Understanding and Learning Discriminant Features based on Multiattention 1DCNN for Wheelset Bearing Fault Diagnosis

Huan Wang, Zhiliang Liu, *Member, IEEE*, Dandan Peng, and Yong Qin, *Member, IEEE*

***Abstract*—Recently, deep-learning-based fault diagnosis methods have been widely studied for rolling bearings. However, these neural networks are lack of interpretability for fault diagnosis tasks. That is, how to understand and learn discriminant fault features from complex monitoring signals remains a great challenge. Considering this challenge, this article explores the use of the attention mechanism in fault diagnosis networks and designs attention module by fully considering characteristics of rolling bearing faults to enhance fault-related features and to ignore irrelevant features. Powered by the proposed attention mechanism, a multiattention one-dimensional convolutional neural network (MA1DCNN) is further proposed to diagnose wheelset bearing faults. The MA1DCNN can adaptively recalibrate features of each layer and can enhance the feature learning of fault impulses. Experimental results on the wheelset bearing dataset show that the proposed multiattention mechanism can significantly improve the discriminant feature representation, thus the MA1DCNN outperforms eight state-of-the-arts networks.**

***Index Terms*—Attention mechanism, convolutional neural network (CNN), deep learning, fault diagnosis, wheelset bearing.**

I. Introduction

ROLLING bearings are widely used in various industrial equipment. However, under complex conditions such as high speed, high workload, and strong impact for a long time, rolling bearings are prone to occur wear, spalling, or other faults, which lead to performance degradation of the equipment and even lead to safety accidents. Therefore, accurate fault diagnosis of rolling bearings has always been an important engineering topic.

In recent years, intelligent fault diagnosis based on machine learning has attracted wide attention, which mainly includes two steps: fault feature extraction and fault mode classification. For the former step, fault features in time domain, frequency domain, and time–frequency domain [1] are extracted from monitoring signals by Hilbert–Huang transform [2], empirical mode decomposition [3], or local mean decomposition [4]. For the latter step, these features are fed into machine learning models to obtain diagnosis results, such as k-nearest neighbor [5], random forest [6], or support vector machine [7].

However, there are the following three disadvantages in the aforementioned conventional intelligent fault diagnosis methods. First, they heavily rely on specific domain knowledge. Effective features are hardly extracted to represent complex dynamic behavior of faulty rolling bearings. Second, they mostly adopt shallow networks, which hardly deal with complex working conditions, such as strong noise interference and time-variant working conditions. Finally, they have difficulty in learning with large-scale datasets.

Deep learning [8], as a fused feature extraction and classification method, is expected to provide an automatic and effective end-to-end solution for rolling bearing fault diagnosis. The core of deep learning is feature learning, which aims to learn discriminant fault feature from the input data by modeling high-level abstraction of data with deep networks. It can address the aforementioned three disadvantages in the conventional intelligent fault diagnosis. Therefore, denoising autoencoder [9], deep belief network [10], and convolutional neural network (CNN) [10] have been studied for rolling bearing fault diagnosis. In particular, the CNN has achieved remarkable success in fault diagnosis [12]–[20], speech recognition [21], and image processing [22] due to its good characteristics of local weight sharing, local receptive field, and multiconvolution kernels. According to network input, the CNN-based rotating machinery fault diagnosis can be grouped into the following two categories: one is based on one-dimensional (1-D) CNN (1DCNN) [12]–[14], [16], [17], [19], [23], [24]; the other converts the 1-D signals into the two dimensional (2-D) images, and then, uses 2-D CNN (2DCNN) to implement fault diagnosis [15], [18], [20], [25]–[27]. Compared with the 2DCNN, the 1DCNN has the following advantages. First, the 1-D temporal

Manuscript received July 2, 2019; revised October 9, 2019; accepted November 9, 2019. Date of publication November 25, 2019; date of current version May 26, 2020. This work was supported in part by the National Natural Science Foundation of China under Grant 61833002, in part by the State Key Laboratory of Rail Traffic Control and Safety under grant RCS2018K002, in part by Beijing Jiaotong University, and in part by the State Key Laboratory of Traction Power, Southwest Jiaotong University under Grant TPL1608. Paper no.TII-19-2988. *(Corresponding authors: Zhiliang Liu; Yong Qin.)*

H. Wang, Z. Liu, and D. Peng are with the School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: wh.huanwang@gmail.com; zhiliang_liu@uestc.edu.cn; dandanpeng2@gmail.com).

Y. Qin is with the State Key laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing 100044, China (e-mail: yqin@bjtu.edu.cn).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TII.2019.2955540

1551-3203 © 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: University of Georgia. Downloaded on September 21,2023 at 03:55:21 UTC from IEEE Xplore. Restrictions apply.

signal is directly collected from data acquisition system, so it is more natural to use the 1-D input. Second, the 2DCNN requires additional 1-D-to-2-D conversion process (e.g., time–frequency representation method) that may lose some useful information related to faults due to irreversible conversion. Finally, the 2-D input usually has a higher dimension than the 1-D input, which makes the CNN more complex. For example, the input layer of the 2DCNN in [27] is $224 \times 224 \times 3$. Therefore, the 1DCNN is used in this article.

In addition, the previous works mainly focus on improving the automatic feature learning ability of the CNN. It still has the following challenges.

1) Lacking fault signal learning mechanism. For example, impulse segments often reflect intrinsic behavior of faulty rolling bearings. However, the conventional CNN treats the impulse segments the same with the rest fault-irrelevant segments in the temporal signals.
2) Lacking discriminant feature learning mechanism. Different features in the CNN are of different importance for fault diagnosis tasks. The conventional CNN cannot adaptively focus on learning more discriminant features and ignore useless features.
3) Poor network interpretability. The conventional CNN is still a "black box." This "black box" characteristic greatly affects the CNN's development in the field of rolling bearing fault diagnosis, because interpretability is very important to fully control this technology.

To deal with the aforementioned challenges, this article explores the CNN improvement with attention mechanism. We propose the following three feature optimization modules for rolling bearing fault diagnosis: 1) channel attention module (CAM), 2) excitation attention module (EAM), and 3) joint attention module (JAM). The CAM uses global information to adaptively enhance more discriminant features and to suppress irrelevant features by explicitly modeling the interdependence between convolution feature channels. The EAM is dedicated to locating fault impulses and to optimizing feature response of the proposed network, so as to enhance fault-related temporal features and to ignore irrelevant temporal features. Finally, the JAM optimizes feature mapping by using both channel and excitation attention mechanisms, and then, obtain the output by fusing features of the two modules.

Based on the aforementioned attention modules, a multiattention 1DCNN (MA1DCNN) is proposed to diagnose wheelset bearing faults. This network adaptively optimizes feature mapping by embedding attention modules in different network depths. This advantage can be accumulated in the whole network by adding multilayer attention modules. In addition, attention mechanism can learn the relation between the target output and the temporal input, explore and find the most relevant activation maps and temporal signal segments for fault diagnosis tasks, so as to enhance relevant features and to suppress irrelevant features. Therefore, in this way, we can explore the potential relation between the input and the output, and can expect to learn more explanatory knowledge, so as to improve the interpretability of the CNN.

The contributions of this article are summarized as follows:

1) This article explores attention mechanism in improvement direction of 1DCNN, and proposes three attention modules that physically connect to rolling bearings: the CAM, the EAM, and the JAM. The proposed attention modules provide an effort to understand the "black box" mechanism of the CNN.
2) The MA1DCNN that uses attention mechanism is proposed as an end-to-end approach to diagnose wheelset bearing faults. To the best of our knowledge, this is the first time to use multiattention-mechanism-based 1DCNN in rotating machinery fault diagnosis.
3) Compared with eight state-of-the-arts fault diagnosis methods, the proposed MA1DCNN achieves better diagnosis results on the wheelset bearing dataset.

This article is organized as follows. In Section II, the proposed MA1DCNN is described in detail. Section III verifies effectiveness and superiority of the MA1DCNN. Section IV provides discussions on the proposed MA1DCNN. Section V summarizes this article.

II. Methodology

The convolution layer in the 1DCNN, whose input is a feature representation, fuses temporal information and channel information in the local receptive field to generate a new set of feature representations as output. It consists of a set of activation maps, each of which represents different feature types. The activation maps are connected along a nontime dimension (namely channel dimension), to form feature representation. Here, we define a set of the 1-D convolution transforms, $F : X \mapsto X', X \in \mathbb{R}^{W \times C}, X' \in \mathbb{R}^{W' \times C'}$, where W is the length of activation maps, with C and C' being the input and output channels, respectively. Each convolution operation encodes features among adjacent temporal signals in the input feature X, and output features that have greater information perception ability. Therefore, by superimposing multilayer convolution operations, the network can encode a more global feature representation from low-level features. Researchers often construct a deeper CNN [22] or a new network structure [14] to achieve better fault diagnosis performance. However, insufficient attention has been paid to how to effectively improve the CNN's performance to obtain more discriminant fault features and to explore the potential feature learning mechanism of the CNN.

With the help of attention modules, the MA1DCNN pays different attention to different activation mappings and temporal signal segments, thus enhancing the learning ability for fault-related features. In this way, the network can extract more discriminant features, and its learning mechanism is more explanatory. In the following, we introduce the three attention modules and the corresponding fault diagnosis method in detail.

A. Channel Attention Module

Different activation maps recognize fault features in different degrees, so some features may not be related to fault information or even indicate false information. In this article, the CAM is introduced to adaptively enhance fault-related activation maps and to suppress irrelevant or even false information, so as

Authorized licensed use limited to: University of Georgia. Downloaded on September 21,2023 at 03:55:21 UTC from IEEE Xplore. Restrictions apply.

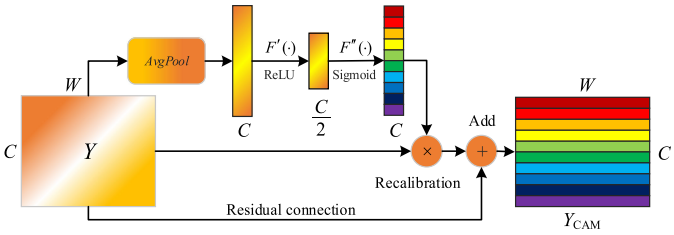

Fig. 1. Basic structure of the CAM.

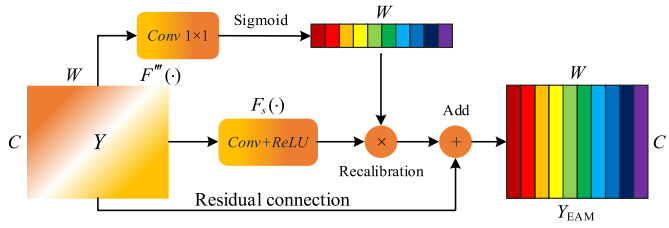

Fig. 2. Basic structure of the EAM.

to learn more discriminant fault features from the input 1-D signals.

The basic structure of the CAM is shown in Fig. 1. In convolution operation $F(\cdot)$, the correlation of each channel is entangled with the temporal correlation learnt by convolution kernels. Therefore, the CAM aims to improve the network sensitivity to different features by explicitly modeling the interdependence among channels, thus adaptively distinguishing the importance of features. Suppose that the input $(Y = [y_1, y_2, \ldots, y_c])$ of the CAM is a combination of channels $(y_i \in \mathbb{R}^{w\times 1})$. First, the CAM compresses global temporal information into a channel descriptor by using a global average pooling layer, and generates channel-wise statistics vector z $(z \in \mathbb{R}^{1\times C})$. The ith element of z is calculated by

$$z_i = \text{Avgpool}\,(y_i) = \frac{1}{1\times W}\sum_{j=1}^{W} y_i\,(j). \tag{1}$$

This operation embeds the global temporal information in z. Next, the CAM adopts a simple gating mechanism to fully capture channel-wise dependencies and generates the channel recalibration vectors $z^{'}$, defined as

$$z' = \sigma\left(F''\left(\delta\left(F'(z)\right)\right)\right), \tag{2}$$

where δ is a ReLU activation function; $F^{'}$ and $F^{''}$ represent convolution operations with channel number 1 and convolution kernel size 1×1, respectively, which encodes the channel-wise dependencies; and σ is the Sigmoid function, which compresses the dynamic range of the input activation vector to [0,1], in order to obtain the channel recalibration vector $z^{'}$. The value of z_i' indicates the importance of the ith channel. The channel recalibration vector is used to recalibrate feature Y to

$$M = [m_1, m_2, \ldots, m_C] = Y\cdot z' = [y_1 z'_1, y_2 z'_2, \ldots, y_C z'_C]. \tag{3}$$

Therefore, the obtained features M fully take into account the guidance of global information, and can effectively highlight more discriminant feature information. However, the range of vector z' is [0, 1]. Repeated feature recalibration operations lead to the reduction of the response value of deep features, which affects the diagnostic performance. Therefore, we use the idea of residual learning [8] and introduce the residual connection to improve the feasibility of optimization while retaining the original information. The final output of the CAM is: $Y_{\text{CAM}} = Y + M$.

B. Excitation Attention Module

When a local fault occurs in a rolling bearing, the fault location produces impulse excitation on other parts of the contacted rolling bearing, and excites the whole system to produce high-frequency attenuation vibration with resonance frequency. Therefore, the fault excitation signal segments of vibration signals centrally reflect the intrinsic properties of faults. In this article, the EAM is proposed to enable the network to learn the signal segments related to faults. It can not only effectively improve efficiency and reliability of feature learning, but also make the feature learning mechanism of the network more explanatory.

The basic structure of the EAM is shown in Fig. 2. The relative importance of different temporal signal segments is reflected by activation maps in convolution layers. Therefore, the EAM aggregates the feature information of all activation maps across channels with a convolution layer to locate the fault-related temporal signal segments. Specifically, assume that the input feature Y is represented as $Y = [y^1, y^2, \ldots, y^W]$, where $y^j \in \mathbb{R}^{1\times C}$ corresponds to the jth temporal signal location and $j = 1, 2, \ldots, W$. First, the EAM gets the projection of feature Y on temporal signals through a 1×1 convolution layer with one channel, which is $s = F'''(Y)$. Then, the temporal weight vector is obtained through the Sigmoid function, namely $s' = \sigma(s)$, $s \in \mathbb{R}^{1\times W}$. $F'''(\cdot)$ aggregates features of all activation maps in the input Y across channels, so s'_j indicates the importance of the jth time-series point. The temporal weight vector $s^{'}$ is used to recalibrate feature Y to

$$N = [n^1, n^2, \ldots, n^W] = F_s\,(Y)\cdot s'. \tag{4}$$

Before recalibrate Y, the EAM uses a convolution layer $F_s(\cdot)$ to encode feature information between local temporal signal segments to prevent overfocusing on excitation impulse signal segments. Similar to the CAM, the residual connection is also introduced to prevent the reduction of feature response value in the EAM. The final output of the EAM is $Y_{\text{EAM}} = Y + N$.

C. Joint Attention Module (JAM)

The JAM is a combination of the EAM and the CAM. It can give different weights to channel features and time-series features of the input Y. The basic structure of the JAM is shown in Fig. 3. The input feature Y is adaptively optimized by the EAM and the CAM in turn from different angles. Therefore, the output of the JAM can be expressed as

$$Y_{\text{JAM}} = F_{\text{CAM}}\left(F_{\text{EAM}}(Y)\right). \tag{5}$$

Authorized licensed use limited to: University of Georgia. Downloaded on September 21,2023 at 03:55:21 UTC from IEEE Xplore. Restrictions apply.

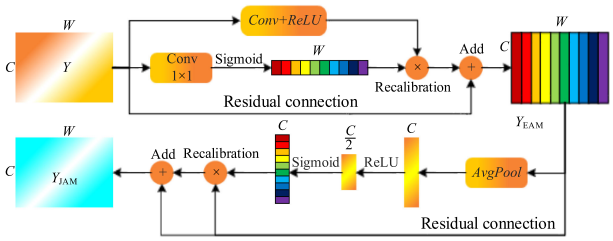

Fig. 3. Basic structure of the JAM.

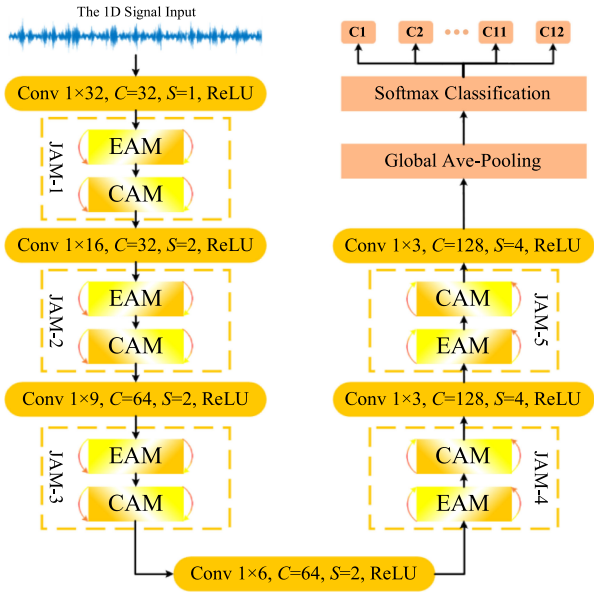

Fig. 4. Overall architecture of the MA1DCNN (Conv denotes convolution operation; C denotes the number of channels; and S denotes the stride of convolution).

A location (i, j) of the input feature Y is given higher activation when it gets high importance from both channel rescaling and time-series rescaling. This recalibration encourages the network to learn more discriminant features.

D. Multiattention 1DCNN

The attention modules proposed previously can be simply integrated into the conventional 1DCNN to improve the feature learning ability and fault diagnosis performance. Based on the proposed JAM, the MA1DCNN is proposed for rolling bearing fault diagnosis. The overall architecture of the MA1DCNN is shown in Fig. 4. The MA1DCNN consists of a backbone network, multiple JAMs, and fault classification layers. The backbone network includes six convolution modules, each of which contains a convolution layer and a ReLU function. The JAM is simply embedded behind each convolution module to adaptively optimize features. Finally, the MA1DCNN uses the global average pooling layer to replace the conventional full connection layer to avoid the overfitting problem. The softmax function is used in the classification layer to give the final diagnosis results. Assuming that there are n classes for the input samples, the output probability Q_k for the class k is calculated as (6). The diagnosis output is the fault label corresponding to the largest Q_k.

$$Q_k = \frac{\exp\left(\theta^{(k)} x\right)}{\sum_{k=1}^{K} \exp\left(\theta^{(k)} x\right)}, \quad k = 1, 2, \ldots, K, \tag{6}$$

where $\theta^{(k)}$ is the model parameter; and x denotes the input of the network.

TABLE I
NETWORK CONFIGURATION OF THE MA1DCNN ARCHITECTURE

Layer	Type	Kernel	Channel	Stride	Padding	Output
0	Intput	-	-	-	-	2048×1
1	Conv	32×1	32	1	Yes	2048×32
2	EAM	1×1	1	-	Yes	2048×32
3	CAM	-	-	-	-	2048×32
4	Conv	16×1	32	2	Yes	1024×32
5	EAM	1×1	1	-	Yes	1024×32
6	CAM	-	-	-	-	1024×32
7	Conv	9×1	64	2	Yes	512×64
8	EAM	1×1	1	-	Yes	512×64
9	CAM	-	-	-	-	512×64
10	Conv	6×1	64	2	Yes	256×64
11	EAM	1×1	1	-	Yes	256×64
12	CAM	-	-	-	-	256×64
13	Conv	3×1	128	4	Yes	64×128
14	EAM	1×1	1	-	Yes	64×128
15	CAM	-	-	-	-	64×128
16	Conv	3×1	128	4	Yes	16×128
Global Average Pooling						
Softmax						
C1	C2	C3	…	C10	C11	C12

E. Architectures of the MA1DCNN

The detailed structure of the MA1DCNN is summarized in the Table I. To ensure that the input signal sample contains complete periodic signal information, the length of the input signal sample of the MA1DCNN is 2048 × 1. In the MA1DCNN, we use strided convolutions instead of max-pooling to aggregate feature information and to reduce feature dimension without information loss. In addition, the core of the EAM is selective learning of the fault impulse that is related to the fault feature. To make the EAM fully learn the region of interest in the signal, we adopt a special down-sampling operation, that is, the first layer convolution stride is 1, then the convolution stride is gradually increased, finally, the feature signal length is compressed to 16 × 1. However, most other methods adopt a larger convolution stride or pooling layer in the first layer, which greatly reduces the dimension of the feature signal, thus making the network lose the learning of the fault impulses. Inspired by [12], we adopted convolution kernel of different sizes in the network to learn the long-term and the short-term features of the input signal. The number of convolutional channels determines the number of feature signals learned in a convolutional layer. However, a larger number of channels bring a larger number of parameters. Therefore, in the MA1DCNN, the number of convolution channels gradually increases from 32 to 128.

III. EXPERIMENTAL VALIDATION

In this section, effectiveness and superiority of the proposed MA1DCNN is validated on the wheelset bearing dataset.

Authorized licensed use limited to: University of Georgia. Downloaded on September 21,2023 at 03:55:21 UTC from IEEE Xplore. Restrictions apply.

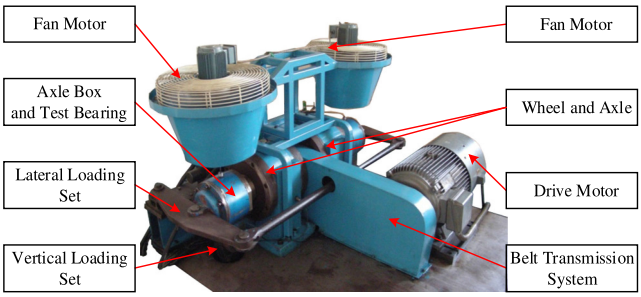

Fig. 5. Wheelset bearing test rig.

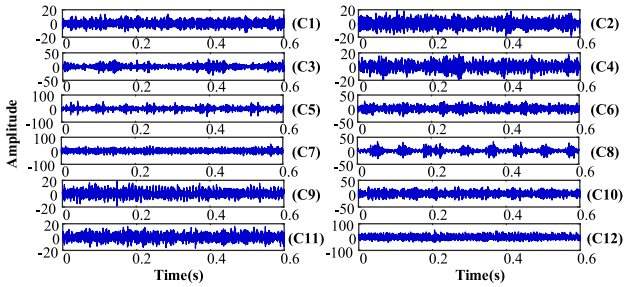

Fig. 6. Raw vibration signals from 12 health conditions of wheelset bearings.

TABLE II
WHEELSET BEARING SPECIFICATIONS

Specification	Value
Inner race diameter	130 mm
Outer race diameter	230 mm
Pitch diameter	180 mm
Top diameter of rolling elements	22.81 mm
Base diameter of rolling elements	24.74 mm
Height of rolling elements	47.46 mm
Number of rolling elements in per cage	20

TABLE III
TWELVE HEALTH CONDITIONS OF WHEELSET BEARINGS

Fault Mode	Label
Normal	C1
Inner race pitting	C2
Rolling element pitting	C3
Rolling element flaking with a size of 3 mm×35 mm	C4
Inner race flaking with a size of 3 mm×45 mm	C5
Rolling element cracking	C6
Mixed fault with outer race flaking and rolling element pitting (the flaking size is 10 mm×45 mm)	C7
Inner race flaking with a size of 10 mm×45 mm	C8
Outer race flaking with a size of 10 mm×30 mm	C9
Rolling element flaking with a size of 1 mm×1 mm	C10
Cage cracking	C11
Outer race flaking with a size of 10 mm×45 mm	C12

A. Wheelset Bearing Dataset

As shown in Fig. 5, the wheelset bearing test rig is mainly composed of a drive motor, a belt transmission system, a vertical loading set, a lateral loading set, two fan motors, and a control system. The vertical and the lateral loading sets are designed to mimic 2-D loads in real train operation. An axle and its two supporting bearings are assembled to the test rig.

The experimental wheelset bearing adopts double-row taper roller bearing, whose specifications are summarized in Table II. Considering that various fault modes may occur in the real operation of wheelset bearings, 12 typical health conditions of wheelset bearings are tested. The fault information with respect to the test bearings is listed in Table III, where the labels are C1, C2, . . . , C12, respectively. The bearing faults are formed naturally in real operation lines. The bearing faults are distributed in inner race, outer race, rolling element, and cage. Vibration data come from four acceleration channels that are fixed in the direction of 9 o'clock and 12 o'clock of the axle box. The sampling frequency is 5120 Hz.

To simulate various complex working conditions of wheelset bearings during their operation as much as possible, under each health condition, the vertical loads of 56, 146, 236, and 272 kN are set, and two lateral loads (0 and 20 kN) are set. The train running speed is divided into five speed levels from low to high speed, namely, 60, 90, 120, 150, and 180 km/h. In total, each health condition includes 40 kinds of working conditions. Fig. 6 shows an example of temporal vibration signals from 12 health conditions of wheelset bearings under the vertical load of 56 kN, the lateral load of 0 kN, and the speed of 120 km/h.

B. Validation Setup

Deep learning needs a large amount of data for training, so data augmentation is a practical technique to increase the sample number. This article adopts the same sliding segmentation approach as [12] for data augmentation. The length of each sample and the step size of sliding segmentation are set to 2048 and 256, respectively. Finally, the sliding segmentation approach is used to generate 188 088 samples.

The MA1DCNN is implemented in the Keras library and Python 3.5. The MA1DCNN Python codes are available at [28]. Network training and testing are performed on a workstation with Ubuntu 16.04 operating system, an Intel Core i7-6850K CPU, and a GTX 1080Ti GPU. Each sample accelerates the convergence speed of the network by subtracting mean and dividing variance. Compared with the standard normalization method of subtracting mean and dividing by standard deviation, we found that our normalization method performs better in the wheelset bearing dataset. Finally, during the training, we adopt the cross-entropy loss function and the Adam optimization algorithm with a learning rate of 0.0001 and with a batch size 196. We also adopted a learning rate decay operation, that is, when loss of more than 5 epochs did not decrease, the learning rate was multiplied by 0.9. The initialization of network weights follows the Glorot normal distribution initialization method [29]. The parameters are generated from the normal distribution with zero mean and standard deviation of $sqrt(2/(in + out))$, where in is the number of input elements of the weight tensor and out is the number of output elements of the weight tensor.

Authorized licensed use limited to: University of Georgia. Downloaded on September 21,2023 at 03:55:21 UTC from IEEE Xplore. Restrictions apply.

TABLE IV
RESULTS OF DIFFERENT NUMBER OF THE JAM IN THE MA1DCNN (SNR $= -6$ DB)

Indicators	MA1DCNN-1	MA1DCNN-2	MA1DCNN-3	MA1DCNN-4	MA1DCNN-5	MA1DCNN-6
Accuracy	76.64±0.82	78.64±1.34	82.86±1.16	82.97±0.83	**83.21**±0.61	82.46±1.03
Precision	74.62±0.98	77.05±1.46	82.22±1.48	81.87±0.75	**82.50**±0.52	81.67±1.02
Recall	73.46±1.14	75.79±1.31	80.91±1.31	81.19±0.73	**81.39**±0.58	80.72±1.23

In this article, three performance indicators, which are accuracy, precision, and recall, are adopted to indicate the diagnosis performance with four-fold cross validation. The three indicators are commonly comprehensive metric measuring the classification performance, defined as

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + FP + TN} \times 100\% \tag{7}$$

$$\text{Precision} = \frac{TP}{TP + FP} \times 100\% \tag{8}$$

$$\text{Recall} = \frac{TP}{TP + FN} \times 100\% \tag{9}$$

where *TP*, *FP*, *TN*, and *FN* refer to the number of true positive samples, false positive samples, true negative samples, and false negative samples, respectively. The accuracy, the recall, and the precision range from zero to one. A larger value means a better fault diagnosis performance.

We add white Gaussian noise into the raw vibration signals to stimulate strong noise disturbance of wheelset bearings in the real operation. The signal-to-noise ratio (SNR) is defined as

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

where P_{signal} and P_{noise} are the power of the signal and the noise, respectively.

C. Influence of the Number of Attention Modules

This section explores the influence of the number of the JAM in the MA1DCNN on the fault diagnosis performance with SNR $= -6$ dB. This experiment sets up six network structures: MA1DCNN-1, MA1DCNN-2, …, and MA1DCNN-6, where the ending number in their names represents the number of the JAM. The mean and the standard deviation of four-fold cross-validation results are summarized in Table IV. Fig. 7 shows the same accuracy, the same precision, and the same recall for the six varieties of the MA1DCNN. Obviously, when the number of the JAM increases from one to five, network performance also increases. This shows that the optimization effects of multiple attention modules can be accumulated in the network, thus continuously improving the network performance. Although more attention modules increase the network parameters, the accuracy, the precision, and the recall of the MA1DCNN-5 increase by 6.57%, 7.88%, and 7.93%, respectively, comparing with the MA1DCNN-1, so a small increase in network parameters is acceptable. Network performance of the MA1DCNN-6 is slightly lower than that of the MA1DCNN-5, this is because more attention modules may lead to the overfitting problem.

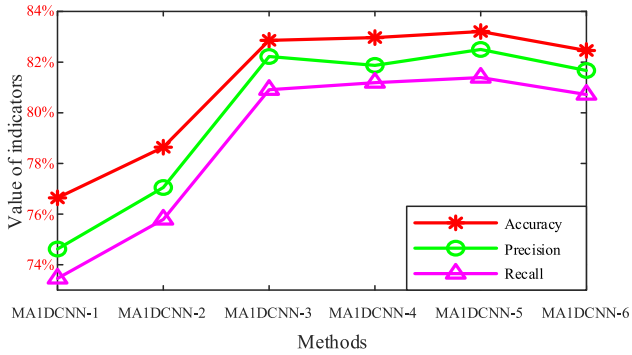

Fig. 7. Performance of the MA1DCNN changing with the number of the JAM (SNR $= -6$ dB).

TABLE V
RESULTS OF THE 1DCNN, MA1DCNN-EAM, MA1DCNN-CAM, AND MA1DCNN-JAM (SNR $= -6$ DB)

Methods	Accuracy	Precision	Recall
1DCNN	70.46±1.55	68.19±1.74	67.16±1.51
MA1DCNN-EAM	**74.90**±0.42	**73.13**±0.32	**71.99**±0.59
MA1DCNN-CAM	**81.67**±0.37	**80.69**±0.10	**79.67**±0.45
MA1DCNN-JAM	**83.21**±0.61	**82.50**±0.52	**81.39**±0.58

Therefore, we use five JAMs in the proposed MA1DCNN, which explains multiattention in the name of the MA1DCNN.

D. Effectiveness of the EAM

This section verifies the effectiveness of the EAM under SNR $= -6$ dB. In this experiment, two network structures, i.e., the 1DCNN (considering as a baseline without using any attention modules) and the MA1DCNN using only the EAM (denoted by MA1DCNN-EAM), are set up. The experimental results are summarized in Table V. Fig. 8 shows the accuracy, the precision, and the recall of the first 100 epochs. The MA1DCNN-EAM has better fault diagnosis performance. In particular, the accuracy, the precision, and the recall of the MA1DCNN-EAM are improved by 4.44%, 4.94%, and 4.83%, respectively, comparing with the 1DCNN. This proves that the EAM can effectively improve the feature learning ability of networks by enhancing the learning of the excitation signal segment, so as to obtain better performance.

To further understand the feature learning mechanism of the EAM, we visualize the temporal weight vectors in the five EAMs of the MA1DCNN-EAM, as shown in Fig. 9. The visualization of all temporal weight vectors has the same length as the input signal to align with it. There are obvious excitation impulses in

Authorized licensed use limited to: University of Georgia. Downloaded on September 21,2023 at 03:55:21 UTC from IEEE Xplore. Restrictions apply.

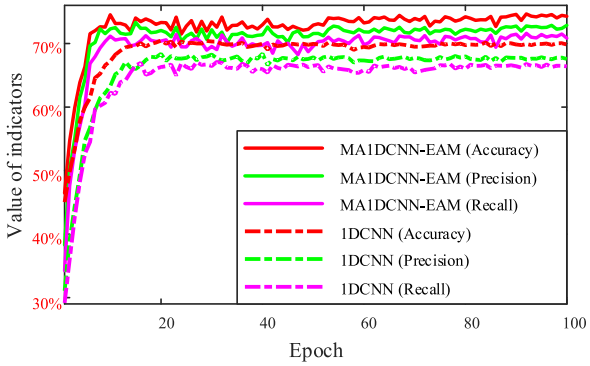

Fig. 8. Performance comparison of the 1DCNN and the MA1DCNN-EAM (SNR $= -6$ dB).

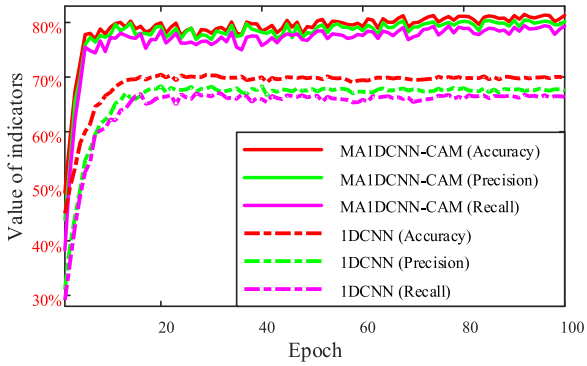

Fig. 10. Performance comparison of the 1DCNN and the MA1DCNN-CAM (SNR $= -6$ dB).

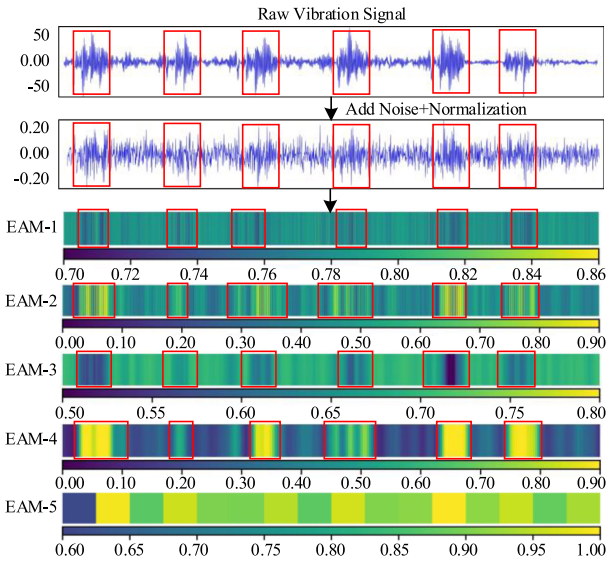

Fig. 9. Visualization of the temporal weight vectors in the five EAMs of the MA1DCNN-EAM.

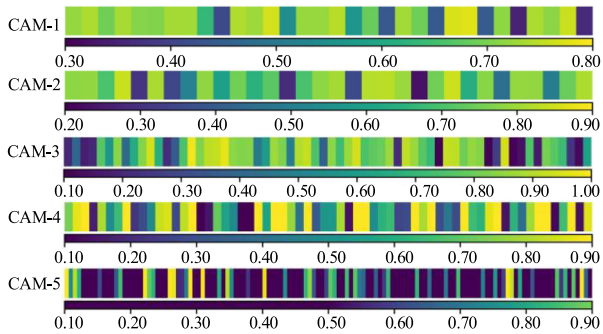

Fig. 11. Visualization of the channel recalibration vectors in the five CAMs of the MA1DCNN-CAM.

the input signal. When noise is added, these excitation impulse signals with the most discriminant features are submerged as shown in the second subfigure. Fortunately, the five EAMs can still adaptively locate the right fault-related excitation impulse signal segments, and can make the network pay more attention to these signal segments. Obviously, this adaptive feature selection mechanism gives the network the ability to selectively learn more important feature information, thus can improve efficiency and ability of the feature learning. This ability also brings considerable improvement to the fault diagnosis accuracy of the CNN. The experimental results also prove that it is effective to pay attention to the impulse signal segments.

E. Effectiveness of the CAM

In this section, the validity of the CAM is verified under SNR $= -6$ dB. In this experiment, two network structures, i.e., the 1DCNN and the MA1DCNN using only the CAM (denoted by MA1DCNN-CAM), are set up. The experimental results are summarized in Table V. Fig. 10 shows the accuracy, the precision, and the recall for the first 100 epochs. Obviously, compared with the 1DCNN, the MA1DCNN-CAM has tremendous advantages in network optimization speed and fault diagnosis performance. Specifically, the accuracy, the precision, and the recall of the MA1DCNN-CAM are improved by 11.21%, 11.9%, and 12.51%, respectively, comparing with the 1DCNN. This shows that the CAM can effectively learn more discriminant fault features by adaptively optimizing the features of different activation maps, so as to obtain better diagnosis results. Moreover, smaller standard deviation in Table V also proves that the CAM can effectively improve the stability of the network.

To understand the feature learning mechanism of the CAM, we visualize the recalibration vectors in the five CAMs of the MA1DCNN-CAM, as shown in Fig. 11. The recalibration vectors in the five CAMs encode the relative weights among the activation maps so that the network can adaptively enhance the activation maps related to faults, so as to learn more discriminant fault features. Unlike the EAM, the CAM is more concerned with the selection of channel feature. From the CAM-1 to the CAM-5, we can see more and more redundant and unimportant feature channels in the CNN. In particular, the CAM-5 has only selected a few feature channels, but the network performance has been greatly improved. This further proves necessity of the feature selection and effectiveness of the proposed CAM.

F. Effectiveness of the JAM

This section verifies the effectiveness of the JAM under SNR $= -6$ dB. In this experiment, two network structures,

Authorized licensed use limited to: University of Georgia. Downloaded on September 21,2023 at 03:55:21 UTC from IEEE Xplore. Restrictions apply.

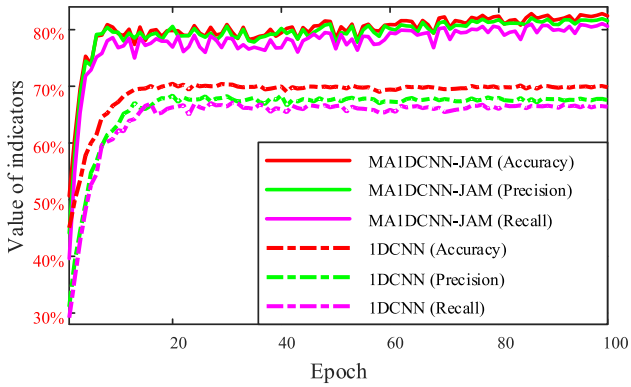

Fig. 12. Performance comparison of the 1DCNN and the MA1DCNN-JAM (SNR = −6 dB).

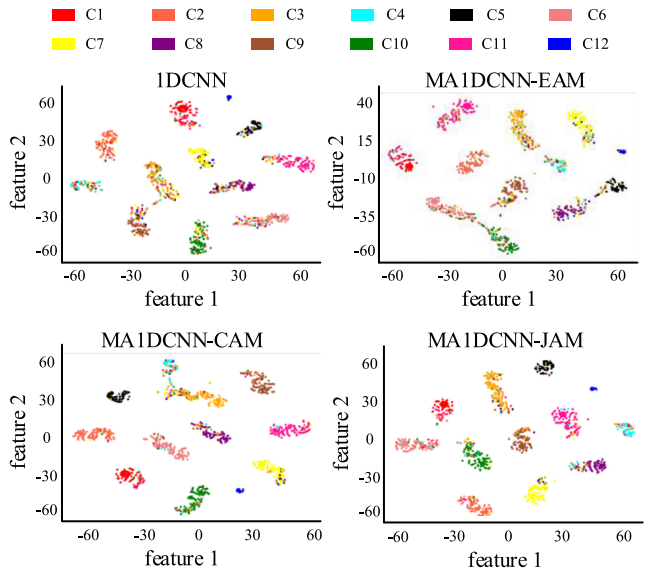

Fig. 13. Scatter plots of the 1DCNN, the MA1DCNN-EAM, the MA1DCNN-CAM, and the MA1DCNN-JAM in 2-D space.

i.e., the 1DCNN and the MA1DCNN using the JAM (denoted by MA1DCNN-JAM), are set up. The experimental results are summarized in Table V. Fig. 12 shows the accuracy, the precision, and the recall for the first 100 epochs. Obviously, the JAM can effectively improve the fault diagnosis performance of networks. Compared with the 1DCNN, the MA1DCNN-JAM improves the accuracy, the precision, and the recall by 12.75%, 14.31%, and 14.23%, respectively. According to Table V, the JAM is better than the CAM and the EAM in improving the network performance. This shows that the CAM and the EAM can reinforce with each other, so as to improve fault diagnosis performance better.

We use the t-SNE technology [30] to visualize the final output distribution of the 1DCNN, MA1DCNN-EAM, MA1DCNN-CAM, and MA1DCNN-JAM in a 2-D space. The visualization results are shown in Fig. 13, where different colors represent different health conditions of wheelset bearings. Obviously, the output distribution of the MA1DCNN-JAM has the best discrimination, followed by the MA1DCNN-CAM and the MA1DCNN-EAM. The 1DCNN has the worst discrimination, which shows that the proposed multiattention modules can effectively improve the network's feature learning ability.

TABLE VI
RESULTS OF THE WDCNN AND THE WDCNN–JAM (SNR = −6 dB)

Methods	Accuracy	Precision	Recall
WDCNN [17]	72.50±0.71	70.36±0.83	69.79±0.83
WDCNN-JAM	**77.89**±0.83	**76.14**±0.91	**75.43**±0.57

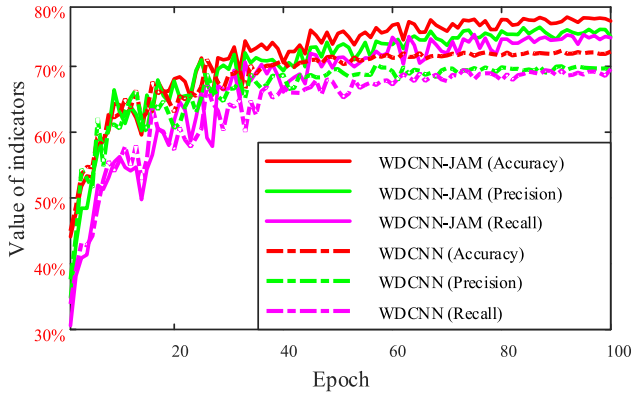

Fig. 14. Performance comparison of the WDCNN and the WDCNN-JAM (SNR = −6 dB).

G. Generalization of Attention Modules in the Existing Network

This section explores effectiveness of attention modules in the existing network under SNR = −6 dB. In this experiment, taking the WDCNN as an example, two network structures, i.e., the WDCNN (considering as a baseline without using any attention modules) and the WDCNN-JAM (Improvement by including the JAM) are set up. The experimental results are summarized in Table VI. The accuracy, the precision, and the recall for the first 100 epochs are shown in Fig. 14. The WDCNN-JAM improves the accuracy, the precision, and the recall by 5.39%, 5.78%, and 5.64%, respectively, comparing with the original WDCNN. This example demonstrates that the proposed JAM has an excellent generalization ability for the existing networks and can improve their fault diagnosis performance.

H. Comparison With State-of-the-Arts Networks

In this section, the diagnostic performance of the MA1DCNN and eight state-of-the-arts networks are validated under three SNR scenarios (−6 dB, 0 dB, and 6 dB) that are used to simulate three noise conditions.

The eight deep learning networks are Wen-CNN [18], adaptive deep convolution neural network (ADCNN) [26] based on the 2DCNN, deep convolutional neural networks with wide first-layer kernels (WDCNN) [17], residual convolution neural network (ResCNN) [23] based on the 1DCNN, gated recurrent unit (GRU), recurrent neural network (RNN), stacked autoencoder (SAE), and the five-layer back propagation neural network (BPNN) with the same structure as [27]. Bruin *et al.* [31] applied

Authorized licensed use limited to: University of Georgia. Downloaded on September 21,2023 at 03:55:21 UTC from IEEE Xplore. Restrictions apply.

TABLE VII
PERFORMANCE OF NINE COMPARISON NETWORKS UNDER THE THREE SNR SCENARIOS

Noise (SNR)	Indicators	MA1DCNN	Wen-CNN [18]	ADCNN [26]	WDCNN [17]	ResCNN [23]	GRU	RNN	SAE	BPNN [33]
−6 dB	Accuracy	**83.21**±0.61	71.04±1.00	57.11±1.22	72.50±0.71	67.28±1.69	80.47±0.13	52.89±2.31	43.73±0.51	43.74±0.75
	Precision	**82.50**±0.52	68.79±0.83	52.89±1.22	70.36±0.83	65.72±1.39	78.93±0.18	43.65±2.58	34.58±0.54	35.10±0.28
	Recall	**81.39**±0.58	67.84±0.90	50.80±1.37	69.79±0.83	62.46±2.72	79.54±0.17	44.34±2.57	36.15±0.96	34.10±0.58
0 dB	Accuracy	**97.63**±1.20	93.57±0.24	77.51±0.97	93.18±1.03	89.15±1.18	94.47±0.17	75.72±1.86	67.37±0.58	67.98±1.21
	Precision	**96.88**±0.18	93.21±0.09	75.43±0.79	92.71±1.14	88.52±0.84	94.03±0.13	71.30±2.19	62.81±0.72	64.30±1.08
	Recall	**96.76**±0.26	92.99±0.36	74.83±1.01	92.55±1.11	88.06±1.56	94.27±0.24	72.48±2.22	63.98±0.52	63.30±1.23
6 dB	Accuracy	**99.39**±0.06	98.26±0.20	86.65±0.34	97.90±0.18	96.45±0.39	97.62±0.13	86.30±0.83	78.86±0.22	80.51±0.89
	Precision	**99.34**±0.08	98.17±0.24	85.06±0.53	97.73±0.20	96.13±0.46	97.37±0.12	83.96±1.04	75.91±0.32	78.10±0.82
	Recall	**99.30**±0.11	98.08±0.22	84.72±0.36	97.65±0.19	95.99±0.42	97.41±0.19	84.15±0.89	76.36±0.25	77.82±0.85

the RNN to railway track circuit fault diagnosis. Zhao *et al.* [32] proposed a local feature-based gated unit and achieved good performance in three machine health monitoring tasks. Similarly, the RNN and the GRU constructed in this experiment have two-layer RNN cell or GRU cell with the softmax classification layer, where 64 is the length of time steps and 32 is the dimension of input data (input size). Finally, the SAE has eight hidden layers and a full connection classification layer with softmax. The experimental results are summarized in Table VII.

First, we discuss performance of the MA1DCNN and the four CNN-based networks. Obviously, the accuracy, the precision, and the recall of the MA1DCNN are better than the four CNN-based networks in all SNR scenarios. In particular, when SNR = −6 dB, the MA1DCNN still obtains 83.21% diagnostic accuracy. It has almost 10.71% improvement compared with the WDCNN, which is the best in the four CNN-based networks. This shows that the MA1DCNN has strong antinoise ability without any additional denoising preprocessing. What is more, with the increase of noise, fault diagnosis performance decreases. For example, when the SNR varies from 0 to −6 dB, the noise intensity increases by 3.98 times compared with the raw vibration signal. The diagnostic accuracy of the Wen-CNN, ADCNN, WDCNN, and ResCNN decreased by 22.53%, 20.40%, 20.68%, and 21.87%, respectively, but the MA1DCNN only decreases by 14.42%, which further shows that the MA1DCNN has better antinoise ability than the other four networks.

Second, we discuss performance of the MA1DCNN and the rest four non-CNN based networks. According to Table VII, it can be found that the GRU achieves 80.47% accuracy under the condition of strong noise of −6 dB, which is far better than the other non-CNN based networks, indicating that the GRU can effectively learn the feature relation between long temporal signals, so as to obtain better antinoise performance. Moreover, the MA1DCNN is still about 2% better than the GRU by adaptively capturing the features related to the fault information. Due to the simple unit structure of the RNN and the limited performance of the network structure of the SAE and the BPNN, it is difficult for them to obtain acceptable accuracy.

To analyze the classification accuracy of each fault mode and the precision and the recall in detail, the confusion matrix of the proposed MA1DCNN under SNR = −6 dB and SNR = 6 dB are provided in Figs. 15 and 16, respectively. In the two figures, row and column represent predicted label and true label, respectively. The diagonal cells are accuracies of the 12 fault modes. The bottom row shows the precision of each fault mode. The rightmost column represents the number of testing samples in each fault mode. It can be seen that even if SNR = −6 dB, the precision of the MA1DCNN in the normal category is 91.01%, indicating that the model can still distinguish normal samples from fault samples with relatively high accuracy under strong noise. The main error comes from the wrong classification between fault modes. In addition, under SNR = 6 dB, the recall and the precision of the network in most fault modes are very close to 100%, which indicates that the MA1DCNN has good potential in practical application.

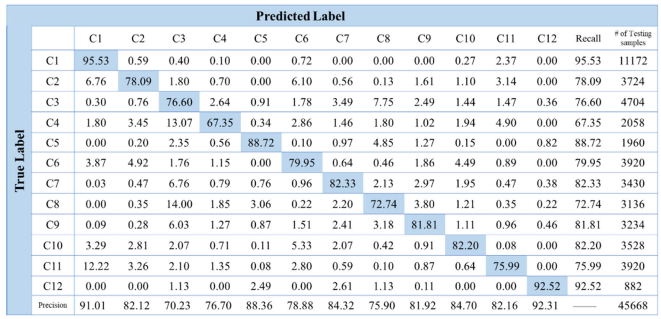

Fig. 15. Confusion matrix of the proposed MA1DCNN (SNR = −6 dB).

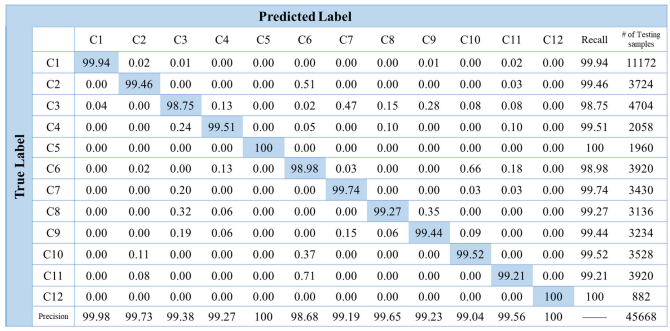

Fig. 16. Confusion matrix of the proposed MA1DCNN (SNR = 6 dB).

Authorized licensed use limited to: University of Georgia. Downloaded on September 21,2023 at 03:55:21 UTC from IEEE Xplore. Restrictions apply.

IV. Discussions

A. Assumptions and Limitations of the MA1DCNN

The MA1DCNN has made the following three assumptions.

1) Feature space should be the same for the testing samples and the training samples, that is, they are the 1-D vibration signal with the same length and the same sampling frequency.
2) Label space should be the same for the testing samples and the training samples.
3) Working condition should be the same for the testing samples and the training samples, that is, they have the same feature distribution.

The limitation of the MA1DCNN is that it does not consider fault severity. In fault diagnosis, it is important to find the fault at the early stage. In fact, C5 and C8, C4 and C10, and C9 and C12 in the wheelset bearing dataset are the same fault mode at different stages. One solution is to merge these modes and to modify the network output so that the network can learn more robust features of these fault modes. Another solution may be to interpret the output of the CNN as probability by directly using (6). We can set different thresholds to adjust the network's sensitivity for early fault diagnosis.

B. Significance and Advantages of the MA1DCNN

Effectiveness of the MA1DCNN has been proved in Section III. We next discuss significance and advantages of the MA1DCNN.

1) Reinforcing Fault Signal Learning Mechanism: The local faults of the surface form a sudden shock force on the contact surface of rolling bearings. The vibration signal caused by this force can be divided into two categories. One is the low-frequency vibration caused by the repeated impact between the fault and the contact surface. It generates periodic fault impulse components, whose frequency is the fault characteristic frequency. It is the main evidence of determining the bearing fault. The second is the high-frequency resonance vibration of the mechanical system caused by the fault shock. It can be found that the periodic fault impulse components mainly reflect the intrinsic characteristics of such fault. The EAM can locate fault impact segments and pay more attention to them so as to extract more fault information. Starting from the characteristics of the data itself, the EAM improves the learning efficiency of the network from the input signals, which is of great significance to improve the performance of the CNN.

2) Exploring Discriminant Feature Learning Mechanism: Feature representation in the 1DCNN consists of a set of activation maps connected by channel dimensions. These mappings are made up of different filters convolved with the input signals. The information captured by different filters is of different importance for fault diagnosis tasks. The CAM can adaptively determine the importance of different channels, so as to enhance the channel features with more discriminative features. It is of great significance to enhance the optimization ability of the CNN and to explore more effective filter parameters.

3) Increasing Network Interpretability: Figs. 9 and 11 show that the features learned by the network are very redundant, and treating all features equally is detrimental to performance of the network. They also show that in the field of fault diagnosis, it is beneficial to enhance the learning of fault impact area. These phenomena provide further explanation for the interpretability of the CNN.

V. Conclusion

This article proposed the MA1DCNN that introduced attention mechanism into an end-to-end deep learning approach for wheelset bearing fault diagnosis. By using the attention modules, the proposed network could model the interdependencies between channels by the CAM to adaptively optimize features of each layer, and it also encode the relative importance of temporal signal segments by the EAM to selectively enhance the learning of fault impulse segments, so as to obtain more discriminant features. Conclusions are summarized as follows.

1) The proposed attention mechanism had demonstrated its effectiveness for understanding and learning discriminant features of the 1-D signals. It provided a solution deal with the three challenges mentioned in Section I.
2) The proposed MA1DCNN was applied to wheelset bearing fault diagnosis and showed significant advantages in the accuracy, the precision, and the recall over eight state-of-the-arts networks.
3) The proposed attention modules could also be applied to the existing deep learning networks as feature optimization modules to improve their performance.

In the future, we will study application and improvement of the attention mechanism in speed-domain adaptability or load-domain adaptability tasks, to further explore the feature learning mechanism of the CNN.

References

[1] X. Dai and Z. Gao, "From model, signal to knowledge: A data-driven perspective of fault detection and diagnosis," *IEEE Trans. Ind. Inform.*, vol. 9, no. 4, pp. 2226–2238, Nov. 2013.
[2] Z. K. Peng, P. W. Tse, and F. L. Chu, "A comparison study of improved Hilbert-Huang transform and wavelet transform: Application to fault diagnosis for rolling bearing," *Mech. Syst. Signal Process.*, vol. 19, no. 5, pp. 974–988, 2005.
[3] J. Faiz, V. Ghorbanian, and B. M. Ebrahimi, "EMD-based analysis of industrial induction motors with broken rotor bars for identification of operating point at different supply modes," *IEEE Trans. Ind. Inform.*, vol. 10, no. 2, pp. 957–966, May 2014.
[4] Z. Liu, Y. Jin, M. J. Zuo, and Z. Feng, "Time-frequency representation based on robust local mean decomposition for multicomponent AM-FM signal analysis," *Mech. Syst. Signal Process.*, vol. 95, pp. 468–487, 2016.
[5] T. Jing, C. Morillo, M. H. Azarian, and M. Pecht, "Motor bearing fault detection using spectral kurtosis based feature extraction and K-nearest neighbor distance analysis," *IEEE Trans. Ind. Electron.*, vol. 63, no. 3, pp. 1793–1803, Mar. 2016.
[6] S. Shevchik, F. Saeidi, B. Meylan, and K. Wasmer, "Prediction of failure in lubricated surfaces using acoustic time-frequency features and random forest algorithm," *IEEE Trans. Ind. Inform.*, vol. 13, no. 4, pp. 1541–1553, Aug. 2017.
[7] D. You, X. Gao, and S. Katayama, "Multisensor fusion system for monitoring high-power disk laser welding using support vector machine," *IEEE Trans. Ind. Inform.*, vol. 10, no. 2, pp. 1285–1295, May 2014.
[8] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
[9] M. Meng, C. Sun, and X. Chen, "Deep coupling autoencoder for fault diagnosis with multimodal sensory data," *IEEE Trans. Ind. Inform.*, vol. 14, no. 3, pp. 1137–1145, Mar. 2018.

Authorized licensed use limited to: University of Georgia. Downloaded on September 21,2023 at 03:55:21 UTC from IEEE Xplore. Restrictions apply.

[10] H. Shao, H. Jiang, H. Zhang, and T. Liang, "Electric locomotive bearing fault diagnosis using novel convolutional deep belief network," *IEEE Trans. Ind. Electron.*, vol. 65, no. 3, pp. 2727–2736, Mar. 2018.
[11] T. Ince, S. Kiranyaz, L. Eren, M. Askar, and M. Gabbouj, "Real-time motor fault detection by 1-D convolutional neural networks," *IEEE Trans. Ind. Electron.*, vol. 63, no. 11, pp. 7067–7075, Nov. 2016.
[12] D. Peng, Z. Liu, H. Wang, and Q. Yong, "A novel deeper one-dimensional convolutional neural network with residual learning for fault diagnosis of wheelset bearings in high speed trains," *IEEE Access*, vol. 1, pp. 10278–10293, 2018.
[13] Z. Chen, C. Li, and R. Sanchez, "Gearbox fault identification and classification with convolutional neural networks," *Shock Vibration*, vol. 2015, no. 2, pp. 1–10, 2015.
[14] G. Jiang, H. He, J. Yan, and P. Xie, "Multiscale convolutional neural networks for fault diagnosis of wind turbine gearbox," *IEEE Trans. Ind. Electron.*, vol. 66, no. 4, pp. 3196–3207, Apr. 2019.
[15] J. Senanayaka, K. Huynh, and K. Robbersmyr, "Multiple classifiers and data fusion for robust diagnosis of gearbox mixed faults," *IEEE Trans. Ind. Inform.*, vol. 15, no. 8, pp. 4569–4579, Aug. 2019.
[16] L. Jing, M. Zhao, and P. Li, "A convolutional neural network based feature learning and fault diagnosis method for the condition monitoring of gearbox," *Measurement*, vol. 111, pp. 1–10, 2017.
[17] W. Zhang, G. Peng, C. Li, Y. Chen, and Z. Zhang, "A new deep learning model for fault diagnosis with good anti-noise and domain adaptation ability on raw vibration signals," *Sensors*, vol. 17, no. 2, pp. 425–446, 2017.
[18] L. Wen, X. Li, L. Gao, and Y. Zhang, "A new convolutional neural network-based data-driven fault diagnosis method," *IEEE Trans. Ind. Electron.*, vol. 65, no. 7, pp. 5990–5998, Jul. 2018.
[19] W. Sun, Z. Rui, R. Yan, S. Shao, and X. Chen, "Convolutional discriminative feature learning for induction motor fault diagnosis," *IEEE Trans. Ind. Inform.*, vol. 13, no. 3, pp. 1350–1359, Jun. 2017.
[20] R. Liu, G. Meng, B. Yang, C. Sun, and X. Chen, "Dislocated time series convolutional neural architecture: An intelligent fault diagnosis approach for electric machine," *IEEE Trans. Ind. Inform.*, vol. 13, no. 3, pp. 1310–1320, Jun. 2017.
[21] G. Hinton *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012.
[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 770–778.
[23] W. Zhang, X. Li, and Q. Ding, "Deep residual learning-based fault diagnosis method for rotating machinery," Dec. 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0019057818305202#!
[24] W. Zhang, C. Li, G. Peng, Y. Chen, and Z. Zhang, "A deep convolutional neural network with new training methods for bearing fault diagnosis under noisy environment and different working load," *Mech. Syst. Signal Process.*, vol. 100, pp. 439–453, 2017.
[25] M. Xia, T. Li, L. Xu, and L. Liu, "Fault diagnosis for rotating machinery using multiple sensors and convolutional neural networks," *IEEE/ASME Trans. Mechatronics*, vol. 23, no. 1, pp. 101–110, Feb. 2017.
[26] X. Guo, L. Chen, and C. Shen, "Hierarchical adaptive deep convolution neural network and its application to bearing fault diagnosis," *Measurement*, vol. 93, pp. 490–502, 2016.
[27] S. Shao, S. McAleer, R. Yan, and P. Baldi, "Highly-accurate machine fault diagnosis using deep transfer learning," *IEEE Trans. Ind. Inform.*, vol. 15, no. 4, pp. 2446–2455, Apr. 2019.
[28] *Understanding and Learning Discriminant Features Based on Multi-Attention 1DCNN for Wheelset Bearing Fault Diagnosis*, 2019. [Online]. Available: https://github.com/erphm/MA1DCNN
[29] X. Glorot and B. Bengio, "Understanding the difficulty of training deep feedforward neural networks," *J. Mach. Learn. Res.*, vol. 9, pp. 249–256, 2010.
[30] L. Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 2605, pp. 2579–2605, 2008.
[31] T. de Bruin, K. Verbert, and R. Babuska, "Railway track circuit fault diagnosis using recurrent neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 3, pp. 523–533, Mar. 2017.
[32] R. Zhao, D. Wang, R. Yan, K. Mao, and J. Wang, "Machine health monitoring using local feature-based gated recurrent unit networks," *IEEE Trans. Ind. Electron.*, vol. 65, no. 2, pp. 1539–1548, Feb. 2018.
[33] G. Bin, J. Gao, X. Li, and B. Dhillon, "Early fault diagnosis of rotating machinery based on wavelet packets-Empirical mode decomposition feature extraction and neural network," *Mech. Syst. Signal Process.*, vol. 27, no. 1, pp. 696–711, 2012.

Huan Wang was born in Hunan, China. He received the B.S. degree in mechanical engineering in 2016 from the University of Electronic Science and Technology of China, Chengdu, China, where he is currently working toward the M.S. degree in mechanical engineering.

His research interests include mechanical fault diagnosis, image recognition, deep learning, and machine learning.

Zhiliang Liu was born in Rizhao, China, in 1984. He received the Ph.D. degree from the School of Automation Engineering from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2013.

From 2009 to 2011, he was with the University of Alberta, Edmonton, AB, Canada, as a Visiting Scholar for two years. From 2013 to 2015, he was an Assistant Professor with the School of Mechanical and Electrical Engineering, UESTC, where since 2015, he has been an Associate Professor. He has authored and coauthored more than 60 papers including 20+ Science citation index (SCI)-Indexed journal papers. He currently held more than ten research grants from the National Natural Science Foundation of China, Open Grants of National Key Laboratory, China Postdoctoral Science Foundation, etc. His research interests include fault diagnosis and prognostics of rotating machinery by using advanced signal processing and data mining methods.

Dandan Peng was born in Shaanxi, China, in 1992. She received the B.S. and M.S. degrees in mechanical engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 2016 and 2019, respectively. She is currently working toward the Ph.D. degree in mechanical engineering with Katholieke Universiteit Leuven, Leuven, Belgium.

Her research interests include Hilbert Huang transform, convolutional neural network, machinery condition monitoring, and fault diagnosis.

Yong Qin received the B.Sc. and M.Sc. degrees in transportation automation and control engineering from Shanghai Railway University, Shanghai, China, in 1993 and 1996, respectively, and the Ph.D. degree in information engineering and control from the China Academy of Railway Sciences, Beijing, China, in 1999.

He is a Professor and the Vice Dean with the State Key laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing. He has authored or coauthored more than 100 publication papers Science Citation Index/Engineering Index (SCI/EI), one Essential Science Indicators (ESI) highly cited paper, and five books, and also has 23 patents granted including two USA patents. His research interests include prognostics and health management for railway transportation system, transportation network safety and reliability, and rail operation planning and optimization.

Prof. Qin is a member of the IEEE Intelligent Transportation Systems Society and IEEE Intelligent Transportation Systems Society, and a Senior Member of the IET. He won 11 science and technology progress award of ministry.

Authorized licensed use limited to: University of Georgia. Downloaded on September 21,2023 at 03:55:21 UTC from IEEE Xplore. Restrictions apply.