IOP Publishing Measurement Science and Technology
Meas. Sci. Technol. **31** (2020) 065014 (12pp) https://doi.org/10.1088/1361-6501/ab7280

Fault diagnosis of reciprocating compressor based on group self-attention network

Ganchao Bao, Hongli Zhang, Yuan Wei, Dan Gu and Shulin Liu

School of Mechatronics Engineering and Automation, Shanghai University, Shanghai 200444, People's Republic of China

E-mail: zhang40941@126.com

Received 30 November 2019, revised 16 January 2020
Accepted for publication 3 February 2020
Published 2 April 2020

Abstract

Reciprocating compressors are widely used in the petroleum industry and because of their complex and nonlinear signals, it is difficult to extract fault features. Recently, deep learning has been used in intelligent mechanical fault diagnosis and achieved great success. In the deep learning model, the recursive neural network (RNN) can capture global features, but it is difficult to parallelize and not good at dealing with long sequences. The convolutional neural network (CNN) can capture local features, but its receptive field is limited by the number of layers of the network and the size of the sliding window, resulting in the model not capturing sufficient features. In this paper, we propose a deep learning model without any RNN or CNN structures, called the group self-attention network (GSAN), for fault diagnosis of multisource signals in reciprocating compressors. The GSAN model mainly includes intra-group self-attention, inter-group self-attention and a fusion gate. Among them, intra-group self-attention is used to capture local features within a group, inter-group self-attention is used to capture global features between groups, and the fusion gate finally integrates these features. Experimental results show that compared with other models based on the RNN or the CNNs, the GSAN proposed in this paper not only has higher prediction accuracy, but also better anti-noise performance. In addition, the effectiveness of each part of the model is verified by ablation experiment.

Keywords: reciprocating compressor, fault diagnosis, deep learning, self-attention

(Some figures may appear in colour only in the online journal)

1. Introduction

Due to some parts of reciprocating compressors running at high temperature, high pressure and other adverse factors, once a fault occurs, not only will production be interrupted, but also injuries may be caused. Therefore, fault diagnosis of reciprocating compressors is of great significance. There are many rotating parts and vibration sources in reciprocating compressors, so the fault signal usually presents complex nonlinearity [1,2]. The methods used in other types of machinery are often not applicable to reciprocating compressors. In addition, because the faults in reciprocating compressors are affected by a variety of parameters, in most cases, one signal does not reflect the potential fault [3,4]. The collected signals usually include vibration, temperature, pressure and other parameters. The relationship between these parameters is very complex and they affect each other, so it is necessary to conduct comprehensive fault identification of multi-source signal data.

Early fault diagnosis methods for reciprocating compressor depended on traditional signal processing methods. Yang *et al* [5] used discrete wavelet transform to extract the four orders of statistical features, and used an artificial neural network and a support vector machine to identify the faults in a small reciprocating compressor. Tang *et al* [6] used adaptive peak decomposition to extract the features of vibration signals in a reciprocating compressor in four states. Qi *et al* [7] proposed the sparse code for the operation data of a reciprocating compressor and then used a support vector machine to identify faults. However, traditional feature

extraction methods need a lot of domain knowledge and prior knowledge. In addition, the accuracy of classification results largely depends on the extracted features.

In recent years, deep learning has gained widespread attention and success in many fields, such as computer vision and natural language processing [8–11] . It is an end-to-end recognition method, which can eliminate feature extraction steps and it has employed in mechanical fault diagnosis. For example, Jiang *et al* proposed an improved convolutional deep belief network [12] that combined feature learning with fault diagnosis of compressed sensing, and proposed the deep wavelet auto-encoder (DWAE) [13] based on an extreme learning machine. Mao *et al* [14] integrated an automatic encoder and a multi-layer extreme learning machine for fault diagnosis. Shao *et al* [15] put forward a continuous deep belief network with locally linear embedding that can accurately predict mechanical performance trends. Many experiments show that the deep learning model performs better than traditional algorithms in terms of accuracy and generalization [16–20]. As can be seen from previous literatures, the deep learning model is gradually becoming the mainstream algorithm for fault diagnosis.

In the deep learning model, the convolution neural network (CNN) is designed to handle image or time series data [21–25]. Jing *et al* [26] proposed a CNN to learn features from the original data and directly diagnosed the original vibration signal. Zhang *et al* [19] proposed a one-dimensional deep convolutional neural network (1d-DCNN) that displayed good real-time and generalization performance. The recurrent neural network (RNN) and its associated variants, included long-short-term memory (LSTM) [27] and gated recursive units (GRUs) [28], can capture information on a time series. Pan *et al* [29] proposed the LSTM-CNN combined one-dimensional CNN with LSTM that has the classification ability of the CNN and the temporal coherence representation ability of LSTM. With the development of deep learning, the limitations of RNNs and CNNs began to appear. The RNN model is difficult to parallelize, and it is difficult to capture remote dependencies when the input sequence is too long. The receptive field of the CNN is limited by the number of network layers and the size of sliding window, so the model cannot capture enough features.

The attention mechanism has been used as an aid to help improve the RNN and the CNN, rather than as a single layer of the network. Recently, Vaswani *et al* [30] proposed the first completely attention-based model for machine translation and achieved the best performance. Compared to the CNN and the RNN, the attention-based model is flexible in extracting both remote and local correlations. It turns out that without the CNN or the RNN, the attention-based model hasa powerful ability in feature extraction and performs well in some tasks. However, as far as we know, few attention-based models for fault diagnosis methods have been proposed.

The contributions of our work are briefly outlined as follows:

(a) We propose the group self-attention network (GSAN), a completely attention-based model, for fault diagnosis in reciprocating compressors. In the design of the model, we used the grouping strategy to extract eatures within and between groups step by step. First, intra-group self-attention is used to capture local features within a group. Then, inter-group self-attention is used to capture global features between groups.
(b) We designed a fusion gate to fuse the local and global feature vectors toimprove the anti-noise ability and robustness of the model.
(c) The GSAN model abandons the classical network RNN and CNN, and processes the multi-source signal of reciprocating compressor based on a self-attention network, providing a new solution for some fields of fault diagnosis. Experimental results show that the GSAN has higher accuracy and antinoise ability than the CNN and RNN models for fault diagnosis in a reciprocating compressor.

2. Background

2.1. Attention mechanism

The attention mechanism is essentially a means to calculate the alignment score between elements in sequences. Given a sequence $x = \{x_1, x_2, \ldots, x_n\}$ and a query vector q, the attention mechanism measures the degree of correlation between x_i and q by the compatibility function $f(x_i, q)$.

The compatibility function $f(x_i, q)$ can be calculated in a variety of ways, the most commonly used of which is additive attention [31,32], as shown in the following equation:

$$f(x_i, q) = W^T \tanh\left(W^1 x_i + W^2 q + b^1\right) \tag{1}$$

$$a = [f(x_i, q)]_{i=1}^{n} \tag{2}$$

where $x_i \in \mathbb{R}^{d_k}$, $q \in \mathbb{R}^{d_k}$ are the vectors of input, W^T, W^1, $W^2 \in \mathbb{R}^{d_k \times d_k}$, $b^1 \in \mathbb{R}^{d_k}$ are the parameters, a$\in \mathbb{R}^{n \times n}$ is the alignment score.

Then the softmax function normalizes all attention values of x and converts the alignment score $[f(x_i, q)]_{i=1}^{n}$ into a probability distribution $p(z = i|x, q)$. The output s is the weighted sum of each element x_i, as shown below:

$$p(z = i|x, q) = \text{softmax}(a) = \frac{\exp(f(x_i, q))}{\sum_{i=1}^{n} \exp(f(x_i, q))} \tag{3}$$

$$s = \sum_{i=1}^{n} p(z = i|x, q)\, x_i \tag{4}$$

where $s \in \mathbb{R}^{d_k}$ can be used as the attention vector corresponding to x_i.

2.2. Self-attention

Self-attention is a special case of the above-described attention mechanism, as shown in figure 1. The only difference between self-attention and other attention mechanisms is its compatibility function. It replaces the query vector q with the input

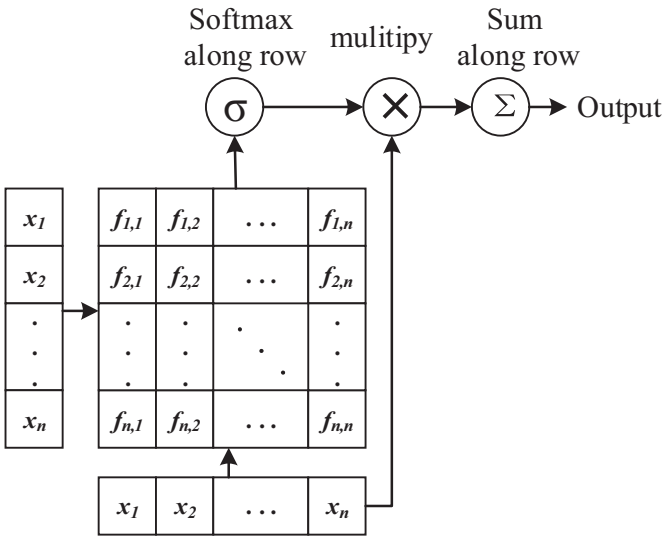


Figure 1. Diagram of self-attention.

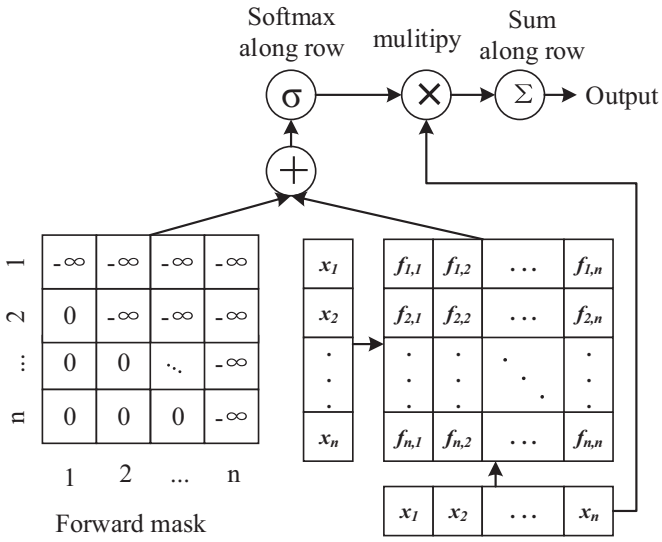


Figure 2. Diagram of self-attention with a forward mask.

sequence itself. Through calculating the attention weight between each pair of elements x_i and x_j, self-attention can correlate information about different positions in a sequence. So, the compatibility function can be described as the following:

$$f(x_i, x_j) = W^T \tanh\left(W^1 x_i + W^2 x_j + b^1\right) \tag{5}$$

where $x_i \in \mathbb{R}^{d_k}$, $x_j \in \mathbb{R}^{d_k}$ are the input elements at different positions.

Self-attention is good at capturing both local and remote features in a sequence. Compared with the RNN, self-attention has faster computing speed and fewer parameters. Compared with the CNN, it has a better ability to extract global features. Recently, we have witnessed its success in some natural language processing tasks, such as neural machine translation (Vaswani *et al* [30]) and reading comprehension (Hu *et al* [33]).

2.3. Two variants of self-attention

In this article, our model will use two variations of self-attention. The first one is masked self-attention [30], as shown in figure 2. When the original self-attention calculates the attention weight of the x_i and x_j, it ignores the sequential relationship between them because the attention weight of x_i to x_j is the same as the attention weight of x_j to x_i. However, the sequence information is very important for vibration signals in fault diagnosis. So, by adding a forward mask and a backward mask to the original attention matrix, masked self-attention can calculate the attention weight from the sequence information in the forward and backward directions.

These two masks are defined as:

$$M_{ij}^{fw} = \begin{cases} 0, & i < j \\ -\infty, & \text{otherwise} \end{cases} \tag{6}$$

$$M_{ij}^{bw} = \begin{cases} 0, & i > j \\ -\infty, & \text{otherwise} \end{cases}. \tag{7}$$

In the forward attention calculation, the attention matrix is added with the forward mask and only the information before this position is considered, as shown in figure 2. Because in the forward mask matrix the value of the upper right triangle is $-\infty$, that means the weighted probability value after logarithm calculation is 0. The backward mask matrix in the backward attention calculation is just the opposite.

The compatibility function after adding the mask to self-attention is as follows:

$$f(x_i, x_j) = W^T \tanh\left(W^1 x_i + W^2 x_j + b^1\right) + M_{ij}. \tag{8}$$

Another variant of self-attention is multi-dimensional self-attention [34], which captures the relationship between x_i and the whole sequence x. Compared with the original self-attention, it calculates the function of attention weight in a different way. The query vector q is deleted from equation (1), so the compatibility function can be shown as below:

$$f(x_i) = W^T \tanh\left(W^1 x_i + b^1\right) + b. \tag{9}$$

In this paper, the two types of self-attention play completely different roles. Masked self-attention can calculate the relationship between the positions of a sequence in the forward and backward directions. Multi-dimensional self-attention can compress a multi-dimensional vector into a one-dimensional vector representation.

3. Proposed GSAN model

3.1. Overall architecture

In this paper, the group self-attention network (GSAN) for fault diagnosis in reciprocating compressors is based on self-attention, as shown in figure 3. The GSAN mainly consists of three core parts: intra-group self-attention, inter-group self-attention and the fusion gate. We will introduce our network model layer by layer, starting with the input of the model.

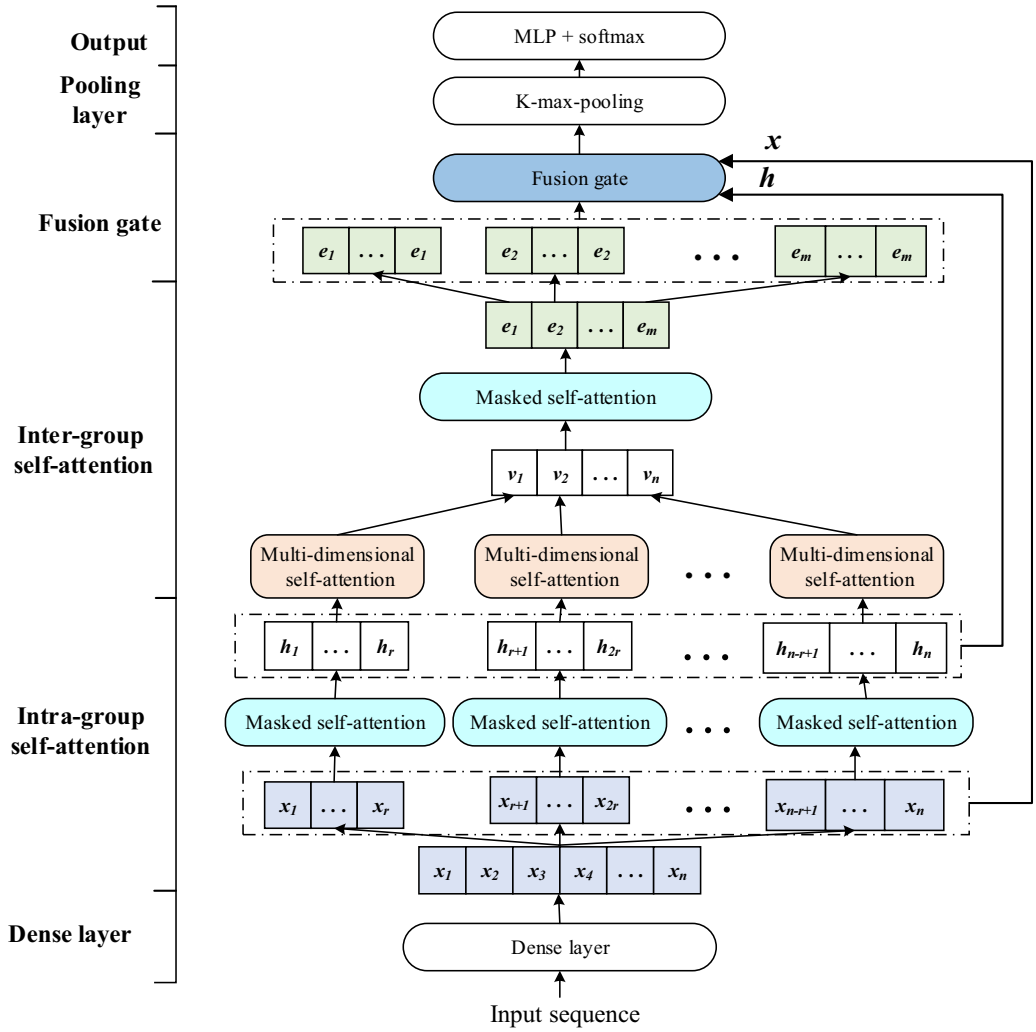


Figure 3. The architecture of the group self-attention network model.

3.2. Dense layer

Since there are eight channels in the multi-source signal of a reciprocating compressor, the data size is $n \times 8$. The dense layer merges information from these channels and transforms the size of the data to $n \ n \times d_e$. In a dense layer, every position in the sequence has the same fully connected calculation and the rectified linear unit (ReLU) activation function. The output can be calculated as the following:

$$\mathrm{D}(x_i) = \max(0, x_i W + b) \tag{10}$$

where $x_i \in \mathbb{R}^{1\times 8}$ represents the value of the eight channels at ith time and $W \in \mathbb{R}^{8\times d_e}, b \in \mathbb{R}^{d_e}$ are the parameters to train.

3.3. Intra-group self-attention

After the dense layer, the sequence x is equally divided into m parts $\{g^1, g^2, \ldots, g^m\}$, where $g^1 = \{x_1, x_2, \ldots, x_r\}$, $g^2 = \{x_{r+1}, x_{r+2}, \ldots, x_{2r}\}$ and $g^m = \{x_{n\text{-}r+1}, x_{n\text{-}r+2}, \ldots, x_n\}$. Each group in the intra-group attention layer has a masked self-attention of shared parameters.

Then, like other attention mechanism, we calculate the compatibility function $f(x_i, x_j) \in \mathbb{R}^{d_e\times n}$. The output of masked self-attention is as follows:

$$h_j = \sum_{i=1}^{n} \mathrm{softmax}(f(x_i, x_j)) \odot x_i \tag{11}$$

where h_j represents the vector of the jth position of output h, and $\odot$ denotes element-wise multiplication. The output is $h = [h_1, h_2, \ldots, h_n] \in \mathbb{R}^{d_e\times n}$.

Given a sequence x, by self-attention with a forward mask and a backward mask, two feature vectors h^{fw}, h^{bw} can be obtained and concatenated as $h = [h^{fw}; h^{bw}] \in \mathbb{R}^{2d_e\times n}$, which is the final output of masked self-attention, as shown in figure 4. This idea of bi-direction is similar to Bi-LSTM [35], which also has a bi-direction structure.

With masked self-attention, the correlation of the sequence information in each intra-group is captured and the new vector representation containing local features is output. Meanwhile, dividing the data into groups can reduce the space complexity and memory occupation of the algorithm and make the model lighter.

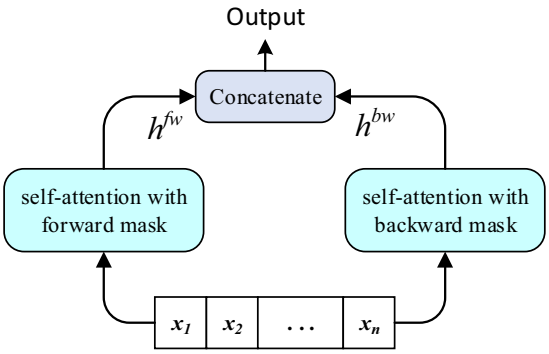


Figure 4. Diagram of masked self-attention.

3.4. Inter-group self-attention

In inter-group self-attention part, we used both multi-dimensional self-attention and masked self-attention. First, multi-dimensional self-attention is applied in each group respectively and it outputs a vector v for each group. According to the compatible function of multi-dimensional self-attention and input h, the output v is calculated as below:

$$v_j = \sum_{i=1}^{n} \mathrm{softmax}(f(h_i)) \odot h_i \tag{12}$$

$$v = \sum_{j=1}^{n} v_j \tag{13}$$

where v is the self-attention output of the sequence in one group.

Multi-dimensional self-attention compresses a piece of sequence in each group into a vector. Then, the masked self-attention calculates the attention weight of vectors of different groups and output global feature vector $e = [e_1, e_2, \ldots, e_m] \in \mathbb{R}^{d_e \times m}$. Then we duplicate e_i for r times to get $[e_i, e_i, \ldots, e_i]$ to represent the vector for each group.

3.5. Fusion gate

After intra-group self-attention and inter-group attention, the network can capture local and global feature information. However, due to the complex work environment of reciprocating compressors, vibration signals in the running process often encounter serious noise, so the input sequence may contain a large amount of useless information. While extracting global or local features in the network, some valuable information may be ignored and irrelevant noise may be amplified. Therefore, it is necessary to design an effective fusion gate to fuse the output from different layers and retain more valuable information with less noise. The structure of the update gate in a GRU [28] can effectively fuse information between the current input a_t and the output of the previous step y_{t-1}. Thus, the update gate can be regarded as a filter that removes noise and retains valuable feature information. As shown in figure 5, the update gate has two main parts: the sigmoid part determines which value should be updated, and the tanh part creates

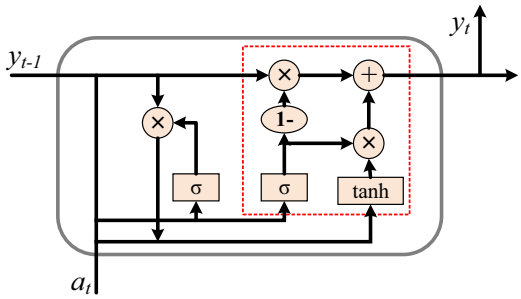


Figure 5. Diagram of a cell of the GRU. The red part is the update gate.

a new candidate vector y_t. The fusion gate in our model has a similar structure to the update gate, which can fuse information about three kinds of feature vectors.

Based on the structure of an update gate, we designed a fusion gate. The fusion gate differs from the update gate in that it has three inputs, including the original input sequence x, the local feature vector h and global feature vector e, as shown below:

$$F = \tanh\left(W^{f1}[x;h;e] + b^{f1}\right) \tag{14}$$

$$G = \sigma\left(W^{f2}[x;h;e] + b^{f2}\right) \tag{15}$$

$$u = G \odot F + (1 - G) \odot h \tag{16}$$

where σ is the sigmoid activation function, G is a value between 0 and 1, $\odot$ denotes element-wise multiplication, and $u = [u_1, u_2, \ldots, u_n] \in \mathbb{R}^{d_e \times n}$ is the output of the fusion gate, consisting of attention representations of n elements.

3.6. K-max-pooling

After the sequence has been subjected to local and global attention calculations, we can get a new vector representation u of the sequence. Then k-max-pooling layer compresses the sequence u and retains the important information. The k-max-pooling is extended form max-pooling, which extracts the elements of the top k largest value in the dimension of time and outputs a one-dimensional vector. The k-max-pooling retains more important information than max-pooling and reduces information loss. Finally, the sequence enters the output layer with a dense layer and the softmax function.

4. Experiment

4.1. Data description

Multisource signal data with eight channels was collected from the No. 1 working unit of the south district compressor station of the Daqing Natural Gas Company. Thus, how to fully integrate the information from these channels is very important and is somthing that our model GSAN is good at. In order to collect experimental data under different working conditions, the second cylinder of No. 1 working unit of

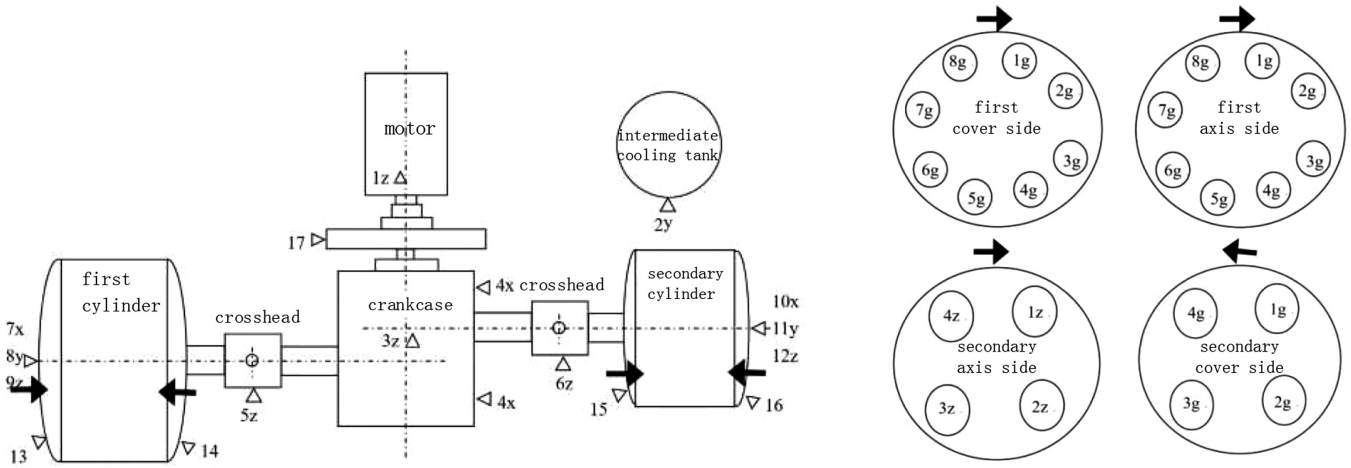


Figure 6. Structure schematic diagram of reciprocating compressor and measuring point layout of first valve, secondary valve.

Daqing Natural Gas Company was equipped with valves in four states, including normal state, valve piece fracture, spring damage and valve piece notch. During the experiment, the secondary gas outlet pressure is 1040 kPa, the inlet pressure is 310 kPa, the outlet temperature is 104 °C, the inlet temperature is 32 °C and the gas flow is 3611 m^3 h^{-1}. The acquisition system includes an INV306U-6660 intelligent data acquisition processing analyzer and INV-1021 program-controlled multifunctional signal disposal apparatus from the Beijing Oriental Institute of Vibration and Noise Technology. The sampling frequency is set to 20 kHz. The collected data has eight channels, including acceleration 1, acceleration 2, acceleration 3, pressure 1, pressure 2, pressure 3, pressure 4 and key-phase. As shown in figure 6, the three acceleration sensors are respectively arranged on the secondary valve 4g, 3g and 3z, and the four pressure sensors are respectively arranged on the first valve 3g, 6z and the secondary valve 4z, 1g. 120 000 data points for each channel in the normal valve state, and 80 000 data points are collected in the other three valve fault states. Figure 7 shows the collected data for eight channels of a valve piece fracture.

Since the GSAN model requires enough samples with labels to train, if there are not enough training samples, it is easy to fall into the trap of overfitting. In order to obtain a large number of training and test samples, the data acquisition process uses a partial overlap cutting method, as shown in figure 8. When the length of the original signal is constant, the number of samples depends on the sliding length and the length of each sample. If the sliding stride is too small or too large, the information between the samples will be highly redundant or the number of samples may not be enough. In this paper, sliding length is set to 50 data points and each sample length is 1024 data points; a total of 7120 samples are obtained. The ratio of training set, verification set and test set is 70%, 20% and 10%, respectively, as shown in the table 1.

Table 1. Statistics of training set, validation set and test set.

Fault type	Train	Valid	Test
Valve piece fracture	1106	316	158
Spring damage	1106	316	158
Valve piece notch	1106	316	158
Normal	1666	476	238

4.2. Comparison with other models

In order to evaluate the performance of the proposed GSAN model for fault diagnosis of reciprocating compressors, a series of comparative experiments were carried out. All the tested algorithms were coded in Python and executed on a computer with an Intel Core i7-7700 CPU and 16 GB RAM. Table 2 reports the experimental results and structural parameters of different models for fault diagnosis in reciprocating compressors.

As is shown in table 2, all of the models have the same input layer with size of 1024 × 8, which is the shape of fault data for reciprocating compressors. The 1d-DCNN model is mainly composed of multiple one-dimensional convolution and pooling layers. In the all pooling layers, the size of the filter is 2 × 1, and the stride is 2. But the convolution parameters for each layer are slightly different. For example, in the first convolution layers, the size of the filters is 16 × 1 and stride is 1. The LSTM-CNN model is mainly composed of convolution layers and an LSTM layer with 32 cells. Our GSAN model has fewer hyper-parameters than the other two models, and two main ones. Firstly, in group masked self-attention, when the sequence is divided into 16 groups, the model performs best. In the *k*-max-pooling layer, *k* is set to 8 in the experiment. There are four fault types of reciprocating compressors in the experiment, so the number of neurons in the output layer of all models is four. In these models, softmax is used to be their final activate function, the other activation is ReLU, the dropout rate is 0.5 and the batch size is set to 32. All models are optimized using the Adam algorithm. During calculations using the Adam algorithm, the exponential moving means of the gradient are calculated and the decay rate of these moving averages are controlled by hyperparameters β_1 and β_2. In these models, $\beta_1 = 0.9$, $\beta_2 = 0.999$. The initial learning rate is 0.001.

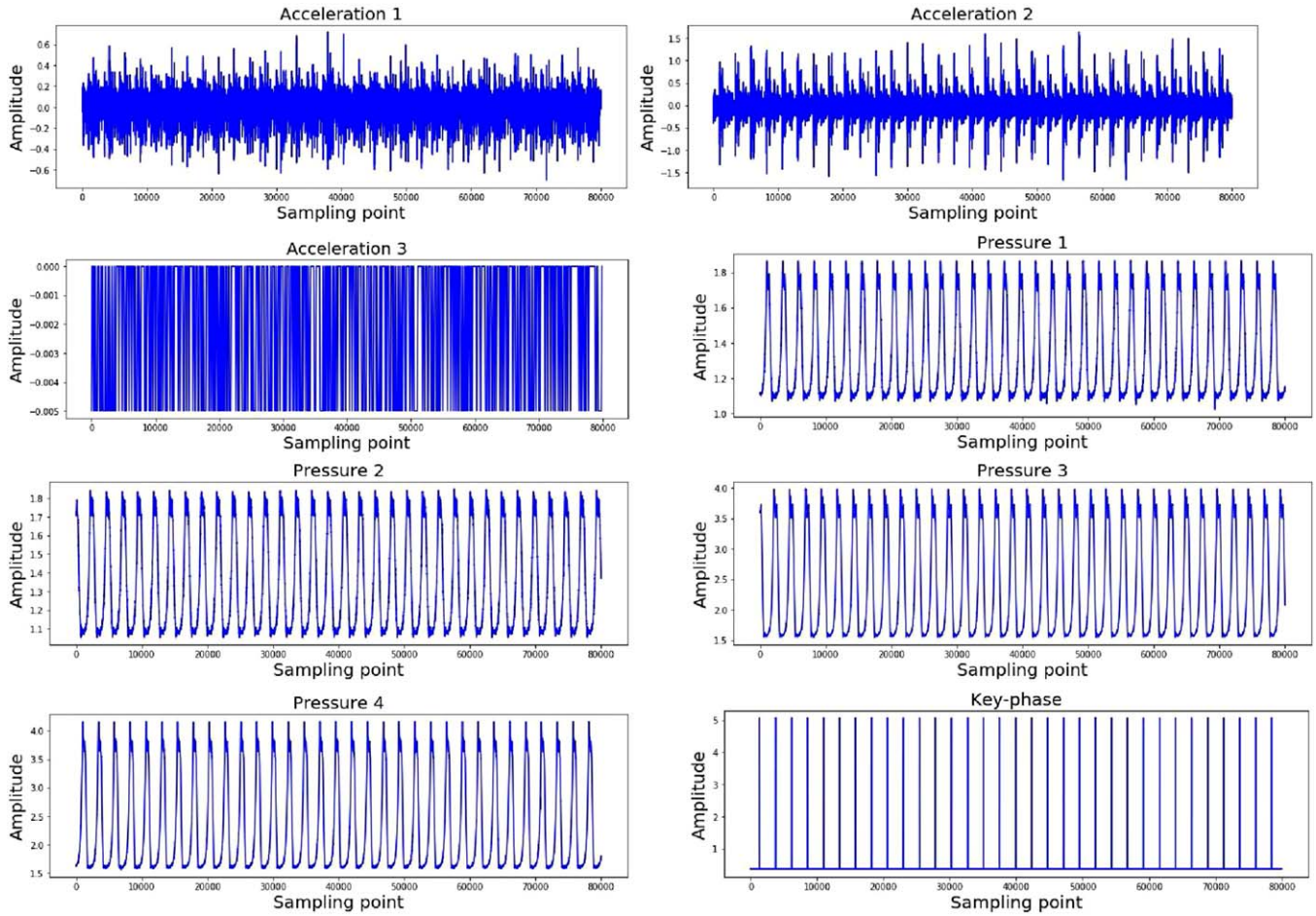


Figure 7. Diagram of collected data for eight channels of a valve piece fracture.

Table 2. Experimental results and structural parameters of different models for fault diagnosis in reciprocating compressors.

Model	1d-DCNN	LSTM-CNN	GSAN
Input layer	1024 × 8	1024 × 8	1024 × 8
Hidden layer	Conv1d (16 × 1, 16) Pooling (2 × 1, 2) Conv1d (3 × 1, 32) Pooling (2 × 1, 2) Conv1d (3 × 1, 64) Pooling (2 × 1, 2) Conv1d (3 × 1, 64) Pooling (2 × 1, 2) Flatten Dense (128)	Conv1d (32 × 1, 16) Pooling (2 × 2, 2) Conv1d (16 × 1, 32) Pooling (2 × 2, 2) LSTM (32) Flatten Dense (128)	Dense (32) Group masked self-attention (16) Multi-dimensional self-attention Masked self-attention Fusion gate K-max-pooling (8) Dense (128)
Output layer	Dense (4) + softmax	Dense (4) + softmax	Dense (4) + softmax
Test accuracy	0.988	0.971	0.993

Figures 9–11 show the training processes of the training and validation sets, and the confusion matrix of the test set under the three models. It can be seen that, after about 40 training epochs, the classification accuracy of all models almost reaches convergence and shows that the deep learning method is very effective for fault diagnosis in reciprocating compressors. However, compared with the 1d-DCNN and the LSTM-CNN, the proposed GSAN model performs more stably. The GSAN not only achieves a very high validation accuracy, but also achieves a 99.3% accuracy rate in the test

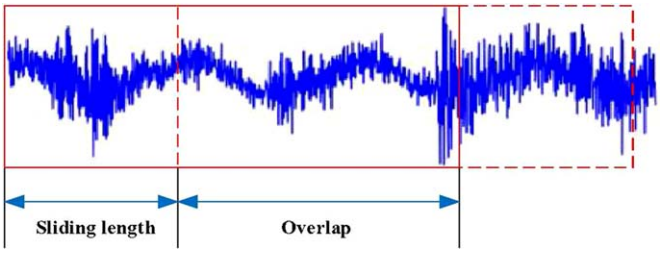


Figure 8. Schematic diagram of the cutting method of the collected data.

set. Meanwhile, the accuracy of results based on the 1d-DCNN and the LSTM-CNN in the test set is slightly lower at 98.8% and 97.1%.

The depth of the three models is almost the same, but the final training results are different. That means the CNN, the RNN and self-attention networks do not have equal feature extraction capabilities and performances for fault diagnosis signals. It can be seen from the above three models that the feature extraction ability of the GSAN is the best, the CNN is the second and the RNN is the worst. Theoretically, the CNN is limited by its fixed sliding window, which causes it to extract only local signal information. Generally, the range of the receptive field of the CNN can be expanded by increasing the depth of the network, but the global feature cannot be completely extracted. The LSTM used in our experiments is a classic RNN whose cyclic structure is capable of extracting global features from sequence signals, but it does not give good experimental results. One possible reason is that the LSTM cannot extract long-distance dependencies and features from long signals, although it has a gating mechanism and a memory cell. In our model GSAN, local features of the signal are caught by the intra-group attention and global features is caught by the inter-group attention. Then these two kinds of feature are combined through the fusion gate. The GSAN is more comprehensive and flexible in extracting features, and fully integrates multisource information from the reciprocating compressor, so the classification results are better than those from the other two models.

4.3. Performance under noise

In practical applications, the working environment of the machine is different. In order to simulate the performance of the model under different working environments and verify the generalization ability of the models, we added noise to the signals in the test set, tested them under different signal-to-noise ratios (SNR)s. The definition of SNR is as follows:

$$\mathrm{SNR} = 10\log_{10}\left(\frac{P_{\mathrm{signal}}}{P_{\mathrm{noise}}}\right). \tag{17}$$

In this paper, we test the noise-added signals of 2 dB, 4 dB, 6 dB and 8 dB respectively. Figure 12 shows the waveforms of the noisy signal with original signal, noise and noise-added signals of 4 dB.

Table 3. Result of different ablation models.

Model	Accuracy
GSAN (full model)	0.993
Without intra-group self-attention	0.709
Without inter-group self-attention	0.942
Without fusion gate	0.915

Figure 13 shows the comparison results of the accuracy of the three models for the reciprocating compressor valve diagnosis under differing test conditions with different SNR. It can be seen that the classification accuracy of each model decreases significantly as SNR decreases. When SNR = 8 dB, the accuracy of the three models is above 95%, and the difference is small. However, with the decline of SNR, the gap between the GSAN model and the other two models becomes larger. When the SNR value of the test set was reduced from 8 dB to 2 dB, the accuracy of the 1d-DCNN model decreased the most, from 98.8% to 73.4%. The accuracy of the LSTM-CNN model also dropped by 21.7%. The accuracy fluctuation of the GSAN model is minimal, down 13.5%. On test sets with different SNR values, the proposed GSAN model achieved the highest classification accuracy. It shows that the GSAN has good denoising ability and generalization ability, while the 1d-DCNN model has large decrease in the case of high noise.

From the result, we can know that the proposed GSAN model has a stronger ability to extract features and avoid local false features caused by environmental noise. In order to visually understand the ability of the proposed GSAN model to extract features, we extract the output of the last fully connected layer on the test set data of different noise-added signals and visualize it with t-SNE [36], as shown in figures 14–17. It can be seen that the GSAN can still extract useful feature information in a variety of SNR signals. When SNR = 8 dB, the feature vectors of samples extracted by the GSAN can be clearly divided into four categories. When SNR = 4 or 6 dB, some samples of valve piece notch will be close to the range of normal samples, and some samples of spring damage will be close to the range of samples of valve piece fracture, but the count is very small. When SNR = 2 dB, the boundaries of four samples of different categories start to become unclear, leading to a decrease in the accuracy of model fault identification. It can be seen that the GSAN model has a good anti-noise ability in the reciprocating compressor signal with noise. When the SNR is larger than 4 dB, it can clearly divide the boundaries of various samples and achieve high accuracy.

4.4. Ablation analysis

In order to demonstrate the effectiveness of the different components in our GSAN model, ablation experiments were designed. Intra-group self-attention, inter-group self-attention or the fusion gate were removed from the original model. All models were trained for 50 epochs and tested respectively. We compared the results of three ablation models and the GSAN (full model), as shown in table 3.

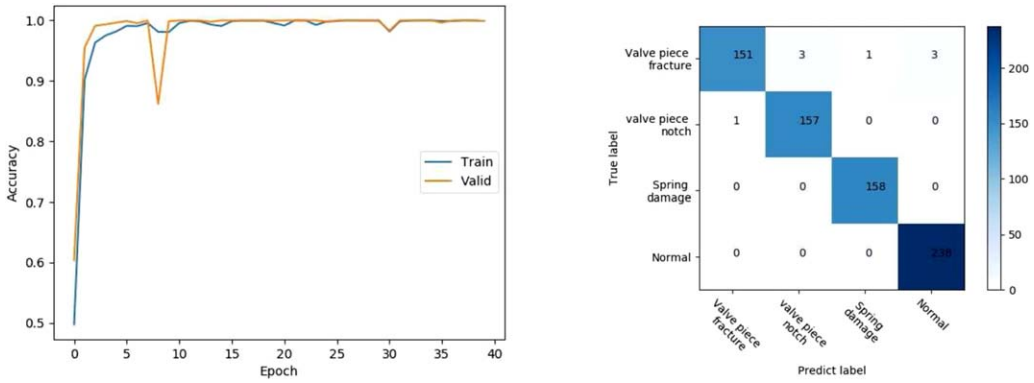


Figure 9. Training process and confusion matrix of the 1d-DCNN.

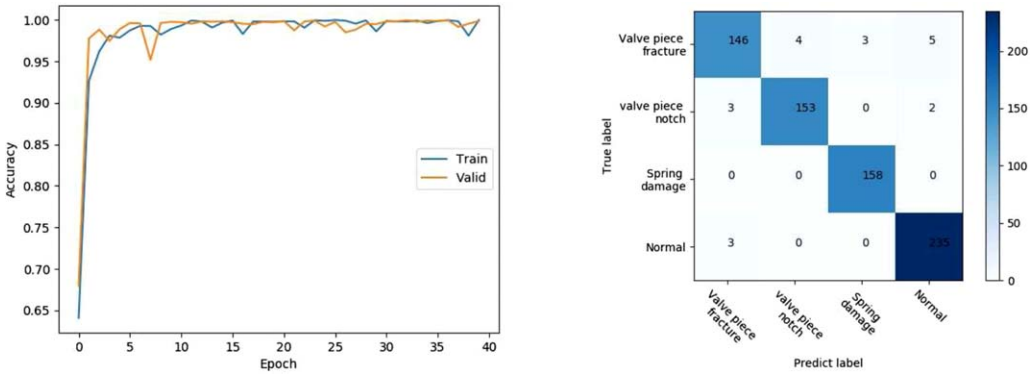


Figure 10. Training process and confusion matrix of the LSTM-CNN.

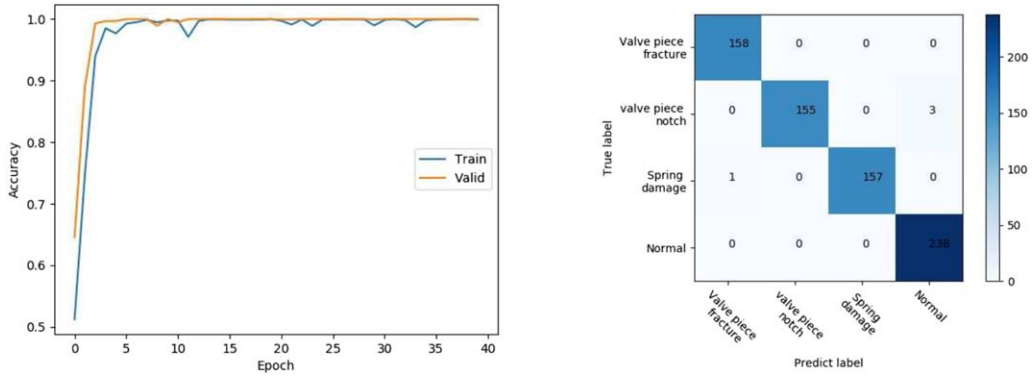


Figure 11. Training process and confusion matrix of the GSAN.

Without an intra-group self-attention layer, the fusion gate only needs to merge the vectors x and e. The influence is significantly large, causing the accuracy to decrease by 28.4%. Intra-group self-attention is used to extract local features within a group and, when this part is removed, the accuracy becomes the lowest of the four models. This indicates that in the fault signal of the reciprocating compressor, the local feature contains most important fault information.

Without an inter-group self-attention layer, the fusion gate only needs to merge the vectors x and h. The inter-group self-attention is used to capture the global features. When it is removed, the accuracy decreases by 5.1%. This indicates that the global features of the fault signal also contain some fault information that affects the fault identification results of the model.

Without a fusion gate, the vector e will replace the previous output of fusion gate and go to the next layer of the network. The accuracy decreased by 7.8%. This illustrates that the integration of the local feature vector h, the global feature vector e and the original vector x by the fusion gate is essential to the model.

Through ablation analysis, the different function of each component in the model can be observed. In addition, it proves that in fault signals from a reciprocating compressor,

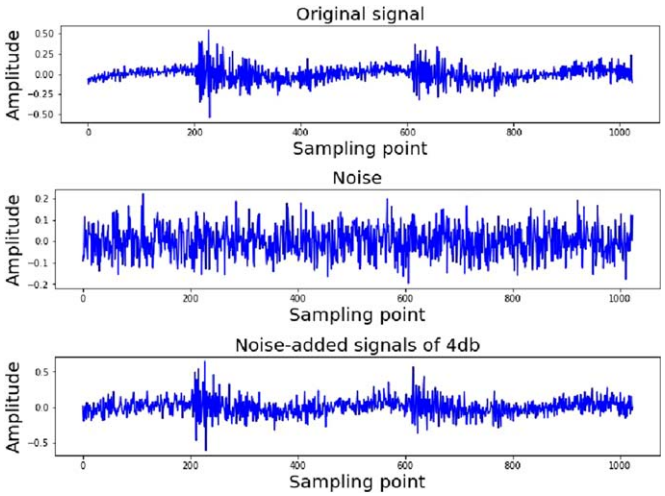
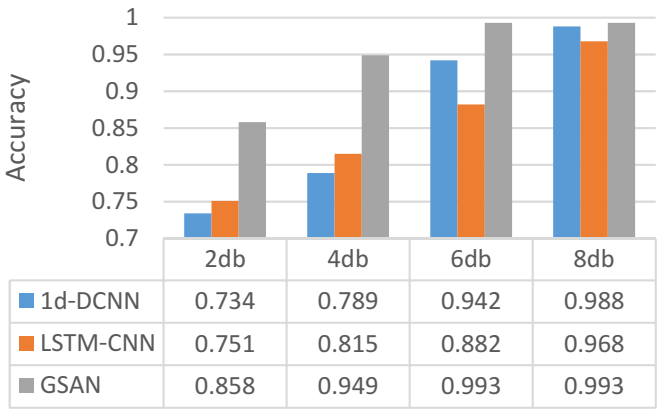


Figure 12. Diagrams from top to bottom are original signal, noise and noise-added signals of 4 dB.



	2db	4db	6db	8db
1d-DCNN	0.734	0.789	0.942	0.988
LSTM-CNN	0.751	0.815	0.882	0.968
GSAN	0.858	0.949	0.993	0.993

Figure 13. Test results of three models with different noise-added signals.

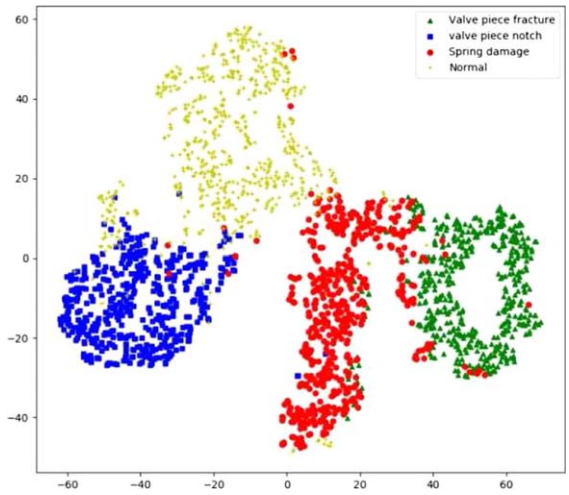


Figure 14. Feature vector visualization of noise-added signals of 2 dB.

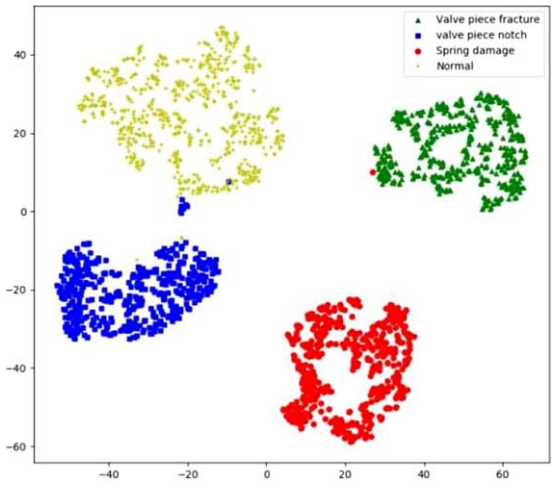


Figure 15. Feature vector visualization of noise-added signals of 4 dB.

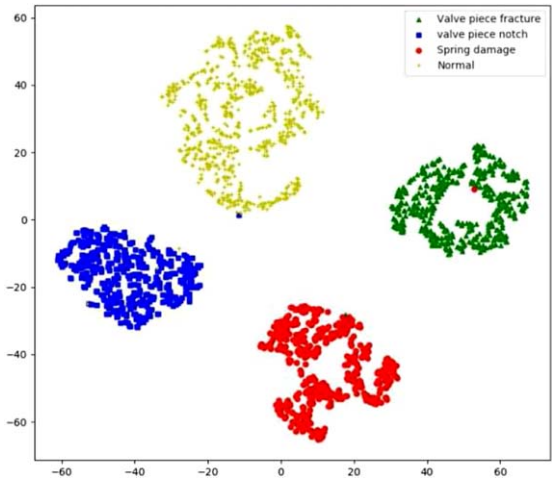


Figure 16. Feature vector visualization of noise-added signals of 6 dB.

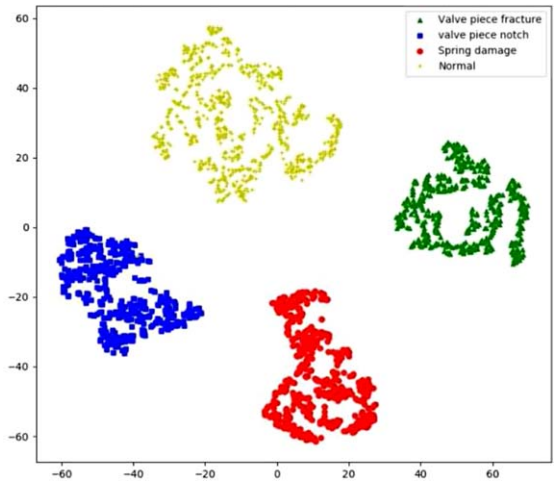


Figure 17. Feature vector visualization of noise-added signals of 8 dB.

inter-group self-attention and intra-group self-attention extracted different fault information, which can complement each other.

5. Conclusion

In this paper, a completely attention-based model GSAN for fault diagnosis in a reciprocating compressor is proposed. The GSAN uses multisource raw signals collected by several sensors as input and adopts the grouping strategy to realize intra-group self-attention and inter-group self-attention, which can extract local and global features respectively, and finally integrate them by using a fusion gate. The experiment shows that, compared with other models, the intelligent fault diagnosis method based GSAN has high accuracy and anti-noise ability. The prediction accuracy of the GSAN can reach 99.3% and the prediction accuracy in different noise states can reach above 95%.

At the same time, the results of our experiments also prove that the attention-based network has a stronger ability to extract features than the CNN and the RNN if the structure is reasonable. The attention-based network can also be used as an effective intelligent fault diagnosis algorithm and not only in the field of natural language processing. In the future work, we will consider improving the GSAN and designing other attention-based models for fault diagnosis.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant Nos. 61603238, 11802168) and a project funded by the China Postdoctoral Science Foundation (No. 2019M661458).

ORCID iDs

Ganchao Bao https://orcid.org/0000-0002-8579-4019
Yuan Wei https://orcid.org/0000-0002-9776-3615

References

[1] Liu Q, Miao W and Li C 2019 Effects of trailing-edge movable flap on aerodynamic performance and noise characteristics of VAWT *Energy* **190** 58–79
[2] Xiao S, Liu S, Jiang F, Song M and Cheng S 2019 Nonlinear dynamic response of reciprocating compressor system with rub-impact fault caused by subsidence *J. Vib. Control* **25** 1737–51
[3] Wei Y and Liu S 2019 Numerical analysis of the dynamic behavior of a rotor-bearing-brush seal system with bristle interference *J. Mech. Sci. Technol.* **33** 3895–903
[4] Li D, Liu S and Zhang H 2017 A method of anomaly detection and fault diagnosis with online adaptive learning under small training samples *Pattern Recognit.* **64** 374–85
[5] Yang B S 2005 Condition classification of small reciprocating compressor for refrigerators using artificial neural networks and support vector machines *Mech. Syst. Signal Process.* **19** 371–90
[6] Tang Y, Liu S, Lei N and Jiang R 2013 Adaptive peak decomposition approach for the fault diagnosis of reciprocating compressor based on general frequency *Inf. Technol. J.* **12** 287–96
[7] Qi G, Zhu Z, Erqinhu K, Chen Y, Chai Y and Sun J 2018 Fault-diagnosis for reciprocating compressors using big data and machine learning *Simul. Model. Pract. Theory* **80** 104–27
[8] Pan B, Shi Z and Xu X 2017 R-VCANet: a new deep-learning-based hyperspectral image classification method *IEEE J. Sel. Topics Appl. Earth Obs. Remote Sens.* **10** 1975–86
[9] Fang B, Li Y, Zhang H and Chan J C-W 2018 Semi-supervised deep learning classification for hyperspectral image based on dual-strategy sample selection *Remote Sens.* **10** 574
[10] Brunetti A, Buongiorno D, Trotta G F and Bevilacqua V 2018 Computer vision and deep learning techniques for pedestrian detection and tracking: a survey *Neurocomputing* **300** 17–33
[11] Nguyen V N, Jenssen R and Roverso D 2018 Automatic autonomous vision-based power line inspection: a review of current status and the potential role of deep learning *Int. J. Elect. Power Energy Syst.* **99** 107–20
[12] Shao H, Jiang H, Zhang H, Duan W, Liang T and Wu S 2018 Rolling bearing fault feature learning using improved convolutional deep belief network with compressed sensing *Mech. Syst. Signal Process.* **100** 743–65
[13] Shao H, Jiang H, Lin Y and Li X 2018 A novel method for intelligent fault diagnosis of rolling bearings using ensemble deep auto-encoders *Mech. Syst. Signal Process.* **102** 278–97
[14] Mao W, Feng W, Liang X and Zhang X 2019 A novel deep output kernel learning method for bearing fault structural diagnosis *Mech. Syst. Signal Process.* **117** 293–318
[15] Shao H, Jiang H, Li X and Liang T 2018 Rolling bearing fault detection using continuous deep belief network with locally linear embedding *Comput. Ind.* **96** 27–39
[16] Zhu Z, Peng G, Chen Y and Gao H 2019 A convolutional neural network based on a capsule network with strong generalization for bearing fault diagnosis *Neurocomputing* **323** 62–75
[17] Lu C, Wang Z, Qin W and Ma J 2017 Fault diagnosis of rotary machinery components using a stacked denoising autoencoder-based health state identification *Signal Process* **130** 377–88
[18] Zhang W, Peng G, Li C, Chen Y and Zhang Z 2017 A new deep learning model for fault diagnosis with good anti-noise and domain adaptation ability on raw vibration signals *Sensors* **17** 425
[19] Zhang W, Li C, Peng G, Chen Y and Zhang Z 2018 A deep convolutional neural network with new training methods for bearing fault diagnosis under noisy environment and different working load *Mech. Syst. Signal Process.* **100** 439–53
[20] Chen L, Wang Z and Zhou B 2017 Intelligent fault diagnosis of rolling bearing using hierarchical convolutional network based health state classification *Adv. Eng. Inform.* **32** 139–51
[21] Zhao D F, Liu S L, Gu D, Sun X and Wang L 2020 Enhanced data-driven fault diagnosis for machines with small and unbalanced data based on variational auto-encoder Meas. Sci. Technol. **31** 035004
[22] Ma P, Zhang H L and Fan W H 2019 A novel bearing fault diagnosis method based on 2D image representation and transfer learning-convolutional neural network *Meas. Sci. Technol.* **30**

[23] Wang J R, Li S M, Han B K, An Z H and Xin Y 2019 Construction of a batch-normalized autoencoder network and its application in mechanical intelligent fault diagnosis *Meas. Sci. Technol.* **30**

[24] Tran D T, Iosifidis A and Gabbouj M 2018 Improving efficiency in convolutional neural network with multilinear filters *Neural Netw.* **105** 328–39

[25] Mohammed Y, David J K, Emmett J I and Carl S 2018 A multi-scale fully convolutional network for semantic labeling of 3D point clouds *ISPRS J. Photogramm. Remote Sens.* **143** 191–204

[26] Jing L, Zhao M, Li P and Xu X 2017 A convolutional neural network based feature learning and fault diagnosis method for the condition monitoring of gearbox *Measurement* **111** 1–10

[27] Hochreiter S and Schmidhuber J 1997 Long short-term memory *Neural Comput.* **9** 1735–80

[28] Cho K *et al* 2014 Learning phrase representations using RNN encoder-decoder for statistical machine translation (arXiv:1406.1078)

[29] Pan H, He X, Tang S and Meng F 2018 An improved bearing fault diagnosis method using one-dimensional CNN and LSTM *J. Mech. Eng.* **64** 443–52

[30] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser L and Polosukhin I 2017 Attention is all you need *CoRR* abs/1706.03762

[31] Bahdanau D, Cho K and Bengio Y 2015 Neural machine translation by jointly learning to align and translate *Int. Conf. on Learning Representations*

[32] Shang L, Lu Z and Li H 2015 Neural responding machine for short-text conversation *ACL-IJCNLP*

[33] Hu M, Peng Y and Qiu X 2017 Reinforced mnemonic reader for machine comprehension (arXiv:1705.02798)

[34] Lin Z *et al* 2017A structured self-attentive sentence embedding (arXiv:1703.03130 [Cs])

[35] Graves A, Jaitly N and Mohamed A-R 2013 Hybrid speech recognition with deep bidirectional LSTM *Automatic Speech Recognition and Understanding (ASRU) 2013 IEEE Workshop on* pp 273–8

[36] Maaten L and Hinton G E 2008 Visualizing high-dimensional data using t-SNE *J. Machine Learn. Res.* **9** 2579–605