Scholars Journal of Engineering and Technology

OPEN ACCESS

Abbreviated Key Title: Sch J Eng Tech
ISSN 2347-9523 (Print) | ISSN 2321-435X (Online)
Journal homepage: https://saspublishers.com

Fault Detection Method based on Local and Global Attention Mechanisms

LU Zhen Jie[1*]

[1]School of Information Engineering, Shenyang University of Chemical Technology, Tiexi, Shenyang, China, 110142

DOI: 10.36347/sjet.2023.v11i08.004 | **Received:** 13.07.2023 | **Accepted:** 18.08.2023 | **Published:** 21.08.2023

***Corresponding author:** LU Zhen Jie
School of Information Engineering, Shenyang University of Chemical Technology, Tiexi, Shenyang, China, 110142

Abstract **Review Article**

In recent years, due to the wide application of Distributed control system, a large number of production process data can be collected and stored, which provides a solid data foundation for process monitoring technology based on deep learning. The Transformer model is a fully connected attention mechanism model that captures the global dependencies of data by calculating the correlation between any two items. This paper proposes a Transformer model based on local and global attention mechanisms. Firstly, after the data is standardized to eliminate the impact of different dimensions, positional encoding is used to mark the position information. Then, the data is divided into two equal parts from the feature dimension. One part enters the standard attention mechanism to capture the global information of the sequence, and the other part enters the local attention mechanism to capture the local information of the sequence, Then, the captured local information and global information are fused to reduce computational complexity and compensate for the shortcomings of the Transformer model in capturing local information. By applying the model proposed in this paper to the penicillin fermentation process for fault detection, it has been experimentally verified that the proposed model has an improved fault detection accuracy compared to the standard Transformer model.

Keywords: Fault detection, Transformer, attention mechanism, local attention mechanism, penicillin fermentation process

INTRODUCTION

In complex large-scale industrial processes, the production process generally includes multiple production units, and each unit contains a large number of sensors and controllers. Due to the high dimensionality and nonlinearity of process data in such industrial production, traditional process monitoring methods based on data-driven methods are difficult to achieve good monitoring results [1]. How to extract important information from hundreds or thousands of dimensional process data and monitor the operational status of the process is an urgent research topic.

In recent years, with the development of deep learning, some process monitoring technologies based on deep learning have been used to solve high-dimensional complexity problems in large-scale industrial production processes. The Transformer model [2] is a fully connected attention mechanism model that captures global dependencies of data by calculating the correlation between any two items. The model was initially applied to machine translation tasks [3] and achieved good results, so more and more researchers began to study the operation mechanism and principle of the Transformer model and improve it to adapt to their respective research fields. A large number of literature has shown that the Transformer model has not only achieved good results in the field of machine translation, but also in images the audio and video fields have also achieved the best experimental results currently available.

Although the Transformer model has achieved good results in dealing with long-distance dependencies of time series data, there are still some shortcomings. Firstly, such as the Transformer model's poor ability to extract local information; Secondly, due to Transformer is a model based on a fully connected self-attention mechanism that calculates the correlation between any two items, achieving good performance requires a significant amount of computation as a cost. When there are complex data characteristics such as large-scale high-dimensional and nonlinear data, traditional data-driven process monitoring methods are difficult to achieve good fault detection results. So there is an urgent need for an effective deep learning based method

to unify modeling and fault detection of large-scale data collected in modern industrial production processes.

Improved Transformer Model:

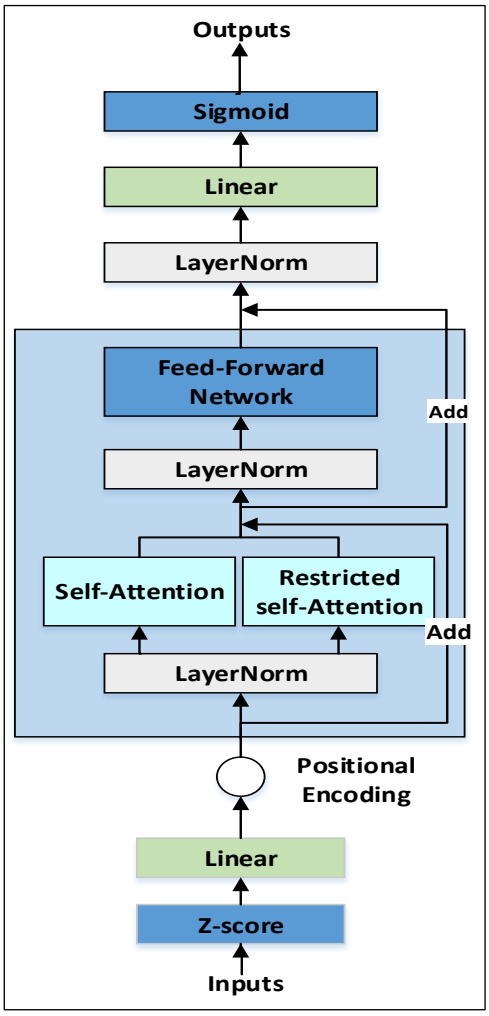

Figure 1: Overall flowchart of improved Transformer model

The overall flowchart of the improved Transformer model is shown in Figure 1. Firstly, the model applies the input data $X \in \mathrm{R}^{N\times M}$ undergoes Z-score standardization to eliminate the impact of different dimensions, $X_{\text{Z-score}} \in \mathrm{R}^{N\times M}$ is obtained. Then, through the fully connected layer, the feature dimension M of the data is mapped to the D dimension to obtain its high-dimensional representation, where $X_{\text{Z-score-d}} \in \mathrm{R}^{N\times D}$. The above process is shown in Equations (1) - (3).

$$y_i = \frac{x_i - \bar{x}}{S} \quad \text{..... (1)}$$

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \quad \text{..... (2)}$$

$$S = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2} \quad \text{..... (3)}$$

Where $\bar{x}$ is the average value of all samples, S is the variance, x_i is the i-th sample, $i \in [1, N]$, y_i is the data that the i-th sample after Z-score standardization processing.

Secondly, the position information of the sequence is marked using Positional Encoding (PE), and the Positional Encoding equation is shown in (4) - (5) [2]. In order to accelerate the convergence rate of the model, this paper uses the Pre Layer Normalization structure [4], first performing layer normalization operations on the data.

$$\mathrm{PE}_{(pos,2i)} = \sin(pos / 10000^{2i/d_{model}}) \quad \text{..... (4)}$$

$$\mathrm{PE}_{(pos,2i+1)} = \cos(pos / 10000^{2i/d_{model}}) \quad \text{..... (5)}$$

$$X_{\text{Z-score-d}} = \mathrm{Layer}Norm(\mathrm{Re}\,\mathrm{Lu}(X_{\text{Z-score-d}})) \quad \text{..... (6)}$$

Thirdly, divide the normalized data $X_{\text{Z-score-d}} \in \mathrm{R}^{N\times D}$ into two equal parts in terms of dimensions to obtain $X_{\text{local}} \in \mathrm{R}^{N\times D/2}$ and $X_{\text{global}} \in \mathrm{R}^{N\times D/2}$, X_{global} enters the self-attention mechanism of the standard Transformer to capture global information of the sequence, while X_{local} enters the local attention mechanism [5] to capture local information of the sequence.

$$Q = X_{gloabl} \cdot W^Q,\ K = X_{gloabl} \cdot W^K,\ V = X_{gloabl} \cdot W^V \quad \text{.....(7)}$$

$$\mathrm{MultiAttention}(Q, X_{gloabl}) = Concat(a_1, ...a_k) \cdot W^O \quad \text{...(8)}$$

$$a_i = \mathrm{Attention}(Q \cdot W^Q, K \cdot W^K, V \cdot W^V) \quad \text{..... (9)}$$

$$\mathrm{Attention}(Q, K, V) = Softmax(\frac{Q \cdot K^T}{\sqrt{d_k}}) \cdot V \quad \text{..... (10)}$$

In the standard attention mechanism, for vector sequence $X_{\text{global}} \in \mathrm{R}^{N\times D/2}$, first use Equation (7) [2] to map the vector X_{gloabl} into three different parameter matrices: Query vector, Key vector and Value vector. These three matrices are further decomposed into h (the number of multi-head attention mechanism heads) subspaces, with each head paying attention to the information of the $D/2/h$ dimensions in the input vector. At this point, the three parameter matrices of each head have dimensions, and the attention matrix of each head is obtained by scaling the Dot-Product self-attention mechanism. Then, the attention matrix of each head is concatenated and mapped with the parameter matrix to obtain the final output result of the standard self-attention layer.

Where $\sqrt{d_k}$ represents the dimension of the key vector. The purpose of scaling the attention score by dividing the similarity calculation result by $\sqrt{d_k}$ is to avoid its value being too large in the SoftMax function and prevent the gradient from disappearing. $\mathrm{W^Q}$, $\mathrm{W^K}$, $\mathrm{W^V}$ and $\mathrm{W^O}$ are the learnable parameter matrices of the model; Concat represents the concatenation function; k represents the number of multi-heads; a_i represents the result calculated by the i-th head in the multi-head attention mechanism.

$$C_i^t = \mathrm{Concat}[h_{i-1}^{t-1}; h_i^{t-1}; h_{i+1}^{t-1}; h_i; S^{t-1}] \quad \text{..... (11)}$$

$$h_i^t = \mathrm{MultiAttention}(h_i^{t-1}, C_i^t) \quad \text{..... (12)}$$

$$\text{MultiAttention}(q, X_{local}) = Concat(a_1,...a_k)\cdot W^O \text{ .....(13)}$$

$$a_i = \text{Attention}(q\cdot W^Q, K\cdot W^K, V\cdot W^V) \text{ ..... (14)}$$

$$\text{Attention}(q,K,V) = Softmax(\frac{q\cdot K^T}{\sqrt{d_k}})\cdot V \text{ ..... (15)}$$

In the local attention mechanism, for a vector sequence $H=X_{\text{local}} \in \mathrm{R}^{N\times D/2}$, $H=[h_1,..., h_N]$, where $h_i \in \mathrm{R}^{1\times D/2}$ represents the high-dimensional embedding of the i-th sample node. Introducing $S \in \mathrm{R}^{1\times D/2}$ represents a virtual sample data, with an initialization state of S^0=average(H) and $H^0=H$. Assuming that the encoder in this paper has a T-layer, where S^t and H^t represent the states of the t ($t \in [1, T]$) layer sample data and virtual data. The state of each sample data h_i is updated from its adjacent data, mainly including the state h of the upper layer on the left and right data $h_{i-1}^{t-1} \in \mathrm{R}^{1\times D/2}$ and $h_{i+1}^{t-1} \in \mathrm{R}^{1\times D/2}$. State $S^{t-1} \in \mathrm{R}^{1\times D/2}$ of the layer above the virtual data. State h of the previous layer on this node $h_i^{t-1} \in \mathrm{R}^{1\times D/2}$ and the initial state of this data $h_i \in \mathrm{R}^{1\times D/2}$. Splice into C through Concat function $C_i^t \in \mathrm{R}^{5\times D/2}$, as shown in Equation (11) [5]. Calculate the state of h_i in this layer using the multi head attention mechanism.

Where C_i^t represents the context information of the i-th sample data, $q \in \mathrm{R}^{1\times d_{\text{model}}}$ represents the high dimension embedding of each sample data.

Finally, the captured local information X_{local} and global information X_{global} are fused to obtain $X_{\text{trm}} \in \mathrm{R}^{N\times D}$. This achieves the goal of reducing computational complexity and can compensate for the shortcomings of the Transformer model in capturing local information. Then, the layer normalization and Feedforward neural network operation are used to filter out the information that is conducive to fault detection. Finally, through the full connection layer, the D dimension of the data is reduced to one dimension to get $X_{\text{trm}} \in \mathrm{R}^{N\times 1}$. Then use the sigmoid function to convert the final detection results of each sample into values between [0-1], and obtain the detection results of the samples. where y represents the prediction result of the model.

$$X_{trm} = \text{LayerNorm}(\text{ReLu}(X_{trm})) \text{ .....(16)}$$

$$\mathrm{S}(X_{trm}) = \frac{1}{1+e^{X_{trm}}} \text{ ..... (17)}$$

$$y = \begin{cases} 1, & \mathrm{S}(X_{trm}) \geqslant 0.5 \\ 0, & \mathrm{S}(X_{trm}) < 0.5 \end{cases} \text{ ..... (18)}$$

Simulation of improved Transformer model:

Dataset Introduction:

Penicillin is the first natural antibiotic, which is mainly used for the treatment of pathogenic microbial infections. So far, penicillin and its Semisynthesis antibiotics are still the most widely used antibiotics. The production process mainly includes three stages: the first stage is the bacterial growth stage, which is relatively short in time. After a short period of adaptation, the bacterial body begins to grow, develop, and reproduce until the concentration of the bacterial body in the culture medium reaches the critical concentration for penicillin synthesis. At this time, the fermentation process transitions to the second stage, which is the penicillin synthesis stage, where the penicillin synthesis rate reaches its maximum value. The third stage is the bacterial autolysis stage, at this stage, the bacterial body begins to age and die, and the rate of penicillin synthesis decreases [6]. The dataset used in this paper was collected from the PensimV2.0 simulation system [7], which was developed by the Process Modeling, Monitoring, and Control Research Group led by Cinar at the Illinois Institute of Technology from 1998 to 2002. This simulation system is specifically designed for the penicillin fermentation process, and relevant research has shown the practicality and effectiveness of this simulation platform. Therefore, it has become an internationally influential penicillin simulation platform. The specific simulation system structure is shown in Figure 2. The simulation platform can not only simulate the actual penicillin fermentation process under normal Initial condition, but also set several common faults and simulate under this condition.

There are a total of 17 variables involved in the penicillin fermentation process, as detailed in Table 1. Currently, PensimV2.0 can set three types of faults: 1. Aeration rate; 2. Agitator power; 3. Substrate feed rate. There are two types of fault disturbances: step and slope, and the amplitude, introduction time, and termination time of the two disturbances can be further set. This feature enriches the functionality of the simulation platform, and literature [8] conducted preliminary research on process monitoring based on PensimV2.0. The parameter values of this simulation are set as the default settings for initial condition, set points and temperature controller types. The variables causing the failure are aeration rate and agitator power. The fault types are all step type. The training data adds 6% step disturbance at the 100th hour, and the simulation platform runs for 400h; All test datasets were subjected to a 6% step disturbance at the 85th hour, and the simulation platform runs for 400h. The sampling interval is 0.2h. The specific description is shown in Table 2.

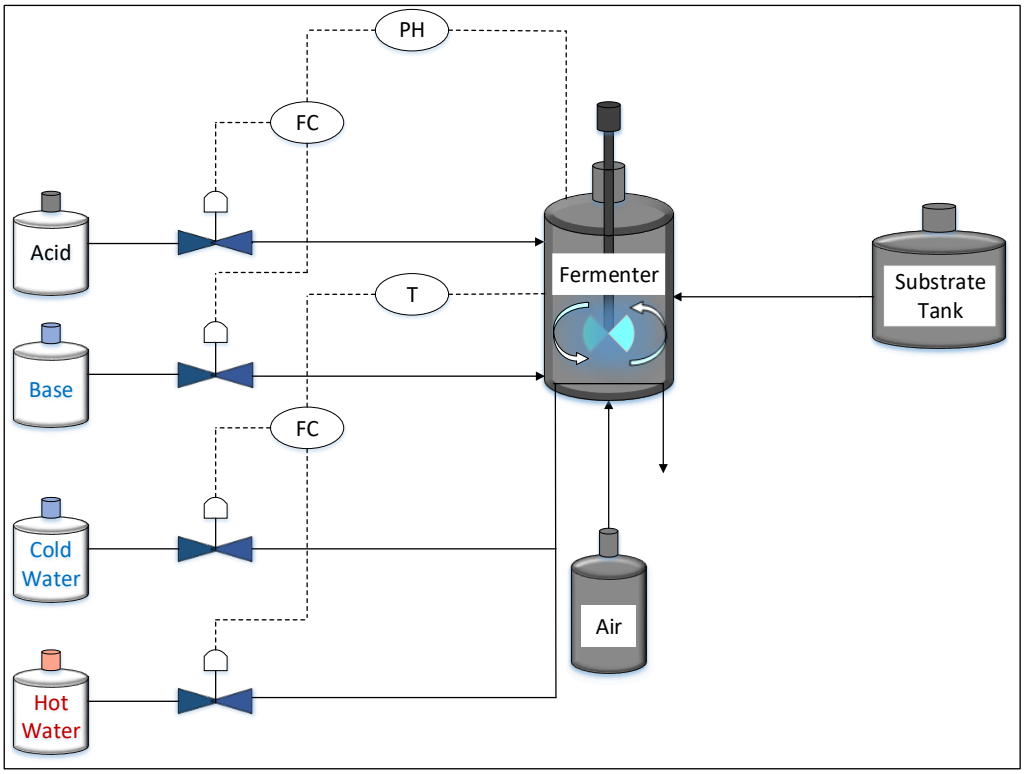

Figure 2: Reaction process of penicillin simulation platform

Table 1: Introduction to Variables in Penicillin Fermentation Process

NO.	Variable Description	NO.	Variable Description
1	Aeration rate	10	CO_2 concentration
2	Agitator power	11	PH
3	Substrate feed rate	12	Temperature
4	Substrate feed temperature	13	Generated heat
5	Substrate concentration	14	Acid flow rate
6	DO saturation	15	Base flow rate
7	Biomass concentration	16	Cold water flow rate
8	Penicillin concentration	17	Hot water flow rate
9	Culture volume		

Table 2: Explanation of Fault Types in the Dataset of Penicillin Fermentation Process

No.	Fault Variable	Fault Type	Magnitude	Occurrence Time	Termination Time
1	Aeration rate	Step	6%	85h	400h
2	Agitator power	Step	6%	85h	400h

Analysis of simulation results:

In the research, we use accuracy (ACC) and Confusion matrix to evaluate the quality of the model. Accuracy is the proportion of all samples whose categories are correctly classified. It is the most basic evaluation index to evaluate the quality of a model. The accuracy calculation is shown in Equation (19). Where, *TL* represents the total number of test set samples, *TP* represents the number of samples with model prediction of 0 and label of 0, and *TN* represents the number of samples with model prediction of 1 and label of 1. The Confusion matrix represents the specific situation of each classification situation. Assuming that the classification task has *N* categories in total, the Confusion matrix is an $N \times N$ matrix, where the values in the *i*-th row and *j*-th column represent the number of *i*-class samples predicted as *j*-class. The values on the diagonal are predicted correctly. Through the Confusion matrix, we can clearly see how many samples of each type are classified correctly, and which categories these samples are classified into among the samples with wrong classification. T-SNE is a dimensionality reduction technique used to represent high-dimensional datasets in two-dimensional or three-dimensional low dimensional spaces, thereby visualizing them.

$$\mathrm{ACC} = \frac{TP + TN}{TL} \quad \text{..... (19)}$$

Table 3: Experimental results of different models on two types of faults

Fault	MLP [9]	LSTM [10]	Transformer [2]	improved-Transformer
1	83.29%	84.80%	96.40%	97.25%
2	94.48%	87.26%	96.50%	97.55%

As shown in Table 3, through comparative experiments, it was found that the detection accuracy of different models for fault 1 in the penicillin fermentation process was 83.29%, 84.80%, 96.40% and 97.25%, respectively. The detection accuracy rates for Fault 2 are 94.48%, 87.26%, 96.50% and 97.55%, respectively. The standard Transformer has a lower model detection rate than the improved Transformer model due to insufficient local information extraction capability. The improved Transformer model proposed in this paper performs best in both fault types due to the fusion of local information extracted by local attention mechanism and global information extracted by standard attention mechanism, which is 0.85% and 1.05% higher than the standard Transformer model, respectively.

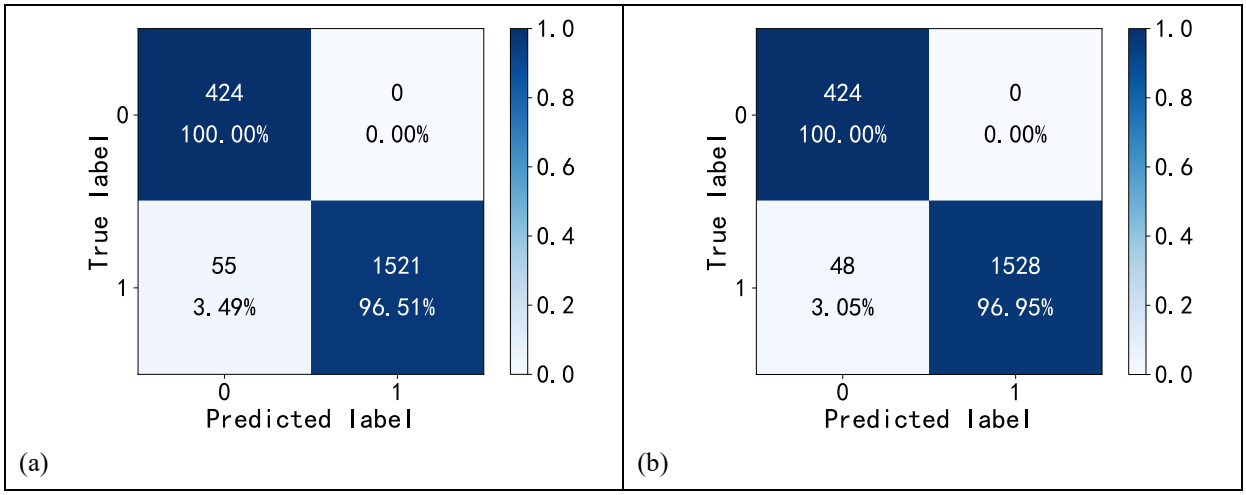

Figure 3: Confusion matrix of improved model in fault 1 and fault 2 datasets (a) Fault 1; (b) Fault 2.

In the confusion matrix, the number above the matrix in *i*-th row and *j*-th column represents the number of class *i* samples predicted by the model as class *j*, while the number below represents the percentage of class *i* samples predicted by the model as class *j* to the total number of class *i* samples. Therefore, the diagonal represents the correct number and accuracy of class *i* classification. The visualization results of the Confusion matrix of the two fault types are shown in Figure 3(a) and 3(b), we can observe that the improved Transformer model can distinguish the normal samples of the two fault types well. So the improved Transformer model proposed in this paper has achieved good detection results on both types of fault datasets.

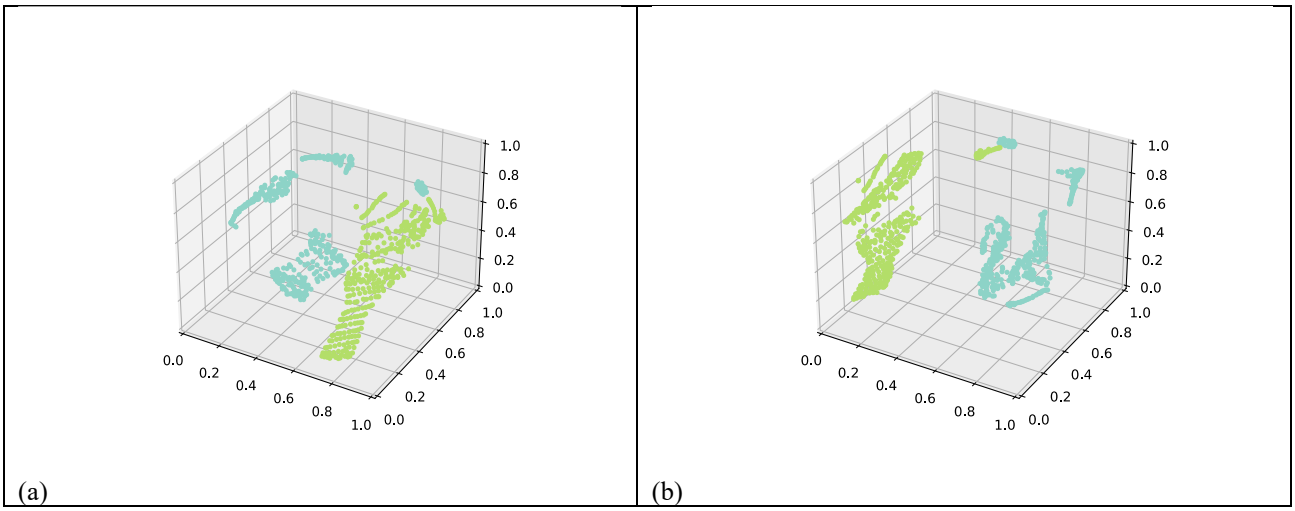

Figure 4: T-SNE diagram of Fault 1 dataset before and after processing by encoder layer (a) before, (b) after

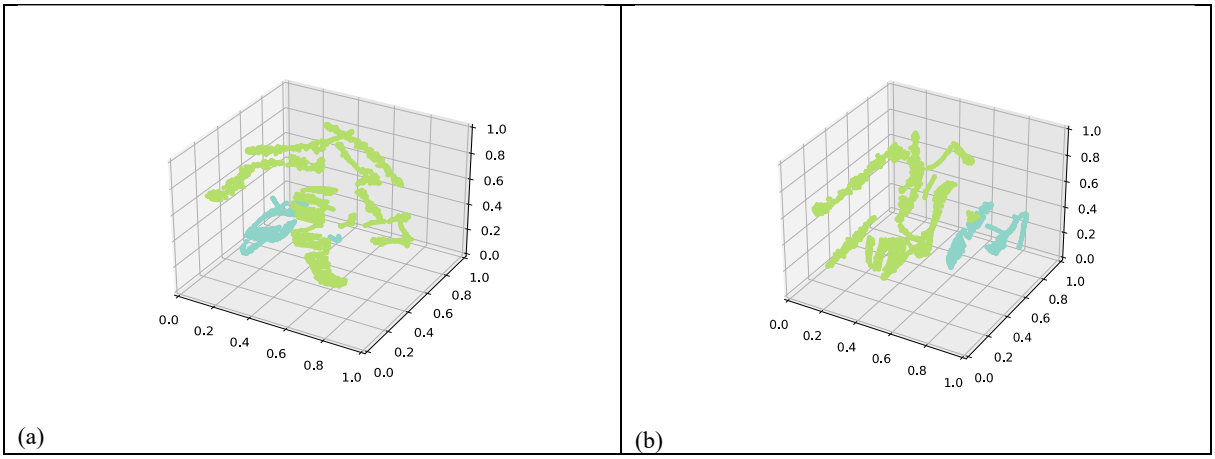

Figure 5: T-SNE diagram of Fault 2 dataset before and after processing by encoder layer (a) before, (b) after

As shown in Figures 4(a) and 5(a), it can be observed that the t-SNE distribution of two types of fault samples and normal samples is disorderly and some data overlaps before being processed by the encoder layer, which will have a certain negative impact on the fault detection results; As shown in Figures 4(b) and 5(b), after being processed by the encoder layer, the overlapping parts become almost non-existent, and samples of the same category become more and more concentrated, that is, the separability of features becomes stronger. This indicates that after being processed by the encoder layer, local and global information between data can be effectively extracted, and the detection ability of the test set is improved.

SUMMARIZE

Due to the high dimensionality and nonlinearity of process data in industrial production, traditional process monitoring methods based on data-driven methods are difficult to achieve good monitoring results. This paper proposes a Transformer model based on local and global attention mechanisms. Firstly, the data is bisected in the feature dimension. One part of the data uses the local attention mechanism to capture the sequence to obtain local information, while the other part uses the standard attention mechanism to capture the global information of the sequence. Then, the extracted local information is fused with the global information. The above methods can not only reduce the computational complexity of the model to a certain extent, but also make up for the lack of local information extraction capability of Transformer model. Finally, by using the PensimV2.0 simulation platform to collect data on the penicillin fermentation process for simulation experiments, the experimental results showed that the improved Transformer model had improved fault detection efficiency compared to the current mainstream algorithms MLP, LSTM and standard Transformer, confirming the feasibility and effectiveness of the improved model in the field of fault detection.

REFERENCES

1. Abbasi, M. A., Khan, A. Q., Mustafa, G., Abid, M., Khan, A. S., & Ullah, N. (2021). Data-Driven Fault Diagnostics for Industrial Processes: An Application to Penicillin Fermentation Process. IEEE Access, 2021.
2. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.
3. Ranathunga, S., Lee, E. S. A., Prifti Skenduli, M., Shekhar, R., Alam, M., & Kaur, R. (2023). Neural machine translation for low-resource languages: A survey. *ACM Computing Surveys*, *55*(11), 1-37.
4. Liu, L., Liu, X., Gao, J., Chen, W., & Han, J. (2020). Understanding the difficulty of training transformers. *arXiv preprint arXiv:2004.08249*.
5. Qipeng, G., Xipeng, Q., Pengfei, L., Yunfan, S., Xiangyang, X., & Zheng, Z. (2019). Star-Transformer North American Chapter of the Association for Computational Linguistics.
6. Yu, J. (2012). Multiway discrete hidden Markov model-based approach for dynamic batch process monitoring and fault classification. *AIChE journal*, *58*(9), 2714-2725.
7. Liu, Y., & Wang, H. Q. (2006). Pensim simulator and its application in penicillin fermentation process. *Journal of System Simulation*, *18*(12), 3524-3527.
8. Ündey, C., Tatara, E., & Çınar, A. (2004). Intelligent real-time performance monitoring and quality prediction for batch/fed-batch cultivations. *Journal of Biotechnology*, *108*(1), 61-77.
9. Pinkus, A. (1999). Approximation theory of the MLP model in neural networks. *Acta numerica*, *8*, 143-195.
10. Yu, Y., Si, X., Hu, C., & Zhang, J. (2019). A review of recurrent neural networks: LSTM cells and network architectures. *Neural computation*, *31*(7), 1235-1270.