Novel Transformer Based on Gated Convolutional Neural Network for Dynamic Soft Sensor Modeling of Industrial Processes

Zhiqiang Geng, Zhiwei Chen, Qingchao Meng, and Yongming Han

***Abstract*—Industrial process data are usually time-series data collected by sensors, which have the characteristics of high nonlinearity, dynamics, and noises. Many existing soft sensor modeling methods usually focus on dominant variables and auxiliary variables at a single time point while ignoring the timing characteristics of industrial process data. Meanwhile, the soft-sensing methods considering timing characteristics based on the deep learning are usually faced with gradient vanishing and the difficulty in parallel computing. Therefore, a novel Gated Convolutional neural network-based Transformer (GCT) is proposed for dynamic soft sensor modeling of industrial processes. The GCT encodes short-term patterns of the time series data and filters important features adaptively through an improved gated convolutional neural network (CNN). Then, the multihead attention mechanism is applied to modeling the correlation between any two moments. Finally, the prediction results are obtained through a linear neural network layer with the highway connection. In this article, the experiments in the dynamic soft sensor modeling of polypropylene and purified terephthalic acid industrial processes show that the proposed method achieves state-of-the-art comparing with the back propagation neural network, the extreme learning machine, the long short-term memory (LSTM) and the LSTM based on the CNN.**

***Index Terms*—Deep learning, dynamic soft sensor, gated convolutional neural network (CNN), polypropylene, purified terephthalic acid (PTA), transformer.**

I. Introduction

With the rapid development of data-measuring technologies and the wide use of the distributed control systems (DCS), recording and collecting data from industrial processes have become more and more easy [1]. As a result, a series of soft sensor modeling methods have been developed, which are playing an increasingly important role in improving the production efficiency and the product quality, and providing guidance for the optimal control of the reaction process [2].

Generally, soft sensor modeling methods are mainly divided into two categories with first-principle models and data-driven models. The first-principle models are usually established by modeling energy balance, force balance, and physical and chemical reaction mechanism in the reaction process, relying on the mastery of the domain knowledge [3]. However, due to the complicated reaction process, it is difficult to guarantee system stability at any time, which limits the quality control in actual industrial processes [4]. In contrast, data-driven modeling methods build predictive models based on easily measurable auxiliary variables (such as temperature, pressure, and flow) [5]. Because of the rapid response and flexibility of data-driven models, they have been widely used for online monitoring of key properties of various industrial products.

Traditional data-driven soft sensor modeling methods are mainly based on artificial neural networks (ANN) [6], principal component analysis (PCA) [7], and partial least squares (PLS) [8]. Shang *et al.* [9] have proved that the ANN is competent to establish soft sensors, and a large number of ANNs are the feed-forward networks, such as multilayer perceptron [5] and radial-basis function network [10]. However, the real industrial process data are much more complex that was characterized by multimode, high-dimensional, corrupted and unbalanced data, which cause great difficulties to the ANNs [11], [12]. As a feature learning method, the PCA converts the original input into a linear combination of features, which can effectively reduce the complexity of soft sensor model and solve the high-dimensionality and collinearity problems in the industrial process data to a certain extent [13]. Compared with the PCA, the feature learning process of the PLS is subject to additional constraints of the target space, so it can effectively extract important information related to the target variable. Based on the improved PLS framework, Xie *et al.* [8] handle the outliers of the process data, and realizes prediction and fault diagnosis related to key quality indicators. However, ANNs are still facing the problems of learning slowly and easily falling into local optimum. To accelerate the learning procedure, Huang *et al.* [14] designed an extreme learning machine (ELM) with only a single hidden layer to simplify the training and generalization process

Manuscript received November 30, 2020; revised May 5, 2021; accepted June 1, 2021. Date of publication June 7, 2021; date of current version December 6, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 21978013, in part by the Science and Technology Major Project of Guizhou Province (Guizhou Branch [2018]3002), and in part by the Fundamental Research Funds for the Central Universities under Grant XK1802-4. Paper no. TII-20-5370. *(Corresponding author: Yongming Han.)*

The authors are with the College of Information Science and Technology, Beijing University of Chemical Technology, Beijing 100029, China, and also with the Engineering Research Center of Intelligent PSE, Ministry of Education in China, Beijing 100029, China (e-mail: zhiqianggeng@mail.buct.edu.cn; 2020400187@mail.buct.edu.cn; Jochen_M@163.com; hanym@mail.buct.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TII.2021.3086798.

Digital Object Identifier 10.1109/TII.2021.3086798

of traditional ANNs and achieved a breakthrough in the global optimization problem. He *et al.* [15] proposed an improved ELM combined with the PCA to establish a soft-sensing model for the acetic acid content at the top of the purified terephthalic acid (PTA) solvent dehydration tower. Han *et al.* [16] developed a resource optimization model using ELM with t-distributed stochastic neighbor embedding for complex industrial processes modeling. Due to the delay of reaction control and the retention of various materials and products in the device, there is a certain degree of dependence and comparability among the historical data at different times [17]. However, limited by the structure of the networks, these aforementioned methods usually only consider current status and do not fully explore the potential features in the historical data.

In recent years, with the improvement of computing ability, many deep learning methods have been developed for the soft sensor modeling [18]. Moreover, deep neural networks established as latent variable models have stronger capabilities in learning and representation over traditional methods, which can help to describe highly correlated process variables [19]. To capture quality-relevant features and predict key indicators, Yuan *et al.* [13] proposed a stacked quality-driven autoencoder. Han *et al.* [20] used the long short-term memory neural network (LSTM) to evaluate whether the PTA process is optimized. A spatiotemporal attention-based LSTM (STA-LSTM) is introduced to identify both variables and temporal importance in samples [21]. However, the forward and backward calculations of the recurrence neural network (RNN) are both sequential, which bring the problems of gradient vanishing and gradient explosion [22]. Although the LSTM solves the problem of short-term gradient vanishing to a certain extent, it is still powerless to deal with long-term dependence and gradient explosion. Meanwhile, the sequential calculation manner of the RNN limits the parallel computing ability of the model. Convolutional neural network (CNN) is another popular feature extractor. And the core operation of the CNN is the convolution calculation, which is usually utilized to capture local features effectively. Because there is no recursive relationship, the CNN models can be trained in parallel, and the training speed is usually several orders of magnitude faster than the RNN. Yuan *et al.* [23] designed a dynamic CNN (DCNN) strategy to learn hierarchical local nonlinear dynamic features for soft sensor modeling, whose effectiveness is verified on an industrial hydrocracking process. In order to take advantage of the correlations between process variables, Wang *et al.* [24] developed two CNN-based soft sensor models to utilize abundant process data and integrate finite impulse response, respectively.

In 2017, Vaswani *et al.* [25] proposed a transformer model based on the attention mechanism, and has achieved significant improvements in natural language parsing [26]. Transformer reduces the distance between any two positions in the sequence to a constant to easily capture the relationship between any two moments, and introduces residual connections between layers, both of which help to solve the problem of gradient vanishing of the RNN. Moreover, because the attention mechanism is not sequential structure, Transformer has good parallelism and can be accelerated on GPU, which is of great significance for the dynamic soft sensor modeling of complex industrial processes with high real-time requirements.

In this article, a novel Gated CNN-based Transformer (GCT) model is proposed to build the dynamic soft sensing model of complex industrial processes. The Gated CNN Unit (GCU) can extract short-term patterns and local dependencies among variables while adaptively filtering the hidden features. Then, the multihead attention mechanism is utilized to capture multilevel long-term dependencies. Finally, the predictive results of the GCT are obtained through the fully connected linear layer with the highway connection.

The main contributions of this article are shown as follows.

1) An improved gated CNN is used to capture short-term patterns and local dependencies among variables while adaptively selecting important features in the data.
2) The multihead attention mechanism is used to find the dependence between any two moments in the time series data, which effectively solves the problem of gradient vanishing and the difficulty in parallelization of the RNN.
3) A dynamic soft-sensing model is proposed to make full use of the historical information and the current observable state to predict key quality indicators at present or at a certain time in the future.
4) The experimental results on the actual industrial production process data of polypropylene and PTA show that the proposed method achieved state-of-the-art performance compared with the BP, the ELM, the LSTM, and the LSTM based on the CNN (CNN+LSTM) in terms of the root relative squared error (RSE), the mean absolute percentage error (MAPE), and the empirical correlation coefficient (CORR) metrics.

II. Problem Formulation

The dynamic soft sensor modeling problem based on time series data can be defined by giving a series of industrial process data as follows:

$$\{X_{t-w+1}, \ldots, X_t; Y_{t-w+1}, \ldots, Y_{t-1}\} \tag{1}$$

where $X_t \in R^m$ represents the auxiliary variables observed by the DCS in the real time, m is the dimension of auxiliary variables, $Y_t \in R$ is the dominant variable to be predicted, and w is the window size of the historical observation data. The dominant variable Y_{t+h} at a certain time in the future is predicted, where h is the desirable horizon ahead the current time step. In other words, to predict Y_{t+h}, the variables in (1) are available. When $h = 0$, it means that we predict the dominant variable at the current time step.

In order to make the data meet the input specifications of the time series models, we need to regularize the data into the required format. Specifically, the acquired data samples are organized into the format

$$\text{samples} = \begin{bmatrix} X_{t-w+1} & X_{t-w+2} & \cdots & X_t \\ Y_{t-w+1} & Y_{t-w+2} & \cdots & Y_t \end{bmatrix}; Y_{t+h} \tag{2}$$

where the left part is input, and the right part is target value. Note that, for time t, since the value of the dominant variable

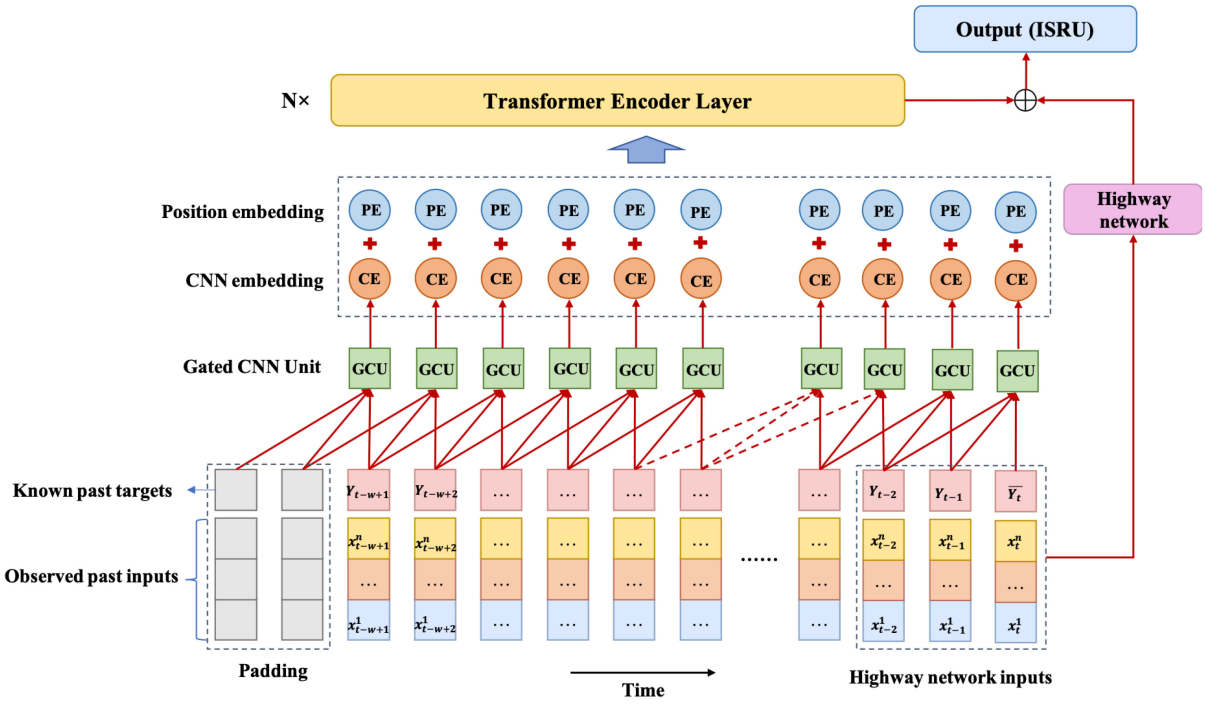

Fig. 1. Overall architecture of the GCT (GCU: Gated CNN Unit; PE: Position Embedding; CE: CNN Embedding; ISRU: Inverse Square Root Unit).

Y_t is unknown, it is set as the normalized mean of the dominant variables of the training data.

III. Proposed Methodology

The overall architecture of the proposed model is shown in Fig. 1. First, the gated CNN layer encodes the input data. The gated signal and the feature signal are generated by two types of convolution kernels, which are then multiplied positionwise to obtain the final CNN embedding. Then, the CNN embedding and the position embedding are combined as the input of the transformer layers. The dependence between any two moments in the time series data is captured through the multihead attention mechanism. Finally, through the highway network and the fully connected linear layer, final prediction results are obtained.

A. Transformer

The transformer adopts the encoder–decoder architecture, and its encoder and decoder are both stacked by multiple independent feature extractors [25]. In the proposed model, only the transformer encoder is used, and the core of the encoder is the multihead attention mechanism, as illustrated in Fig. 2. Similar to the CNN, multiple heads can learn relevant information in different representation subspaces.

In the self-attention layer with multiple attention heads, $query$, key, and $value$ are mapped to multiple heads through linear transformation, respectively. The proposed model performs self-attention calculations in each projection part in parallel to obtain the correlation between any two moments. And then, all of attention values are concatenated together

$$\text{MultiHead}(Q,K,V) = \text{Concat}\left(\text{head}_1,\ldots,\text{head}_h\right)W^O \quad (3)$$

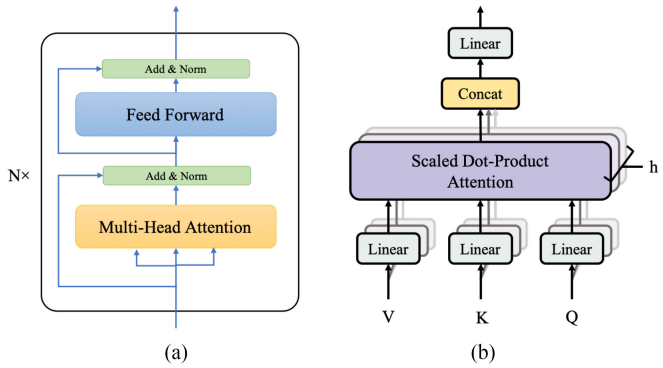

Fig. 2. (a) Transformer encoder structure. (b) Multihead attention mechanism.

$$\text{head}_i = \text{SelfAttention}\left(QW_i^Q, KW_i^K, VW_i^V\right) \quad (4)$$

$$\text{SelfAttention}\left(Q_h,K_h,V_h\right) = \text{softmax}\left(\frac{Q_h K_h^T}{\sqrt{d_k}}\right)V_h \quad (5)$$

where $W^O \in R^{d_{\text{model}}*d_{\text{model}}}$ is the parameter matrix of output projection, $W_i^Q \in R^{d_{\text{model}}*d_k}$, $W_i^K \in R^{d_{\text{model}}*d_k}$ and $W_i^V \in R^{d_{\text{model}}*d_v}$ are trainable projection parameter matrices of $query$, key, and $value$, respectively.

Since transformer contains no recurrence, it will lose timing information. Therefore, the position embedding is introduced to make full use of the order of the time series data [26].

B. Gated CNN Unit (GCU)

Industrial process data have the characteristics of noise and dynamics. Inspired by Oord *et al.* [27], a CNN based on a gating

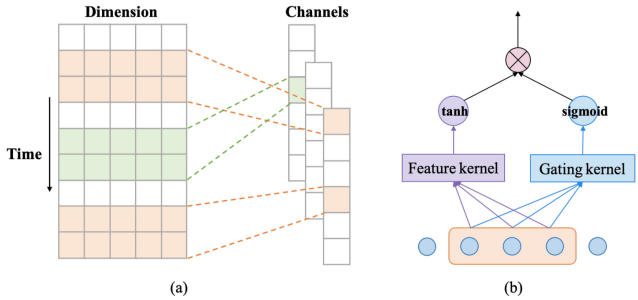

Fig. 3. (a) 1-D CNN along the time dimension. (b) GCU.

mechanism is introduced to smooth the data and capture the short-term patterns of time series data and the local dependence between variables.

The gated CNN layer in the proposed model consists of a set of GCUs. The architecture of the GCU is illustrated in Fig. 3(b). Each GCU consists of two convolution kernels with the same size, but the parameter weights are not shared. One of the convolution kernels is used as a traditional feature extractor, and the convolution result is nonlinearly transformed by the tanh activation function. The convolutional result of another convolution kernel is used as the input of the gating mechanism, and the activation value between 0–1 is obtained through the sigmoid activation function. By multiplying the two activation values, the final convolution result is obtained

$$H_g = g\left(W_g \cdot X + b_g\right) \tag{6}$$

$$H_f = f\left(W_f \cdot X + b_f\right) \tag{7}$$

$$H_c = \text{sigmoid}\left(H_g\right) \cdot \tanh\left(H_f\right), c = 1, 2, \ldots, C \tag{8}$$

where H_g and H_f represent the output of the gating kernel and the feature extractor kernel, respectively. And W_g, W_f, b_g, and b_f are corresponding parameters to be trained. H_c is the final output of each GCU, and C is the number of GCUs in the first convolutional layer.

Unlike standard CNN kernel sliding in both the height and width directions, the width of each kernel of the proposed method is the same as the dimension of inputs, which ensures that the convolution kernels can process complete information of a few time steps at a time, as shown in Fig. 3(a). By sliding the convolution kernel sequentially along the time dimension, the hidden embedding is obtained as follows:

$$H_{GCU} = \begin{bmatrix} H_{1,1} & \cdots & H_{t,1} \\ \vdots & \ddots & \vdots \\ H_{1,C} & \cdots & H_{w,C} \end{bmatrix} \tag{9}$$

whereas w and C are the window size of input historical data and the number of GCU in the first convolutional layer, respectively.

C. Highway Network and Output Layer

Intuitively, recent state of production devices are supposed to have more important reference value for dynamic soft sensor modeling. Therefore, a highway network is introduced [28] to make the GCT pay more attention to the most recent time steps. The essence of the highway network is to accelerate the flow of information, so that when the network is deepened, good performance can also be ensured. Since the self-attention mechanism is able to capture the relationship between any two moments, the output hidden state at the last time step of the transformer is kept and combined with the output of the highway network to obtain the prediction result through the fully connected linear layer.

In this article, the inverse square root activation function (ISRU) [29] is used as the activation function in the output layer. The ISRU has the advantages of symmetry about the origin and self-adapting activation value. Compared with the sigmoid, the ISRU has a larger derivative, which can avoid the problem of gradient vanishing in deep neural networks. The mathematical form of the ISRU is shown as follows:

$$\text{ISRU}(x) = \frac{x}{\sqrt{1 + \alpha \cdot x^2}} \in \left(-\frac{1}{\sqrt{\alpha}}, \frac{1}{\sqrt{\alpha}}\right). \tag{10}$$

The value of α is automatically determined according to the normalized target value of the training set. Specifically, $\alpha = 1/y_{\max}^2$, which keeps the predicted value within the effective range.

The final soft sensing result can be expressed as

$$\hat{Y} = \text{ISRU}\left(\text{fc2}\left(\text{fc1}\left(H_t\right) + \text{highway}\left(H_h\right)\right)\right) \tag{11}$$

where fc1 and fc2 are two fully connected linear layers, H_t is the hidden state of transformer at the last time step, and H_h represents the input data at the most recent h time steps.

D. Objective Function and Optimization Strategy

In soft sensor modeling field, the loss function with the mean square error (MSE) is usually used as the objective function. In this article, an L2 regularization term is added to the objective function to avoid overfitting. Therefore, the final form of the objective function can be formalized as

$$J = \sum \left\| Y_i - \widehat{Y}_i \right\|^2 + 1/2 \|\Theta\|_2^2, i = 1, 2, \ldots, n \tag{12}$$

where Θ is the parameter set to be learned, $\|\cdot\|^2$ is the Frobenius norm.

The stochastic gradient descent (SGD) algorithm with a momentum term is utilized to optimize the parameters of the model. The momentum term can help the model accelerate the convergence and avoid falling into local optimum.

E. Metrics

Three commonly used performance evaluation metrics are used to judge the pros and cons of each model, including the RSE, the MAPE, and the CORR

$$\text{RSE} = \frac{\sqrt{\sum_{i=1}^{N} \left(y_i - \hat{y}_i\right)^2}}{\sqrt{\sum_{i=1}^{N} \left(y_i - \bar{y}_i\right)^2}} \tag{13}$$

$$\text{MAPE} = \frac{1}{N} \sum_{i=1}^{N} \frac{|y_i - \hat{y}_i|}{|y_i|} \tag{14}$$

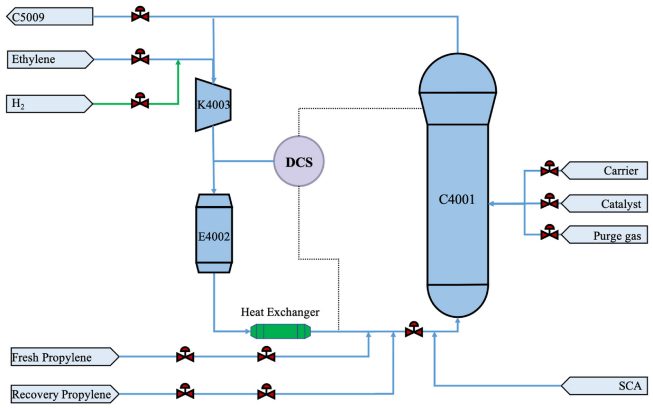

Fig. 4. Schematic of the polypropylene production process.

TABLE I
INPUT VARIABLES OF THE MI SOFT SENSOR MODELING

Input variables	Variable description
v_1	temperature of reactor bed
v_2	pressure of reactor
v_3	density in the top of the reactor bed
v_4	density in the middle of the reactor bed
v_5	density in the bottom of the reactor bed
v_6	reactor material level
v_7	ratio of hydrogen to propylene
v_8	reactor catalyst carrier feed flow
v_9	catalyst feed flow
v_{10}	T2 propylene carrier flow
v_{11}	electrons flow
v_{12}	total propylene feed
v_{13}	pressure of propylene feed
v_{14}	propylene flow into the vaporizer

$$\mathrm{CORR} = \frac{1}{N}\sum_{i=1}^{N}\frac{(y_i-\bar{y}_i)\left(\hat{y}_i-\bar{\hat{y}}_i\right)}{\sqrt{(y_i-\bar{y}_i)^2\left(\hat{y}_i-\bar{y}_i\right)^2}} \tag{15}$$

where y_i is the target value, $\hat{y}_i$ is the predict value of the model, and N is the number of samples in the testing set.

The RSE and the MAPE represent the error between the predict value and the ground truth, respectively. The CORR represents the linear correlation between the predict value and the ground truth. For the RSE and the MAPE, lower value is better, whereas for the CORR, higher value is better.

IV. CASE STUDIES

In order to verify the effectiveness of the proposed method, relevant experiments on actual production process data of polypropylene and PTA are conducted by comparing the soft sensor performance of the BP, the ELM, the LSTM, the CNN+LSTM, and the GCT model. All experiments are performed on a GPU server with OS (CentOS v7.0, 64 b), CPU (Intel(R) Xeon(R) Silver 4210@2.20 GHz) and GPU (TITAN XP, 12 GB).

A. Polypropylene Melt Index Soft Sensor Modeling

The production process of polypropylene can be roughly categorized into three types: 1) gas phase process, 2) slurry process, and 3) bulk process [30]. The modeling object studied in the article is a gas-phase fluidization reaction process, which has been put into production by a domestic petrochemical enterprise, as illustrated in Fig. 4.

The core of the polypropylene polymerization reaction system is a reaction loop composed of a gas-phase fluidized bed reactor (C4001), a circulating gas cooler (E4002), and a circulating gas compressor (K4003). The liquid phase propylene, the T2 catalyst, and the electron donor enter the reactor continuously through the circulating gas loop. The concentration of H_2, ethylene, and propylene in the circulating gas is monitored by the online analyzer, and the raw material flow is automatically controlled to ensure the composition of the circulating gas. The polymer powder produced by the polymerization reaction is discharged alternately and intermittently through product discharge system, and the discharged powder is directly sent to the product degassing chamber (C5009).

Polypropylene melt index (MI, g/10 min) [31] refers to the mass of the melt flowing through the standard aperture (diameter (2.0950 mm±0.0005 mm), length(8.00 mm±0.02 mm)) within 10 min under a certain temperature and pressure, which is an important indicator of the flow characteristics of the plastic melt. In the production process, the grades of polypropylene are mainly divided by the MI, and the MI value can reflect the molecular weight of polypropylene.

To establish a soft-sensing model for the MI based on the polypropylene production process data, some easily measurable auxiliary variables are selected as the input of the model as shown in Table I, and the output variable is MI.

The DCS data of a domestic petrochemical company from 2020-03 to 2020-06 with a time interval of 2 h is obtained, whereas the corresponding MI data are obtained through manual sampling and offline laboratory analysis. During these four months, the production is a continuous process and there is only one production batch. Finally, a total of 1152 samples are retained after preprocessing and regularization, and the dataset is separated into training, validation, and testing sets according to the ratio of 6:2:2.

In order to verify the performance of the ELM and the BP, samples of single time step and multitime steps are both used as the input of them, and they are expressed as ELM(m) and BP(m), respectively. For the ELM, ELM(m), BP, and BP(m), the number of hidden units is determined in the range of {16, 24, 32, 64, 128} with cross-validation on the validation set, and finally 128, 64, 64, and 64 are selected.

For the LSTM, the CNN+LSTM, and the GCT, the number of LSTM hidden units is set to 128, the number of LSTM layers to 4, the CNN convolution kernel size to 3 and the number of convolution kernels or GCUs to 32. For the GCT, the number of transformer encoder layers is set to 4, the embedding dimension to 128, and the number of heads of multihead attention to 8.

To determine the best window size, pre-experiments are conducted. Experiment results are shown in Fig. 5, which indicate that the GCT model performs best when $w = 12$. For all models, corresponding experiments with horizon in {0, 3, 6} are

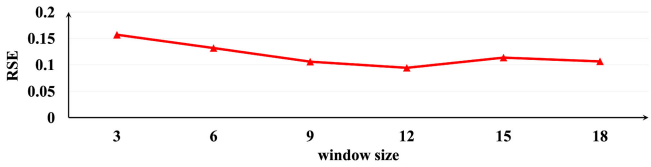

Fig. 5. RSE of GCT model with different window sizes on MI dataset.

TABLE II
PREDICTION RESULTS OF THE MI SOFT SENSOR MODELING

metrics	h	ELM	ELM(m)	BP	BP(m)	LSTM	CNN+LSTM	GCT
RSE	0	0.3169	0.2123	0.1449	0.1350	0.1380	0.1132	**0.0943**
	3	0.3180	0.2374	0.1908	0.1683	0.1467	**0.1174**	0.1184
	6	0.3334	0.2329	0.2100	0.2059	0.1735	**0.1184**	0.1198
MAPE	0	0.0329	0.0640	0.0153	0.0150	0.0091	0.0076	**0.0060**
	3	0.0475	0.0744	0.0181	0.0185	0.0114	0.0073	**0.0062**
	6	0.0246	0.0736	0.0199	0.0194	0.0105	0.0073	**0.0063**
CORR	0	0.3818	0.3458	0.7365	0.7893	0.7903	0.8395	**0.8477**
	3	0.2520	0.1175	0.6154	0.6464	0.7498	0.8216	**0.8380**
	6	0.1551	0.1732	0.6094	0.5248	0.7239	**0.8210**	0.8177

These bold entities mean the optimal performance of all models on the corresponding index.

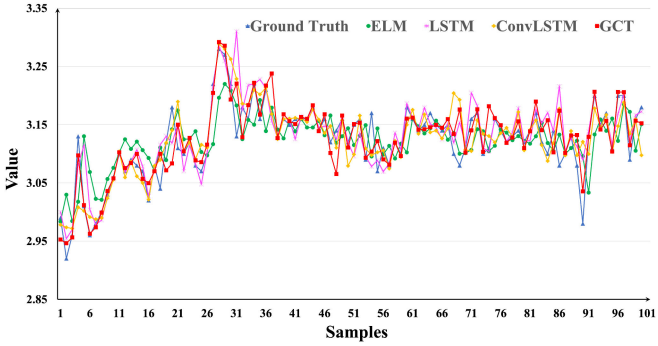

Fig. 6. Predict curves of the ELM, the LSTM, the CNN+LSTM, and the GCT.

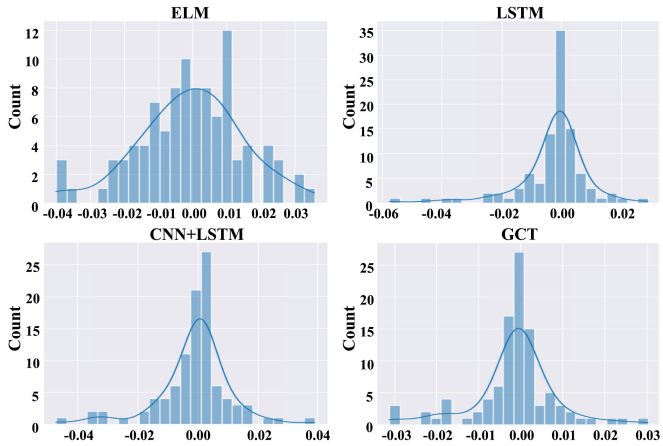

Fig. 7. Histogram of the predict errors of the ELM, the LSTM, the CNN+LSTM, and the GCT.

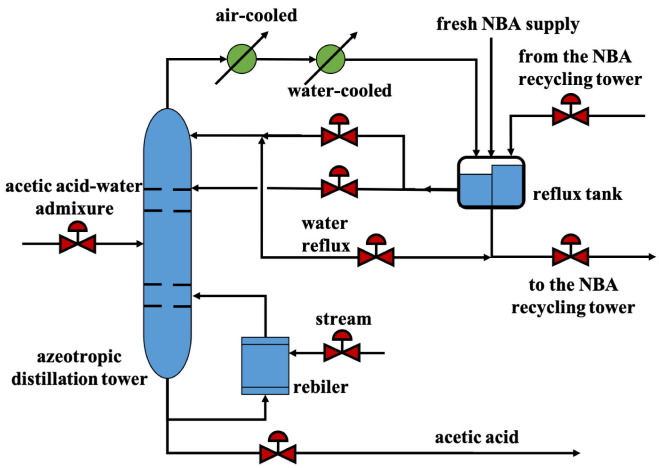

Fig. 8. Schematic of the PTA production process.

conducted. The SGD algorithm with learning rate of 0.001, momentum term factor of 0.5 is used for training.

The prediction results on the testing set of polypropylene MI are shown in Table II and Figs. 6 and 7. The table shows that the GCT has the best fitting capability with an increase of the CORR by up to 20% and improvements of the RSE and the MAPE by at least 16.70% and 0.16%, respectively. And the performance of traditional methods (such as ELM and BP) is terrible. Meanwhile, it is still difficult for the ELM(m) and the BP(m) to mine the effective information even if multistep time series data is given, so they cannot obtain a significant performance improvement. In contrast, deep neural networks, such as the LSTM, the CNN+LSTM, and the GCT, which considering the influence of time factors on predictive indicators, are more capable of capturing time patterns in the sequence and more stable as shown in Fig. 6. Compared with the LSTM series models, the GCT is more flexible to the instantaneous mutation of the variable, so it is able to get a better fitting performance, as shown at time step 6 and time step 91 in Fig. 6. We think it is the credit of the attention mechanism, which not only use current state, but also fully mine information as similar moments in historical data. As can be seen from the table and Fig. 7, the GCT has the smallest range of prediction errors and the errors obey an almost unbiased normal distribution with a sharp peak near 0, which reflects the best prediction performance.

B. PTA Acetic Acid Content Soft Sensor Modeling

PTA solvent system is the key part of the PTA production process, which is composed of three parts: 1) the solvent dehydration tower, 2) the reflux tank, and 3) the reboiler as shown in Fig. 8. The production of PTA takes xylene as the raw material and acetic acid as the solvent. Under the action of the Co–Mn–Br catalyst, by controlling the appropriate temperature and pressure, it reacts with oxygen in the air to produce crude terephthalic acid and then PTA can be obtained by refining it. Acetic acid, as a heat-transfer medium, can be consumed during the process of oxidation, wastewater carrying in the top of the recovery tower, removing the impurities in the membrane evaporator, and recovering acetic acid. The consumption of acetic acid is a key indicator of the PTA solvent system efficiency [32]. However, in the PTA production process, there is no direct method to measure the acetic acid content in the solvent system. Therefore,

TABLE III
INPUT VARIABLES OF THE PTA SOFT SENSOR MODELING

Input variables	Variable description
v_1	water reflux
v_2	steam flow
v_3	feed quantity
v_4	NBA side reflux
v_5	feed temperature
v_6	NBA main reflux
v_7	reflux tank level
v_8	feed composition
v_9	reflux temperaturew
v_{10}	temperature of top tower
v_{11}	product quantity of top tower
v_{12}	temperature point above the 35th tray
v_{13}	tray temperature near the up sensitive plate
v_{14}	tray temperature near the low sensitive plate
v_{15}	temperature point between the 35th tray and 40th tray
v_{16}	temperature point between the 44th tray and 50th tray
v_{17}	temperature point between the 53th tray and 58th tray

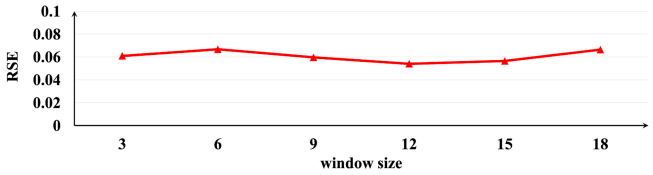

Fig. 9. RSE of GCT model with different window sizes on PTA dataset.

an accurate dynamic soft-sensing model can be established to indirectly analyze the acetic acid content in the device.

The conductivity of the solution at the top of the solvent dehydration tower corresponds to the content of acetic acid, and there are total 17 factors that affect the conductivity of the solution, which are listed in Table III. Therefore, these 17 factors are chosen as the input variables and the conductivity of the tower top as the target variable of the dynamic soft sensor model. In this experiment, a total of 260 samples of PTA production process arranged in chronological order are obtained. After imputation and standardization, the dataset is constructed according to the specific needs of the models.

This experiment maintains almost the same experimental setting in the polypropylene MI soft sensor modeling experiment, including that the window size is set to 12, as illustrated in Fig. 9. The difference is that for the ELM, ELM(m), BP, and BP(m), and the number of hidden units is determined to 32, 128, 32, and 64. For the LSTM, the CNN+LSTM, and the GCT, the number of the LSTM hidden units is set to 64.

The prediction results on the testing set are shown in Table IV and Figs. 10 and 11. It can be seen from Table IV that the GCT has achieved the best or approximately the best prediction performance on various experimental settings. Compared with other models, the GCT has achieved optimal results in terms of the RSE, the MAPE, and the CORR, where the GCT obtains an improvements of 10.45%, 0.03%, and 0.69%, respectively. Fig. 10 shows that the GCT has the best fitting ability, and its scatter data points are closely distributed near the line $y = x$ with the correlation coefficient of 0.9791. As can be seen from Fig. 11, since the ELM only focus on the current state signal, it is very sensitive to outliers, resulting in the performance of the

TABLE IV
PREDICTION RESULTS OF THE PTA SOFT SENSOR MODELING

metrics	h	ELM	ELM(m)	BP	BP(m)	LSTM	CNN+LSTM	GCT
RSE	0	0.1696	0.2727	0.1706	0.1610	0.0657	0.0603	**0.0540**
	3	0.1728	0.2768	0.1912	0.2316	0.0864	**0.0823**	0.0846
	6	0.2482	0.2897	0.2005	0.2593	0.1012	0.1114	**0.1006**
MAPE	0	0.0074	0.0543	0.0068	0.0062	0.0023	0.0021	**0.0020**
	3	0.0066	0.0549	0.0069	0.0083	0.0031	0.0026	**0.0025**
	6	0.0107	0.0723	0.0085	0.0092	0.0036	0.0036	**0.0033**
CORR	0	0.8082	0.3726	0.7520	0.7989	0.9720	0.9722	**0.9791**
	3	0.7958	0.4548	0.6224	0.6735	0.9446	**0.9537**	0.9516
	6	0.5649	0.2620	0.6277	0.6246	0.9246	0.9033	**0.9256**

These bold entities mean the optimal performance of all models on the corresponding index.

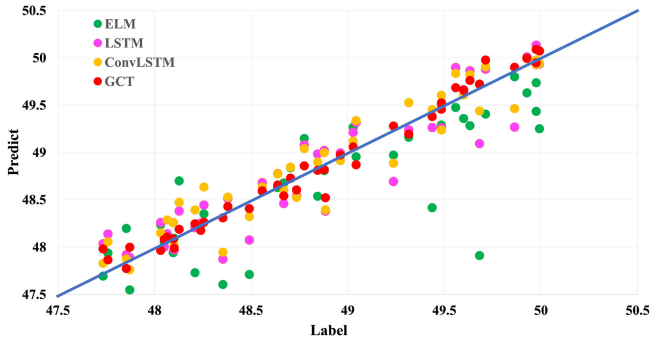

Fig. 10. Scatter plot of predict values and ground truth of the ELM, the LSTM, the CNN+LSTM, and the GCT.

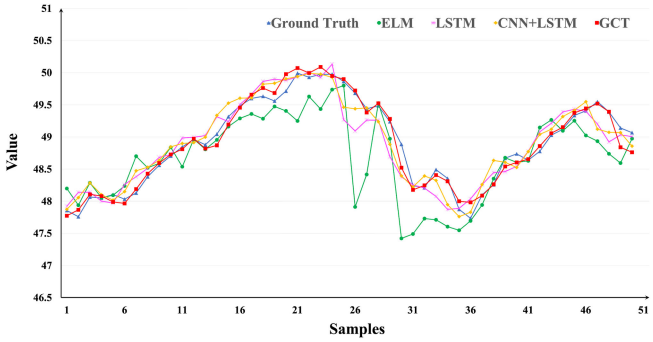

Fig. 11. Predict curves of the ELM, the LSTM, the CNN+LSTM, and the GCT.

ELM is not stable enough. Compared with the ELM and the BP, the LSTM series models, and the GCT that considering timing characteristics have achieved a very significant improvement, which fully indicates the importance of historical information for dynamic soft sensor modeling. As shown at time 31–36 and 45–50 in Fig. 11, the CNN+LSTM performs better than the LSTM, which proves the role of the CNN in smoothing data. By comparing the performance of the CNN+LSTM and the GCT at those two stages, it can be found that the multihead attention mechanism helps the GCT learn time patterns existing in historical data.

C. Ablation Study

In order to verify the contribution of each component in the GCT model, ablation experiments on polypropylene and PTA

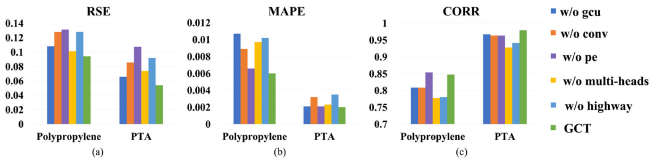

Fig. 12. Ablation study results in terms of (a) RSE, (b) MAPE, and (c) CORR on polypropylene and PTA industrial processes.

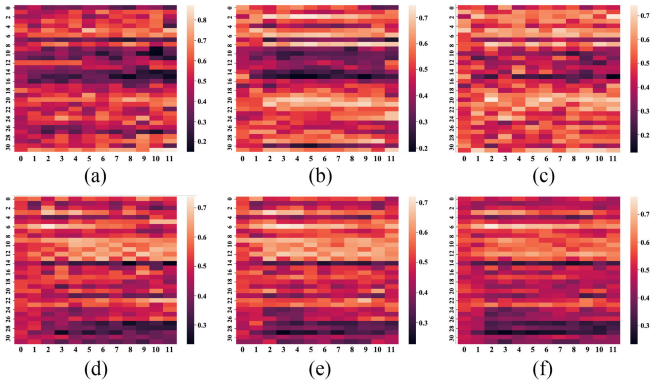

Fig. 13. Visualization of the output of the gated convolutional kernel of the GCU.

process data are conducted. Specifically, the GCU, the convolutional layer, the multihead attention, the position embedding, and the highway network are removed for comparison experiments with the complete model. The experimental results are shown in Fig. 12, which indicate that removing any component will lead to a decrease in the overall performance of the proposed model. Therefore, each component of the proposed model has a certain contribution to the overall performance and is indispensable in the overall architecture.

D. GCU Analysis

To study the ability of filtering important features of the GCU component, the outputs of the gated kernels of different samples are visualized in Fig. 13. The horizontal axis represents time, and the vertical axis represents feature dimensions. The lighter the color, the higher the gate activation value, and vice versa. Extremely similar gated activation value matrix can be easily seen from the results. That is, for the results of polypropylene, the GCU has a higher pass rate for the eigenvalues on the upper and bottom sides, and a lower pass rate for the middle eigenvalue for each moment, as shown in Fig. 13(a)–(c). While for the results of PTA, the completely different pass rates of the upper and lower parts can be seen, as shown in Fig. 13(d)–(f), which proves the effectiveness of the GCU in term of filtering important features.

E. Complexity Analysis

The deep learning models in this article are mainly composed of components such as the RNN, the CNN, and the self-attention mechanism. The complexity analysis of these components is shown in Table V, where s, d, k, and c represent the length of the input sequence, the hidden vector dimension, the size of the convolution kernel, and the number of convolution kernels, respectively.

TABLE V
COMPLEXITY OF DEEP NEURAL NETWORK COMPONENTS

Network type	Computational complexity	Time complexity	Maximum path length
RNN	$O(s \cdot d^2)$	$O(s)$	$O(s)$
CNN	$O(c \cdot k \cdot s \cdot d)$	$O(1)$	$O(log_k(s))$
Self-Attention	$O(s^2 \cdot d)$	$O(1)$	$O(1)$

TABLE VI
COMPLEXITY ANALYSIS AND COMPUTING COST OF EACH METHODS

Method	Computational complexity	Time complexity	Computing cost per epoch (MI/PTA)
BP	$O(n \cdot d)$	$O(n)$	0.0317s/0.0080s
BP(m)	$O(n \cdot d)$	$O(n)$	0.0285s/0.0077s
LSTM	$O(n \cdot s \cdot d^2)$	$O(n \cdot s)$	1.2355s/0.3188s
CNN+LSTM	$O(c \cdot k \cdot s \cdot d + n \cdot s \cdot d^2)$	$O(n \cdot s)$	1.2608s/0.3213s
GCT	$O(c \cdot k \cdot s \cdot d + n \cdot s^2 \cdot d)$	$O(n)$	1.1777s/0.3003s

Therefore, the complexity of each soft sensor method and the actual computing cost on polypropylene and PTA datasets are shown in Table VI, where n represents the number of network layers and h represents the prediction step size. It can be seen from Table VI that the time overhead of the LSTM-based models are proportional to the sequence length s. For the GCT model, because CNN and the attention mechanism can be calculated in parallel, the time overhead depends only on the number of layers and has nothing to do with the sequence length. From the perspective of computational complexity, the length of the input sequence is usually much smaller than the hidden dimension, so the computational complexity of the GCT model is much smaller than that of the LSTM-based models. Although there are fully connected feedforward layers in transformer that increase the complexity of the model to a certain extent, the GCT model still achieves the best performance with minimal time overhead.

V. CONCLUSION

This article proposed a novel GCT for dynamic soft sensor modeling of industrial processes. An improved gated CNN was utilized to capture short-term time patterns and adaptively filtered important features from industrial process data. Through the multihead attention mechanism, the correlation between any two moments can be captured in a parallel manner. The highway network makes the GCT model pay more attention to the information of the most recent time steps. Experimental results of the BP, the ELM, the LSTM, the CNN+LSTM, and the GCT on actual production data of polypropylene and PTA showed that the proposed model can achieve state-of-the-art performance. In general, the proposed model is superior to most of the existing models, and its efficiency is of significance for dynamic soft sensor modeling in actual industrial processes. In the future, we will explore the time patterns in the time series data more fully. Considering that in some scenarios, we are not only concerned with a single predictive value, but more concerned with the possible range of key indicators; therefore, we will try to model the probability of key indicators.

REFERENCES

[1] Z. Q. Ge, "Review on data-driven modeling and monitoring for plant-wide industrial processes," *Chemometrics Intell. Lab. Syst.*, vol. 171, pp. 16–25, Sep. 2017.

[2] X. F. Yuan, B. Huang, Y. L. Wang, C. H. Yang, and W. H. Gui, "Deep learning-based feature representation and its application for soft sensor modeling with variable-wise weighted SAE," *IEEE Trans. Ind. Informat.*, vol. 14, no. 7 pp. 3235–3243, Feb. 2018.

[3] A. Mogilicharla, K. Mitra, and S. Majumdar, "Modeling of propylene polymerization with long chain branching," *Chem. Eng. J.*, vol. 246, pp. 175–183, Jun. 2014.

[4] X. F. Yuan, Y. L. Wang, C. H. Yang, Z. Q. Ge, Z. H. Song, and W. H. Gui, "Weighted linear dynamic system for feature representation and soft sensor application in nonlinear dynamic industrial processes," *IEEE Trans. Ind. Electron.*, vol. 65, no. 2, pp. 1508–1517, Feb. 2018.

[5] S. Yin, X. W. Li, H. J. Gao, and O. Kaynak, "Data-based techniques focused on modern industry: An overview," *IEEE Trans. Ind. Electron.*, vol. 62, no. 1, pp. 657–667, Jan. 2015.

[6] W. W. Yan, D. Tang, and Y. J. Lin, "A data-driven soft sensor modeling method based on deep learning and its application," *IEEE Trans. Ind. Electron.*, vol. 64, no. 5, pp. 4237–4245, May 2017.

[7] Z. Ge and Z. Song, "Distributed PCA model for plant-wide process monitoring," *Ind. Eng. Chem. Res.*, vol. 52, no. 5, pp. 1947–1957, Jan. 2013.

[8] X. C. Xie, W. Sun, and K. C. Cheung, "An advanced PLS approach for key performance indicator related prediction and diagnosis in case of outliers," *IEEE Trans. Ind. Electron.*, vol. 63, no. 4, pp. 2587–2594, Apr. 2016.

[9] C. Shang, F. Yang, D. Huang, and W. X. Lyu, "Data-driven soft sensor development based on deep learning technique,"*J. Process Control*, vol. 24, no. 3, pp. 223–233, Jan. 2014.

[10] W. Kong and J. Yang, "Prediction of polypropylene melt index based on RBF neural networks," *J. Chem. Ind. Eng.*, vol. 54, no. 8, pp. 1160–1163, 2003.

[11] K. K. Huang, Y. Wu, C. Wang, Y. F. Xie, C. H. Yang, and W. Gui, "A projective and discriminative dictionary learning for high-dimensional process monitoring with industrial applications," *IEEE Trans. Ind. Informat.*, vol. 17, no. 1, pp. 558–568, Jan. 2021.

[12] K. K. Huang, H. F. Wen, C. Zhou, C. H. Yang, and W. Gui, "Transfer dictionary learning method for cross-domain multimode process monitoring and fault isolation," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 11, pp. 8713–8724, Nov. 2020.

[13] X. F. Yuan, J. Zhou, B. Huang, Y. L. Wang, C. H. Yang, and W. H. Gui, "Hierarchical quality-relevant feature representation for soft sensor modeling: A novel deep learning strategy," *IEEE Trans. Ind. Informat.*, vol. 16, no. 6, pp. 3721–3730, Jun. 2020.

[14] G. B. Huang, H. M. Zhou, X. J. Ding, and R. Zhang, "Extreme learning machine for regression and multiclass classification," *IEEE Trans. Syst., Man., Cybern.*, vol. 42, no. 2, pp. 513–529, Apr. 2012.

[15] Y. L. He, X. Wang, and Q. X. Zhu, "Soft sensing of PTA content based on PCA improved limit learning machine method," *Control Theory Appl.*, vol. 32, pp. 80–85, 2015.

[16] Y. M. Han *et al.*, "Resource optimization model using novel extreme learning machine with t-distributed stochastic neighbor embedding: Application to complex industrial processes," *Energy*, vol. 225, 2021, Art. no. 120255, doi: 10.1016/j.energy.2021.120255.

[17] L. G. Feng, C. H. Zhao, and Y. X. Sun, "Dual attention-based encoder-decoder: A customized sequence-to-sequence learning for soft sensor development," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 99, pp. 1–12, Aug. 2020.

[18] Q. Q. Sun and Z. Q. Ge, "Probabilistic sequential network for deep learning of complex process data and soft sensor application," *IEEE Trans. Ind. Informat.*, vol. 15, no. 5, pp. 2700–2709, May 2019.

[19] P. Kadlec, B. Gabrys, and S. Strandt, "Data-driven soft sensors in the process industry," *Comput. Chem. Eng.*, vol. 33, no. 4, pp. 795–814, Jan. 2009.

[20] Y. M. Han, C. Y. Fan, M. Xu, Z. Q. Geng, and Y. H. Zhong, "Production capacity analysis and energy saving of complex chemical processes using LSTM based on attention mechanism," *Appl. Thermal Eng.*, vol. 160, Sep. 2019, Art. no. 114072.

[21] X. F. Yuan, L. Li, Y. A. W. Shardt, Y. L. Wang, and C. H. Yang, "Deep learning with spatiotemporal attention-based LSTM for industrial soft sensor model development," *IEEE Trans. Ind. Electron.*, vol. 68, no. 5, pp. 4404–4414, May 2021.

[22] N. Wu, B. Green, X. Ben, and S. O'Banion, "Deep transformer models for time series forecasting: The influenza prevalence case," 2020, *arXiv:2001.08317*.

[23] X. F. Yuan, S. B. Qi, Y. L. Wang, and H. B. Xiao, "A dynamic CNN for nonlinear dynamic feature learning in soft sensor modeling of industrial process data," *Control Eng. Pract.*, vol. 104, Aug. 2020, Art. no. 104614.

[24] K. Wang, C. Shang, L. Liu, Y. Jiang, D. Huang, and F. Yang, "Dynamic soft sensor development based on convolutional neural networks," *Ind. Eng. Chem. Res.*, vol. 58, pp. 11521–11531, May 2019.

[25] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process, Syst.*, 2017, pp. 5998–6008.

[26] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2019, *arXiv: 1810.04805v1*.

[27] A. Oord, S. Dieleman, H. Zen, K. Simonyan, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," 2016, *arXiv: 1609.03499v2*.

[28] R. Srivastava, K. Greff, and J. Schmidhuber, "Highway networks," 2015, *arXiv: 1505.00387*.

[29] B. Carlile, G. Delamarter, K. Paul, A. Marti, and B. Whitney, "Improving deep learning by inverse square root linear units (ISRLUs)," 2017, *arXiv: 1710.09967*.

[30] S. Yang, J. Guo, and L. Wang, "Progress in world polypropylene production technology," *Gansu Sci. Technol.*, vol. 12, pp. 31–32, 2010.

[31] J. B. Jiang, J. L. Wei, and Y. M. Xu, "Online calculation and engineering application of polypropylene melt index," *Comput. Eng. Appl.*, vol. 38, no. 18, pp. 220–222, 2002.

[32] B. L. Zhang, Y. M. Han, B. Yu, and Z. Q. Geng, "Novel nonlinear auto regression with external input integrating PCA-WD and its application to dynamic soft sensor," *Ind. Eng. Chem. Res.*, vol. 59, pp. 15697–15706, Aug. 2020.

Zhiqiang Geng received the B.Sc. degree in process equipment and control engineering and the M.Sc. degree in chemical machinery from Zhengzhou University, Zhengzhou, China, in 1997 and 2002, respectively, and the Ph.D. degree in control theory and control engineering from the College of Information Science and Technology, Beijing University of Chemical Technology, Beijing, China, in 2005.

He is currently a Professor with the College of Information Science and Technology, Beijing University of Chemical Technology. His research interests include neural networks, intelligent computing, data mining, knowledge management, and process modeling.

Zhiwei Chen received the M.Sc. degree in chemical engineering from the Wuhan Institute of Technology, Wuhan, China, in 2014. He is currently working toward the Ph.D. degree in control science and engineering with the Beijing University of Chemical Technology, Beijing, China.

His research interests include artificial neural networks, process modeling, and optimization.

Qingchao Meng received the B.Sc. degree in 2018 from the Beijing University of Chemical Technology, Beijing, China, where he is currently working toward the M.Sc. degree, both in computer science and technology.

His research interests include deep learning, data mining, and process modeling

Yongming Han received the B.Sc. degree in computer science and technology and the Ph.D. degree in control theory and control engineering from the Beijing University of Chemical Technology, Beijing, China, in 2009 and 2014, respectively.

He is currently a Professor with the College of Information Science and Technology, Beijing University of Chemical Technology. His research interests include knowledge graph analysis, neural networks, intelligent computing, data mining, and optimization.