Based on Improved CNN Bearing Fault Detection

Yaning Xu1, Lei Yang2,*

1School of Computer and Electronic Information, Guangxi University, No.100, University East Road, Nanning, Guangxi, China, 530004
2Guangxi Academy of Sciences, No.98, Tangdaling Road, Xixiang, Nanning, Guangxi, China, 530007
*email: leiy2004@126.com

Abstract. In recent years, the problem about the fault detection of rolling bearings in mechanical equipment has gradually become an important research direction. Among them, the diagnostic method based on vibration signal analysis is widely used in the fault detection of rolling bearings. Since the one-dimensional convolutional neural network (1D-CNN) has certain limitations on the processing of vibration signal data, the solution to this problem in this paper is to integrate the attention mechanism and bi-directional long and short-term memory neural network (BiLSTM) on the basis of the one-dimensional convolutional neural network, using the attention mechanism to give different weights to different feature dimensions in the sample data and extract key and important information, thus the sample data can be further optimized. On the other hand, BiLSTM can automatically extract the deep information of the bearing vibration signal, which makes up for the deficiency of artificial extraction features to a certain extent, and strengthens the discriminative property of high-level features. Subsequently, the improved CNN bearing fault detection model was experimentally validated using the Case Western Reserve University bearing dataset, and it was concluded that the attention mechanism acting on the model obtained by adding BiLSTM to the 1D-CNN could achieve a fault identification accuracy of about 98.9% and the loss degree was reduced to about 0.17%, thus achieving an effective diagnosis of the fault state.

1. Introduction

As an essential part of the rotating mechanical equipment, the proportion of faults in the operation process is up to 30%, and the economic losses caused should not be underestimated, so the fault detection problem for rolling bearings is urgent[1].

Since 2006, the rise of deep learning has opened up a new field of fault diagnosis, under the combination of artificial intelligence, based on deep learning and neural network fault methods have also been discovered one after another. For example, one-dimensional based convolutional neural networks and long-short memory combined with attention mechanism models were recently used to predict the amount of municipal solid waste[2]. Zhangjiali from Shanghai University had conducted research on deep learning based bearing fault identification[3]. In his experiments, the two models of long and short memory network and convolutional neural network processed sequence data respectively, and came up with the result that the one-dimensional convolutional neural network is more effective and less computationally cost. Based on this, this paper employs an improved CNN
model for bearing fault detection, which enables the experimentally verified accuracy and loss to be further improved over the former.

Compared to other neural networks, convolutional neural networks have the function of feature self-extraction, and the internal hierarchical structure can effectively reduce the data processing volume and processing links, thus being more conducive to the model learning and training of data. However, since convolutional neural networks usually process two-dimensional data, but the rolling bearing raw vibration signal is one-dimensional data, a rolling bearing fault diagnosis model using one-dimensional convolutional neural networks is needed. Unlike the 2D convolution that we are often exposed to, the 1D convolution retains only a sliding window and sums the width or height direction of the feature graph, instead of performing the convolution operation in both directions simultaneously. The 1-dimensional convolutional neural network includes a filtering stage and a classification stage, the filtering stage is used to extract signal features, and the classification stage is used to classify the features extracted from the filtering stage, the two stages of the network work together to directly use the original vibration signal as the input signal for end-to-end fault diagnosis[4], which not only effectively avoids errors caused by manual signal processing and feature extraction, but also improves the accuracy of fault diagnosis. BiLSTM, as an enhanced version of the recurrent neural network, can better capture two-way semantic dependencies. On the other hand, any data has more or less redundant information, for one-dimensional vibration signal, noise inevitably becomes a key problem in fault detection, but the attention mechanism can help the model to give different weights to various parts of the input information, and extract relatively more important information to make more accurate judgments.

Therefore, the bearing fault detection model used in this paper adds the attention mechanism and the BiLSTM network to the 1D-CNN base network structure compared to zhangjiali’s test model. BiLSTM is added to the original 1D-CNN network structure to extract features from time series information in the data set. Then the attention mechanism is used to assign weight to the model. Finally, carry out classification and investigation. The improved model makes the experimental results more satisfying to some extent.

2. Based on the improved CNN bearing fault detection model

2.1. Data preprocessing

This data set is a one-dimensional vibration signal for fault diagnosis. In order to meet the training requirements of the improved CNN model, the data set enhancement method with overlapping sampling is adopted[5]. In other words, each sample overlaps with the last sample during training sample sampling of the original signal, and the sampling method is shown in Figure 1. Assuming that the selected sample length is L, step is S offset, if the data set has N data, then \((N-L)/S+1\) training samples can be obtained. Moreover, using this method to expand the amount of data can not only meet the training requirements of improved CNN, but also avoid the loss of fault features at the edges of the signal due to the non-overlapping interception method.

![Fig. 1 Overlapping sampling method](Image)
2.2. Model structure
Due to the problem that CNN cannot acquire global features of text and BiLSTM cannot focus on local features of text, the bearing fault detection model in this paper integrates the CNN network and BiLSTM network while introducing the attention mechanism to screen the extracted features.

Aiming at the limitation of feature extraction in traditional fault diagnosis methods, this paper mainly deals with the feature dimension. Different features are usually obtained after the original failure data enters the convolutional layer, and attention mechanisms are introduced to measure the importance of these different features. In this mechanism, the sample element is treated as a data pair composed of < Key, Value >, by calculating the correlation between an element Query in a given sample and keys in all samples, the weight coefficient $\alpha_i$ of each Key corresponding to the Value is obtained. Then all the weight coefficients and values are weighted and summed, and the final value $C$ is the Attention value. The principle can be expressed as:

$$C = \sum_{i=1}^{L} \alpha_i * Value_i$$  \hspace{1cm} (1)

Where $L$ is the length of the sample.

This mechanism assists the model in capturing important information and ignoring noise and redundancy in the input by scanning the global picture, which is essentially a mechanism for assigning weight parameters. The model uses the convolutional layer to extract the main features and connects the BiLSTM to extract the timing information. The main framework contains a convolutional layer, a pooling layer, a bidirectional LSTM cycle layer, an attention mechanism layer and a classification layer, the network framework structure is shown in Figure 2.

![Network framework structure](image)

Fig. 2 Network framework structure

In which the convolutional layer processes the feature mapping of the input layer can be divided into three steps: convolutional operation, batch normalization operation (BN) and activation function unit (ReLU). In the 1D-CNN network, given a vibration signal, a one-dimensional array is used as the convolution kernel, and the feature map of the one-dimensional convolution can be expressed as:

$$x_i^k = f[\sum_{i\in M} (x_{i-1}^k \cdot k_{j}^h) + b_j^h]$$  \hspace{1cm} (2)

Where, $k$ is the $k$-th layer network; $y$ is the output of the current layer; $x$ is the output of the previous layer and also the input of the current layer; $b$ is the standard deviation of the $j$-th neuron in the $k$-th layer; $f$ is the activation function ReLU function, which can be expressed as:

$$f = 1/(1 + e^{-x})$$  \hspace{1cm} (3)

The essence of the pooling layer is down sampling. To ensure the invariance of the mapping, the maximum pooling can be used, which can be expressed as:
\[ y_{i}^{k} = \max \text{pooling}(x_{i}^{k-1}, s_{\text{scale}}, s_{\text{stride}}) \]  
(4)

Where, \( y \) is the output value of the k-th layer; \( s_{\text{scale}} \) is the scale of the pooling area; \( s_{\text{stride}} \) is the step size of the pooling area.

BN is used to accelerate network training and improve the performance of deep neural networks to prevent gradient explosion and improve model accuracy[6]. The pooling feature information layer is then connected to the BiLSTM network to further improve the classification performance of the model. Subsequently, the attention module optimizes the weight parameters for the different scale features after splicing the feature matrix and outputs them to the classification layer. The classification layer consists of two fully connected layers: the first layer expands the output of the attention module to form a one-dimensional sequence. The second layer is Softmax classifier, where the experiment inputs the original vibration bearing signal into the network, automatically extracts features using the convolutional network, and then uses the unique thermal coding to set labels for the fault types and inputs them to the Softmax layer for classification, and outputs the fault type of the bearing vibration signal[7], so as to achieve the predicted target class output.

2.3. Model training

The degree of neural networks non-fit to training data is commonly represented by mean square error (MSE) and cross-entropy error (CEE). Here, the training model is the MSE loss function[8], which can reduce the Loss value as much as possible, while the lower Loss, the closer the output value of the model to the true value. In addition to this, the training model employs the Adam optimizer to optimize the model to make the neural network model converge further.

3. Experimental verification

3.1. Data set

The data of this experiment is mainly used the rolling bearing fault vibration data set which is publicly available in the Bearing Data Center of CWRU, the system sampling frequency is 12kHz, motor speed is 1797r/min. This experiment combines the rolling bearing fault by bearing diameter, inner ring fault, rolling body fault, outer ring fault, obtained 10 fault states (label 10 in normal state), as shown in Table 1:

| Fault_label | Motor Load(HP) | Inner race | Ball | Out race(centered @6:00) |
|-------------|----------------|------------|------|-------------------------|
| 0.007''     | 0/1/2/3        | 1          | 2    | 3                       |
| 0.014''     | 0/1/2/3        | 4          | 5    | 6                       |
| 0.021''     | 0/1/2/3        | 7          | 8    | 9                       |

3.2. Experimental analysis

The experimental software platform is to build the improved CNN network model in the Keras deep learning library environment with TensorFlow as the backend, and the data in the dataset is divided into training set, validation set and test set according to a certain ratio. After initializing the model parameters, the data from the training set can be used to test the accuracy and loss in each round, and finally the data of the test set can be used to evaluate the trained model. The training model parameters are shown in Table 2:

| network layer | Lead Acting Parameters | Output parameters |
|---------------|------------------------|-------------------|

The trends of loss and accuracy obtained from the model after training are shown in Figures 3:

**Table 1**

| Layer Type          | Convolution kernel size | Pooling kernel size | BiLSTM layer activation | Return sequences | Dense layer |
|---------------------|-------------------------|--------------------|--------------------------|------------------|-------------|
| Convolution layer   | 64*16, step size 16     | 128*64             | `relu`                   | True             | 8192        |
| Pool layer          | 1*2                     | 64*64              |                          |                  | 10          |
| BiLSTM layer        |                         |                    | `relu`                   |                  |             |
| Attention mechanism | module                 |                    |                          |                  |             |
| Flatten layer       |                         |                    |                          |                  |             |
| Dense layer         |                         |                    |                          |                  |             |

**Fig. 3** Accuracy and Loss of the new model's training set and validation set

**Fig. 4** Accuracy and Loss of training set and validation set of the 1D-CNN+BiLSTM model

It can be concluded from Figure 3 that the model has a negligible minimum loss rate of zero in the training set and an accuracy of up to 100%; the minimum loss rate of 0.1% in the validation set and an accuracy of 99%. Moreover, the loss degree of the test result of the evaluation standard test set of this model is about 0.17%, and the accuracy is about 98.9%. As the number of iterations increases, the loss will keep decreasing and the accuracy keeps improving.

In order to make a clearer understanding of the effect of the new model, this paper also adds a group of training effect diagrams based on the comparison model of 1D-CNN+BiLSTM. The experiment adheres to the principle of maintaining a single variable, and the network structure only removed the attention mechanism, while other parameters remained unchanged. Figure 4 is the training effect diagram.

The comparison between Figure 3 and Figure 4 shows that the accuracy of the new model and the 1D-CNN+BiLSTM model only improves slightly after training, while the loss degree decreases...
significantly. It can be seen that the new model combined with attention mechanism is more accurate and perfect in fault detection.

Even though the training time of the model is relatively longer, the lossiness is greatly reduced and the accuracy is not low compared to traditional neural networks. Therefore in future it is required to keep giving improvements to the model to make it fast running, highly accurate and very low consumption.

4. Conclusion
In this paper, a deep network model of bearing fault diagnosis is proposed, which combines the attention mechanism with the convolutional neural network and the recurrent neural network, and use adding BiLSTM and attention mechanism on the 1D-CNN infrastructure to jointly build a deep network model for bearing fault diagnosis. The experimental analysis shows that both the accuracy and loss degree of classification have been improved to some extent, and the main advantage of this model is that the high-precision discrimination can be done without manual feature extraction, which gets rid of the complexity of traditional neural networks for feature extraction. In addition to this, the use of BiLSTM has better classification results, higher accuracy under variable load, and certain generalization ability.

Acknowledgments
The research work in this paper is supported by Guangxi’s innovation-driven development Special Top "R&D and Industrialization Application Demonstration of ‘Internet+’ Engine Intelligent Manufacturing Platform" (Guike AA20302002) and Guangxi Science and Technology Base and Talent Special Project "Construction of China-Cambodia Joint Laboratory of Intelligent Manufacturing Technology" (Guike AD21076002).

References
[1] Zhang W 2017 Research on bearing fault diagnosis algorithm based on convolutional neural network Harbin Institute of Technology
[2] Lin K, Zhao Y, Tian L, Zhao C, Zhang M, Zhou T 2021 Estimation of municipal solid waste amount based on one-dimension convolutional neural network and long short-term memory with attention mechanism model: A case study of Shanghai Science of the Total Environment p 791
[3] Zhang J 2019 Bearing fault detection and identification based on deep learning https://blog.csdn.net/zhangjiali12011/article/details/90523879
[4] Wang B 2021 Study on Fault Diagnosis of Boost Bearing Based on 1-D Convolutional Neural Network Mine Machinery 49 (09) pp 29-34
[5] Wang Q 2021 Intelligent diagnosis method of rolling bearing fault based on improved one-dimensional convolutional neural network Lanzhou University of Technology
[6] Wu N, Wang Z 2021 Bearing fault diagnosis based on one-dimensional CNN and BiLSTM Modular Machine Tool & Automatic Manufacturing Technique (09) pp 38-41 p 45
[7] Zhao J, Zhao Z, Yang S 2021 Research on rolling bearing fault diagnosis based on residual connection and 1D-CNN Vibration and Shock 40(10)
[8] Zhang L, Peng S, Xiong GL, Wang L, Huang J, Hu JF 2021 Intelligent diagnosis method of locomotive wheel-to-wheel bearing based on MSE and PSO-SVM Journal of Railway Science and Engineering 18(09) pp 2408-2417