An improved deep convolutional neural network with multiscale convolution kernels for fault diagnosis of rolling bearing

Liangcheng Fu¹, Li Zhang¹ and Junyong Tao¹

¹ Laboratory of Science and Technology on Integrated Logistics Support, National University of Defense Technology, Changsha 410073, China

E-mail: f1768186115@163.com

Abstract. With the help of multiple layers of nonlinear mapping capabilities, deep neural network models can adaptively extract fault features and diagnose faults, which improve the efficiency and accuracy of fault diagnosis. Based on Deep Convolutional Neural Networks with Wide First-layer Kernels (WDCNN), this paper proposes an improved model named Deep Convolutional Neural Networks with Multiscale First-layer Convolution Kernels (MDCNN). The proposed method uses 1D convolution kernels of different sizes to extract multiscale features from original vibration signals. Afterwards, to achieve feature fusion, different feature maps are reduced to the same size through adaptive convolution operations. Finally, through the learning of the multi-layer network, intelligent fault diagnosis from the signal to the health state is realized. A test based on CWRU dataset is performed to verify the accuracy of the proposed method for rolling bearing fault diagnosis. Results indicate that MDCNN shows higher performance than WDCNN.

1. Introduction

Rolling bearings are indispensable components of rotating machinery. Once the failure of rolling bearing occurs, it will seriously damage the stability of the machine and even cause huge economic losses. Apart from factors such as manufacturing and assembly, there are two main sources of vibration in working devices with rolling bearings. One is the natural vibration caused by the bearing's own elastic characteristics, and the other is the impact shock vibration associated with operational status and the failure or damage of bearings. In order to prevent possible damage, data-driven method is commonly used to monitor the running state of the machine with the vibration signals under different working conditions.

Intelligent diagnosis method has proved to be efficient and well-performed in classifying fault modes based on fault vibration data. Generally, the intelligent fault diagnosis can be divided into feature extraction [1] and fault classification [2]. The vibration signal collected from rotational machine is original time signal, which contains useful health state information and relatively useless noise information. Common signal processing techniques used to extract representative features from the original signal include time-domain statistical analysis [3], wavelet transform [4], and spectrum analysis [5]. Besides, it is essential to screen obtained features with significant fault-related characteristics to improve the diagnosis accuracy and increase the calculation efficiency, such as principal component analysis (PCA) [6] and singular value decomposition (SVD) [7]. After extracting
and selecting features from the original signal, a classifier should be trained for fault diagnosis. Some common classifiers include classification support vector machines (SVM) [8], autoencoders (AE) [9], restricted Boltzmann machines (RBM) [10], and artificial neural networks (ANN) [11].

With the rapid development of machine learning technology, deep learning is widely applied into bearing fault diagnosis [12]. Shao et al. used the frequency band signal features extracted by the dual-tree complex wavelet packet (DTCWPT) to train the adaptive deep confidence network (DBN), which achieves great classification accuracy and robustness [13]. Shao [14] proposed the improved bidirectional long short-term memory (Bi-LSTM) neural network for fault diagnosis, and the rationality was verified in the experiment with simulation and data. In recent years, some papers have tried to use convolutional neural networks to diagnose the fault of mechanical parts [15]. LeCun [16] proposed that convolutional neural networks (CNN) have two key characteristics: space sharing weights and space pools. In terms of fault diagnosis, Janssens et al. [17] proposed a CNN model for state recognition of rotating machinery whose input is the DFT of two lines of signals collected from two vertically placed sensors. In reference [18], the input of CNN model for motor fault detection is 1D raw time series data, which successfully avoided the time-consuming feature extraction process. Zhang et al. [19] proposed WDCNN (Deep Convolutional Neural Networks with Wide First-layer Kernels) which inputs original vibration signals and adopts a wide kernel in the first convolutional layer for extracting features and suppressing high frequency noise. Although the above methods have the advantage of automatic feature learning in time domain, the fixed-size kernels they used have the same receiving field range, which results in incomplete feature data and interference from background noise.

In this paper, in order to achieve high-accuracy fault diagnosis of rolling bearings under changeable operating conditions, a diagnosis model named Deep Convolutional Neural Networks with Multiscale First-layer Convolution Kernels (MDCNN) is proposed. In the first layer network, multiple one-dimensional (1D) convolution kernels of different sizes are adopted to extract multiscale features from the original vibration signal which voids manual extraction and the fixed feature type. So as to achieve feature fusion, different feature maps are reduced to the same size through adaptive convolution operations.

The arrangement of this paper is organized as follows. A brief introduction of CNN is presented in Section 2. An improved intelligent diagnosis method based on MDCNN is proposed in Section 3. An introduction of CWRU bearing test rig and a training and testing sample acquisition method are shown in Section 4. After that, discussions about the results of tests based on CWRU dataset are presented in Section 5. Conclusions and future work are described in Section 6.

2. A Brief Introduction of CNN
This section gives a brief introduction to convolutional neural networks. Convolutional neural networks are mainly composed of convolutional layers, pooling layers, and fully connected layers. Each layer has multiple feature maps. The single feature map contains the calculation result of the previous layer output and a convolution kernel. Convolutional layers and pooling layers are the core modules to realize the feature extraction function. Similar to the traditional multi-layer perceptron, the fully connected layer corresponds to the hidden layer and logistic regression classifier. The specific operation method of layers will be described below.

2.1. Convolutional layer
The convolutional layer convolves input local regions with the convolution kernel, and each block uses the same kernel to extract features, which mean the achievement of weight sharing. We use $K_i^j$ and $b_i^j$ to represent the weight and bias of $i$-th convolution operation in layer $l$, respectively, and use $x_i$ to denote $j$-th local region in layer $l$. Therefore, the convolution process is described as follows:

$$y_i^{l+1}(j) = K_i^j * x_i^l(j) + b_i^j$$  \hspace{1cm} (1)
After the convolution operation, the rectified linear unit (ReLU) is generally applied as the activation function to enhance the representation ability of the network and make learned features more separable. The formula of ReLU is show as follows:

\[ a_i^{j+1}(j) = f\left(y_i^{j+1}(j)\right) = \max\{0, y_i^{j+1}(j)\} \]  

where \( y_i^{j+1}(j) \) is the output value of convolution operation and \( a_i^{j+1}(j) \) is the activation of \( y_i^{j+1}(j) \).

2.2. Pooling layer
In the structure of CNN, it is common to add a pooling layer after the convolution layer. As a downsampling operation, pooling can reduce the amount of data meanwhile retaining the useful information. Sampling can confuse the specific location of the feature. We only need the relative position of this feature and other features, which can satisfy the situation where the same fault feature signal appears at different locations. The max-pooling is described as follow:

\[ P_{i}^{j+1}(j) = \max_{(j-1)W+1 \leq t \leq jW} \{ a_i^{j}(t) \} \]  

where \( a_i^{j}(t) \) denotes the value of \( t \)-th neuron in the \( i \)-th frame of layer \( l \), \( W \) is the width of the pooling region, \( t \in [(j-1)W+1,jW] \) and \( P_{i}^{j+1}(j) \) denotes the value in layer \( l+1 \) after the pooling operation.

2.3. Fully connected layers
Through mapping the learned distributed features to the sample label space, the fully connected layer acts as a classifier. Actually, the fully connected layer can be realized by a global convolution operation which convolves the output of the previous layer and same dimension kernels. And reduce the number of fully connected kernels until it matches the number of classification labels. Besides, the output of the fully connected layer also needs to be transformed with the activation function.

3. The proposed Diagnosis Method MDCNN
As mentioned in the first section, convolutional neural networks have been widely used in fault diagnosis. Compared with traditional methods, some CNNs have limited exaltation, and exist some problems, such as shallow layers, insufficient fitting ability and single-scale features. The WDCNN proposed by Zhang et al. [19] adopts a deeper network but uses a fixed-length convolution kernel in the first layer. Obviously, under the condition of high load and speed, the scale of the vibration signal containing fault information is difficult to measure.

As shown in the Figure 1, the red part reflects the fault signal. When the convolution kernel is small, the difference in values between adjacent neurons is reduced, resulting in less prominent feature information and lower sparsity. At the same time, if the larger convolution kernel is used, the fault information can be highlighted and the sparsity can be increased. Therefore, the fixed-length convolution kernel has limitations.

![Figure 1. Comparison of features extracted by convolution kernels of different sizes.](image-url)
The above problems can be solved by the proposed MDCNN. In MDCNN, when extracting features from the original signal, convolution kernels and step of different sizes are selected, so that maximum fault information can be extracted. Different convolution kernels result in different feature map sizes, so the networks in the previous layers are multi-channel parallel. For large feature maps, convolution and pooling are used to reduce the size, and for small feature maps, 1x1 convolution [20] is adopted to increase the nonlinearity while maintaining the size. After mapping through multiple layers of networks, feature maps will reach the same size and be merged into a whole which will contain the multiscale features extracted from original vibration signals. Ultimately, we use a deeper convolutional neural network to get the diagnosis result. To sum up, MDCNN has a deeper network, better mapping effect and multiscale features than previous CNN methods, which make fault diagnosis have the superiority. The overall framework of proposed MDCNN is shown in Figure 2.

![Figure 2. The overall framework of MDCNN.](image)

### 3.1. Framework of MDCNN

![Figure 3. Framework of MDCNN.](image)
As shown in Figure 3, an original vibration signal of the rolling bearing is input to MDCNN. In the first layer of network, we selected four convolution kernels of different sizes to extract feature maps. The difference with the general CCN is in the first three-layer network, especially the first-layer network whose feature extraction is comprehensive. Besides, the wide size kernel suppresses high frequency noise and the narrow can obtain weak fault features.

3.2. Detailed parameters of MDCNN

As shown in Table 1, the MDCNN contains ten layers. The initial three-layer network contains four parallel convolution operations, which are used to extract multi-scale features and reduce dimension for merging features. The following multiple convolutional layers or pooling layers are used to further improve the generalization ability of the model. Ultimately, the accurate fault judgment is achieved through three fully connected layers.

| Layer | Name            | Kernel size/stride | Output shape |
|-------|-----------------|--------------------|--------------|
| 1     | Conv1_1         | 16x1/8             | 256x8        |
|       | Conv1_2         | 32x1/16            | 128x8        |
|       | Conv1_3         | 64x1/32            | 64x8         |
|       | Conv1_4         | 128x1/64           | 32x8         |
|       | Conv2_1         | 2x1/2              | 128x8        |
|       | Conv2_2         | 2x1/2              | 64x8         |
| 2     | Conv2_3         | 2x1/2              | 32x8         |
|       | Conv2_4         | 1x1/64             | 32x8         |
|       | Conv3_1         | 4x1/4              | 32x16        |
|       | Conv3_2         | 4x1/2              | 32x16        |
| 3     | Conv3_3         | 1x1/1              | 32x16        |
|       | Conv3_4         | 1x1/1              | 32x16        |
|       | FeatureMerging  | -                  | 32x64        |
| 4     | Conv4           | 3x1/1              | 32x128       |
| 5     | MaxPooling1     | 2x1/2              | 16x128       |
| 6     | Conv5           | 3x1/1              | 16x256       |
| 7     | MaxPooling2     | 2x1/2              | 8x256        |
| 8     | FeatureExpansion| -                  | 2048x1       |
| 9     | FC1             | 2048x800           | 800x1        |
| 10    | FC2             | 400x800            | 200x1        |
|       | FC3             | 200x10             | 10x1         |

3.3. Training of MDCNN

In order to identify the type of faults, the softmax function is usually used to convert the output of the last fully connected layer into a probabilistic form. The function is shown as follows:

$$q(z_j) = \frac{e^{z_j}}{\sum e^{z_k}}$$  \hspace{1cm} (4)

where $z_j$ represents the output of the $j$-th neuron, and $q(z_j)$ denotes the probability of this fault type. We encode true labels by one-hot Equation (5), and choose cross entropy as the loss function which is shown as follows:
\[
\text{onehot} : \begin{bmatrix}
[1,0,0,...,0]_n \\
[0,1,0,...,0]_n \\
... \\
[0,0,0,...,1]_n
\end{bmatrix}
\]

(5)

\[
\text{loss} = -\sum_i p(i) \log q(i)
\]

(6)

where \( p(i) \) denotes a real label corresponding to a sample in onehot. With the goal of minimizing loss, the network updates the weights and bias of each layer through a gradient descent algorithm called Adam [21]. In order to improve the diagnosis effect and prevent overfitting, the fully connected layer uses the Dropout method which can disable neurons without participating in training according to a certain probability.

4. Vibration data acquisition for faulty rolling bearings

4.1. CWRU dataset

The vibration data of fault rolling bearings used in this paper is based on CWRU bearing dataset [23]. The bearing fault test rig is shown in the Figure 4. The object of the experiment is the drive end bearing (SKF6205) in Figure 4. The bearing data is collected by the acceleration sensor under loads 0, 1, 2 and 3 horsepower (HP) with a sampling rate of 48 kHz. The vibration signals used in this study include normal state, outer ring failure, inner ring failure and rolling element failure. Because the difference of the outer ring fault location will cause the vibration to change, the data of the fault location at 6 o'clock is selected. The three types of defects are generated by point spark machining, and the failure sizes are 7mils, 14mils, and 21mils. According to different working conditions, data sets are named A, B, C, D, as shown in the Table 2.

![Figure 4. Motor driving mechanical system in CWRU.](image)

**Table 2.** Working condition of the corresponding data set.

| Name | Load(HP) | Rotational speed(rpm) | Failure sizes(mils) |
|------|----------|-----------------------|---------------------|
| A    | 1        | 1772                  | 7,14,21             |
| B    | 2        | 1750                  | 7,14,21             |
| C    | 3        | 1730                  | 7,14,21             |
| D    | 0        | 1797                  | 7,14,21             |
4.2. Data Augmentation

![Schematic diagram of data augmentation methods.](image)

**Figure 5.** Schematic diagram of data augmentation methods.

So as to obtain more samples, this paper uses overlapping segmentation for vibration data, which is shown in Figure 5. Through 240-step overlap sampling, all data sets with a length of 2048 are obtained. Each failure state extracts 1800 training samples and 200 test samples. The specific number of samples is given in Table 3.

**Table 3.** Data sets of rolling bearings in different states.

| Data set | Fault Location | None | Ball | Inner Race | Outer Race |
|----------|----------------|------|------|------------|------------|
| A        | Training       | 1800 | 1800 | 1800       | 1800       |
| A        | Testing        | 200  | 200  | 200        | 200        |
| B        | Training       | 1800 | 1800 | 1800       | 1800       |
| B        | Testing        | 200  | 200  | 200        | 200        |
| C        | Training       | 1800 | 1800 | 1800       | 1800       |
| C        | Testing        | 200  | 200  | 200        | 200        |
| D        | Training       | 1800 | 1800 | 1800       | 1800       |
| D        | Testing        | 200  | 200  | 200        | 200        |

5. Results and discussions

5.1. Validation of the Proposed MDCNN

In order to verify the effectiveness of the MDCNN model, we tested all data sets separately. Considering the influence of the random initial value of the network, the data set was repeatedly trained and tested ten times, and average results are shown in the Figure 6. When the training step reaches to 2000, the accuracy of all data sets exceeds 99.0%, which shows a great diagnostic performance. Data sets C and D require excessive training time. Because of the none load and high load, the fault characteristic information is weak or affected by high noise. The Figure 7 shows the change of the loss value and accuracy when we employ the data set B to train MDCNN model.

![Diagnostic accuracy of each data set under different training times.](image)

**Figure 6.** Diagnostic accuracy of each data set under different training times.
5.2. Diagnostic analysis under variable load
The diagnostic results have a very high accuracy rate under a single working condition, because the training data and the test data have the same distribution. However, under actual working conditions, the state of the machine will fluctuate or change. Therefore, in order to verify the generalization ability of MDCNN, we adopted the cross-test method. A→B means that the model is trained under the A working condition and tested under the B working condition. Similarly, we conducted A→C, B→A, B→C, C→A, C→B experiments, and compared with other models such as SVM [8], Resnet [22], WDCNN, as shown in the Figure 8. Obviously, WDCNN and MDCNN have a great performance in cross-testing. And the proposed MDCNN generally exceeds WDCNN, which shows that the use of multi-size convolution kernels can extract multiscale features and improve the performance of the network.

Figure 7. Index change of the data set B during a training session.

Figure 8. Results of the proposed MDCNN of six domain shifts on the data sets A, B and C, compared with FFT-SVM, Resnet, WDCNN.

In order to further study the diagnostic performance of the model under complex operating conditions, one of the four data sets is used as the training set, and the remaining three sets are served as the testing set. As shown in the Figure 9, when the training set is single and the test set data structure and distribution complexity are high, it is difficult to achieve a high diagnostic effect. Although the effectiveness of diagnosis decreases, the MDCNN based on multiscale feature extraction still has the advantage of fault diagnosis under complex conditions.
Figure 9. Results of MDCNN under complex operating conditions compared with WDCNN.

6. Conclusions
In this paper, a model named MDCNN makes some improvements to the traditional deep convolutional neural networks. The use of 1D convolution kernels of different sizes and the parallel convolution operations improve the ability of feature extraction. In some network layers, 1x1 convolution kernels and Dropout are used to enhance nonlinear fitting and generalization capabilities. Based on CWRU bearing dataset and overlapping sampling, a large number of reliable samples can be used for training and testing. In the experiment under a single working condition, the MDCNN has a very high accuracy rate which exceeds 99%. Furthermore, in the case of changing conditions, the diagnostic performance obtains a greatly promotion compared to WDCNN, and reflects a strong robustness. Due to the limited training data, the performance of MDCNN showed a downturn in study of complex working conditions. Because of multiscale feature extraction, MDCNN still has the advantage of fault diagnosis. For actual engineering applications, MDCNN can be optimized by adding data. Considering the training time, this model requires a choice between the number of multiscale convolution kernels and the computational efficiency of the model. In addition, the use of multiscale convolution kernels will result in redundant features. It is still a technique problem to be optimized in different working scenarios.

References
[1] Haidong S, Hongkai J and Ying L 2018 A novel method for intelligent fault diagnosis of rolling bearings using ensemble deep auto-encoders J. Mechanical Systems and Signal Processing 102 278-97
[2] Wang C, Gan M and Zhu C 2017 Intelligent fault diagnosis of rolling element bearings using sparse wavelet energy based on overcomplete DWT and basis pursuit J. Journal of Intelligent Manufacturing 28(6) 1377-91
[3] Wang X, Zheng Y and Zhao 2015 Bearing fault diagnosis based on statistical locally linear embedding J. Sensors 15 16225-47
[4] Wang D, Zhao Y and Yi C 2018 Sparsity guided empirical wavelet transform for fault diagnosis of rolling element bearings J. Mechanical Systems & Signal Processing 101 292-308
[5] Seryasat O R, Shoorehdeli M A and Honarvar F 2010 Multi-fault diagnosis of ball bearing using FFT, wavelet energy entropy mean and root mean square (RMS) C. IEEE International Conference on Systems Man & Cybernetics
[6] Zheng J, Liu T, Meng R and et al 2018 Generalized composite multiscale permutation entropy and PCA based fault diagnosis of rolling bearings J. Journal of Vibration and Shock 37(20) 61-66
[7] Jiang H, Chen J, Dong G and et al 2015 Study on Hankel matrix-based SVD and its application in rolling element bearing fault diagnosis J. Mechanical Systems & Signal Processing 52-53 338-59

[8] Jun Zhong X, Tao S U, Yang Z and et al 2014 Fault Diagnosis Method of Rolling Bearings Based on Ensemble Empirical Mode Decomposition Energy Entropy and LS-SVM J. Noise and Vibration Control

[9] He Y, Zhang X and Friswell M I 2009 Defect Diagnosis for Rolling Element Bearings Using Acoustic Emission J. Prognostics & System Health Management Conference

[10] Deng S, Cheng Z, Li C and et al 2017 Rolling bearing fault diagnosis based on Deep Boltzmann machines C. Prognostics & System Health Management Conference

[11] Babu N T, Aravind A, Rakesh A and et al 2018 Application of EMD, ANN and DNN for Self-Aligning Bearing Fault Diagnosis J. Archives of Acoustics 43(2) 163-75

[12] Chen Y, Peng G, Xie C and et al 2018 ACDIN: bridging the gap between artificial and real bearing damages for bearing fault diagnosis J. Neurocomputing

[13] Shao H, Jiang H, Wang F and et al 2017 Rolling bearing fault diagnosis using adaptive deep belief network with dual-tree complex wavelet packet ISA J. Transactions

[14] Qiu D, Liu Z, Zhou Y and et al 2019 Modified Bi-Directional LSTM Neural Networks for Rolling Bearing Fault Diagnosis C. ICC 2019-2019 IEEE International Conference on Communications (ICC)

[15] Li X U, Ge L I, Liang Y U and et al 2018 Roller Bearing Fault Diagnosis Based on Wavelet Packet and CNN J. Journal of Sichuan University of Science & Engineering (Natural Science Edition)

[16] Lecun Y and Bottou L 1998 Gradient-based learning applied to document recognition. J. Proceedings of the IEEE 86(11) 2278-324

[17] Janssens O, Slavkovik J, Vervisch B and et al 2016 Convolutional Neural Network Based Fault Detection for Rotating Machinery J. Journal of Sound and Vibration 331-45

[18] Ince T, Kiranyaz S, Eren L and et al 2016 Real-Time Motor Fault Detection by 1-D Convolutional Neural Networks J. IEEE Transactions on Industrial Electronics 63(11):7067-75

[19] W. Zhang, G. Peng and et al 2017 A New Deep Learning Model for Fault Diagnosis with Good Anti-Noise and Domain Adaptation Ability on Raw Vibration Signals J. sensor

[20] Lin M, Chen Q and Yan S 2013 Network In Network J. Computer Science

[21] Kingma D P, Ba J 2014 Adam: A Method for Stochastic Optimization J. Computer Science.

[22] He K M, Zhang X Y and Ren S Q 2016 Deep Residual Learning for Image Recognition C. IEEE Conference on Computer Vision & Pattern Recognition

[23] Loparo K 2012 Bearings Vibration Data Set Case Western Reserve University available online: http://www.eecs.case.edu/laboratory/bearing/welcome_overview.htm