Autonomous Intelligent Systems

ORIGINAL ARTICLE Open Access

Fault diagnosis of industrial robot based on dual-module attention convolutional neural network

Kaijie Lu[1,2], Chong Chen[1*], Tao Wang[1,2], Lianglun Cheng[1] and Jian Qin[3]

Abstract

Fault diagnosis plays a vital role in assessing the health management of industrial robots and improving maintenance schedules. In recent decades, artificial intelligence-based data-driven approaches have made significant progress in machine fault diagnosis using monitoring data. However, current methods pay less attention to correlations and internal differences in monitoring data, resulting in limited diagnostic performance. In this paper, a data-driven method is proposed for the fault diagnosis of industrial robot reducers, that is, a dual-module attention convolutional neural network (DMA-CNN). This method aims to diagnose the fault state of industrial robot reducer. It establishes two parallel convolutional neural networks with two different attentions to capture the different features related to the fault. Finally, the features are fused to obtain the fault diagnosis results (normal or abnormal). The fault diagnosis effect of the DMA-CNN method and other attention models are compared and analyzed. The effectiveness of the method is verified on a dataset of real industrial robots.

Keywords: Fault diagnosis, Industrial robots, Deep learning, CNN, Attention

1 Introduction

With the development of intelligent manufacturing, industrial robots are widely used in automobile manufacturing, welding, handling, assembly and various types of mechanical processing and manufacturing because of their high flexibility, low cost and high work efficiency [1]. In the production line, if an industrial robot fails, it will affect the operation of the entire production line. Therefore, it is very important and meaningful to identify and predict the failure of industrial robots [2]. In recent decades, many researches have been carried out in this direction. Wang et al. [3] proposed a new multi-sensor information fusion technology, which takes the signals of multiple sensors as the input of one-dimensional convolution neural network (DCNN), and realized the fault diagnosis of the industrial robot through the improved convolution neural network. Hong et al. [4] collected the attitude data set of the last joint of the multi-joint robot and trained the depth sparse automatic coding network to establish an intelligent fault recognition model, which diagnoses the fault state of the multi-joint robot. Industrial robots had different structures such as Cartesian coordinates, and parallel and multi-joints. Among them, the multi-joint robot had the characteristics of compact structure and flexible operation [5]. For multi-joint robots, kinematics [6], joint clearance [7] and friction model [8] had been well studied. These research results showed that the reducer was an important part of the multi-joint robot. When the transmission accuracy of the robot decreases, its work efficiency and output product quality would decline. And once the reducer failed, it would cause great losses to production [9]. Therefore, developing an effective fault diagnosis method to detect the state of the industrial robot reducer is an important measure to ensure the working performance of multi-joint industrial robots.

*Correspondence: chenc2021@gdut.edu.cn
[1]Guangdong Provincial Key Laboratory of Cyber-Physical System, Guangdong University of Technology, Guangzhou 510006, China
Full list of author information is available at the end of the article

© The Author(s) 2022. This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Fault diagnosis refers to the processing and analysis of measurement signals by detecting the status information of mechanical equipment under relatively static conditions or in operation. It predicts the operation status of the equipment before the equipment fails, and predicts the abnormality or failure; after the failure occurs, it makes a timely judgment on the location, cause and extent of the failure to determine the maintenance plan. In the existing studies, the fault diagnosis approaches that used mechanism analysis and manual feature extraction are prevailing [10]. However, traditional methods such as wavelet transform [11] were limited by computational capability and manual feature extraction, which required large time consumption and professional expertise. Also, the efficiency and accuracy of the fault diagnosis were not satisfactory [12]. With the development of information communication technology, a large amount of data can be collected from industrial production processes, and data-driven methods had become popular in industry and academia [13]. Data-driven methods such as back-propagation neural networks [14], support vector machines [15], and artificial neural networks [16] had been widely used in fault diagnosis. Classifiers based on general machine learning methods such as artificial neural networks and support vector machines are shallow learning models that cannot yet fully reveal the complex internal relationships between faults and signal features [17]. In recent years, deep learning had been widely used in the field of fault diagnosis with its strong learning ability and feature extraction ability [18]. Deep autoencoder [19], deep belief network [20], convolutional neural network (CNN) [21], recurrent neural network [22] and other deep learning methods had been applied to fault diagnosis and achieved satisfactory performance.

CNNs had attracted the attention of many researchers due to the developability and adjustability of the network [23]. CNN used multiple convolution operations to capture the characteristics of the image from the global sensing field for image description [23]. As the core of the CNN, the convolution kernel is usually regarded as an information aggregate that aggregates the spatial information and the channel-wise information on the local receptive field. However, training a decent network requires a lot of effort, and the challenges come from many aspects. Recently, many studies had been proposed to improve the performance of the network, such as directly transferring the shallow layers of the trained offline CNNs to the online CNN [24]; embedding multi-scale information in the Inception structure [25], aggregating features on a variety of different sensing fields to obtain performance gains. Although these works achieved good performance, they were still not enough to fully utilize the CNN model to mine the fault-related information in the data. These methods generally extracted features from raw sensor data directly and performed fault diagnosis. However, some handcrafted features or auxiliary data with domain knowledge also can provide valued information that can reflect degradation trends. Therefore, designing a data transformation method to obtain feature data and establishing a network structure to fuse the features extracted in different ways can enhance fault diagnosis performance. Such as considering space and channel regions separately in Depth-Wise convolutional networks [26], and build two different CNN branches in parallel to extract time-domain and time-frequency domain features respectively [27]. These methods proved that mining more features from monitoring data through different technologies was an effective way to improve prediction accuracy. However, most methods ignored the differences within the monitoring data when acquiring auxiliary data, which would limit the performance of some diagnostic models. Thus, giving more attention to the fault-related features in the data can reduce the negative impact of individual differences and further promote diagnostic performance. Specifically, Yu et al. [28] used the wavelet transform method to preprocess data of multi-channel information and proposed an MC1-DCNN method combining multi-channel CNN and one-dimensional convolution kernel to investigate feature learning from high-dimensional process signals. Jiang et al. [29] proposed a new multi-scale convolutional neural network architecture to simultaneously extract and classify multi-scale features. Ling et al. [30] proposed an improved CNN using the transfer learning method. The network trained several sub-convolutional neural networks to form a convolutional neural network group for different faults, and then connected with the multilayer fully connected neural network. Although the above work achieved good performance, they all improved CNN in terms of data preprocessing and the structure of CNN itself. On the other hand, trying to add some modules to CNN may contribute to better performance.

Due to the introduction of attention mechanisms, many scholars added attention modules to CNN for fault diagnosis. Since CNN ignored the inter-channel connections of the features and limited the feature extraction capabilities of the CNN, from Squeeze-and-Excitation Networks to Selective Kernel Networks [31, 32], CNN had been optimized to varying degrees. In addition, Hao et al. [33] proposed a multi-scale convolution neural network based on an attention mechanism to enhance fault-related multi-scale features and suppress ineffective multi-scale features. Ye et al. [34] proposed a convolution-based self-attention mechanism (CSAM) module, which effectively integrated the powerful feature processing ability of CNN and the local feature processing ability of the self-attention mechanism. The performance of the original model had been effectively improved by adding this module to CNNs and

RNNs. However, these attention models only pay attention to the channel information or the spatial information of the feature map, which still limit the feature extraction ability of the models in some aspects. Zeng et al. [35] proposed a lightweight and efficient Dual Attention Module based on the self-attention mechanism to extract attention in both channel and spatial dimensions. And adding this module to CNN achieved good performance. Liu et al. [36] proposed a novel general deep architecture named dual attention based Temporal Convolutional Network, in which a Temporal Convolutional Network equipped with a dual attention mechanism was developed, which used two parallel attentions to enhance the feature representation of raw temporal data. Although these attention models pay attention to the local features of feature maps from both channel and space, these works were still insufficient. If the information from both distinct attention mechanisms can be jointly used, a better fault diagnosis results can be obtained.

Based on the above analysis, a fault diagnosis method based on a dual-module attention convolutional neural network (DMA-CNN) for industrial robot reducers is proposed. This method establishes two parallel convolutional neural networks with two different attentions. It pays attention to the spatial dimension and channel dimension of the feature map from different aspects at the same time, which can comprehensively extract fault-related features. Subsequently, the features are fused through the multilayer perceptron to obtain fault diagnosis results. A case study was revealed by using an industrial robotics database to validate the results. The main contributions of this study are shown as follows:

1) A two-module CNN based on the attention mechanism is proposed to comprehensively capture the failure relevant features of the monitoring data from the industrial robot.

2) Dual Attention Model is introduced to adaptively integrate local features with global dependencies horizontally from spatial and channel dimensions. Dual Attention Model constructs parallel spatial attention and channel attention based on the self-attention mechanism, so it can capture the internal relationship between data and features.

3) Convolutional Block Attention Module (CBAM) is introduced to emphasize meaningful features along channel and spatial. CBAM is based on convolution operation to make feature graph pass-through channel attention and spatial attention successively.

The remainder of this article is organized as follows: Sect. 2 introduces the DMA-CNN model proposed by this research. The experimental steps are given in Sect. 3; The experimental results and discussion are given in Sect. 4; The conclusions of this paper are given in Sect. 5.

2 Methodology

The whole proposed fault diagnosis process is shown in Fig. 1. The whole framework is divided into two parts: data pre-process and fault diagnosis. Then fault diagnosis is divided into two parts: model training and state evaluation (normal or abnormal).

In the data pre-processing stage, the key features in the monitoring data of industrial robot are first extracted. In this paper, three types of prevailing time-domain features, which are mean value, standard deviation and kurtosis are extracted for fault diagnosis. Mean, standard deviation and kurtosis are calculated as follows:

$$\text{mean} = \frac{\text{data}_1 + \text{data}_2 + \cdots + \text{data}_n}{n}, \tag{1}$$

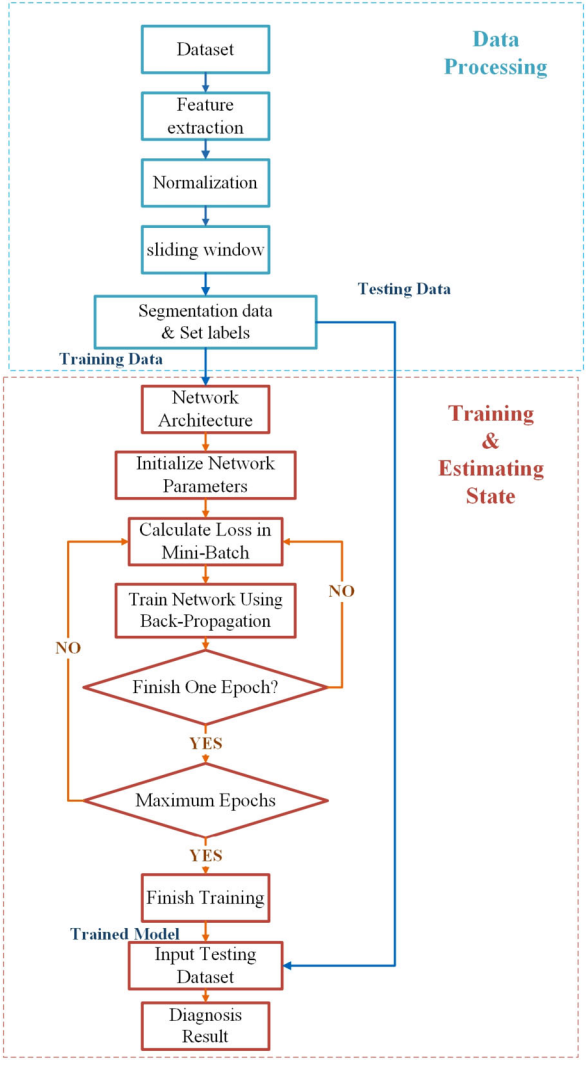


Figure 1 Flow chart of the DMA-CNN based robotic fault diagnosis

$$\text{std} = \frac{1}{n}\sqrt{\sum_{i=1}^{n}(\text{data}_i - \text{mean})^2}, \tag{2}$$

$$\text{Kurtosis} = \frac{[\int_{-\infty}^{+\infty} x(t) - \text{mean}]^4 p(x)\,dx}{\text{std}^4}. \tag{3}$$

Subsequently, data is normalized. Data normalization converts original data into data bounded in a specific range by using the maximum and minimum of variable values, so as to eliminate dimensionality and order of magnitude effects and improve the computational efficiency of the algorithm. Data normalization is as follows:

$$x^{i,j}_{\text{norm}} = \frac{2(x^{i,j} - x^{j}_{\text{min}})}{x^{j}_{\text{max}} - x^{j}_{\text{min}}} - 1, \quad \forall i, j, \tag{4}$$

where $x^{i,j}$ denotes the original ith data point of the jth feature, and $x^{i,j}_{\text{norm}}$ is the normalized value of $x^{i,j}$, x^{j}_{max} and x^{j}_{min} denote the maximum and minimum values of the original measurement data from the jth feature, respectively.

Then, since the dataset is time-series data, Sliding Windows are used to generate samples for capturing more useful sequential information. Let s_3 denote the size of the time window. At each time step, all the past feature data within the time window are collected to form a high-dimensional feature vector and used as the inputs for the network.

Finally, since the entire fault diagnosis model is a binary classification task, the processed data samples are labeled as the normal state (1) and faulty state (0) respectively, and then the scrambled data set is divided into the training set and validation set.

In the fault diagnosis stage, the model is firstly constructed and the CNN is used to extract features. As shown in Fig. 2, in order to make the model pay more attention to fault information, dual attention and CBAM are added to the two parallel networks, and the features extracted by the two convolutional neural networks are fused at the end, then the fused features are input into a fully connected layer to obtain fault diagnosis results. After training the model with the training set, the validation set is used to verify the accuracy of the model.

2.1 Convolutional neural networks

CNN is a deep learning-based supervised algorithm that combines feature extraction and feature classification methods. It is originally used in image processing. CNNs are a very effective technique in large-scale applications due to CNNs' ability to automatically learn high-dimensional features and solve the overfitting problem of the machine learning method [37]. CNNs consist of an input layer, multiple convolutional layers, pooled layers, a fully connected layer, and an output layer. Input data of CNN is typically two-dimensional (2D) data that learn abstract spatial features by alternating overlays of convolutional kernels and pooling operations. Optimization parameters, dropout layers, and batch normalization are also included to help CNNs rely less on training data.

In this paper, the processed data is 2-dimensional, the first dimension is the number of features, and the other dimension is the time series data associated with the feature. Because the relationship between the three features extracted from the data set is not obvious [38] in the data processing stage. Therefore, although the input and corresponding feature maps are 2-dimensional, the convolution kernel in the proposed network is one-dimensional (1D) (Fig. 3) in practice. The processed data samples are fed from the input layer to the convolutional layer, and when generating a set of feature maps, the largest computational task occurs in the convolutional layer. In each convolutional layer, the input data is convolved using a kernel with a local receive domain. Then, a bias term is added to generate an output feature map by a nonlinear activation function, such as Tanh, which is fed into subsequent convolutional layers. Convolution operations can be defined as Eq. (5):

$$z_i = \text{Tanh}\left(W^T x_{i,j+F_l-1} + b\right), \tag{5}$$

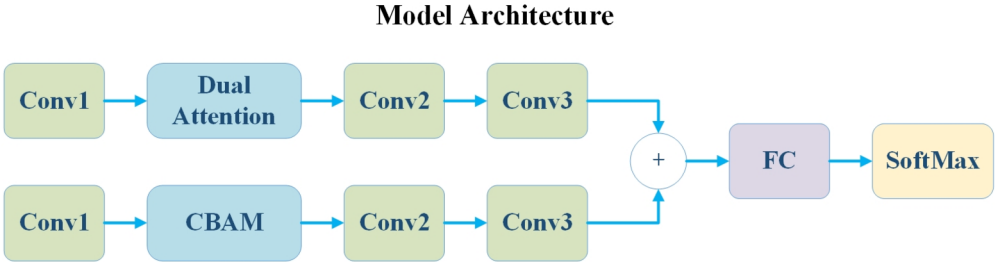


Figure 2 The Model Architecture of DMA-CNN

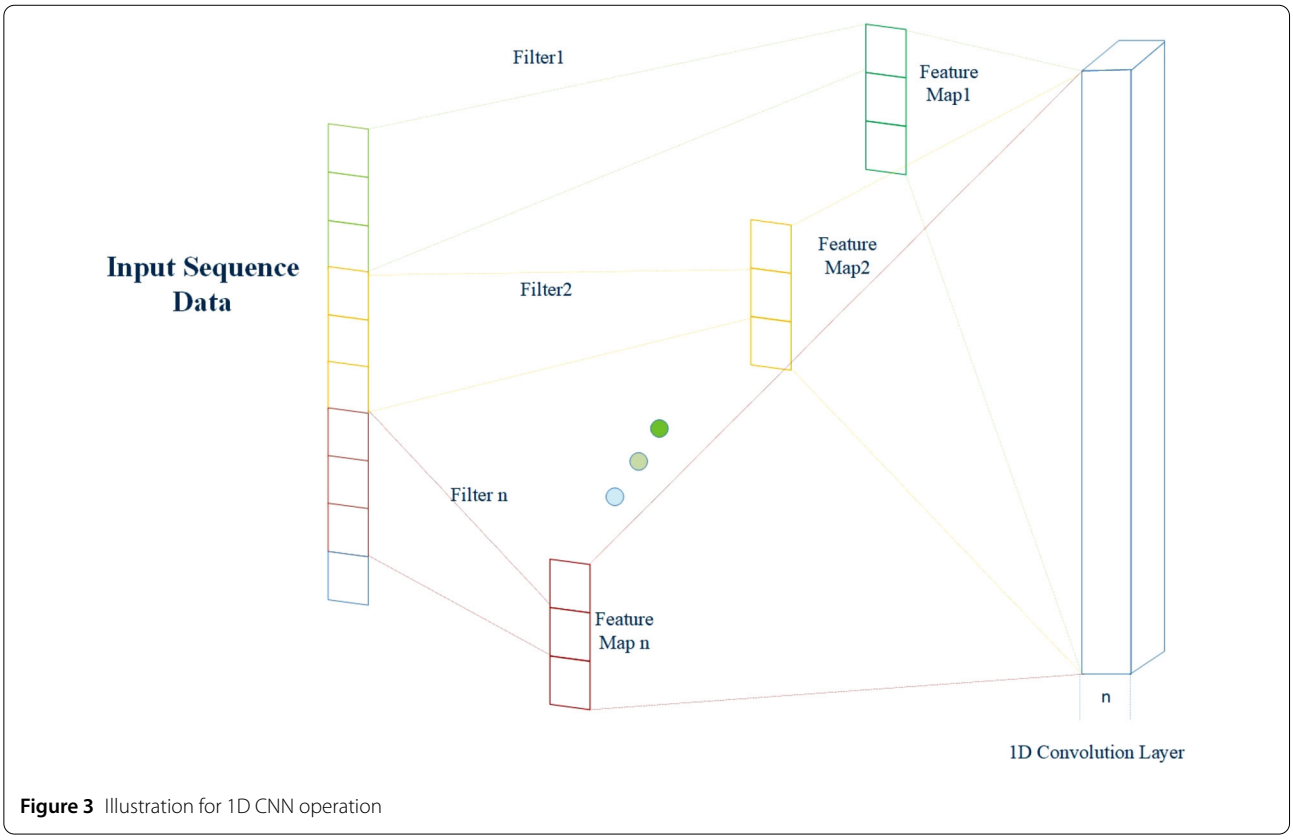


Figure 3 Illustration for 1D CNN operation

where, the Tanh activation function is shown in the following Eq. (6):

$$y = \mathrm{Tanh}(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}, \tag{6}$$

where $w \in R^{F_L}$ represents the convolutional kernel, b represents the bias term, $x_{i,j+F_l-1}$ represents a subsequence of x with the length F_l from the point i, and z_i represents the learned feature.

After the convolutional layer, the Dropout layer is added to the network. Dropout is a technique that can help reduce data overfitting when training neural networks, especially if the training dataset is small [39]. Overfitting of the training data often results in better network performance of the training dataset and poorer network performance of the test dataset. Dropout provides a simple and effective way to solve this problem. In this study, the dropout technique is applied to the proposed network to prevent complex collaborative adaptation of the training data and to avoid repeated extraction of the same features. In fact, dropout is achieved by setting the activation output of some hidden neurons to zero so that those neurons are not included in the forward propagation training process. However, dropout is turned off during the test, suggesting that all hidden neurons are involved in the test. In this way, the robustness of the network is enhanced. Dropout can also be thought of as a simple method of model integration within a network, helping to improve the feature extraction capabilities of a network.

2.2 Attention mechanism

The attention mechanism is originally used for machine translation and has now become an important concept in the field of neural networks. The attention mechanism in deep learning is essentially similar to the selective visual attention mechanism of humans, and the core goal is to select the information that is more critical to the current mission from a large amount of information. The attention mechanism can be divided into spatial domains, channel domains, layer domains, mixed domains and time domains according to the domains of concern. Since convolution operations extract informative features by fusing the cross-channel and spatial information, the attention model based on the channel domain and space domain is suitable for CNN and the effect is obvious [40], both the dual attention model and CBAM used in this paper is based on this.

The Dual Attention Model introduces a self-attention mechanism to capture feature dependencies in the spatial and channel dimensions respectively [41]. Specifically, it contains two parallel attention parts, one is position at-

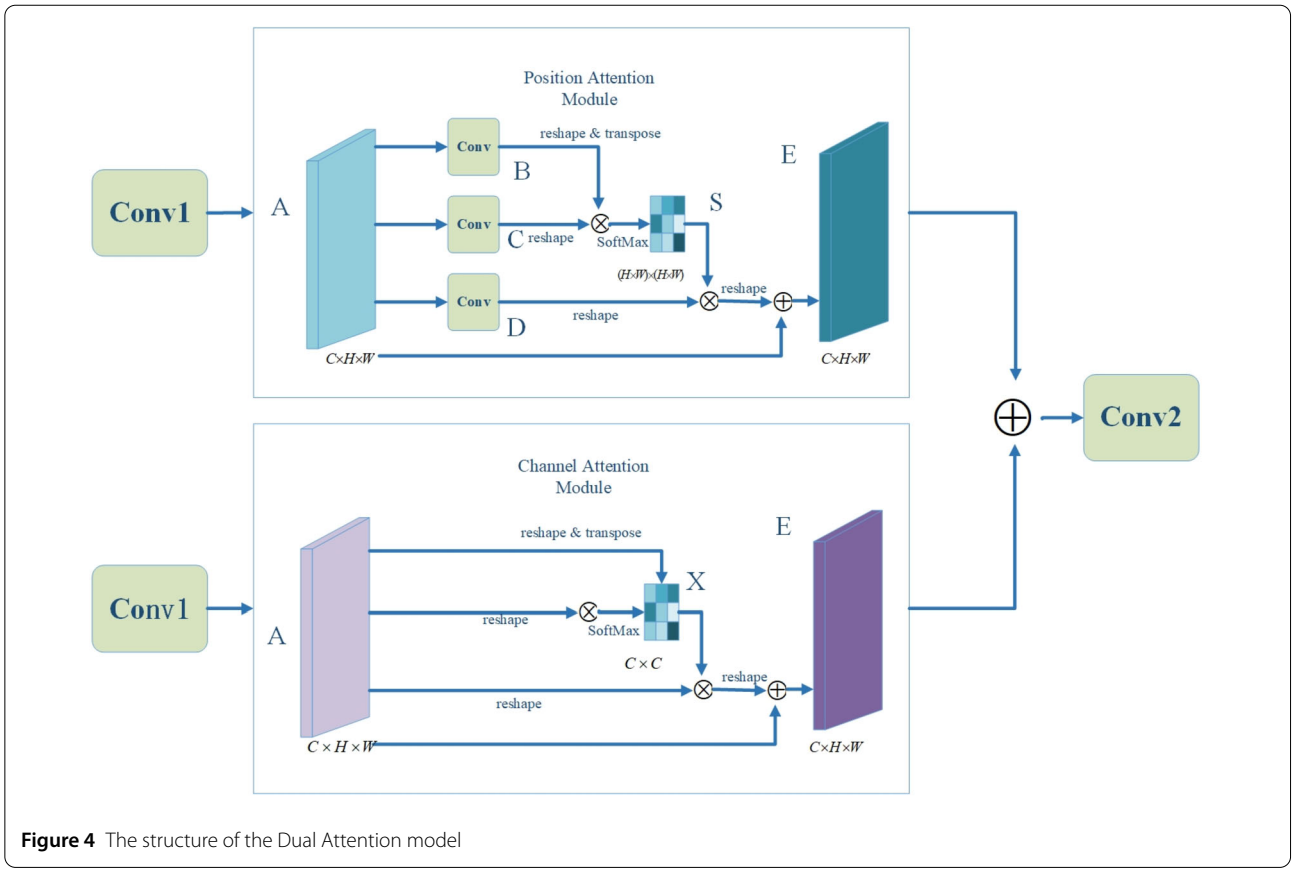


Figure 4 The structure of the Dual Attention model

tention and the other is channel attention. The processing methods of the two attentions are similar (Fig. 4). First, the convolution layer is used to obtain the dimensionality reduction feature map A. These features are then fed into the Location Attention module to generate new spatial features through the following three steps. The first step is to generate a spatial attention matrix that simulates the spatial relationship between any two pixels of a feature. Next, matrix multiplication is performed between the attention matrix and the original feature. Third, element-wise summation of the above multiplied matrix and original features is performed to obtain the final representation. At the same time, the channel attention module is used to capture contextual information on the channel dimension. The process of obtaining channel relationships is similar to the position attention module, but the first step is to calculate the channel attention matrix on the channel dimension. Finally, the outputs of the two attention modules are aggregated to obtain better feature representations for pixel-level predictions.

For positional attention, feeding the feature map $A \in \mathbb{R}^{C\times H\times W}$ to a convolutional layer produces two feature maps B and C, where $\{B, C\} \in \mathbb{R}^{C\times H\times W}$, and then reshaping B and C to $\{B, C\} \in \mathbb{R}^{C\times N}$, where $N = H \times W$, after which do matrix multiplication with the transpose of B and C, applying a SoftMax layer computing spatial attention map $S \in \mathbb{R}^{C\times H\times W}$, as follows:

$$s_{ji} = \frac{e^{B_i C_j}}{\sum_{i=1}^{N} e^{B_i C_j}}, \tag{7}$$

where s_{ji} measures the ith position's impact on jth position. The more similar feature representations of the two position contribute to greater correlation between them.

At the same time, feeding feature map A into another convolutional layer to produce feature map $D \in \mathbb{R}^{C\times H\times W}$, and reshape D to $D \in \mathbb{R}^{C\times N}$, then do matrix multiplication of D and S, reshape the result to $\mathbb{R}^{C\times H\times W}$. Finally, multiply the result by a parameter α and do the element summation operations in the original feature map A to get the final output $E \in \mathbb{R}^{C\times H\times W}$, the formula is as follows:

$$E_j = \alpha \sum_{i=1}^{N} (s_{ji} D_i) + A_j, \tag{8}$$

where α is initialized as 0 and gradually learns to assign more weight. The resulting feature E at each position is a weighted sum of the features across all positions and original features. Therefore, it has a global contextual view and

selectively aggregates contexts according to the spatial attention map.

For channel attention, calculate the channel attention map $X \in \mathbb{R}^{C\times C}$ directly by A, reshape A to $\mathbb{R}^{C\times N}$, with itself to do matrix multiplication, apply a SoftMax layer to calculate the channel attention map $X \in \mathbb{R}^{C\times C}$, as follows:

$$x_{ji} = \frac{e^{A_i A_j}}{\sum_{i=1}^{C} e^{A_i A_j}}, \tag{9}$$

where x_{ji} measures the ith channel's impact on the jth channel.

After that X and A do matrix multiplication, and the result is reshaped to $\mathbb{R}^{C\times H\times W}$. Finally, the result by a parameter β and do the element summation operation in the original feature graph A are multiplied to get the final output $E \in \mathbb{R}^{C\times H\times W}$, the formula is as follows:

$$E_j = \beta \sum_{i=1}^{C} (x_{ji} A_i) + A_j, \tag{10}$$

where β gradually learns a weight from 0. Equation (10) shows that the final feature of each channel is a weighted sum of the features of all channels and original features, which models the long-range semantic dependencies between feature maps. It helps to boost feature discriminability.

CBAM: Convolutional Block Attention Module based on the convolution operation [42]. As shown in Fig. 5, CBAM contains two concatenated attentions, one is channel attention and the other is spatial attention, for channel attention, average pooling and maximum pooling are first used to aggregate the spatial information of feature maps, generating two different spatial feature maps: F^c_{avg} and F^c_{max}, which represent the average pooling feature and the maximum pooling feature, respectively. This feature map is then forwarded to a shared network, generating our channel attention $M_c \in \mathbb{R}^{C\times 1\times 1}$. A shared network consists of a multilayer perceptron and a hidden layer. To reduce parameter overhead, the hidden activation size is set to $\mathbb{R}^{C/r\times 1\times 1}$ where r is the reduction ratio. After applying a shared network to each feature map, an output feature vector is generated using element summation. In short, the channel attention diagram is calculated as:

$$M_c(F) = \sigma\left(\text{MLP}\left(F^c_{\text{avg}}\right) + \text{MLP}\left(F^c_{\text{max}}\right)\right), \tag{11}$$

where σ the sigmoid function is represented.

For spatial attention, two pooling operations are used to aggregate the channel information of the feature map, generating two two-dimensional feature maps: $F^s_{\text{avg}} \in \mathbb{R}^{1\times H\times W}$ and $F^{s\,1\times H\times W}_{\text{max}}$. Represents the average pooling characteristic and the maximum pooling characteristic on the channel, respectively. Then connect the two feature maps into a convolutional layer to produce our 2D spatial attention map. In short, spatial attention is calculated as:

$$M_s(F) = \sigma\left(f^{7\times 7}\left(\left[F^s_{\text{avg}}; F^s_{\text{max}}\right]\right)\right), \tag{12}$$

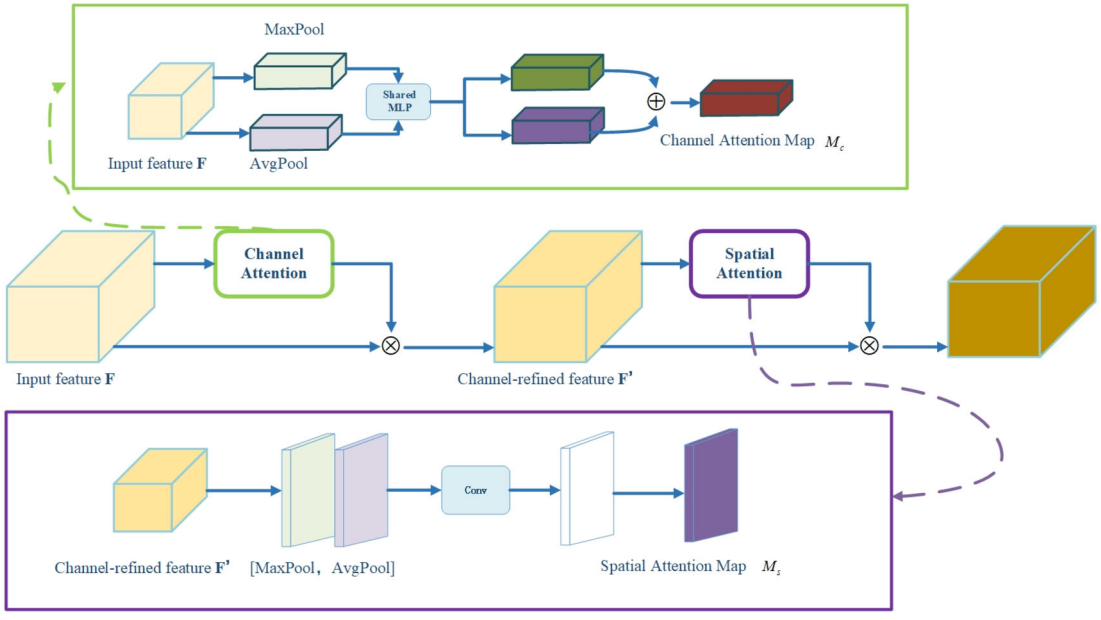


Figure 5 The structure of the CBAM model

where σ represents the sigmoid function. $f^{7\times7}$ represents a convolution 7×7 operation for which the convolution kernel.

2.3 Fault classification performance evaluation indicators

Since the proposed method judges bearing faults from current signals, evaluation parameters such as accuracy, recall rate, F1 and accuracy are used to evaluate classification problems. These parameters can be obtained from Eqs. (13)–(16):

$$Precision \frac{TP}{TP + FP}, \tag{13}$$

$$Recall = \frac{TP}{TP + FN}, \tag{14}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}, \tag{15}$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}. \tag{16}$$

3 Experimental setup

3.1 Data pre-processing

The data used in the experiment in this paper are from the six-axis industrial robot, which has the advantages of high precision, strong acceleration ability and good rigidity. Because failure is rare in the operation, failure data is often difficult to collect. The data used in this paper were obtained by fault injection experiments. Specifically, a faulty reducer was mounted on the sixth axis of an industrial robot to generate the failure data. After the faulty reducer was injected, the robot was used to perform some daily tasks to generate relevant monitoring data. In this case, the current signal was successfully collected and used for further modelling. Therefore, this paper uses the current data of the industrial robot to diagnose the fault state of the reducer.

Firstly, feature extraction was carried out on the data set, sliding s_1 step each time, mean, standard deviation and kurtosis were extracted from the data interval with a step length of s_2, and they were normalized. The normalization formula was shown in Eq. (4). Then, the newly constructed data set was processed with a sliding window step of s_3 so each sample shape was $1 \times 3 \times s_3$. These were used as input to the model.

3.2 Prognostic procedure

As shown in Fig. 2, the proposed model consisted of two parallel modules, one for the CNN with Dual Attention and another for the deep convolutional neural network with CBAM, the entire model contained 6 layers, of which each module contained 4 layers of convolution, the first two layers of convolution had the same configuration, each layer convolution was used F_n kernels, each kernel size was $F_l \times 1$, attention was added after the first layer of convolution, in the last layer of convolution, a convolutional layer with a convolutional kernel was used to aggregate the feature map of each module output, the convolutional kernel size was 3×1, so that each module was obtained an advanced feature map from the original feature separately. Subsequently, the feature maps outputs of the two modules are fused and flattened, and then passed through two linear layers and finally the binary classification results were output by SoftMax.

All convolutional layers used the Tanh function as the activation function, the Xavier normal initializer was used to initialize the network weights, in order to further improve the prediction performance, backpropagation algorithm was used for fine-tuning. The parameters of the proposed model were updated by the cross-entropy loss function, and the Adam was used as the optimizer of the small-batch update. The cross-entropy loss function was as follows:

$$\text{loss}\left(y, y^{\wedge}\right) = -\sum_{j=1}^{q} y_j \log y_j^{\wedge}, \tag{17}$$

where y_j represents the actual label, $y_j^{\wedge}$ represents the model prediction.

80% of the data samples were used as training samples and the remaining 20% of the data samples were used as test samples. For each training epoch, the samples were randomly divided into multiple small batches, each containing 512 samples, and entered into the training system. Based on the cross-entropy loss function of each small batch, the weights of each layer were optimized. It should be noted that the choice of batch size affects the training performance of the network. On the basis of experiments, this paper found a suitable batch size of 512 samples and adopted different learning rates, the maximum training epoch was 150, in the first 100 epochs, the learning rate was set to 0.001 for rapid optimization, and in the last 50 epochs, the learning rate was set to 0.0001 for smooth convergence.

Finally, the test samples were fed into the trained network to predict the failure state. The parameters of the proposed model were shown in Table 1.

Table 1 The DMA-DCNN model parameters

Parameter	Value
s_1	2
s_2	5
s_3	5
Batch size	512
Neurons in fully-connected layer	256
Epoch number	150

3.3 Benchmarking experiment

In order to verify the performance of the proposed method, the results were compared with other similar methods.

DCNN: It is composed of 1D-CNN layers. Compared with the proposed method, there was only one module, and the attention layer is replaced by a convolutional layer; DA-CNN: There is only one Dual Attention module compared to the proposed method; CBAM-CNN: There is only one CBAM module compared to the proposed method; DA-DA-CNN: There are two Dual Attention modules compared to the proposed method; CBAM-CBAM-CNN: There are two CBAM modules compared to the proposed method;

For the above 5 methods, the input layer and output layer were the same as the model proposed in this article, and both used the cross-entropy function as the loss function, and the backpropagation algorithm was used to update the parameters, and the Adam algorithm was used as the optimizer. In order to evaluate the performance of the model, the Precision, Recall, F1 and Accuracy of the results were respectively calculated by running each algorithm 5 times, and their mean values and standard deviations were calculated.

4 Experimental results and discussion

4.1 Experimental results

The experimental results are shown in Fig. 6, DMA-CNN has achieved the best results, and the standard deviation is relatively small, indicating that the results obtained are stable. Specifically, the proposed model achieved an accuracy of 93.5% on the test set, which is not much different from the accuracy rate of 93.1% in the DA-DA-CNN. Compared with the DCNN, the accuracy of the proposed model is improved by 17%. The proposed model achieved 92.2% precession on the test set, which is only a slight advantage over other models that add attention, but the proposed model yields the most stable results in the standard deviation, which had a 20% increase in progress compared to the original DCNN. Moreover, the DMA-CNN reached 94% in recall, which is significantly better than other models and 10% higher than the original DCNN. Finally, on F1_score, it was an improvement of nearly 16% compared to the original DCNN, and it was better than other cases. In summary, compared with the DCNN without attention, the evaluation indicators of the proposed model have been significantly improved, and the experimental results are relatively stable, compared with other DCNNs with attention. The proposed dual-module DMA-CNN performed the best.

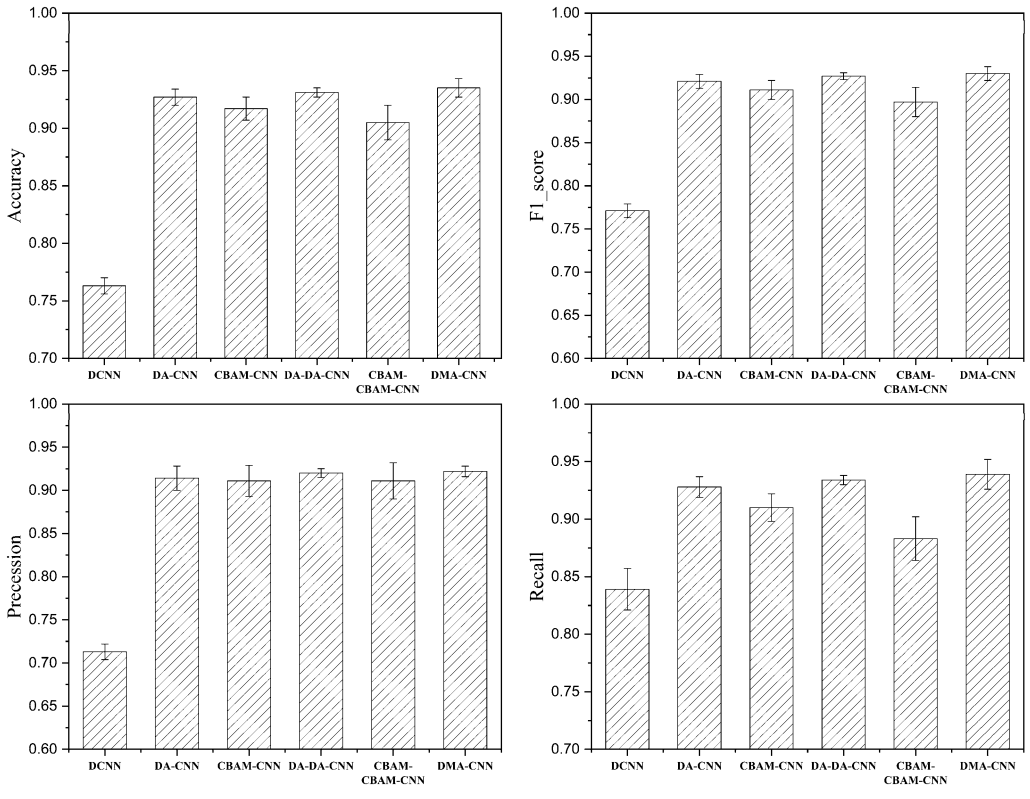


Figure 6 Experimental Results

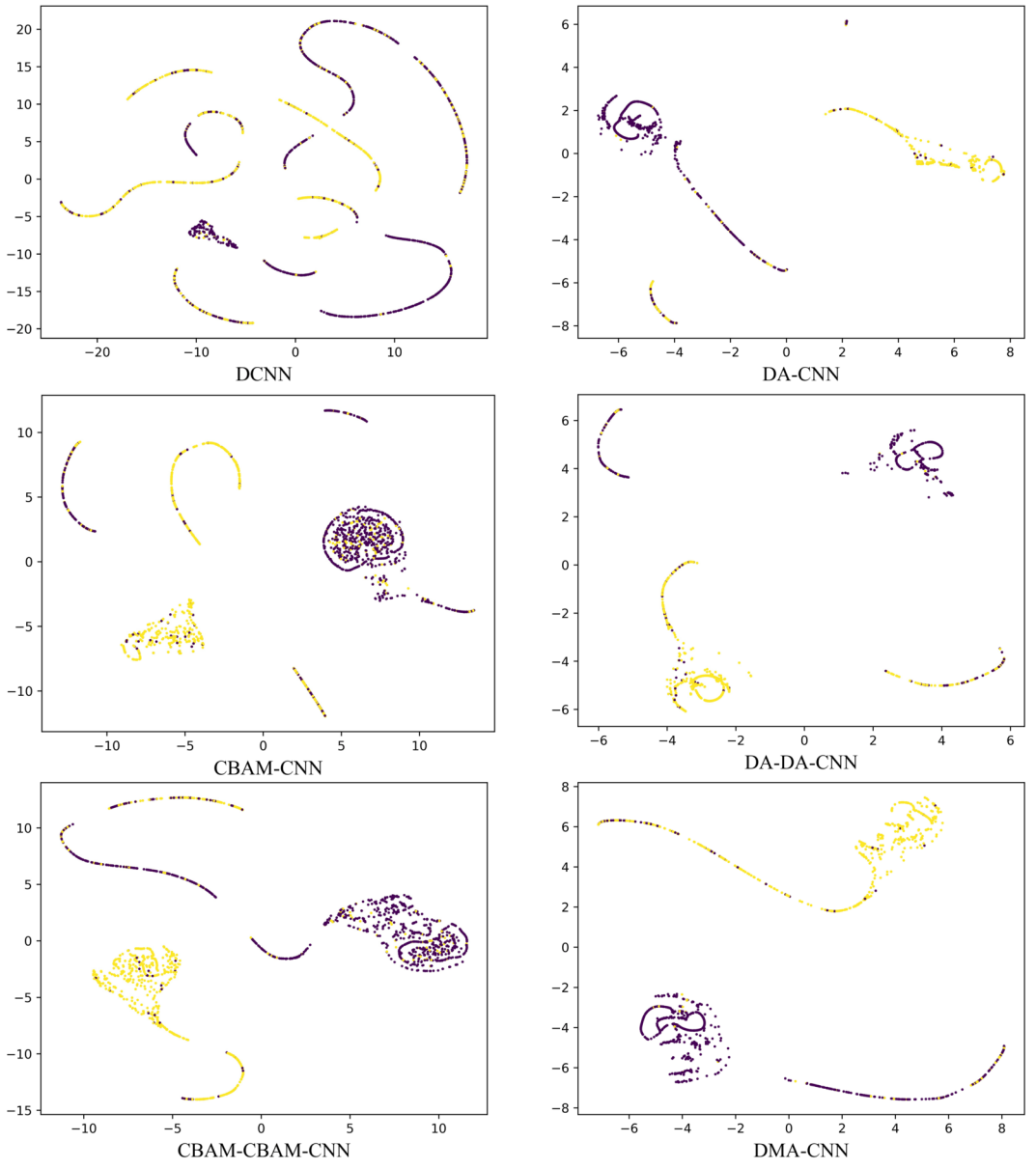


Figure 7 Visualization of experimental results

To present the performance of the proposed model more clearly, the T-SNE technique was used to visualize the distribution of features learned by different models. The visualization results are shown in Fig. 7. It can be seen that the original DCNN classification effect is not satisfactory. The addition of attention mechanism makes the classification boundary clear. In addition, DMA-CNN proposed in this paper has achieved the best classification effect. Compared with other models, the classification plane of this method is the most obvious.

4.2 Discussion

The proposed model extracts the fault-related features in the data sample through two different attention modules. Compared with the traditional attention-based CNN, this model adds two attention modules to their respective CNN to improve the feature extraction ability and adaptability of CNN. Among them, the dual attention model designs parallel channel attention and spatial attention based on the self-attention mechanism, while CBAM designs a series of channel attention and spatial attention based on convolution operation. The two attention modules con-

sider the relationship between features in space and channels from different aspects, in order to pay as much attention to the fault-related features as possible. The network not only reduces the loss of fault-related information but also suppresses unnecessary features to help the flow of information in the network.

Although the proposed fault diagnosis model has achieved good performance, there are still some shortcomings. The model proposed in this paper can only distinguish whether the robot joint is in the faulty stage, but does not judge the type of fault, In the future, the model can be improved for multi-classification problems. Secondly, from the experimental results, the evaluation indicators of the proposed model can only reach about 93% and compared with DA-DA-CNN, the advantages are not obvious. In the future, we can try to further study dual-attention to obtain a better feature map. Meanwhile, in the data preprocessing stage, feature extraction and sliding window processing data with different lengths to obtain better performance needs to be studied. In addition, how to reduce the steps of data preprocessing while maintaining the performance of the model is also the focus of future research. Finally, Industrial robot failures are rare, so it is not easy to collect failure data. And in future work, we will further study the direction of transfer learning.

In addition, since the step length of feature extraction is 5, the length of sliding window is also 5, which means that if the model is applied to online data. Only 25 data are needed to diagnose the equipment once in the practical implications. So as to timely maintain the equipment, we reduce losses and improve industrial production efficiency.

5 Conclusions

Aiming at the problem of fault diagnosis of industrial robots, a DMA-CNN model is proposed, which can locate fault characteristics from different aspects in the data through different attention mechanisms, and improves the characteristics of convolutional neural networks extraction capacity. First, a sample dataset is obtained by data preprocessing, then a DMA-CNN is constructed for feature extraction. Finally, the test set is used for fault diagnosis. The experimental study on the real industrial robot data shows that this model can effectively detect the abnormal state of the industrial robot reducer, and compared with other fault diagnosis models, this method is superior to other models in terms of accuracy, F1_score, precession, recall, which provides a new and effective method for fault diagnosis of industrial robot reducer.

Funding

Our work is supported by multiple funds in China, including the Key Program of NSFC-Guangdong Joint Funds (U2001201), the Natural Science Foundation of Guangdong Province (2020A1515010890, 2020B1515120010), the Dedicated Fund for Promoting High-Quality Economic Development in Guangdong Province (Marine Economic Development Project) GDNRC[2021]44. Our work is also supported by Guangdong Provincial Key Laboratory of Cyber-Physical System (2020B1212060069).

Availability of data and materials

Not applicable.

Code availability

Not applicable.

Declarations

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

KL: Conceptualization, Methodology, Writing – original draft and Validation. CC: Conceptualization, Methodology, Review & Editing. TW: Conceptualization, Resources. LC: Supervision. JQ: Review & Editing. All authors read and approved the final manuscript.

Author details

[1]Guangdong Provincial Key Laboratory of Cyber-Physical System, Guangdong University of Technology, Guangzhou 510006, China. [2]School of Automation, Guangdong University of Technology, Guangzhou 510006, China. [3]Welding Engineering and Laser Processing Centre, School of Aerospace, Transport and Manufacturing, Cranfield University, Cranfield, UK.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 1 April 2022 Revised: 13 May 2022 Accepted: 20 May 2022
Published online: 01 June 2022

References

1. Y. Yao, Y. Li, B. Xie, L. Lei, Y. Wang, Mechanical fault diagnosis based on self-sensing motor drive: review and key issues, in *2021 Global Reliability and Prognostics and Health Management, PHM-Nanjing 2021* (2021). https://doi.org/10.1109/PHM-NANJING52125.2021.9613124
2. Q. Yang, X. Li, Y. Wang, A. Ainapure, J. Lee, Fault diagnosis of ball screw in industrial robots using non-stationary motor current signals. Proc. Manuf. **48**, 1102–1108 (2020). https://doi.org/10.1016/J.PROMFG.2020.05.151
3. J. Wang, D. Wang, X. Wang, Fault diagnosis of industrial robots based on multi-sensor information fusion and 1D convolutional neural network, in *2020 39th Chinese Control Conference (CCC)* (IEEE, 2020)
4. Y. Hong, Z. Sun, X. Zou, J. Long, Multi-joint industrial robot fault identification using deep sparse auto-encoder network with attitude data, in *2020 Prognostics and Health Management Conference (PHM-Besançon)* (IEEE, 2020)
5. J. Long, J. Mou, L. Zhang, S. Zhang, C. Li, Attitude data-based deep hybrid learning architecture for intelligent fault diagnosis of multi-joint industrial robots. J. Manuf. Syst. **61**, 736–745 (2021). https://doi.org/10.1016/J.JMSY.2020.08.010
6. T. Zeng, P. Liu, D. You, Kinematics simulation and operation space analysis of arc welding six-axis robot based on Matlab and Adams, in *2019 2nd World Conference on Mechanical Engineering and Intelligent Manufacturing (WCMEIM)* (IEEE, 2019)
7. S. Cheng, Dynamic analysis of reciprocating compressor with revolution joint clearance and translational joint clearance, in *2021 Global Reliability and Prognostics and Health Management (PHM-Nanjing)* (IEEE, 2021)
8. M.A. Tadese, F. Yumbla, J.-S. Yi, W. Lee, J. Park, H. Moon, Passivity Guaranteed Dynamic Friction Model With Temperature and Load Correction: Modeling and Compensation for Collaborative Industrial Robot. IEEE Access. **9** (2021)
9. A.D. Pham, H.J. Ahn, Rigid precision reducers for machining industrial robots. Int. J. Precis. Eng. Manuf. **22**, 1469–1486 (2021). https://doi.org/10.1007/S12541-021-00552-8/FIGURES/12
10. J. Sun, C. Yan, J. Wen, Intelligent bearing fault diagnosis method combining compressed data acquisition and deep learning. IEEE Trans. Instrum. Meas. **67**, 185–195 (2018). https://doi.org/10.1109/TIM.2017.2759418

11. X. Wang, R. Zhang, A sensor fault diagnosis method research based on wavelet transform and Hilbert-Huang transform, in *Proceedings - 2013 5th Conference on Measuring Technology and Mechatronics Automation, ICMTMA 2013* (2013), pp. 81–84. https://doi.org/10.1109/ICMTMA.2013.32
12. A.D. Spyronasios, M.G. Dimopoulos, A.A. Hatzopoulos, Wavelet analysis for the detection of parametric and catastrophic faults in mixed-signal circuits. IEEE Trans. Instrum. Meas. **60**, 2025–2038 (2011). https://doi.org/10.1109/TIM.2011.2115550
13. Z. Xiao, Z. Cheng, Y. Li, A review of fault diagnosis methods based on machine learning patterns, in *2021 Global Reliability and Prognostics and Health Management, PHM-Nanjing 2021* (2021). https://doi.org/10.1109/PHM-NANJING52125.2021.9612779
14. Y. Liu, J. Wang, Y. Shen, Research on verification of sensor fault diagnosis based on BP neural network, in *2020 11th International Conference on Prognostics and System Health Management (PHM-2020 Jinan)* (IEEE, 2020)
15. H. Zhan, Application of rough set and support vector machine in fault diagnosis of power electronic circuit, in *2010 2nd IEEE International Conference on Information Management and Engineering* (IEEE, 2010)
16. S. Yang, W. Li, C. Wang, The intelligent fault diagnosis of wind turbine gearbox based on artificial neural network, in *2008 International Conference on Condition Monitoring and Diagnosis* (IEEE, 2008)
17. J. Hu, L. Shen, S. Albanie, G. Sun, E. Wu, Squeeze-and-excitation networks, in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42 (2020)
18. D. Zhu, X. Song, J. Yang, Y. Cong, L. Wang, A bearing fault diagnosis method based on L1 regularization transfer learning and LSTM deep learning, in *2021 IEEE International Conference on Information Communication and Software Engineering (ICICSE)* (IEEE, 2021)
19. Z. Xu, W. Mo, Y. Wang, S. Luo, T. Liu, Transformer fault diagnosis based on deep brief sparse autoencoder, in *2019 Chinese Control Conference (CCC)* (IEEE, 2019)
20. J. Jiao, C. Weidong, Research on EMU cable fault diagnosis based on deep belief network and signal domain conversion, in *2021 4th International Conference on Electron Device and Mechanical Engineering (ICEDME)* (IEEE, 2021)
21. X. Zhang, G. Chen, T. Hao, Z. He, Rolling bearing fault convolutional neural network diagnosis method based on casing signal. J. Mech. Sci. Technol. **34**, 2307–2316 (2020). https://doi.org/10.1007/S12206-020-0506-8
22. J. Wang, G. Wu, L. Wan, Y. Sun, D. Jiang, Recurrent neural network applied to fault diagnosis of underwater robots, in *2009 IEEE International Conference on Intelligent Computing and Intelligent Systems* (IEEE, 2009)
23. G. Sun, Y. Wang, C. Sun, Fault diagnosis of planetary gearbox based on signal denoising and convolutional neural network, in *2019 Prognostics and System Health Management Conference (PHM-Paris)* (IEEE, 2019)
24. G. Xu, M. Liu, Z. Jiang, W. Shen, C. Huang, Online fault diagnosis method based on transfer convolutional neural networks. IEEE Trans. Instrum. Meas. **69**, 509–520 (2020). https://doi.org/10.1109/TIM.2019.2902003
25. X. Liu, H. Tian, Z. Dai, Bearing fault diagnosis based on multi-scale convolution neural network and dropout, in *2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)* (IEEE, 2020)
26. Y. Lin, R. Li, W. He, X. Zho, J. He, P. Li, Y. Jiang, L. Liu, N. Wu, C. Shi, A high-speed low-cost CNN inference accelerator for depthwise separable convolution, in *2020 IEEE International Conference on Integrated Circuits, Technologies and Applications (ICTA)* (IEEE, 2020)
27. M. Liang, P. Cao, J. Tang, Rolling bearing fault diagnosis based on feature fusion with parallel convolutional neural network. Int. J. Adv. Manuf. Technol. **112**, 819–831 (2021). https://doi.org/10.1007/s00170-020-06401-8
28. J. Yu, C. Zhang, S. Wang, Multichannel one-dimensional convolutional neural network-based feature learning for fault diagnosis of industrial processes. https://doi.org/10.1007/s00521-020-05171-4
29. G. Jiang, H. He, J. Yan, P. Xie, Multiscale convolutional neural networks for fault diagnosis of wind turbine gearbox. IEEE Trans. Ind. Electron. **66**, 3196–3207 (2019). https://doi.org/10.1109/TIE.2018.2844805
30. Z.-Q. Ling, G.-Z. Cao, Y.-P. Zhang, An improved convolutional neural network for rolling bearing fault diagnosis, in *2021 IEEE 11th Annual International Conference on CYBER Technology in Automation, Control, and Intelligent Systems (CYBER)* (IEEE, 2021)
31. H. Wang, J. Xu, R. Yan, R.X. Gao, A new intelligent bearing fault diagnosis method using SDP representation and SE-CNN. IEEE Trans. Instrum. Meas. **69**, 2377–2389 (2020). https://doi.org/10.1109/TIM.2019.2956332
32. Y. Xu, H. Fan, J. Li, DenseNet matting algorithm based on embedded improved SKNet, in *Proceedings of 2020 IEEE International Conference on Progress in Informatics and Computing, PIC 2020* (2020), pp. 80–84. https://doi.org/10.1109/PIC50277.2020.9350782
33. Y. Hao, H. Wang, Z. Liu, H. Han, Multi-scale CNN based on attention mechanism for rolling bearing fault diagnosis, in *2020 Asia-Pacific International Symposium on Advanced Reliability and Maintenance Modeling (APARM)* (IEEE, 2020)
34. R. Ye, W. Wang, Y. Ren, K. Zhang, Bearing fault detection based on convolutional self-attention mechanism, in *2020 IEEE 2nd International Conference on Civil Aviation Safety and Information Technology (ICCASIT)* (IEEE, 2020)
35. Y. Zeng, X. Guo, H. Wang, M. Geng, T. Lu, Efficient dual attention module for real-time visual tracking, in *2019 IEEE International Conference on Visual Communications and Image Processing, VCIP 2019* (2019). https://doi.org/10.1109/VCIP47243.2019.8965683
36. C. Liu, L. Zhang, R. Yao, C. Wu, Dual attention-based temporal convolutional network for fault prognosis under time-varying operating conditions. IEEE Trans. Instrum. Meas. **70** (2021)
37. S. Tang, S. Yuan, Y. Zhu, Data Preprocessing Techniques in Convolutional Neural Network Based on Fault Diagnosis Towards Rotating Machinery. IEEE Access. **8** (2020)
38. X. Li, Q. Ding, J.Q. Sun, Remaining useful life estimation in prognostics using deep convolution neural networks. Reliab. Eng. Syst. Saf. **172**, 1–11 (2018). https://doi.org/10.1016/j.ress.2017.11.021
39. C. Tingting, X. Jianlin, C. Huafeng, Improved convolutional neural network fault diagnosis method based on dropout, in *2020 7th International Forum on Electrical Engineering and Automation (IFEEA)* (IEEE, 2020)
40. A. Dey, Attention based LSTM CNN framework for sentiment extraction from Bengali texts, in *2020 11th International Conference on Electrical and Computer Engineering (ICECE)* (IEEE, 2020)
41. J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, H. Lu, Dual attention network for scene segmentation, in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (IEEE Comput. Soc., Los Alamitos, 2019), pp. 3141–3149. https://doi.org/10.1109/CVPR.2019.00326
42. S. Woo, J. Park, J.-Y. Lee, I.S. Kweon, *CBAM: Convolutional Block Attention Module* (2018)

Submit your manuscript to a SpringerOpen journal and benefit from:

- ► Convenient online submission
- ► Rigorous peer review
- ► Open access: articles freely available online
- ► High visibility within the field
- ► Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com



https://dspace.org/

Fault diagnosis of industrial robot based on dual-module attention convolutional neural network

Lu, Kaijie; Chen, Chong; Wang, Tao; Cheng, Lianglun; Qin, Jian

2022-06-01
Springer
https://doi.org/10.1007/s43684-022-00031-5
Lu K, Chen C, Wang T, et al., (2022) Fault diagnosis of industrial robot based on dual-module attention convolutional neural network. Autonomous Intelligent Systems, Volume 2, June 2022, Article number 12

Downloaded from a DSpace repository. https://dspace.org/