Input-output Driven Cross-Attention for Transformer for Quality Prediction of Light Naphtha in Industrial Hydrocracking Processes

Ziyi Yang, Xiaofeng Yuan, Kai Wang, Zhiwen Chen, Yalin Wang, Chunhua Yang, Weihua Gui

Central South University, Changsha, China
(Tel: 151-1576-4503; e-mail: 234612143@csu.edu.cn, yuanxf@csu.edu.cn, kaiwang@csu.edu.cn)

Abstract: Monitoring and predicting the key variables is significant in industrial processes. However, it is not effective in extracting temporal features of variables and predicting them accurately due to the high dimensionality and the long term of the input sequence. Therefore, this paper proposes an input-output driven cross-attention for the transformer network (IDCA-Former), which takes historical labels into account as a part of the input sequence. Then, cross-attention is conducted to compute the similarity between historical labels and original input data to capture more potential information. Moreover, sliding windows are designed by setting the input length of historical labels. The proposed IDCA-Former is applied for light naphtha prediction in the hydrocracking process. Extensive experiments show that IDCA-Former performs better in time series forecasting compared to other methods.

Keywords: Historical label; Cross attention; Transformer; Soft sensor; Industrial process;

1. INTRODUCTION

With the fast development of information technologies and the increased demands of customers, the process industry is going to be more intricate and high standard (Sun and Ge, 2021). Thus, how to monitor industrial processes efficiently and accurately is becoming more significant in modern industry. Usually, the stable and safe operations can be reflected by a few key factors or indicators in the production process. To control the entire procedure in real time, on-site staff must focus on the timely monitoring of some key variables. However, under the actual production environment , these variables cannot be monitored in real time, making it challenging to implement the process control and optimization. Hence, it is necessary to establish an efficient and reliable prediction model for these key variables.

Due to the development of data-driven methodologies, many soft-measurement approaches have been applied for process variable prediction. In contrast to physical sensors, soft measurements offer several advantages, like fast response, low maintenance costs, and real-time estimation . For example, to solve the problem of process variables that are far apart in topology but have high correlations (Yuan et al., 2024), a variable correlation analysis-based convolutional neural network (VCA-CNN) was proposed for far topological feature extraction. Moreover, a new quality-driven regularization (QR) was proposed for deep networks to learn quality-related features so that some important information related to quality variables may not be discarded (Ou et al., 2022). However, the data collected from industrial plants are usually complex and nonlinear, which means that the simple models cannot handle these samples well. Currently, various types of methods based on deep learning are utilized in this field. For instance, a multiphase attention-based recurrent neural network (MPA-RNN) was suggested to capture the decomposed information and extract spatial relationships (Geng et al., 2022). To handle dynamic time sequences with heterogeneous sample interval, attention-based interval-aided networks (AIA-Net) were proposed (Yuan et al., 2023), which considered the sampling intervals between sequential samples so that the temporal information could be better represented. Moreover, a channel based on gated recurrent unit (GRU) network was proposed (Zhang et al., 2023) to tackle the problem of inconsistent sampling rates in the process industry.

However, as the time sequences become longer and longer, these models may show poor performance, especially in long prediction tasks. That is mainly because the recursive structures of models like LSTM and GRU lead to long-distance dependency and the current outputs are influenced by the historical state, making it difficult to predict long-term sequences.

Transformer is currently widely used in modeling time series, and its powerful attention mechanism performs well in capturing similarities between data samples of time series. Also, the inputs of transformer do not depend on the state of the former temporal sequences, which is suitable for parallel computing. Therefore, the effectiveness of transformer has been successfully demonstrated in time series prediction. For instance, a frequency enhanced decomposed transformer (FEDformer) was developed to solve the problems of expensive computation and the inability to capture the global view of time series (Zhou et al., 2022). To fully exploit the characteristics of time-series data and avoid some fundamental limitations of transformer, exponential smoothing attention (ESA) and frequency attention (FA) were adopted by replacing self-attention, which improved both the model accuracy and efficiency (Woo et al., 2022). A self-attentive mechanism based on deconstruction and dot product (DDPformer) in transformer was proposed to highlight the features of the partial head in the multi-headed attention mechanism (Xie and He, 2022). However, industrial process variables may have close spatiotemporal relationships among these variables as the dimensionality becomes higher and higher, and these correlations are often difficult to be extracted directly.

Additionally, the aforementioned methods did not consider the connection between input variables and output variables, which often has a prompt effect on the former during the model training process.

In this paper, an input-output driven cross-attention for transformer network (IDCA-Former) is proposed to solve this problem. In IDCA-Former, historical labels are considered as a part of the input sequence to extract features. Then, cross-attention is designed to calculate the similarity between historical labels and input variables. Extensive experiments have demonstrated the outperformance of quality variable prediction in oil hydrocracking processes compared with the other methods.

The remainder of the paper is as follows. In Section II, the structure of the self-attention mechanism and transformer are introduced briefly. Next, the details of the IDCA-Former are thoroughly presented in Section III. Finally, the proposed IDCA-Former is applied to the oil hydrocracking process for experiments in Section IV. The summaries are presented in Section V.

2. ATTENTION MECHANISM AND TRANSFORMER ARCHITECTURE

2.1 Self-Attention mechanism

As the key component of transformer, attention mechanism is designed to calculate the similarity between two vectors. The structure of self-attention is shown in Fig.1(a).

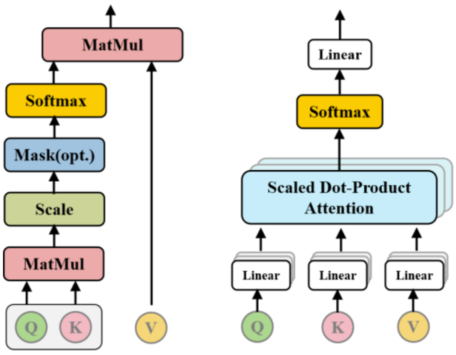


Figure 1. Self-Attention Mechanism:(left)scaled dot-product attention(right)multi-head attention. (Liu et al., 2023)

Assuming that the query matrix (Q), key matrix (K), and value matrix (V) can be obtained by the input sequence X, respectively. Then the dot-product operation is carried out to get the similarity between Q and K. Here, the scale factor $\sqrt{d_m}$ is used to avoid large values of the inner product (Li, Wang and Mcauley, 2022). Finally, after normalizing by $Softmax$ function and multiplying with the value matrix, the output is obtained. The specific calculations are described as:

$$Attention(Q,K,V) = Softmax(\frac{K^TQ}{\sqrt{d_m}})V \quad (1)$$

$$Q = X(W^q)^T \quad (2)$$

$$K = X(W^k)^T \quad (3)$$

$$V = X(W^v)^T \quad (4)$$

where $\sqrt{\mathrm{d_m}}$ represents the dimension of the mapped vector, Softmax(·) represents the normalization function, $\mathrm{W^q}$, $\mathrm{W^k}$, $\mathrm{W^v}$ represent the parameter matrices.

To capture different aspects of features, multi-head attention extends from self-attention mechanism. In this technique, the Q, K, and V matrices are segmented into separate groups to obtain $[q_1, q_2, \dots, q_h]$, $[k_1, k_2, \dots, k_h]$, $[v_1, v_2, \dots, v_h]$, where h is the number of heads. Each sub-matrix performs similar operations as dot-product attention and is subsequently linked by joint function. The calculation proceeds as:

$$MultiHead(Q,K,V) = Concat(head_1, head_2, .., head_h) \quad (5)$$

$$head_i = Attention(Q_i, K_i, V_i),\ i = 1,2..,h \quad (6)$$

where $Concat(\cdot)$ represents joint function to connect vectors, and $head_i$ is the result of each head.

2.2 Transformer model

The components of transformer can be summarized as four modules: input module, encoder module, decoder module and output module. Specifically, they include embedding block, attention block, residual connection layer, feedforward layer and output block (Liu et al., 2023). The overall structure of t transformer is shown in Fig.2.

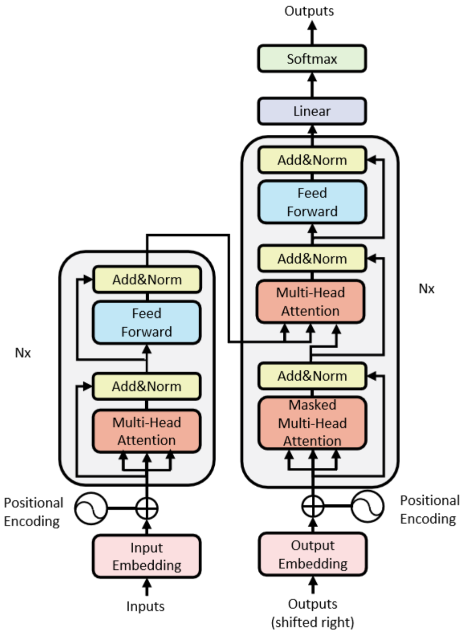


Figure 2. The framework of transformer. (Liu et al., 2023)

Assuming that the input sequence is embedded by the input embedding layer and positional embedding layer. Next, the embedded sequence goes through the encoder module: firstly, the multi-head attention mechanism computes attention scores to get information from the input sequence; Then, the residual connection layer is designed to preserve the original information; Finally, the feed-forward module works, and the output of the encoder is obtained after the residual network.

Decoder module consists of masked multi-head attention mechanism, multi-head attention mechanism, residual connection layer, and feedforward neural network layer. The input sequence of the decoder module first goes through the masked multi-head attention mechanism, which aims to mask the prompt effect of future information. Subsequently, the obtained results are fed into multi-head attention mechanism to provide information about the value matrix, while information of the query matrix and key matrix comes from the output of the encoder. After going through the feedforward neural network and residual connection layer, the output of the decoder is obtained. Finally, the output probabilities are calculated by linear and softmax layers.

3. INPUT-OUTPUT DRIVEN CROSS ATTENTION FOR TRANSFORMER

3.1 Input-output driven cross attention for transformer

In the IDCA-Former network, the correlations between input variable data and historical labels are taken into account. To this end, both of them are combined as input sequences and embedded by the input embedding layer as well as the positional embedding layer, respectively. Moreover, cross-attention is designed to obtain similarity between input variable data and historical labels. The overall framework of IDCA-Former is shown in Fig. 3.

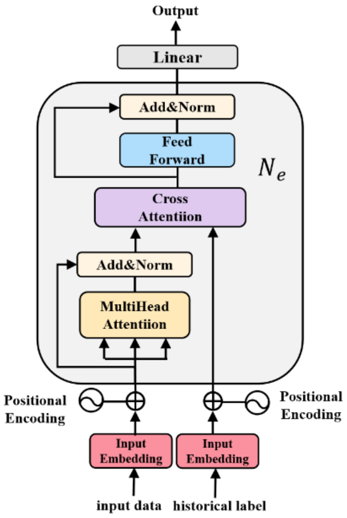


Figure 3. The structure of IDCA-Former.

To capture more information from the combined inputs, the model is fed with the historical labels $\mathrm{X}_y \in \mathbb{R}^{n_y \times D_y}$ and the input data $\mathrm{X}_{data} \in \mathbb{R}^{n_{data} \times D_{data}}$, where n_y and n_{data} represent the length of the historical label and the input variable data, respectively. D_y and D_{data} represent the dimensions of the label output and the input variable data, respectively. Then, they go through the input embedding module. This step is to map the input sequence to the dimensions required by the model. The calculation process is as:

$$X_y^E = Embedding(X_y) = X_y \mathrm{W}_E^y + \mathrm{b}_E^y \tag{7}$$

$$X_{data}^E = Embedding(X_{data}) = X_{data} \mathrm{W}_E^x + \mathrm{b}_E^x \tag{8}$$

where$\mathrm{X}_y^E \in \mathbb{R}^{l \times d_m}$and $\mathrm{X}_{data}^E \in \mathbb{R}^{l \times d_m}$ represent the embedded matrices after linear mapping. l represents the sequence length after sliding windows. d_m is the dimension required by the model. $\mathrm{W}_E^y \in \mathbb{R}^{D \times d_m}$, $\mathrm{b}_E^y \in \mathbb{R}^{D \times d_m}$ represent the weight and bias of mapped linear layers of the label data, respectively. $\mathrm{W}_E^x \in \mathbb{R}^{D \times d_m}$, $\mathrm{b}_E^x \in \mathbb{R}^{D \times d_m}$ represent the weight and bias of mapped linear layers of the input data, respectively.

To further represent the positional relationship of each sequence component, positional embedding is introduced. The calculation is as:

$$X_y^P = X_y^P + PE(X_y^P) \tag{9}$$

$$X_{data}^P = X_{data}^P + PE(X_{data}^P) \tag{10}$$

$$\mathrm{PE}(pos, 2i) = \sin(pos/10000^{2i/d_m}) \tag{11}$$

$$\mathrm{PE}(pos, 2i+1) = \cos(pos/10000^{2i/d_m}) \tag{12}$$

where $\mathrm{X}_y^P \in \mathbb{R}^{l \times d_m}$ and $\mathrm{X}_{data}^P \in \mathbb{R}^{l \times d_m}$ represent the corresponding positional embedding matrices, pos represents the position of every vector. $2i$ and $2i$+1 represent the even and base dimensions in d_m, which ensures that each dimension has an independent corresponding sin-cos component.

Then, $\mathrm{X}_{data}^{Att1} \in \mathbb{R}^{l \times d_m}$ is obtained after X_{data}^P going through the multi-head attention layer and residual connection layer. The multi-head attention mechanism is based on the attention mechanism that uses multiple heads to deal with different information separately. The calculation is as:

$$X_{data}^{Att} = MultiHead(Q, K, V) \tag{13}$$

$$MultiHead(Q, K, V) = Concat(head_1, head_2, .., head_h) \tag{14}$$

$$head_i = Attention(Q_i, K_i, V_i),\ i = 1,2.., h \tag{15}$$

where $W^O \in \mathbb{R}^{dv \times d_m}$ represents output matrix, $Concat(\cdot)$ represents the joint function.

Then X_{data}^{Att} and X_y^P participate in the computation of cross attention. In this mechanism, X_{data}^{Att} provides information about the query matrix Q_d, and X_y^P provides the information about key matrix K_l and value matrix V_l. The specific process is shown in Fig. 4.

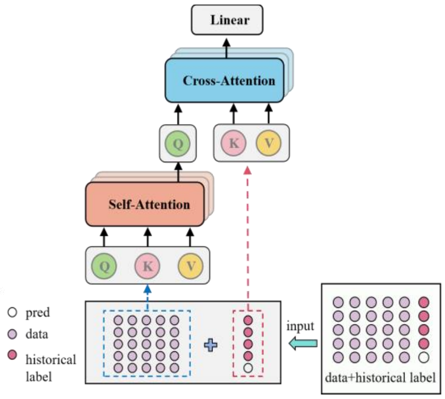


Figure 4. Detailed attention of model IDCA-Former.

In cross attention, the computational procedure is shown below:

$$\mathrm{Cross}Attention(\mathrm{Q}_d, K_l, V_l) = Softmax(\frac{K_l{}^T \mathrm{Q}_d}{\sqrt{d_m}})V_l \quad (26)$$

$$\mathrm{Q}_d = \mathrm{X}_{data}^{Att}(W^q)^T \quad (17)$$

$$K_l = \mathrm{X}_y^P(W^k)^T \quad (18)$$

$$V_l = \mathrm{X}_y^P(W^v)^T \quad (19)$$

Finally, the output of the encoder is obtained through feed-forward neural network layer and residual connection layer, which are calculated as follows:

$$\mathrm{X}^{enc} = \mathrm{Norm}(X^{Att} + \mathrm{FeedForward}(X^{Att})) \quad (20)$$

$$\mathrm{FeedForward}(X^{Att}) = \max(0, X^{Att}\mathrm{W}_1 + \mathrm{b}_1)\,\mathrm{W}_2 + \mathrm{b}_2 \quad (21)$$

where W_1, W_2 represent the weight, b_1, b_2 represent the bias, $\mathrm{Norm}(\cdot)$ represents normalization and $\max(\cdot)$ represents the maximization function.

To explore the influence of history labels to variable data, sliding windows are designed to set the number of history labels. For a given sequence with history labels $\mathrm{X}_{label} = [\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_n] \in \mathbb{R}^{n_{label} \times D_{label}}$, setting the sliding window $length = l$, $stride = k$, and then the sequence $\mathrm{X}_l = [\mathrm{X}_{l1}, \mathrm{X}_{l2}, \ldots, \mathrm{X}_{lp}]$, $p = \frac{(n-1)}{k}$ is generate, in which $\mathrm{X}_{l1} = [\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_l]$, $\mathrm{X}_{l2} = [\mathrm{X}_{1+k}, \mathrm{X}_{2+k}, \ldots, \mathrm{X}_{l+k}]$, $\mathrm{X}_{lp} = [\mathrm{X}_{1+pk}, \mathrm{X}_{2+pk}, \ldots, \mathrm{X}_{l+pk}]$, $\mathrm{X}_l \in \mathbb{R}^{p \times D}$. When $stride = 1$, the sliding windows process can be represented in Fig .5.

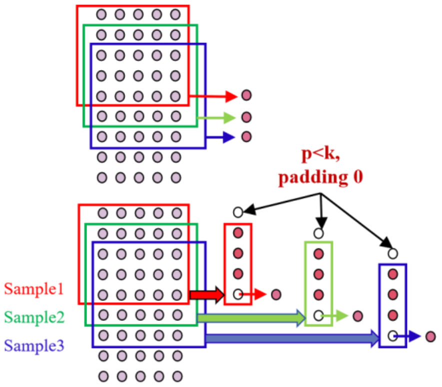


Figure 5. Detailed sliding window of IDCA-Former.

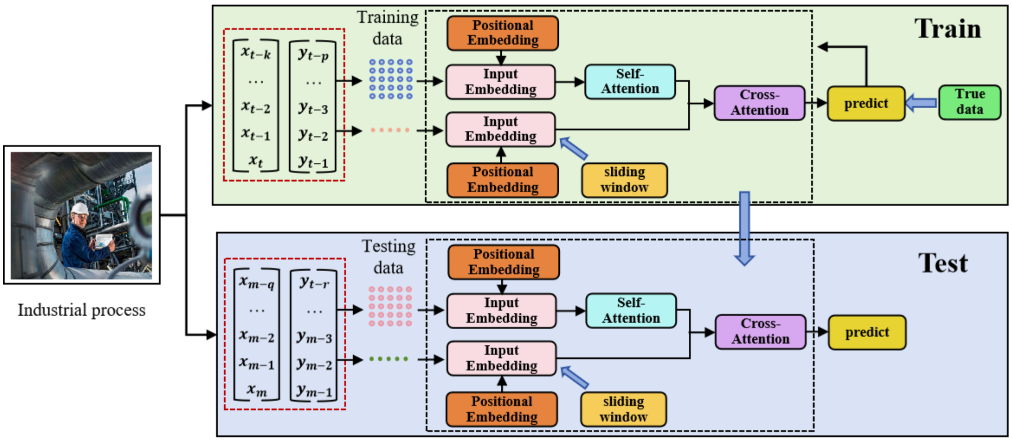


Figure 6. The process of IDCA-Former soft sensor modelling

3.2 IDCA-Formerbased soft sensor modeling

The datasets are first divided into training data and test data. In the training step, variable data and historical labels are generated by data preprocesses. Next, they are embedded by the input embedding layer and positional embedding layer. Then they are fed into IDCA-Former for training to obtain the optimal hyperparameters. In this process, the loss values are computed to modify the weight and bias of the networks. After that, training parameters are tested on the test set. The sliding windows are designed for setting different lengths of historical labels. The entire modeling process is shown in Fig.6. The model evaluation metrics, the mean absolute error (MAE), the mean square error (MSE), the mean absolute percentage error (MAPE), the prediction root mean squared error(RMSE) and coefficient of determination R2 are calculated by the following:

$$\mathrm{MAE} = \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{y_i} - \tilde{\mathrm{y}}_\mathrm{i} \quad (22)$$

$$\mathrm{MSE} = \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{y_i} - \tilde{\mathrm{y}}_\mathrm{i})^2 \quad (23)$$

$$\mathrm{MAPE} = \frac{100\%}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}} \left|\frac{\mathrm{y_i} - \tilde{\mathrm{y}}_\mathrm{i}}{\mathrm{y_i}}\right| \quad (24)$$

$$RMSE = \sqrt{\frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{y_i} - \tilde{\mathrm{y}}_\mathrm{i})^2} \quad (25)$$

$$R^2 = 1 - \sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{y_i} - \tilde{\mathrm{y}}_\mathrm{i})^2 / \sum_{\mathrm{i}=1}^{\mathrm{n}} (\bar{\mathrm{y}}_\mathrm{i} - \tilde{\mathrm{y}}_\mathrm{i})^2 \quad (26)$$

4. INDUSTRIAL APPLICATION

In the petrochemical industry, the hydrocracking link is one of the petroleum refining processes. Carbon 5 (C5) is the main by-product in the naphtha cracking process, with a capacity of 14%-20% of the cracked ethylene industry. It is also a valuable resource for the comprehensive utilization of the chemical industry.

At present, one of the most efficient ways to increase the economic efficiency of the petrochemical industry is by fully separating and utilizing C5 fraction resources. It is also significant for the development of petroleum industry (Yuan et al., 2021a). So it is necessary to monitor and predict C5.

The datasets used in this paper is from a chemical and refinery plant in China, and the data were recorded from 2016 to 2018. A total of 2580 data were recorded as samples in the datasets, which were divided into a training set and a test set in the ratio of 8:2. After model training, the optimal hyperparameters are obtained as shown in Table 1.

Table 1 The hyper parameters

Symbol	Description	Value
L_{SW}	the length of sliding window	9
d_{model}	the dimension of model	1024
n_{heads}	the number of attention heads	6
n_{en}	the number of encoder block	2
batch	the batch size	32
d_{ff}	the dimension of feedforward layer	2048
kernel	the size of the convolving kernel	8

Comparing LSTM, GRU, and Logtrans models, the following experimental results are obtained as shown in Table 2.

Table 2 Comparison performances with different models for predicting C5 in hydrocracking

Method	MAE	MSE	RMSE	MAPE	R^2
LSTM	0.2973	0.1879	0.4335	0.5285	**0.8151**
GRU	0.2944	0.1881	0.4337	0.5355	**0.8388**
LogTrans	0.3370	0.1838	0.4287	0.1527	**0.8604**
IDCA-Former	0.1832	0.0741	0.2722	0.0868	**0.9555**

From Table 2, it can be found that LSTM and GRU, although feasible in time series prediction, are not very effective to extract features in the long-term sequence. This is mainly because their recursive structure lead to serious time-dependence problem. Moreover, the result of R^2indicates the predicted value at the current moment depends on the state value at the previous moment. Compared with the method of LSTM, GRU is slightly better in prediction. Since LogTrans (Li et al., 2019) has an attention mechanism to compute the similarity between vectors and extract features, the performance of prediction becomes better. However, IDCA-Former is found to reach a more ideal prediction accuracy after introducing sliding windows with history labels for feature learning and is more adaptable to the task of predicting long-term sequences.

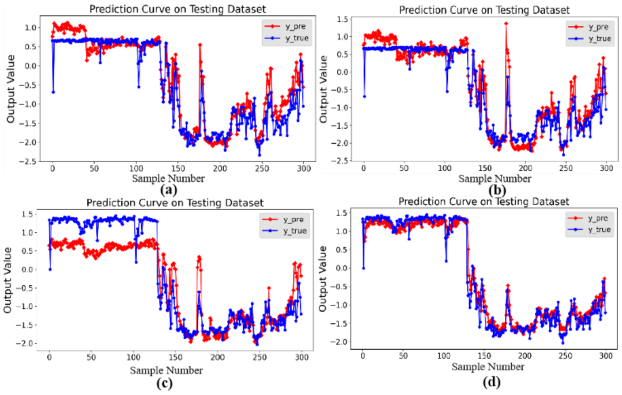


Fig. 7. Prediction results for C5 in hydrocracking:(a)LSTM;(b)GRU;(c)LogTrans;(d)IDCA-Former

To exhibit the results more intuitively, Fig. 7 plots the curves between the real and predicted values of each model. It can be seen that the trends of LSTM and GRU generally go with the ground truth, but they predict not well in some areas. LogTrans shows a similar overall trend between real values and predicted values. However, the prediction effect of LogTrans is not good enough in some key turning points, while IDCA-Former outperforms better than these methods.

In addition, to further explore how sliding window size will affect the history labels on the prediction performance, different sliding window sizes are designed to train the model. We set the length of label sliding windows from 0 to 7, and the input sequence number is 8. After extensive experiments, the results of prediction metrics under different sliding window sizes are shown in Table 3.

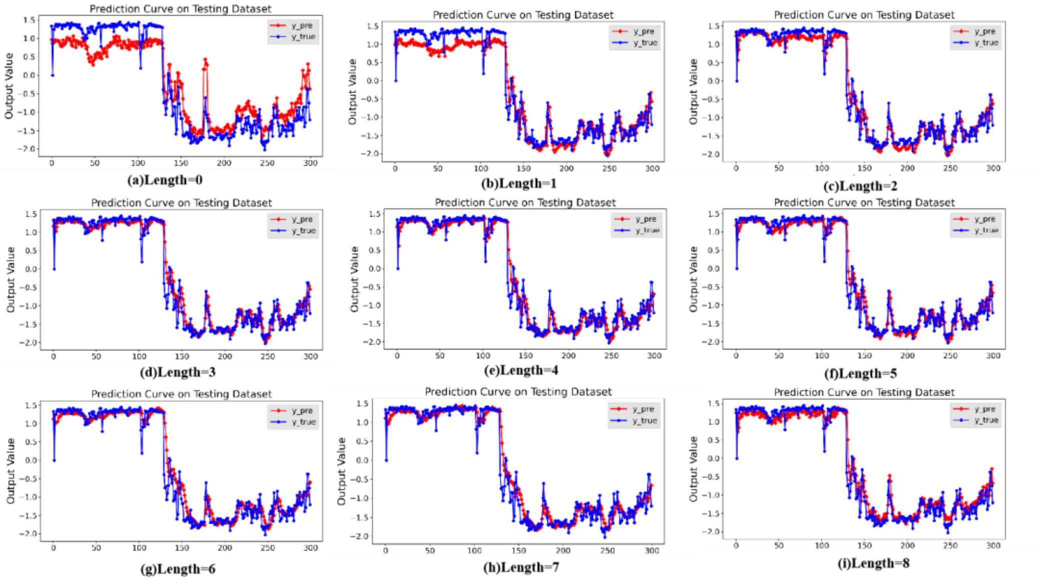
Table 3 Comparison performances with different window sizes for predicting C5 in hydrocracking

Window size	MAE	MSE	RMSE	MAPE	R^2
0.0	0.4232	0.2624	0.5123	0.1896	**0.7535**
1.0	0.2134	0.0931	0.3052	0.1006	**0.9442**
2.0	0.2010	0.0878	0.2963	0.0957	**0.9528**
3.0	0.1820	0.0807	0.2842	0.0873	**0.9544**
4.0	0.1675	0.0779	0.2792	0.0823	**0.9567**
5.0	0.1871	0.0779	0.2792	0.0890	**0.9578**
6.0	0.1725	0.0709	0.2662	0.0831	**0.9612**
7.0	0.1809	0.0710	0.2664	0.0861	**0.9615**

When the sliding window length is set to 0, it indicates that the history label has no prompting effect on the input, so the performance is not good. When the sliding window size is set to 1, the result of R^2 increases significantly, denoting that the history label shows a prompted effect on the input. As the sliding window size continues to increase, it can be seen that the model performs better than before, but the improvement is less. The curves between the true and predicted values are also plotted, as shown in Fig .8.

5. CONCLUSION

In this work, an input-output driven cross-attention for the transformer network (IDCA-Former) is proposed to represent the potential relationships between sampling variables recorded from industrial plants that often show high dimensionality and dynamics. In IDCA-Former, historical labels are taken into account to train the model. In this way, variable data and historical labels are combined as input sequence and sent into the embedding layer. Then cross attention mechanism is introduced to compute the similarity between them. Additionally, sliding windows are designed to explore the prompt effect of historical labels length on the prediction performance. Extensive experiments are conducted in hydrocracking process and the effectiveness of the proposed IDCA-Former are evaluated.



(a)Length=0 (b)Length=1 (c)Length=2 (d)Length=3 (e)Length=4 (f)Length=5 (g)Length=6 (h)Length=7 (i)Length=8

Fig.8 Prediction results for C5 in hydrocracking:(a)Length=0;(b)Length=1;(c)Length=2;(d)Length=3; (e)Length=4;(f)Length=5;(g)Length=6;(h)Length=7;(i)Length=8;

REFERENCES

Geng, J.X., Yang, C.H., Li, Y.G., Lan, L.J. and Luo, Q.W. (2022). "MPA-RNN: A Novel Attention-Based Recurrent Neural Networks for Total Nitrogen Prediction," in IEEE Transactions on Industrial Informatics, vol. 18, no. 10, pp. 6516-6525.

Li, S.Y., , Jin, X.Y., Xuan, Y., et al. (2019). Enhancing the Locality and Breaking the Memory Bottleneck of Transformer on Time Series Forecasting[J].

Li, J.C., Wang, Y.J., Mcauley, J.L. (2022). Time Interval Aware Self-Attention for Sequential Recommendation[C]//WSDM '20: The Thirteenth ACM International Conference on Web Search and Data Mining.ACM.

Liu, D.J., Wang, Y.L., Liu, C.L., Yuan, X.F., Yang, C.H. and Gui, W.H. (2023). "Data Mode Related Interpretable Transformer Network for Predictive Modeling and Key Sample Analysis in Industrial Processes," in IEEE Transactions on Industrial Informatics, vol. 19, no. 9, pp. 9325-9336.

Ou, C., Zhu, H.Q., Shardt, Y., Ye, L.J., Yuan, X.F., Wang, Y.L., Yang, C.H. (2022) "Quality-Driven Regularization for Deep Learning Networks and Its Application to Industrial Soft Sensors," in IEEE Transactions on Neural Networks and Learning Systems, doi: 10.1109/TNNLS.2022.3144162.

Sun, Q.Q. and Ge, Z.Q. (2021). "A Survey on Deep Learning for Data-Driven Soft Sensors," in IEEE Transactions on Industrial Informatics, vol. 17, no. 9, pp. 5853-5866.

Woo, G., Liu, C.H., Sahoo, D.Y., et al. (2022). ETSformer: Exponential Smoothing Transformers for Time-series Forecasting[J].

Xie, H.S. and He, J.W. (2022). "Transformer Based on Deconstruction and Dot Product for Time Series Forecasting of Air Quality," International Conference on Computers and Artificial Intelligence Technologies (CAIT), Quzhou, China, pp. 31-37.

Yuan, X.F., Feng, L., Wang, K., Wang, Y.L. and Ye, L.J. (2021). "Deep Learning for Data Modeling of Multirate Quality Variables in Industrial Processes," in IEEE Transactions on Instrumentation and Measurement, vol. 70, pp. 1-11.

Yuan, X.F., Xu, N. Ye, L.J., Wang, K., Shen, F.F., Wang, Y.L., Yang, C.H., Gui, W.H. (2023) "Attention-Based Interval Aided Networks for Data Modeling of Heterogeneous Sampling Sequences With Missing Values in Process Industry," in IEEE Transactions on Industrial Informatics, doi: 10.1109/TII.2023.3329684.

Yuan, X.F., Wang, Y.C., Wang, C., Ye, L.J., Wang, K., Wang, Y.L., Yang, C.H., Gui, W.H. and Shen, F.F. (2024) "Variable Correlation Analysis-Based Convolutional Neural Network for Far Topological Feature Extraction and Industrial Predictive Modeling," in IEEE Transactions on Instrumentation and Measurement, vol. 73, pp. 1-10, 2024, Art no. 3001110, doi: 10.1109/TIM.2024.3373085.

Zhou, T., Ma, Z.Q., Wen, Q.S., Wang, X., Sun, L., Jin, R. (2022) .FEDformer: Frequency Enhanced Decomposed Transformer for Long-term Series Forecasting[J]. doi:10.48550/arXiv.2201.12740.

Zhang, K.M., Lu, F., Yu, C.L., Dai, C., Huayun, Z.J., Zhang, T.H., Chen, X., Lu, J.Q., Lin, Z.Z. (2023). "Short-Term Electrical Load Forecasting Based on Attention-GRU Networks," 2023 IEEE 6th International Electrical and Energy Conference (CIEEC), Hefei, China, pp. 3338-3343.