Bearing Fault Diagnosis Using Structure Optimized Deep Convolutional Neural Network under Noisy Environment

Guoqiang Jin, Dongjie Li, Ye Wei, Wenhui Hou, Huai’an Chen, Yi Jin and Chang’an Zhu

Department of Precision Machinery and Precision Instrument, University of Science and Technology of China, Hefei, Anhui, 230026, P.R.China.

jinyi08@ustc.edu.cn, jgq@mail.ustc.edu.cn

Abstract. In recent years, deep learning has been gradually applied in bearing fault diagnosis thanks to its powerful learning and representation ability. However, deep learning-based methods are usually computationally intensive and have redundant parameters. In this paper, we propose an end-to-end structure optimized deep convolutional neural network for bearing fault diagnosis using the local sparse structure. The width and depth of the network are increased and the redundant parameters are reduced while maintaining the representation capabilities and performance of the network. The local sparse structure increases the computational efficiency and reduces the risk of overfitting. The proposed method directly employed the raw signals as input without extra feature extraction with the ability that could distinguish both the bearing fault types and the corresponding severity. Experiments show that the proposed method achieves similar performance compared to the original method by using 46.47% parameters as used in the original study, even under the noisy environment.

1. Introduction

Rolling bearings are the critical parts in rotating machinery and their failure will lead to serious consequences [1]. Bearing fault diagnosis plays a key role in maintaining the normal operation of the machine. In the past few decades, bearing fault diagnosis based on data-driven methods have been widely carried out [2]. Early signal processing and learning-based methods have shown their weakness in dealing with signals in complex scenes. These methods require heavily handcrafted preprocessing procedures and expert knowledge. In recent years, deep learning has been gradually applied for fault diagnosis in bearings [3]. Deep learning has a good feature recognition and representation capability, and can automatically extract features without complex handcrafted preprocessing procedures.

The recent deep learning approaches in bearing fault diagnosis focus on extracting features from the raw signals without manual features extraction operation, providing an end-to-end classification method. In [4], the proposed adaptive overlapping convolutional neural network (AOCNN) took the augmented raw signals as the input and used the adaptive convolutional layer with overlapping and root-mean-square pooling layers to overcome the shift variant and marginal problems. Chen et al. [5] proposed a deep Inception net with atrous convolution (ACDIN) method that employed the Inception structure with the raw signal as input to classify the real bearing failure after training on the signals which were collected from the artificially damaged bearings. Zhang et al. [6] employed the convolutional neural network (CNN) as the feature extractor to mine features from the raw signal without complicated preprocessing. The CNN found to be a powerful feature exactor and classifier. In [7], the same author proposed the wide first-layer kernels convolutional neural network (WDCNN) to
classify the raw signal under noisy condition. However, these aforementioned methods are computationally intensive. They require a considerable large number of parameters. Inspired by the Inception structure [8], we presented a local sparse structure-based neural network for fault diagnosis in bearings which is computational efficiency and requires fewer parameters. Therein, the local sparse structure is used to replace the computationally intensive layers in the original deep neural network, especially the fully connected layer. The proposed method is an end-to-end neural network that took the raw signal as the input without handcrafted feature selection procedure and could classify the bearing fault types and the corresponding severity.

2. The basic theory of the neural network

The convolutional neural network [9] was inspired by the biological neural system that combines layers composed of multiple neurons. Each neuron in a layer denotes as a kernel that mainly focuses on a small reception area which greatly reduces the number of parameters compared to the fully connected network. The basic operation in CNN is the convolution operation. The CNN could extract features with variable length of the input and its computation cost is relatively low due to the weight sharing policy among the kernels in the same filter layer. The batch normalization (BN) is proposed to solve the internal covariate shift problems by normalizing the input distribution of a layer [10]. The BN layer is usually added after a convolutional layer before the activation function. The BN could accelerate the training process and improve the stability of the deep neural network.

The pooling uses the overall statistical characteristics of the adjacent regions to replace the output of the network in that region [11]. The pooling has the local invariant characteristic to the input and it reduces the parameters of the output. The maximum pooling and average pooling are two common pooling operations. The maximum pooling uses the maximum value in a local region to represent that region while the average pooling uses the average value of the local region.

3. The proposed structure optimized neural network

3.1. The Inception structure

The Inception structure was originally proposed to increase the representation capability, i.e. the depth and width of the network, while keeping the computational consumption. Directly increasing the depth of the neural network by stacking the layers would increase the network size dramatically and would lead to overfitting more likely. A reasonable way to solve the problem is to replace the computationally intensive layers by the sparse ones. Assuming the projection of some region in the input to a cluster of neurons in the later layers with high correlation, then we could use a layer of 1x1 convolutions to represent this cluster [12]. Similarly, if there would be a more spatially spread out clusters, then the clusters can be represented by a larger kernel in the convolutions, i.e. the 3x1 convolutions. The concatenation of these layers would have a similar feature representation as using the normal deep structure. One of the proposed local sparse structure is shown in figure 1(a).

![Figure 1. The proposed Inception module: (a) the naive version, (b) the dimensionality reduction version.](image-url)
In order to improve the representation ability of network and to reduce the dimension of the concatenated layer, a parallel path for pooling operation followed by 1x1 convolutions is added. The 1x1 convolutions are employed to compress the high dimensional information, and are used as the nonlinear activation function after the pooling operation. The dimensionality reduction structure is depicted in figure 1(b). The stride of the 1x1 and 3x1 convolutions on the right is 2 in order to reduce feature map size. Specifically, the number of the filters in the independent 1x1 convolutions and 3x1 convolutions are 32 and 16 respectively for both the naive and dimensionality reduction versions. The number of filters in the 1x1 convolutions after the maximum pooling is 8. The Inception module improves network representation by increasing the width of the network while avoiding a dramatic increase in computational complexity.

3.2. The architecture of the proposed neural network
The CNN architecture has proven to be an effective method for the bearing fault diagnosis as reviewed above. However, due to large numbers of parameters to be calculated, the deep neural network based methods suffer the redundant calculation problem. We proposed a structure optimized network based on deep convolutional neural network using the local sparse structure to reduce the number of parameters and increase the computational efficiency while keeping the performance. The proposed architecture is shown in figure 2.

Figure 2. The architecture of the proposed neural network.

This proposed model is based on the WDCNN with the Inception structure improvement. Except for the last two local sparse structures and the average pooling operation in the network, the remaining parameters are the same as in the original network. The core idea of the proposed method is to replace the high dimensional convolutional layer and the fully connected layer in the deep neural network with the Inception structures as shown in figure 1. For convenience, we refer the structure in figure 1(a) as A, the structure in figure 1(b) as B, the structure before B shown in figure 2 is called the Base, and the convolutional block in the original WDCNN replaced by block B in figure 2 is called C. The proposed method is mainly composed of the Base block, the B block, and the A block. The effects of different combinations of these blocks are discussed in section 4.3. The Max and Avg modules in the figure represent the maximum pooling and average pooling, respectively. The © mark represents the concatenation operation to concatenate the features from different paths. Since there is no fully connected layer in the proposed network, the number of the network parameters has been reduced a lot compared to the original network, thus reduces the risk of overfitting.

4. Validation of the proposed neural network
4.1. Dataset description
The bearing fault dataset we used is from the Case Western Reserve University (CWRU) Bearing Data Center [13]. The vibration data is collected at the drive end with the sample rate of 48000Hz, and the motor load is 3HP. There are four types of health conditions: (1) normal condition; (2) ball fault; (3) inner race fault; (4) outer race fault. The fault sizes are 7 mils, 14 mils, and 21 mils (1 mil = 0.001 inches) for ball fault, inner race fault, and outer race fault types respectively. There are 10 different classes of data with 289 samples in each class, totally 2890 samples. We randomly select 80% samples
as the training set and the rest as the test set. The input length of the network is 1670 which is calculated from the number of points collected by the motor with one rotating revolution. The amplitude of the raw signal is scaled to [-1, 1]. Noise is widespread, and the Gaussian white noise is one of the most representative and easily quantifiable noises. In order to evaluate the performance of the proposed neural network under noise condition, we add the additive white Gaussian noise with different levels of signal-to-noise ratio (SNR) into the raw data. The SNR is defined as follows:

$$\text{SNR}_{\text{dB}} = 10 \log_{10} \left( \frac{P_{\text{signal}}}{P_{\text{noise}}} \right)$$  \hspace{1cm} (1)$$

where $P_{\text{signal}}$ is the power of the original signal and $P_{\text{noise}}$ is the power of the noise signal. We test the proposed method with the SNR of noisy signals ranging from $-10$ dB to 10 dB.

### 4.2. Training details

We compared the proposed network with the original WDCNN. Both methods used the same scaled raw signal as the input without further feature extraction procedure. The Adam [14] algorithm was used as the optimizer. The batch size was fixed to 64. We trained the original WDCNN and the proposed methods on signals without adding noise and test on signals with different levels of noise. Ten trials were conducted on each method to get the final result. The number of trainable parameters in the proposed method is 22356 while in the original WDCNN the number is 48112. In other words, our proposed method used 46.47% parameters as used in the original study. In order to study the effects of different structures on the network performance and computational cost as mentioned before, we also conducted experiments on different combinations of blocks and finally selected the best one.

### 4.3. Experiment result

The classification results for several different structures are shown in table 1. From the table, it can be seen that the A block has more parameters than the B block and is more suitable as the last block in the network. The proposed method (Base+B+A) has the best accuracy with small standard deviation and a relatively small amount of parameters compared to the other combinations of structures, indicating that the proposed structure optimized network is efficient in reducing parameters while keeping the performance of the neural network.

#### Table 1. Comparison of accuracy with different structures at $-4$ dB SNR.

| Structures     | Original       | Base+C+B     | Base+C+A     | Base+A+B    | Base+B+A     |
|----------------|----------------|--------------|--------------|-------------|--------------|
| Accuracy (%)   | 79.64 ± 4.70   | 77.44 ± 4.03 | 79.53 ± 3.10 | 78.81 ± 4.24 | 80.50 ± 2.91 |
| Parameters     | 48112          | 28452        | 29196        | 22068       | 22356        |

![Figure 3. Comparison of classification accuracy with different SNR values.](image)
The classification result is shown in figure 3. The error bar on each column indicates the standard deviation of each group of 10 trials. From the figure, it is clear that the classification accuracy drops and the standard deviation rises as the level of the noise increases. It is due to the reason that it is difficult to extract useful information from the noisy signals. The classification results indicate that the proposed method has almost the same performance as the original method, even slightly better in high noise conditions. The proposed structure optimized method reduces the number of parameters in the high-level layers of the network and reduces the risk of overfitting. After extracting features from low-level layers in the network, high-level layers networks with local sparse structure are easier to learn global information, thus improve the network representation.

5. Conclusion
In this paper, a structure optimized deep convolutional neural network with reduced parameters was proposed to diagnose the fault in rolling bearings. Experiments proved that by replacing the high dimensional convolutional layer and the fully connected layer with the local sparse structure can greatly reduce the number of network parameters without compromising the performance. The proposed deep convolutional neural network could distinguish not only the bearing fault types but also the corresponding severity. The proposed method is an end-to-end signal identification and classification neural network, which is a promising architecture to apply to other signal processing scenarios.

Acknowledgments
This work was supported by the National Natural Science Foundation of China (No. 51605464).

6. References
[1] Tandon N and Choudhury A 1999 A review of vibration and acoustic measurement methods for the detection of defects in rolling element bearings. Tribol. Int. 32 469-480
[2] Cerrada M, Sánchez R, Li C, Pacheco F, Cabrera D, Valente de Oliveira J and Vásquez R E 2018 A review on data-driven fault severity assessment in rolling bearings. Mech. Syst. Signal. Pr. 99 169–196
[3] Liu R, Yang B, Zio E and Chen X 2018 Artificial intelligence for fault diagnosis of rotating machinery: A review. Mech. Syst. Signal. Pr. 108 33–47
[4] Qian W, Li S, Wang J, An Z and Jiang X 2018 An intelligent fault diagnosis framework for raw vibration signals: adaptive overlapping convolutional neural network. Meas. Sci. Technol. 29 095009
[5] Chen Y, Peng G, Xie C, Zhang W, Li C and Liu S 2018 ACDIN: Bridging the gap between artificial and real bearing damages for bearing fault diagnosis. Neurocomputing. 294 61–71
[6] Zhang W, Peng G and Li C 2016 Advances in intelligent information hiding and multimedia signal processing (New York: Springer International Publishing) pp 77–84
[7] Zhang W, Peng G, Li C, Chen Y and Zhang Z 2017 A new deep learning model for fault diagnosis with good anti-noise and domain adaptation ability on raw vibration signals. Sensors. 17 425
[8] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V and Rabinovich A 2015 IEEE Conf. on computer vision and pattern recognition (New Jersey: IEEE)
[9] LeCun Y, Bottou L, Bengio Y and Haffner P 1998 Gradient-based learning applied to document recognition. P. IEEE. 86 2278–2324
[10] Ioffe S and Szegedy C 2015 Proc. of the 32nd Int. Conf. on machine learning vol 37, ed F Bach and D Blei (Lille: PMLR) pp 448–456
[11] Boureau, Y-Lan, Jean P and LeCun Y 2010 Proc.of the 27th Int. Conf. on machine learning (USA: Omnipress) pp 111–118
[12] Min L, Chen Q and Yan S 2014 Network in network CoRR abs/1312.4400
[13] Lou X and Loparo KA 2004 Bearing fault diagnosis based on wavelet transform and fuzzy inference. Mech. Syst. Signal. Pr. 18 1077–95
[14] Kingma D P, and Ba J 2015 3rd Int. Conf. for learning representations (New York: Ithaca)