1, no. 2, pp. 541–549.
[12] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Trans. Neural Netw.*, vol. 5, no. 2, pp. 157–166, Mar. 1994.
[13] Y. Goldberg, "Neural network methods for natural language processing," *Synth. Lect. Hum. Lang. Technol.*, vol. 10, no. 1, pp. 301–309, Apr. 2017.
[14] L. Lu, X. Zhang, and S. Renais, "On training the recurrent neural network encoder-decoder for large vocabulary end-to-end speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 5060–5064.
[15] Y. Wang, M. Huang, X. Zhu, and L. Zhao, "Attention-based LSTM for aspect-level sentiment classification," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 606–615.

[16] Y. Ding, Y. Zhu, J. Feng, P. Zhang, and Z. Cheng, "Interpretable spatio-temporal attention LSTM model for flood forecasting," *Neurocomputing*, vol. 403, pp. 348–359, Aug. 2020.
[17] Y. Qin, D. Song, H. Chen, W. Cheng, G. Jiang, and G. Cottrell, "A dual-stage attention-based recurrent neural network for time series prediction," 2017, *arXiv:1704.02971*.
[18] Y. Liang, S. Ke, and J. Zhang, "GeoMAN: Multi-level attention networks for geo-sensory time series prediction," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, 2018, pp. 3428–3434.
[19] Y. Liu, C. Gong, and L. Yang, "DSTP-RNN: A dual-stage two-phase attention-based recurrent neural network for long-term and multivariate time series prediction," *Expert Syst. Appl.*, vol. 143, Apr. 2020, Art. no. 113082.
[20] Y. Srivastava, V. Murali, and S. R. Dubey, "PSNet: Parametric sigmoid norm based CNN for face recognition," in *Proc. IEEE Conf. Inf. Commun. Technol.*, Dec. 2019, pp. 1–4.
[21] P. Swietojanski, J. Li, and T. J. Huang, "Investigation of maxout networks for speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 7649–7653.
[22] J. Hu and W. Zheng, "Transformation-gated LSTM: Efficient capture of short-term mutation dependencies for multivariate time series prediction tasks," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, 2019, pp. 1–8.
[23] S. De Vito, E. Massera, M. Piga, L. Martinotto, and G. D. Francia, "On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario," *Sens. Actuators B, Chem.*, vol. 129, no. 2, pp. 750–757, 2008.
[24] D. Krueger, T. Maharaj, J. Kramár, M. Pezeshki, N. Ballas, and R. N. Ke, "Zoneout: Regularizing RNNs by randomly preserving hidden activations," in *Proc. ICLR*, 2017, pp. 1–11, doi: 10.13140/RG.2.1.1027.2889.
[25] R. Häbner, M. Steinhauser, and C. Lehle, "A dual-stage two-phase model of selective attention," *Psychol. Rev.*, vol. 117, no. 3, pp. 759–784, 2008.
[26] L. M. Candanedo, V. Feldheim, and D. Deramaix, "Data driven prediction models of energy use of appliances in a low-energy house," *Energy Buildings*, vol. 140, pp. 81–97, Apr. 2017.
[27] J. Hu and W. Zheng, "A deep learning model to effectively capture mutation information in multivariate time series prediction," *Knowl.-Based Syst.*, vol. 203, Sep. 2020, Art. no. 106139, doi: 10.1016/j.knosys.2020.106139.
[28] Q. Ma, Z. Zheng, and W. Zheng, "Echo memory-augmented network for time series classification," *Neural Netw.*, vol. 133, pp. 177–192, Dec. 2021.
[29] *Traffic Data From MN Department of Transportation*. [Online]. Available: https://www.dot.state.mn.us/

SHUFANG FENG received the B.S. degree in electrical engineering and automation from the Changzhou Institute of Technology, China, in 2019. He is currently pursuing the master's degree in control science and engineering with the School of Information Science and Engineering, East China University of Science and Technology, Shanghai, China. His research interests include chemical process modeling, time series modeling, and fault diagnosis.

YONG FENG received the B.S. degree in automation from the East China University of Science and Technology, Shanghai, China, in 2019, where he is currently pursuing the master's degree in control science and engineering with the School of Information Science and Engineering. His research interests include time series forecasting and chemical process modeling.