Measurement 187 (2022) 110242

Contents lists available at ScienceDirect

Measurement

journal homepage: www.elsevier.com/locate/measurement

Fault diagnosis for small samples based on attention mechanism

Xin Zhang [a], Chao He [b], Yanping Lu [b], Biao Chen [b], Le Zhu [c], Li Zhang [b,*]

[a] School of Materials Science and Engineering, Northeastern University, Shenyang 110819, China
[b] School of Information, Liaoning University, Shenyang 110036, China
[c] School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China

ARTICLE INFO

Keywords:
Convolutional neural network
Bidirectional gated recurrent unit
Attention mechanism
Rolling bearings
Small samples
Fault diagnosis

ABSTRACT

Aiming at the application of deep learning in fault diagnosis, mechanical rotating equipment components are prone to failure under complex working environment, and the industrial big data suffers from limited labeled samples, different working conditions and noises. In order to explore the problems above, a small sample fault diagnosis method is proposed based on dual path convolution with attention mechanism (DCA) and Bidirectional Gated Recurrent Unit (DCA-BiGRU), whose performance can be effectively mined by the latest regularization training strategies. BiGRU is utilized to realize spatiotemporal feature fusion, where vibration signal fused features with attention weight are extracted by DCA. Besides, global average pooling (GAP) is applied to dimension reduction and fault diagnosis. It is indicated that DCA-BiGRU has exceptional capacities of generalization and robustness by experiments, and can effectively carry out diagnosis under various complicated situations.

1. Introduction

With the development of industrial Internet of Things, the manufacturability, integration and precision of rotating machinery system are constantly improving, but complexity, nonlinearity and uncertainty are also significantly enhanced, which has become a huge challenge [1]. During the long-term running, rotating machinery will be affected by material degradation, loads, temperature and humidity, leading to the breakdown of key components easily, which will depress plant benefits, or lead to casualties and ecological pollution. Therefore, it is of great significance to monitor the status of rotating machinery.

In the past few years, fault diagnosis methods based on signal analysis, swarm intelligence evolution and machine learning have continued to emerge [2–4]. However, it is too dependent on prior knowledge of experts and features are extracted by manual, which makes it difficult to process big data and learn advanced features. Additionally, swarm intelligence is a heuristic algorithm and the optimized result is hard to be stable because of randomness. Furthermore, related algorithms with a quite high time complexity cannot guarantee to figure out the global optimum. Finally, in the face of complex and changeable industrial data, it is difficult for vanilla shallow models to achieve ideal results.

In recent years, with the development of deep learning, it has made remarkable achievements in image classification, semantic segmentation, target detection and natural language processing [5–8]. Similarly, it also provides some directions of settling the problems encountered above in fault diagnosis [9]. Accordingly, a series of studies for fault diagnosis have set off a research upsurge, which include convolutional neural network, autoencoder, generative adversarial network, deep belief network, recurrent neural network and capsule network etc [10–16]. Implementation of these methods usually requires to design novel and efficient structures or improve deep optimized algorithms. Alternatively, the distribution features of signals are required to analyze from multiple perspectives. For example, Zhou et al. [17] added a data generation and filtering strategy into autoencoder-generative adversarial networks (AE-GAN) for unbalanced data, where autoencoder was utilized to learn features of unbalanced samples, and the discriminator aimed to filter out unqualified generated samples. Kumar et al. [18] adopted a Deep CNN model based on AdaGrad, which fused multiple sensor data to generate images for fault diagnosis.

Furthermore, small sample fault diagnosis has become a new research focus. Zhang et al. [19] put forward a method for small samples based on siamese neural network, and the same or different sample pairs were input to calculate L_1 distance of feature vectors, judging whether to belong to the same class to train, and then support sets and query sets as pairs were calculated similarity to realize fault diagnosis. On this basis, Wang et al. [20] proposed a comparison diagnosis model which applied the full connected layer as the similarity measure of feature pairs to judge whether they belonged to a certain type, and meanwhile regularization methods were added to improve performance. Wu et al. [21] compared small sample transfer learning among

* Corresponding author.
E-mail address: zhang_li@lnu.edu.cn (L. Zhang).

https://doi.org/10.1016/j.measurement.2021.110242
Received 29 June 2021; Received in revised form 7 September 2021; Accepted 26 September 2021
Available online 7 October 2021

feature transfer, fine-tuning and meta relation network, and concluded that under small samples or the similarity between source domain and target domain was large, the meta relation transfer was dominant. On the contrary, the advantage of feature transfer was gradually obvious. Saufi et al. [22] came up with a small sample fault diagnosis method based on spectral kurtosis filtering and particle swarm optimization stacked sparse autoencoder, where a high diagnostic accuracy can be achieved when the number of per fault training samples is 100. Han et al. [23] applied bidirectional long short-term memory (BiLSTM) and capsule network to design a small sample fault diagnosis method, which proved that capsule network had a satisfying performance after denoising and fusion signals by BiLSTM. Li et al. [24] developed a conditional Wasserstein generative adversarial network (CWGAN), where vast similar samples were generated by training CWGAN with vast source domain samples, and pre-trained CWGAN was fine-tuned to achieve transfer learning under target domain with limited samples.

For small samples, they either utilize regularization technologies and feature extraction advantages of models, or generate substantial high-quality samples based on the distribution of real samples, or apply emerging machine learning technologies such as meta-learning and transfer learning.

The design of big convolution kernels is beneficial to enhance robustness [25], while that of deep small convolution kernels effectively extract abstract features. Also, time-step information cannot be ignored in vibration signals. Compared with CNN, RNN can just meet requirements.

To learn temporal and hidden features in different locations, an effective strategy is to employ a gated RNN structure, LSTM or GRU. LSTM has an excellent time modeling capability while has many parameters, which easily leads to overfitting under small samples. Similarly, it is inappropriate to assume that signals only propagate information forward, so BiGRU with similar performance to BiLSTM, fewer parameters and propagating forth and back is a terrific choice. Zhao et al. [26] put forward a method of combining Manifold Embedded Distribution Alignment(MEDA) and BiGRU for fault diagnosis. The noises of original signals were removed by spectrum information, and BiGRU was utilized to learn features, then MEDA was used to align auxiliary and unlabeled samples. However, the method utilizes artificial prior knowledge for denoising and does not analyze the impact of small samples and time complexity. Yang et al. [27] proposed a fault diagnosis method based on BiGRU and attention. BiGRU was utilized to gain advanced expressions from features extracted by CNN, then attention vectors were realized to diagnose each segment. However, Ref. [27] does not discuss the influence of small samples, and the means of training is relatively conventional, and the performance of model has not been further mined. In addition, the number of training samples of DCA-BiGRU is 60% of that of Ref. [22] with more difficult diagnosis.

Although previous methods have achieved relatively satisfactory results, deep learning models often require plenty of samples to achieve the ideal generalization. However, due to the relatively small labeled data, models are often unable to fully learn the various effective features of the limited samples and prone to overfitting, which increases learning difficulties [28]. Besides, the latest activation functions and gradient descent back propagation algorithms of all sorts have not been deeply comparative explored in fault diagnosis under small samples. Ultimately, due to the interference of different working conditions, the efficiency is difficult to be guaranteed, which puts forward higher requirements.

Therefore, aiming to regularization technologies and feature extraction advantages of models, a new fault diagnosis method for small samples based on dual path convolution with attention mechanism and BiGRU is proposed. The convolution layer aims to extract high–low frequency features of signals. Meanwhile attention mechanism that can be regarded as a cost sensitive learning method [28] values the fused features by allocated weights and sensitive information selection, pouring attention to the main spectra. Then, BiGRU can get the hidden information of different time sequence position. In addition to strengthening the connection between channels and reducing parameters, GAP and big kernels have more robust than capsule network on model by increasing receptive fields [29]. Moreover, the latest regularization methods further improve capacity of generalization on DCA-BiGRU, where label smoothing regularization (LSR) is introduced to balance the distribution differences between the labeled samples and calibrate DCA-BiGRU. Improved AMSGrad accelerator (AMSGradP) can be utilized to realize adaptive gradient optimization, and 1D-Meta-ACON (activate or not) can adaptively activate neurons, and adaptive batch normalization (AdaBN) enables DCA-BiGRU to have stronger transfer performance.

Fig. 1. CNN for fault diagnosis.

The main contributions of the paper are as follows:

1. For small sample fault diagnosis, a novel method based on designed attention mechanism and BiGRU is proposed from the regularization and model structure, and the effects of LSR, activation functions and back propagation algorithms are explored for the first time. Also, the proposed method has a higher test accuracy.
2. The sensitivities of attention mechanism and BiGRU to the ratio of training samples are discussed, where the proposed attention mechanism can capture the channel and spatial information of vibration signals. Then, designing GAP after BiGRU is beneficial for improving diagnostic performance. Also, visualization techniques are utilized to gain a better understanding of blocks in DCA-BiGRU.
3. For the noises contained in practical industrial data, a small sample transfer diagnosis framework based on pre-training is proposed. The experimental results prove that it has excellent capacities of generalization, adaptability and robustness compared to other bearing and gearbox diagnosis models under complex working conditions.

The rest of other parts in this paper is as follows. Section 2 is mainly about the basic theoretical model for fault diagnosis. DCA-BiGRU and latest regularization training strategies will be introduced in detail in Section 3. Section 4 presents some comparative experiments and analysis to prove the excellent performance of the proposed model. In Section 5, it will draw the conclusion and prospect for the future research.

2. Methodologies

2.1. Convolutional neural network

CNN generally consists of two modules: one filter block including convolution and pooling and the other classification block including full connection. The general CNN in fault diagnosis is shown in Fig. 1.

In signal processing, 1D-CNN is utilized to calculate delay accumulation of signals with the same kernel. The output is shown in Eq. (1).

$$y = ReLU\left(\sum_{w=1}^{W} k_w x_{t-w+1} + b_w\right) \tag{1}$$

where k_w and b_w are weight and bias matrix, respectively. x_{t-w+1} are input signals.

Pooling layer selects features and decreases parameters to accelerate convergence. The reason why maximum pooling is often utilized in

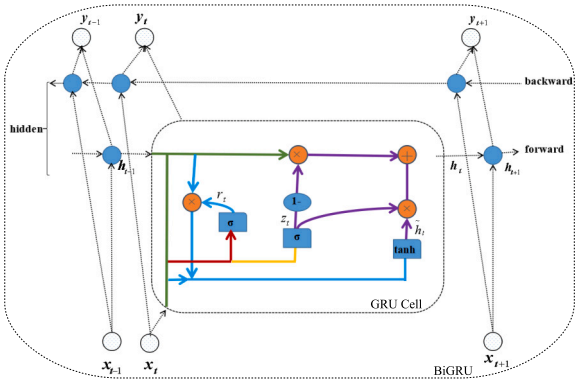


Fig. 2. The core structures of GRU cell and BiGRU.

fault diagnosis is that it can filter out insignificant information, as shown in Eq. (2).

$$y_i = \max_{j \in i} x_j \tag{2}$$

where y_i is representations, and j is neurons in the ith layer.

Batch Normalization (BN) can not only solve the internal covariate migration and improve training efficiency, but also act as a regularization trick because of batch selection randomly, which can enhance generalization instead of Dropout.

Activation functions can enhance learning capacity of neural network, improving the computational efficiency.

The distributed feature representations of vibration signals are mapped to the sample label space through full connection layer. Finally, SoftMax is applied for fault diagnosis.

2.2. Bidirectional gated recurrent unit

As shown in Fig. 2, gated recurrent unit (GRU) consists of an update gate z_t and a reset gate r_t. z_t is applied to control the extent to which h_{t-1} enters h_t. The higher values are, the more information h_t is entered. r_t is utilized to control the extent to which h_{t-1} enters $\widetilde{h}_t$. The smaller values are, the less $\widetilde{h}_t$ entry information. z_t and r_t are calculated at t moment as shown in Eq. (3)~(7).

$$r_t = \sigma[W_r \otimes cat(h_{t-1}, x_t)] \tag{3}$$

$$z_t = \sigma[W_z \otimes cat(h_{t-1}, x_t)] \tag{4}$$

$$\widetilde{h}_t = \tanh[W_{\widetilde{h}_t} \otimes cat(r_t \otimes h_{t-1}, x_t)] \tag{5}$$

$$h_t = (1 - z_t) \otimes h_{t-1} + z_t \otimes \widetilde{h}_t \tag{6}$$

$$y_t = \sigma(W_o \otimes h_t) \tag{7}$$

where $W_r, W_z, W_{\widetilde{h}_t}$ is the weight matrix, $cat()$ means that eigenvectors are connected. σ is $sigmoid$; $\otimes$ means element-wise product; the cell hidden state is h; $\widetilde{h}_t$ means candidate content in the current state, which controls the degree of receiving new information.

For Bidirectional gated recurrent unit (BiGRU), the forward $\overrightarrow{h_t}$ and backward $\overleftarrow{h_t}$ state without sharing parameters of signals are connected through different hidden layers, which together act on results h_t to express ampler features, as shown in Eq. (8)

$$\begin{aligned} &\overrightarrow{h_t} = GRU(x_t, \overrightarrow{h_{t-1}}), \quad \overleftarrow{h_t} = GRU(x_t, \overleftarrow{h_{t-1}}), \\ &h_t = w_t \overrightarrow{h_t} + v_t \overleftarrow{h_t} + b_t \end{aligned} \tag{8}$$

where w_t and v_t are weights corresponding to the forward or backward state of BiGRU respectively, and b_t is bias.

3. The proposed fault diagnosis method

3.1. Fault diagnosis procedure

In intelligent machine fault diagnosis, multiple structures and deep optimized algorithms can be integrated to achieve an amazing effect, where CNN–RNN has been applied to some extent [30,31]. However, as mentioned in Section 1, under small samples, the performance of CNN–RNN has not been further discussed, and deep optimization algorithms and training modes are conventional, whose potentiality has not been further explored.

Besides, in fault diagnosis, at the current moment, BiGRU makes the output state determined by the state of the previous and next moments conjointly. Of course, the last hidden neuron output is generally taken as the final hidden feature for diagnosis, for the reason that it has the most abundant features. Nevertheless, the strategy ignores signal features learned by other GRU cells.

Therefore, an intelligent fault diagnosis method called DCA-BiGRU has been proposed, which is composed of data enhancement, dual path convolution, attention mechanism, BiGRU, GAP and diagnosis layer, as shown in Fig. 4.

As shown in Fig. 3, in practical application, the specific steps of fault diagnosis based on DCA-BiGRU are as follows:

(1) Obtain the original signals and realize data segmentation and standardization.
(2) Divide signals into training, verification and test samples.
(3) Propose the model structures and diagnostic method.
(4) Offline training: use the training set and regularization strategies to train and save the optimal parameters.
(5) Online diagnosis: apply the test set to verify the model performance or load pre-training parameters and fine-tune the whole model to utilize parameter sharing transfer learning to realize timely training and fault diagnosis.

3.2. Dual path convolution and feature fusion

The dual convolution layer adopts two paths to extract the high–low frequency features of signals. On one path, two larger convolution kernels are utilized to learn low-frequency features. As described in Section 2.1, larger convolution kernels can enhance robustness against noises. On the other path, small convolution kernels are adopted to deepen neural network, which integrates four nonlinear activation layers to promote the discriminant capability. A combination of both widens the model and extract multiscale features, which provides a foundation for BiGRU to further learn advanced features. Finally, features are fused through element-wise product, where each channel contains abundant features.

To enhance the adaptability to DCA-BiGRU in different domains, AdaBN is leveraged to replace BN, where statistical information from source domain to target domain is adjusted to improve capacity of generalization [32].

3.3. The proposed attention mechanism of signals

Attention mechanism and LSR can be regarded as cost sensitive learning methods, and 1D-Meta-ACON can be seen as a means of meta-learning. For small samples, these regularization methods will make contributions to generalization and domain adaptability on model.

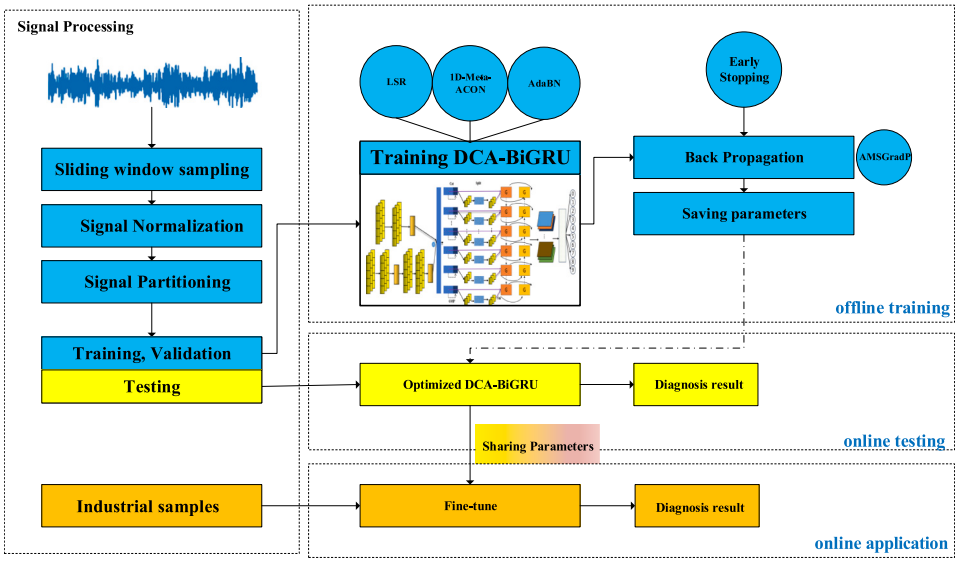


Fig. 3. Fault diagnosis framework based on sharing parameters.

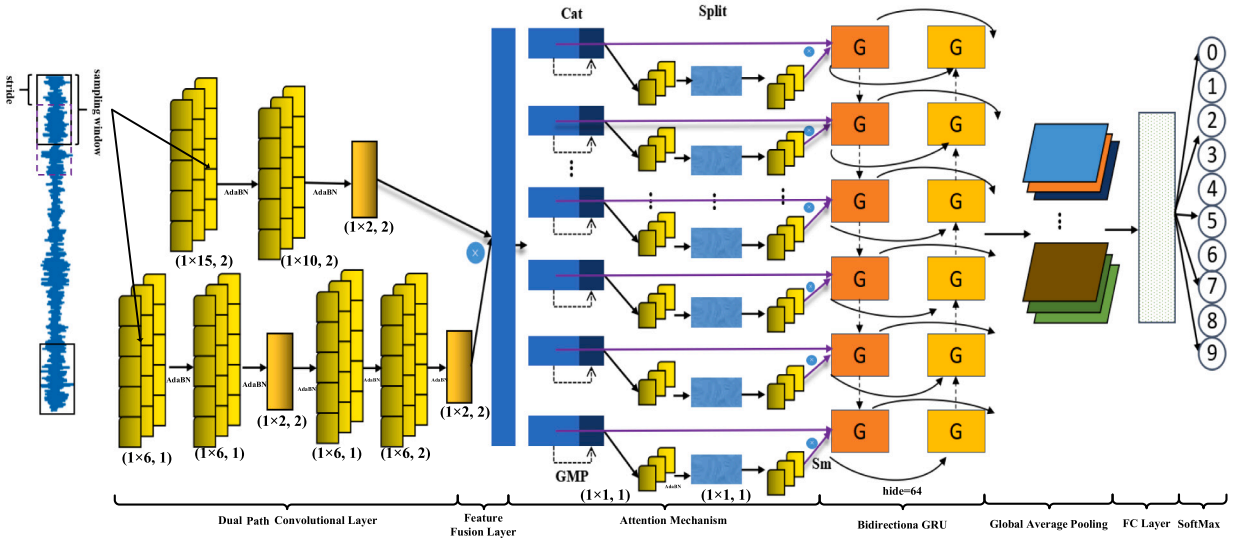


Fig. 4. Overall schema for the proposed network architecture of DCA-BiGRU.

3.3.1. Label smoothing regularization

Cross entropy loss (CE, l_0) tends to focus on one direction, leading to poor regulating capability. Consequently, smoothing coefficient ε are added to increase the correct diagnosis and reduce wrong diagnosis, which contributes to countering overconfidence of models and enhances learning capability. LSR(l) can not only upgrade generalization, but also calibrate models. It is mostly used in the field of image recognition, but rarely studied in fault diagnosis.

Supposing $p(k)$ is predicted distribution, $q(k)$ is real distribution, real distribution after label smoothing is $q'(k)$ with coefficient ε and category K, and label distribution is set to uniform distribution $\mu(k) = 1/K$. The relationship between l_0 and l is succinctly deduced, as shown in Eq. (9).

$$\begin{aligned} l &= -\sum_{k=1}^{K} \log(p(k))q'(k) \\ &= -\sum_{k=1}^{K} \log(p(k))[(1-\varepsilon)q(k) + \frac{\varepsilon}{K}] \\ &= (1-\varepsilon)[-\sum_{k=1}^{K} \log(p(k))q(k)] + \varepsilon[\frac{-\sum_{k=1}^{K}\log(p(k))}{K}] \\ &= (1-\varepsilon)l_0 + \varepsilon[\frac{-\sum_{k=1}^{K}\log(p(k))}{K}] \end{aligned} \tag{9}$$

By learning smooth labels instead of real labels to alleviate overfitting, so we argue that LSR has potential advantages in dealing with small samples in fault diagnosis.

3.3.2. The proposed 1D-signal attention mechanism

In Fig. 5, a 1D-signal attention mechanism is proposed, which can tell us what models demand to focus on about original signals.

To calculate attention between channels, it is indispensable to squeeze the dimension of input feature matrix, and global pooling

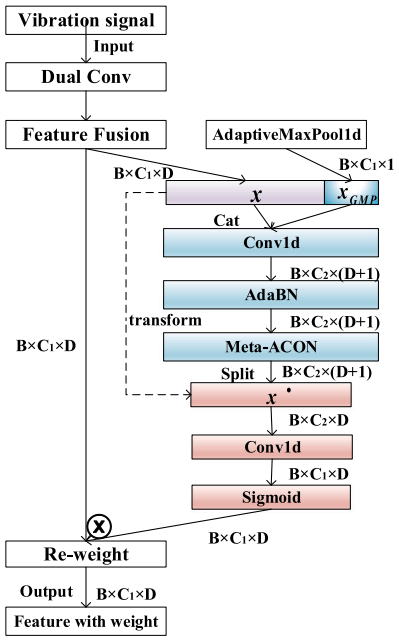


Fig. 5. The architecture of the proposed Attention Block.

is generally adopted. Furthermore, compared with GAP that focuses on the overall information, we argue that global max pooling (GMP) provides the crucial pulses (x_{GMP}) for the signal characteristic matrix (x), and in theory, it is the decisive pulses that is regarded as the main distinguishing criterion for fault diagnosis, so GMP is more suitable than GAP for the proposed attention block, which will be verified by experiment.

The cth channel GMP will be calculated as in Eq. (10).

$$x_{GMP}^{c} = \underset{0\le j<d}{\text{Max}}\, x_c(1,j) \tag{10}$$

Besides, in order to capture the spatial position information, it is perfect to establish the relationship between x_{GMP} and x, so they are concatenated together and sent into a convolution mapping function F_1 that shares 1×6. The dependency relationship is encoded to yield the intermediate characteristic connection matrix f as shown in Eq. (11).

$$f = \delta(F_1[cat(x, x_{GMP})]) \tag{11}$$

where δ is 1D-Meta-ACON activation function.

Then, f is split into x' and others. For the reason that the transformed original characteristic matrix x' has not only information of the critical pulse spectra, but also original signal characteristics x, just x' is retained. Another 1×1 convolution mapping function F_2 transforms x' to the same number of channels as x, as shown in Eq. (12).

$$g = \sigma[F_2(f_{x'})] \tag{12}$$

Finally, the output y_c is shown in Eq. (13).

$$y_c = x_c \otimes g_c \tag{13}$$

3.3.3. The improved 1D-Meta-ACON

Aiming at the nonlinearity of vibration signals, in the proposed attention block, a new activation function, Meta-ACON is applied [33]. Neither ReLU nor Swish, but both are considered and generalized to a general form. It is a form that can learn whether to activate.

Whether or not to activate neurons is determined by the smoothing coefficient β_c, so as to dynamically and adaptively eliminate inessential information. This is similar to the idea of the proposed 1D-signal attention mechanism, focusing on the central part in signals, which can conduce to improving capacity of generalization and transmission performance. Inspired by this, it is transformed into β_c suitable for 1D-signals, 1D-Meta-ACON, as shown in Eq. (14).

$$\beta_c = \sigma[F_4(F_3(\frac{1}{D}\sum_{d=1}^{D} x_{c,d}))] \tag{14}$$

where in forward propagation, β_c is calculated initially. The eigenvector x is calculated the mean value on D dimension. After $F_3, F_4(1 \times 1$ convolution) transform, β_c between (0,1) is obtained through Sigmoid, which is applied to control whether or not to activate or activation degree, where 0 means inactive. Finally, adaptive variables p_1 and p_2 are set, and supposing $p = p_1 - p_2$, return activation output (f_a) obtained by Eq. (15), and p_1 and p_2 are adaptively adjusted by back propagation.

$$f_a = p \times x_{c,d} \times \sigma[\beta_c \times p \times x_{c,d}] + p_2 \times x_{c,d} \tag{15}$$

1D-Meta-ACON is a general form, which not only solves the dead neuron problem, but also requires only a few parameters to learn to whether to activate. The research will explore if it can make a difference in small sample fault diagnosis.

3.3.4. AMSGradP

AdaBN contributes to improving capacity of generalization and scale invariance on model as same as BN. However, Heo et al. pointed out the gradient descent with momentum (GDM) will lead the effective step to decreasing rapidly during back propagation, resulting in slower convergence or even sharp minimizers, so AdamP [34] was proposed, which can just alleviate the puzzle by dropping the radial component during optimized update, regulating growth of weight norm, retarding the decay of the effective step size, thus training the model in a barrierless speed.

In this study, it is easy for small samples to converge to the local optimum. Unfortunately, the author has not given the improvement of more advanced AMSGrad. Inspired by this, the idea of Ref. [34] are introduced into AMSGrad called AMSGradP. In Appendix, Algorithm 1 outlines the pseudocode of AMSGradP.

3.4. BiGRU and GAP in fault diagnosis

LSTM has been described about in Section 1. In addition, by merging the forget gate and the input gate into the update gate, GRU has simpler structures, approximately 3/4 parameter quantity than LSTM, while it has the similar performance to LSTM in various tasks [35]. Apparently, GRU is more suitable for small samples. At the same time, it is argued that signals only have a deep correlation in one direction, which is not appropriate. As mentioned in Section 2.2, BiGRU is more suitable for the research.

FC layer with many parameters can greatly increase the risk of overfitting, while GAP will not produce extra parameters, and retain the partial spatial coding information from signals. In addition, as described in Section 2.2, we consider not only output of the last GRU cell, but also outputs of entire GRU cells, and GAP just fulfills the above requirements, preserving features learned by other GRU cells.

Lastly, we hold the view that the feature matrix has gathered the critical spectra from original signals, whose global information should be focused on, so GAP is preferred instead of GMP. The structures of DCA-BiGRU in detail is shown in Table 1, where a smaller number of parameters will facilitate the small sample fault diagnosis.

4. Result analysis and discussion

The proportion of each kind of training samples (α%) regards as the evaluation criteria. We argue if $\alpha < 0.5$, it can be called small samples [36]. Firstly, the superiorities of new regularization training methods proposed will be verified. Then, when $\alpha = 0.1 \sim 0.5$ (around 20~100 training samples), the small sample learning capacity of different models will be verified, and the performance will be evaluated

Table 1
The structures of DCA-BiGRU.

Type	Kernel/Stride	Unit	Activation	AdaBN	Input	Output	Parameter
Conv1d_1	18/2&10/2	/	1D-Meta-ACON	YES	(−1,1,1024)	(−1,30,248)	19036
Maxpool_1	2/2	/	1D-Meta-ACON	/	(−1,30,248)	(−1,30,124)	/
Conv1d_21	6/1&6/1	/	1D-Meta-ACON	YES	(−1,1,1024)	(−1,40,1014)	15816
Maxpool_21	2/2	/	1D-Meta-ACON	/	(−1,40,1014)	(−1,40,507)	/
Conv1d_22	6/1&6/2	/	1D-Meta-ACON	YES	(−1,40,507)	(−1,30,249)	14976
Maxpool_22	2/2	/	1D-Meta-ACON	/	(−1,30,249)	(−1,30,124)	/
Attention	1/1	/	1D-Meta-ACON	YES	(−1,30,124)	(−1,30,124)	666
BiGRU	/	128	Tanh	/	(−1,30,124)	(−1,30,128)	72960
GAP	/	/	/	/	(−1,30,128)	(−1,30,1)	/
FC	/	10	SotfMax	/	(−1,30)	(−1,10)	310
							Total: 123764

Table 2
Description of experimental parameters.

Settings	Value
Batch_size	32
Maximum epochs	150
Optimizer	AMSGradP
Learning rate	0.001
Weight decay(except bias)	0.0001
Early Stopping(patience)	10
AMSGradP(Nesterov)	True
1D-Meta-ACON(reduction)	16
Attention Block($C_1/C_2/D$)	30/6/124

under different working conditions and noises. Finally, parameter sharing that is applied to the small sample transfer learning to a new data set will be discussed, and meanwhile visual interpretations of DCA-BiGRU will be discussed. All experiments are performed under the same random seed, and the settings about experiments are shown in Table 2.

The experiment is implemented in PyTorch 1.8.0, Python 3.8.5, running on Intel(R) Core i7-6700HQ CPU @3.40 GHz (8G RAM), GTX970M GPU. The flow chart shown in Fig. 3 illustrates the overall framework for fault diagnosis. It has proved that fine-tuning the model can obtain more accurate diagnosis results, and the time cost is affordable [36,37]; hence the paper will adopt to fine-tuning the whole DCA-BiGRU for the anti-noise experiment.

4.1. Data enhancement

Data enhancement aims to generate more samples from vibration signals, prevent ANN from learning irrelevant features. As shown in Fig. 4, assuming that the sliding window is l, a sample is generated starting from the ith with an interval l, where the adjacent samples are set with an overlap value.

Assuming the sliding step size is m, N is the signal length, and the quantity of samples $n = \left\lfloor \frac{N-l}{m} \right\rfloor + 1 (m = 400, l{=}1024)$.

4.2. Model evaluation and metrics method

Diagnosis performance can be formulated by a confusion matrix, where it has two valuable indicators.

In multi-class case, this is the average of F1-score of each class with weighting depending on the average parameter, where sensitivity (recall) and precision are the key performance, which can be calculated as Eq. (16)~(17).

$$precision = \frac{TP}{TP+FP}, sensitivity = \frac{TP}{TP+FN} \tag{16}$$

$$F_\beta = \frac{(1+\beta^2)(precision + sensitivity)}{\beta^2 \times precision + sensitivity} (\beta = 1) \tag{17}$$

where True Positive (TP) is an outcome where the model correctly predicts the positive class. False Positive (FP) is an outcome where the model incorrectly predicts the positive class. False Negative (FN) is an outcome where the model incorrectly predicts the negative class. The weight of sensitivity is β times of precision.

Geometric mean (G-mean) tries to maximize the accuracy on each of classes while keeping these accuracies balanced. For multi-class problems it is a higher root of the product of sensitivity for each class, as shown in Eq. (18).

$$G - mean = \sqrt[n]{\prod_{n=1}^{N} sensitivity_n} \tag{18}$$

Table 3
Partition of CWRU data sets.

Data	Loads	Locations	FD(mm)	Label	α%
A/B/C/D	0/1/2/3	N	0.118/0.356/0.533	0	0.1~0.5
		IF		1/2/3	
		OR		4/5/6	
		BF		7/8/9	

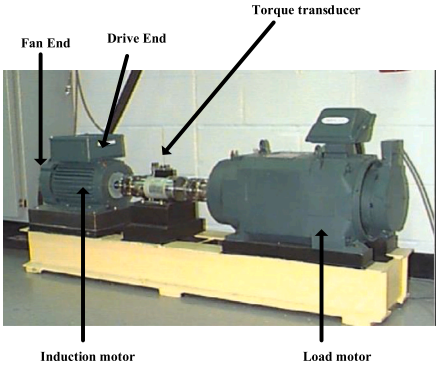


Fig. 6. Bearing fault diagnosis model test-bed.

4.3. Case 1: Data from CWRU

4.3.1. Description and division of data

The drive end rolling bearing data provided by Case Western Reserve University [38] is acquired by the device as shown in Fig. 6, where the single point faults (inner ring, outer ring, rolling element) are caused by electrical discharge machining (EDM), and the sampling frequency is 12 kHz, with 0~3HP loads and three types of damage degrees (0.118/0.356/0.533 mm). The acceleration sensor that is located at the drive end of the motor housing collects acceleration data. According to different loads, signals are divided into four data sets: A, B, C, D, as shown in Table 3.

4.3.2. The discussion of batch_size

A larger batchsize can shorten the training time of each epoch, but it may also reduce capacity of generalization, so a balance should be struck between both. For this reason, under $\alpha = 0.3$ with data set B, only batchsize changes, and the results are shown in Table 4.

Table 4
Comparison of training results between different batch_size.

batch_size	Early stopping	eval_loss	eval_acc	Time/s
8	34	0.5532	100%	311.74
16	40	0.5684	99.64%	226.82
32	75	0.5577	99.93%	249.14
64	73	0.5706	99.64%	244.77
80	126	0.5804	99.71%	367.85
100	100	0.5955	99.21%	296.53
128	80	0.6219	98.50%	223.33

Table 5
G-mean of DCA-BiGRU under different loads.

α (%)	G-mean			
Data set	B→ A	B→ B	B→ C	B→ D
0.1	95.30	96.78	99.21	93.24
0.2	98.60	99.41	99.09	98.32
0.3	99.48	99.71	99.60	98.56
0.4	99.21	100	99.86	98.83
0.5	99.74	100	100	98.97

As is seen to us, the training difficulties with different batchsizes are not consistent, resulting in different epochs of early stopping. Apparently, it can achieve similar performance in batchsize = 8 or 32 (100%, 99.93%), but the latter takes less time, so batch_size = 32.

4.3.3. Ablation comparative experiment

The ablation experiments regarding DCA-BiGRU(M_5) are carried out on four data sets A, B, C and D. The contrast models are PCA-SVM(M_1), DCNN-BiGRU (without attention, M_2), DCNN (without attention and BiGRU, M_3) and DCA (without BiGRU, M_4), which take G-mean as the index. In order to avoid the random influence, each experiment repeats five times to get error bars as shown in Fig. 7, and A→A represents training set→test set. The X-axis shows the proportion of the training(α). At the same time, the running time of different models, different loads in different α is recorded until early stopping, as shown in Table 6.

From Fig. 7, Table 6, as α augments, models learn more features and G-mean gains an increase. Due to the lack of elaborate processed of original signals, SVM cannot effectively deal with high-dimensional signals. Also, by comparing M_4 and M_5, it can be illustrated that BiGRU has advantages in coping with small samples, which generates hidden features, and contributes to the performance of the model to increase by 21%~36%. From M_4 and M_5 in Fig. 7(c), when $\alpha = 0.3$, $M_4 = 0.9031$, while $M_3 = 0.6431$, which demonstrates that attention mechanism also has a promising generalization for small samples, because it can guide models to pour attention to critical pulses, and only add 666 parameters. With a combination of both, $M_5 = 0.9822$. On the whole, both are conducive to performance of the model for small samples.

Furthermore, the trend of running time increases with the increase of α, where the advanced model requires more time. In total, DCA-BiGRU has the highest diagnostic efficiency.

4.3.4. Experiment under different loads

In general, the capability to deal with unlabeled samples from other loads is low when training with one data set. Therefore, it is indispensable to evaluate the migration versatility on DCA-BiGRU in fault diagnosis when the load changes. A, B, C and D have different loads and different signal distributions. In the past, most of the methods used to test the generality under $\alpha = 0.7$. The paper will explore the generality of the proposed model in small samples. Applying the model under training with Data set B, and the statistical results are displayed in Table 5.

As we can see, when G-mean < 0.99, the migration versatility enhances with the increase of α. For Data set D with load 3, although the signal distribution changes comparatively obviously, the performance has not decreased dramatically (average G-mean = 0.97). When G-mean > 0.99, the performance is slightly different due to random values. In addition, when load 0 with $\alpha = 0.1$, under small samples with inapparent fault pulses, G-mean > 0.95. In this case, DCA-BiGRU still achieve high performance, which fully indicates that it has a pleasant migration versatility.

Table 6
Time of test under different loads.

Models	α (%)	Time (s)			
		A	B	C	D
PCA-SVM	0.1	0.09	0.01	0.14	0.20
	0.2	0.16	0.02	0.34	0.15
	0.3	0.23	0.61	0.67	0.23
	0.4	0.31	0.12	0.56	0.29
	0.5	0.33	0.16	1.06	0.48
DCNN-BiGRU	0.1	21.31	19.40	15.44	14.57
	0.2	21.42	13.96	27.94	22.90
	0.3	30.84	21.02	43.89	54.01
	0.4	34.27	17.81	33.15	56.73
	0.5	48.68	26.76	63.85	60.34
DCNN	0.1	60.75	171.66	103.42	65.79
	0.2	103.07	123.74	193.69	96.90
	0.3	221.55	118.44	126.34	264.87
	0.4	144.39	244.19	123.14	292.19
	0.5	223.74	290.73	162.09	197.97
DCA	0.1	53.91	71.75	55.26	87.30
	0.2	148.07	161.38	110.96	126.35
	0.3	160.83	113.39	180.93	488.10
	0.4	224.96	140.84	332.17	256.32
	0.5	212.41	272.37	327.20	394.66
Ours	0.1	171.40	151.00	164.88	132.43
	0.2	403.03	270.81	193.33	387.71
	0.3	237.93	338.11	301.05	304.21
	0.4	249.03	242.73	430.96	317.33
	0.5	201.99	410.71	345.82	375.74

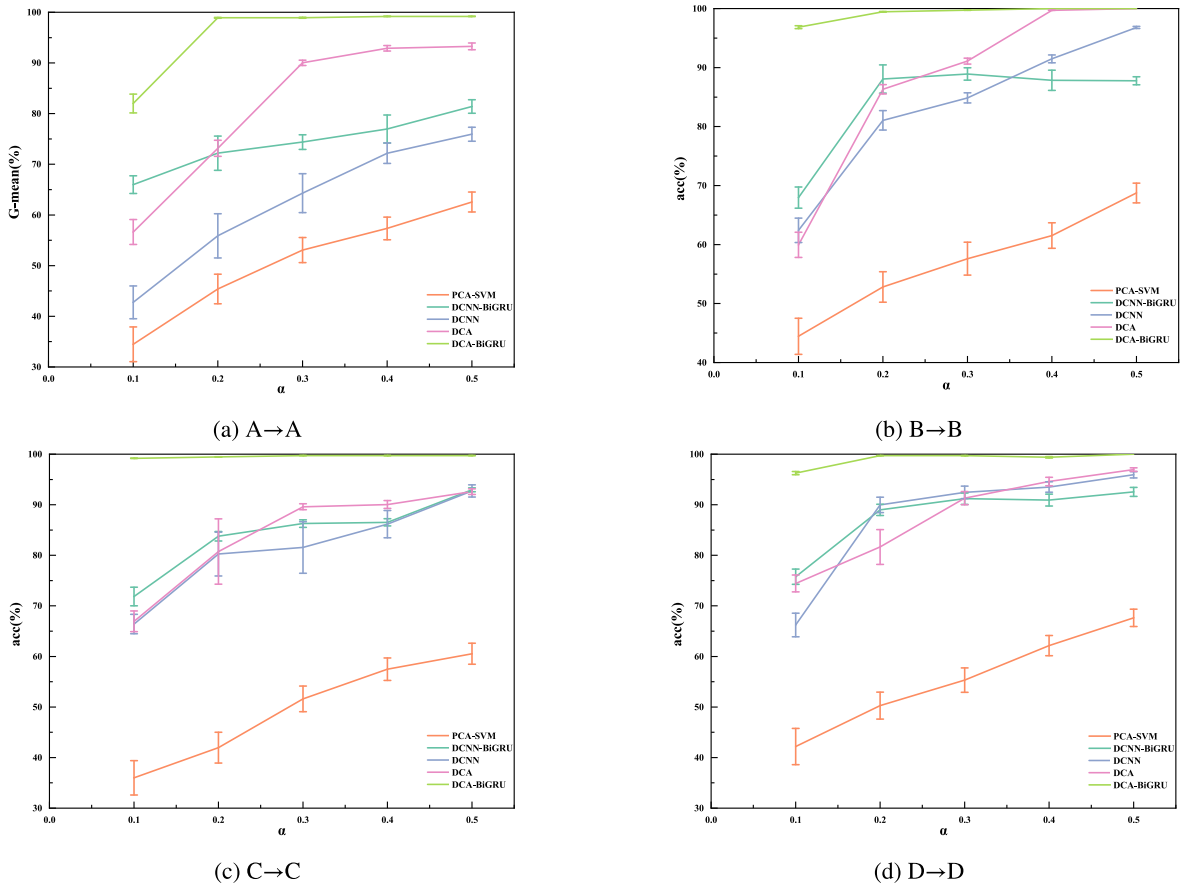
4.3.5. Analysis of regularization means

Vibration signals are distributed nonlinearly, while neural networks belong to linear calculation. In order to avoid vanishing gradient, the nonlinear non-saturating activation function is generally applied. In recent years, some latest activation functions have been widely utilized, but the improvement of them to the method has not been explored carefully in fault diagnosis. 1D-meta-ACON applied in this paper with only 1098 parameters combines the advantages of linear and nonlinear activation functions. One of them can be preferred by referring to the performances of them.

The related activation functions are shown in Fig. 8, where the gradient of Mish is smoother than that of ReLU, and Swish has the features of lower bound without upper bound, smoothness and non-monotonicity, which can be regarded as a smoothing form between linear and ReLU.

All results are carried out under Data set B with $\alpha = 0.3$, and the training loss and the accuracy of transfer are obtained, as shown in Figs. 9 and 10. It can be seen that all models can converge, and Softplus = 0.5856 with the maximum loss triggers early stopping earliest. When epoch = 114, early stopping is triggered on ELU. From the stability of convergence, expect for ReLU and Softplus, the other four functions are relatively stable, where the differences of loss are small in later epochs, and the difference of final losses among Swish, ELU, 1D-Meta-ACON is about 0.0003, while 1D-Meta-ACON has less epochs, with fastest convergence.

In addition, as shown in Fig. 10, Mish, Swish and 1D-Meta-ACON have a better migration generality, reaching 97.15%, 98.84% and 99.09% respectively under B→D. Meanwhile, ReLU and Softplus are poor under B→D, and the performance of 1D-Meta-ACON is the best, which improves by 0.25%. Regardless of extreme accuracy, one of the three activation functions can be chosen according to the reality.



(a) A→A (b) B→B

(c) C→C (d) D→D

Fig. 7. G-mean values of test under different loads.

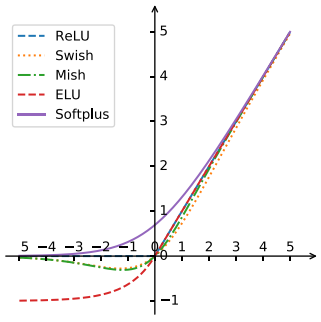


Fig. 8. Different activation functions.

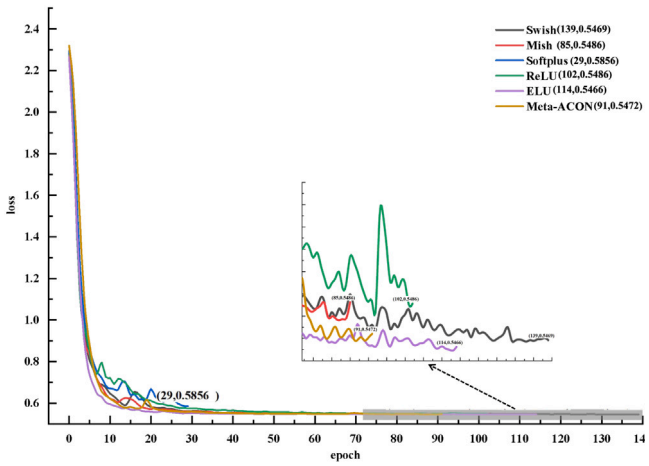


Fig. 9. Losses under different activation functions.

Similarly, the effects of different adaptive optimization gradient algorithms are compared. For a certain neural network, they are utilized to optimize the objective functions, and parameters are continuously updated in a negative direction until an optimal solution. The closer solution is to the global optimum, the neural network has better generalization.

Optimizers:SGDM(0.576), AMSGrad(0.569), AadmW(0.565), AdaBelief(0.561), AdaBound(0.565), AdamP(0.562), Adam(0.579), AMSGradP(0.555). The experimental results of verification set are shown in Fig. 13. It can be seen that the accuracy of several optimization algorithms reach more than 99%. Adam has the maximum oscillation amplitude, and when epoch = 142, it triggers early stopping. Compared with AMSGrad(99.64%), AMSGradP(99.86%) improves by 0.22%. Also, SGDM reaches 99.86%, yet it requires more epochs. From the point of view of convergence speed and value, SGDM, Adabelief and Adam converge slowly, but AMSGradP has the fastest convergence speed and highest validation accuracy, which indicates that adding radial component, AMSGradP retards the reduction of effective step, so that the algorithm reaches the vicinity of optimal point with a relatively appropriate effective step, and constantly updates nearby, converging to 0.555. Except for these, the speed of other algorithms is not much different. The above analyses fully indicate that adding radial components and adjusting norm growth can effectively improve results for gradient descent algorithms in fault diagnosis.

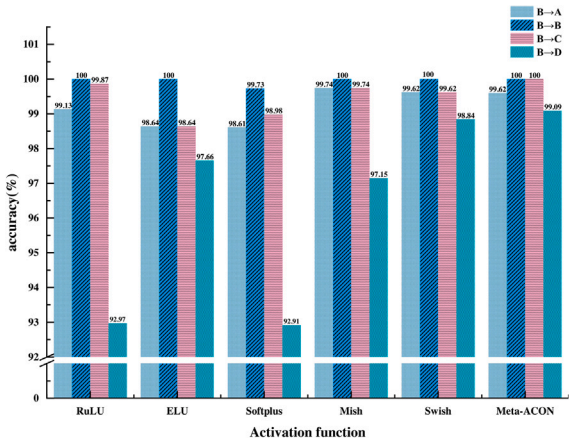


Fig. 10. Generality under different activation functions.

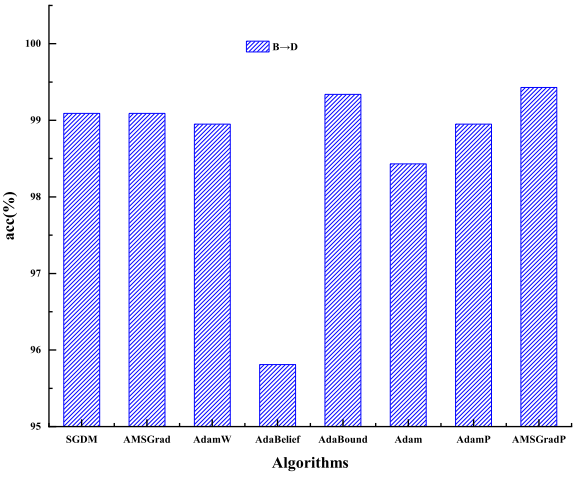


Fig. 11. Performance and time under different algorithms.

Besides, the generality of transfer of each algorithm is evaluated in Fig. 11, which displays the performance of DCA-BiGRU trained under $\alpha = 0.3$ with Data set B.

It can be found that Adabelief has the worst generalization in the rolling bearing task, which is only 96.76% under B→D. AMSGradP and Adabound have similar performance, whereas AMSGradP is more stable for migration because of a smaller error and has the acceptable training time. Considering comprehensively, AMSGradP is more superior.

Based on the above argumentum, a benchmark model can be trained applying AMSGradP and fine-tuned employing AdamW with fastest converge.

Eventually, the effects of different optimized strategies on the model are compared. (W:AdaBN, GHMC:gradient harmonizing mechanism for classification, FL:Focal Loss, G:GAP, B:BN). As an example, AdaBN, GAP and LSR are applied into DCA-BiGRU(WLSRG). In the field of NLP, GHMC, FL, and LSR acquire more attention for unbalanced distribution, but they have not been contrastively studied in fault diagnosis under small samples.

It displays the influence of different loss functions and optimized strategies in Fig. 12. Obviously, the task named B→D is more difficult. Initially, for WLSR, WFL, WGHMC and WCE, it can be stated that CE has the shortest training time, whose accuracy is only 95.37%. GHMC solves the problems of outliers and parameter joint training existed in FL and improves by 0.44%. Compared with three, LSR with 99.32% has the maximum accuracy. Furthermore, by comparing WLSR and BLSR, there is an improvement of about 1.36% by applying AdaBN.

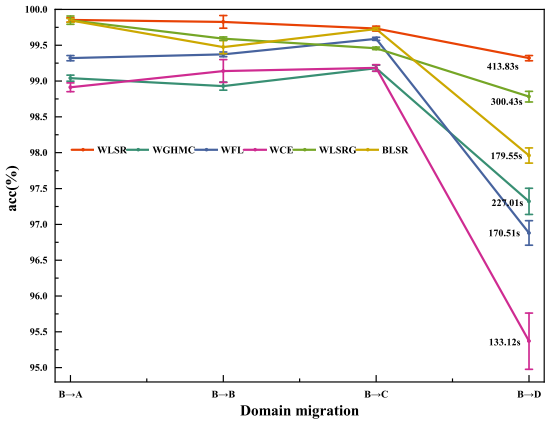


Fig. 12. Accuracy under loss functions and strategies.

Table 7
Time of different α under SNR = −4 dB.

models	Time/s				
α (%)	0.1	0.2	0.3	0.4	0.5
DCNN-BiGRU	16.30	17.44	27.12	43.55	75.82
DCNN	31.67	38.71	54.88	60.88	93.28
DCA	28.68	52.36	75.63	82.83	109.25
DCA-BiGRU	26.01	58.44	68.42	99.94	102.45

Ultimately, as mentioned in 3.3.2, by comparing WLSR and WLSRG, GAP is 98.78%, which descends by 0.54% than GMP in attention block.

In conclusion, the latest training methods make contributions to improve capacity of generalization.

4.3.6. Analysis of anti-noise robustness

Signals mostly contain noises in real situation. Hence, the study will analyze the anti-noise robustness under different signal-to-noise ratio (SNR), which is defined as in Eq. (19).

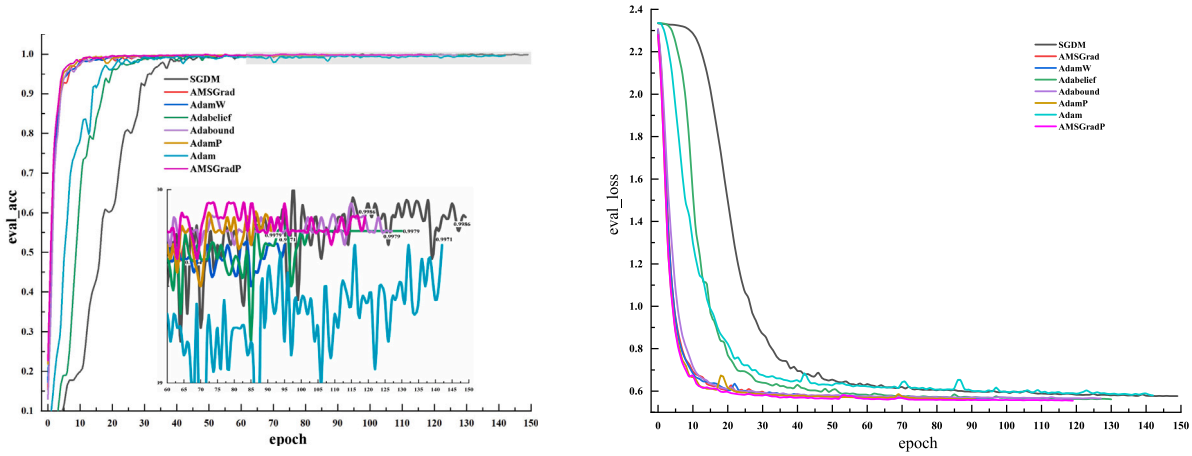
$$SNR_{\mathrm{dB}} = 10\lg\left(\frac{P_{signal}}{P_{nosie}}\right) \tag{19}$$

where, $P_{signal} = \frac{1}{N}\sum_{i=1}^{N} x_i^2$ is original signal power and P_{noise} is noise power.

Different from the previous methods that the model is directly trained by noise signals. The fault diagnosis framework with sharing parameters shown in Fig. 3 will be applied, which consists of off-line pre-training and online. The off-line will utilize AMSGradP and Data set B to obtain the pre-training parameters, while the online mainly aims to fine-tune models to achieve high efficiency, where the training time will be cut down because parameters are close to optimization values, so that noises can be quickly smooth away.

In this study, Gaussian white noises with SNR = −4~6 dB will be added to original signals. AdamW is applied to fine-tune the whole pre-training model. Besides, other settings are the same. Previous studies have declared that with the increase of SNR and α, the accuracy of test is continuously improved. Therefore, a case with SNR = −4 dB and $\alpha = 0.1$ is applied to examine performance. Results regarding training time and G-mean are shown in Table 7 and Fig. 14.

On one hand, with the increase of α, G-mean also increases, but the time cost also increases. However, it is reduced by approximately 2/3, compared with unloaded pre-trained models. On the other hand, DCA-BiGRU still has the highest diagnostic accuracy. Taking $\alpha = 0.3$ as an example, four models are 87.10%, 74.18%, 77.09%, 92.72% in turn, where BiGRU improves 19.92%, and attention mechanism improves 5.62%. All in all, attention mechanism and BiGRU has a strong capacity of robustness and diagnostic efficiency.



(a) Accuracy (b) Loss

Fig. 13. Accuracy and losses of verification set.

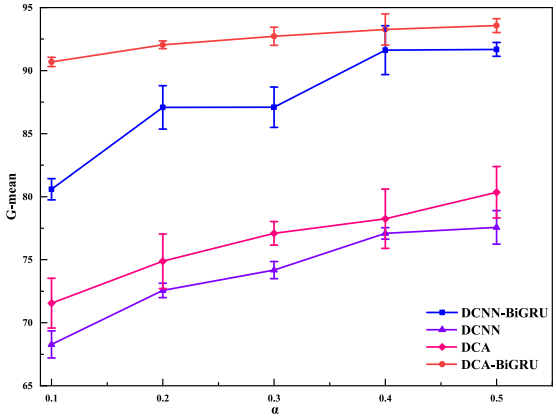


Fig. 14. Fault diagnosis based on sharing parameters.

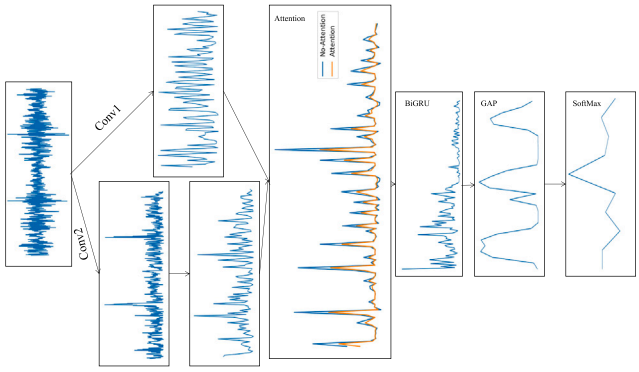


Fig. 15. The signal changes in DCA-BiGRU.

Table 8
Evaluation under $\alpha = 0.1$, Data set B.

SNR	−4	−2	0	2	4	6
G-mean	90.72	92.11	94.84	96.20	98.32	98.32
Time	26.08	60.15	11.09	11.01	11.98	11.00

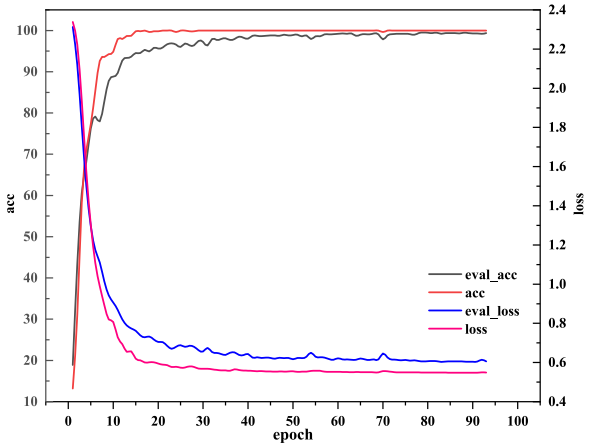


Fig. 16. Training and verification performance.

Another random seed is set to further evaluate DCA-BiGRU under conditions with $\alpha = 0.1$ and SNR = −4~6 dB, as shown in Table 8. With the increase of SNR, G-mean also increases. In addition, SNR = −4 dB or −2 dB requires more time, because it may be that larger noises cause higher learning difficulty, and demands more epochs. Furthermore, DCA-BiGRU achieves G-mean > 0.9 at various SNRs, manifesting that it has an excellent anti-noise performance.

Finally, the changes of original outer ring fault signals in DCA-BiGRU are shown in Fig. 15. With the depth of network, signal features become more abstract, and it is easier to realize diagnosis.

4.4. Case 2: Data from university of connecticut

4.4.1. Description and analysis of data

The data that is shared from University of Connecticut is collected from a two-stage gearbox [39,40], where the acquisition device is shown in Fig. 17, and the acquisition frequency is 20 kHz, and The signals are recorded through a dSPACE system(DS1006 processor board, dSPACE Inc.). The specifications of the accelerometer including frequency range, measure range, and sensitivity are 0.5 Hz–10 kHz, ±50 g, and 100 mV/g, respectively. Nine different gear conditions are introduced to the pinion on the input shaft, including healthy condition, missing tooth, root crack, spalling, and chipping tip with five different levels of severity, and time-domain signals of nine states are showed in Fig. 18.

In original signals, a total of 104 samples with 3600 points are collected for gearbox states. In order to facilitate experiments, all signals in a certain state are integrated into a column, and the training

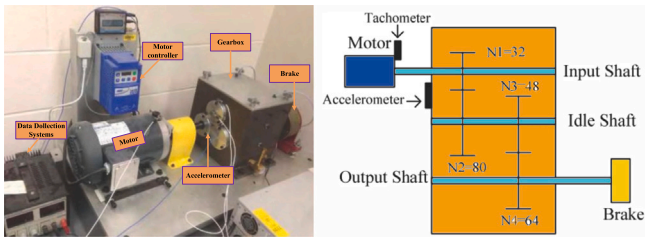


Fig. 17. Gearbox system.

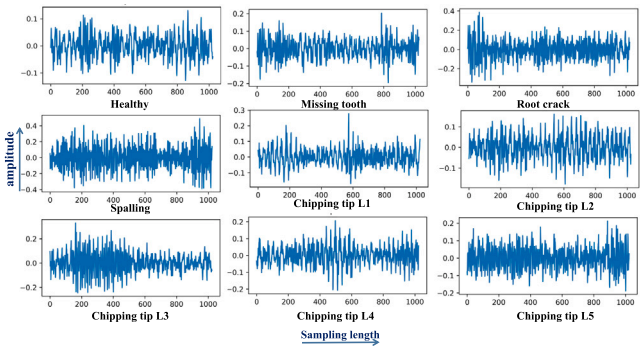


Fig. 18. Vibration signals of nine faults.

set, verification set and test set are obtained by acquisition methods mentioned in Section 4.1. The label of each state is 0~9 as shown in Fig. 18.

4.4.2. Evaluation under different working conditions

In reality, while gearbox system is recorded in a fixed sampling rate, due to speed variations under load disturbance, geometric tolerance, and motor control error etc, the time-domain signals also reflect the changes of different working conditions. And Fig. 16 reflects the change curve of accuracy and loss of training set and verification set at $\alpha = 0.3$, where DCA-BiGRU has an excellent convergence performance. When epoch =93, G-mean = 99.37%.

Figs. 19 and 20 embody the performance and training time of each model with the increase of α. It can be indicated that with the increase of α, G-mean also increases in test set. When $\alpha = 0.1$, DCA-BiGRU has the first-class performance with G-mean = 96.34%, while 79.76% on DCNN-BiGRU. When α=0.5, these models almost always close to 100% except for SVM. Overall analysis displays that when $\alpha < 0.3$, DCA-BiGRU < DCNN-BiGRU < DCA < DCNN, so the combination of attention mechanism and BiGRU can just achieve the optimal performance. Similarly, the cost of high performance is more training time, which requires for loading the pre-training model to save training time.

4.4.3. Visual analysis

In order to further reveal the feature representations, the T-SNE technology is applied to feature visualization, where different colors describe different states. By comparing Figs. 21 and 22, it can be found that DCNN extracts features preliminarily and each state is further separated through the attention mechanism. BiGRU 2 classifies samples by extracting the hidden features at different positions. Finally, parameters of the classifier are reduced by GAP.

BiGRU 1 only gets the output of the last hidden layer. Through the comparison between BiGRU 1 and 2, it can be seen that GAP pays attention to the output of neurons in all hidden layers of BiGRU, which makes fault state separation more obvious and reduces the training pressure of diagnosis layer. In conclusion, DCA-BiGRU can better separate different states, which has a marvelous generalization.

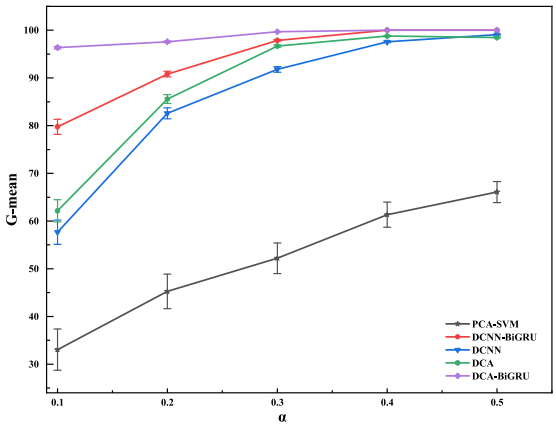


Fig. 19. G-mean in different α.

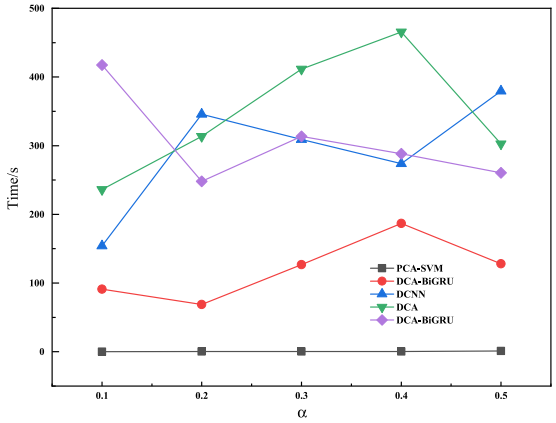


Fig. 20. Time in different α.

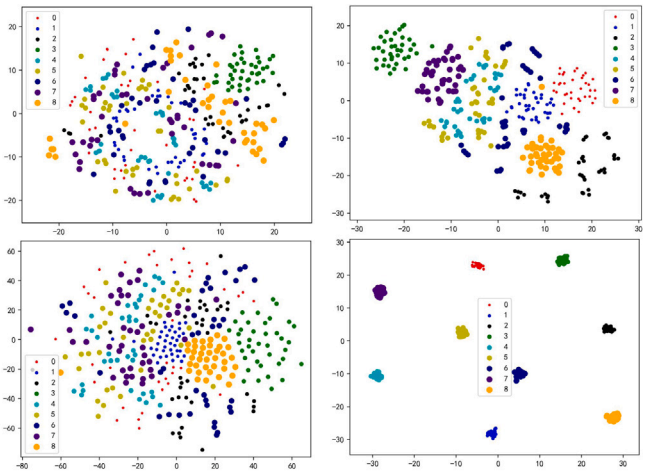


Fig. 21. Feature visualization of different layers.

The visualization of attention mechanism and BiGRU is shown in Fig. 23. The brighter the color, the higher degree of activation. From these, it is observed that the attention mechanism attaches importance to the degree of each channel in signals. In addition, BiGRU 2 further separates the dimensionality reduction signals and extracts more vivid and refined features. Different fault types have different neuron activation areas, so the corresponding features can be extracted from original signals without human intervention.

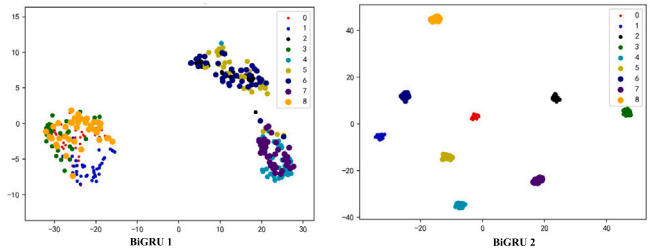


Fig. 22. Visualization of different BiGRU.

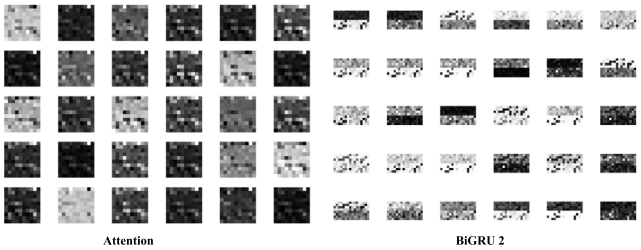


Fig. 23. Weight visualization of attention and BiGRU.

Grad-CAM++ is a widely applied visualization method, whose basic idea is that the weight of the feature map corresponding to a certain classification can be expressed as a gradient, and the global average of the gradient is utilized to calculate the weight. In addition, ReLU and the weight gradient a_i^{kc} are added into the feature map $weight$. Only one back propagation is required to calculate the gradient, which is originally applied to 2D, but is improved and applied to 1D-signals, as shown in Algorithm 2 in Appendix.

Attention mechanism is further explained, and Class Activation Mapping (CAM) is calculated by extracting the convolution kernel feature map of attention mechanism, as shown in Fig. 24. The higher the color level, the higher CAM and the higher the feature distinction. The light blue frame has circled higher parts of CAM. It can be found that the locations of different fault types activated by CAM are different, whose amplitudes are not the same, which fully demonstrates that the attention mechanism can distinguish the fault types without manual preprocessing. For example, Missing tooth and Spalling have two distinct areas of class activation. Besides, Chipping tip with different damage degree has different activation areas, where the impact amplitude is more distinct with the deepening of damage degree.

4.4.4. Anti-noise performance for gearbox

For the Gearbox fault, the learning rate is 0.0009 because of loading pre-training model. AdamW and fault diagnostic framework as shown in Fig. 3 are applied, and other parameters are as same as above.

Under SNR = 6 dB, the anti-noise capacity of models under different α is calculated, as shown in Fig. 25. Besides, the influence of SNR is recorded as shown in Table 9.

When $\alpha = 0.3$, with the improvement of α, G-mean is improving, indicating that the robustness of models is enhanced. Comparison between DCNN and DCNN-BiGRU shows that BiGRU improves performance by 5.31% when $\alpha = 0.1$. For DCA-BiGRU and DCNN-BiGRU, when $\alpha = 0.3$, attention mechanism makes the model increase by 0.86%.

In addition, by comparing whether the pre-training model is loaded or not, it can be found that loading the pre-training model not only improves G-mean, but also saves training time. The greater the noises, the more obvious the advantage of loading pre-training model. As an example SNR = 0 dB, G-mean of loading pre-training parameters is 85.28%, and that of unloading is 78.43%, increasing by 6.85%.

By observing the confusion matrix of both as shown in Fig. 26, DCNN-BiGRU whose sensitivity to Chipping tip L1 and L4 is low misclassifies part of the healthy samples. On the contrary, DCA-BiGRU can correctly distinguish healthy and fault samples, but misclassifies Missing tooth, Chipping tip L2 and L3. In particular, the sensitivity to Chipping tip L3 is low, which requires effective measures to improve performance under noises.

4.5. Comparison studies of diagnostic method

Finally, the rolling bearing data from CWRU is very popular in machinery fault diagnosis researches. Compared with some methods

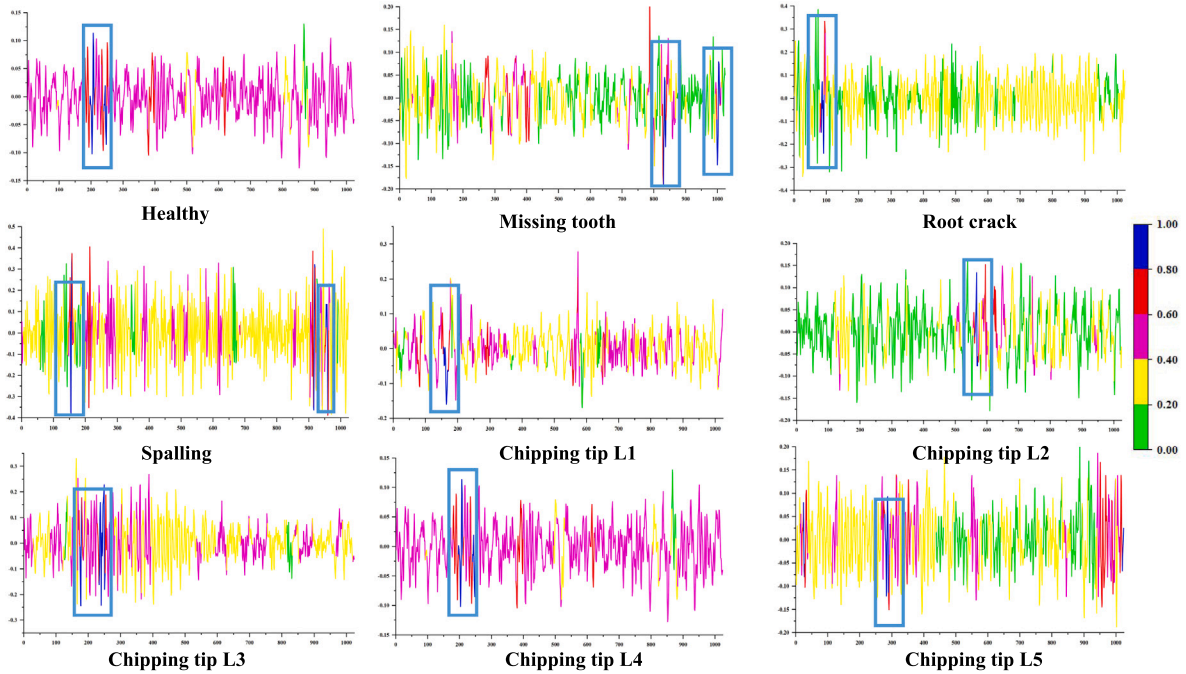


Fig. 24. Visualization of nine fault states under Grad-CAM++.

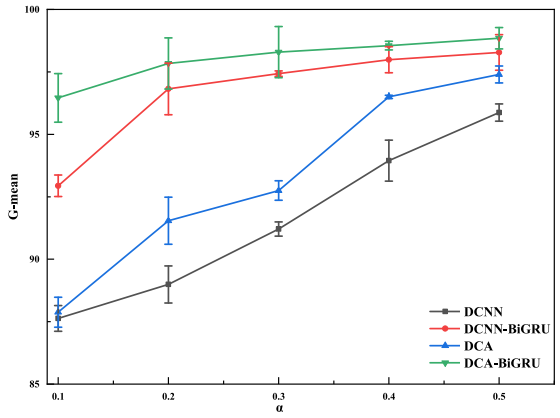


Fig. 25. G-mean of models under SNR=6dB.

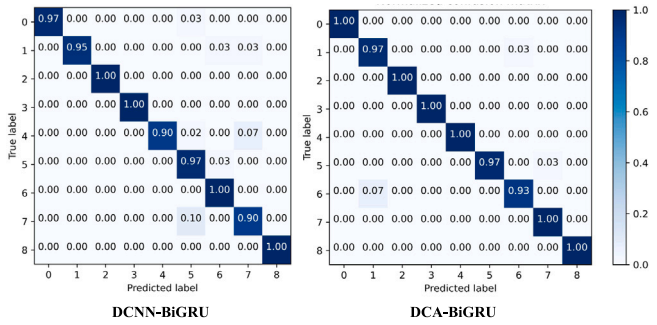


Fig. 26. Confusion matrix under SNR=6 dB, $\alpha = 0.3$.

Table 9
Anti-noise performance of DCA-BiGRU under $\alpha = 0.3$.

Load	Metric	SNR/dB					
		0	2	4	6	8	10
Y	G-mean	85.28	93.56	95.26	98.84	99.42	99.42
	Time	84.47	91.23	81.10	90.89	61.42	81.05
N	G-mean	78.43	87.47	93.01	97.96	98.57	99.13
	Time	103.59	131.34	104.74	195.88	232.62	134.38

listed in Table 10, and DCA-BiGRU still has reach 99.73% diagnostic performance in the case of no human intervention, lower α and less sampling length, and compared with DCA-BiLSTM, DCA-BiGRU increases by 0.17%.

Firstly, the length of sampling points can affect the diagnosis results. The fewer sampling points are, the fewer shock pulse will be contained in one sample. Compared with references listed in Table 10, in the paper, one sample collects 1024 points. Furthermore, although there are fewer sampling points in Ref. [3,22], sample dimension reduction and feature extraction algorithms are applied, most of which contain hyperparameters. In Ref. [3], smart evolution algorithm is adopted to search suitable hyperparameters, with high time complexity, and Ref. [22] is determined by manual experience. Then, compared with the number of training samples, in this paper, there is a lower proportion of training sets and a lower number of training sets. For example, the number of training sets is 60% than Ref. [22]. Finally, DCA-BiGRU also achieved a more interesting diagnostic result under harsher experimental environment and higher diagnostic difficulty. In addition, capsule network also has advantages in small sample fault diagnosis. However, after literature [23] is reproduced, capsule network has about 1.2 million parameters, while the parameters of DCA-BiGRU are about 120 thousand which means that DCA-BiGRU has faster training speed and higher diagnosis efficiency because fewer parameters make faster training speed.

Table 10
Comparison of fault diagnosis of CWRU.

Models	Length	Filtering	α	Accuracy
Ref. [3]	200	MCKD-RCMDE	0.8	99.00%
Ref. [10]	1200	/	0.8	98.36%
Ref. [14]	2000	Wiener filtering	0.7	98.46%
Ref. [22]	784	Fast Kurtogram	(100)	99.00%
ICN-Capsule	3000	Wavelet	0.83	99.96%
DCA-BiLSTM	1024	/	0.3(60)	99.56%
Ours	1024	/	0.3(60)	99.73%

MCKD: Maximum Correlated Kurtosis Deconvolution.
RCMDE: Refined Composite Multiscale Dispersion Entropy.

5. Conclusion

A novel DCA-BiGRU model based on attention mechanism has been proposed to identify the health state of equipment under small samples, where attention mechanism captures the spatial and channel relations of signals. The sensitivities of attention mechanism and BiGRU to the proportion of training set are discussed, and activation functions and gradient descent algorithms of all sorts have been explored. AMSGradP, 1D-Meta-ACON and other novel technologies are introduced to further improve capacities of generalization and robustness. Subsequently, DCA-BiGRU based on transfer learning, is verified on two different test rigs that are CWRU motor bearing data sets (Case 1) and University of Connecticut gearbox data sets (Case 2) respectively. Variety of visualization means are applied to initially reveal working principle of DCA-BiGRU, which shows that DCA-BiGRU has advantages in terms of diagnostic efficiency under different working conditions for small samples.

It can be noted that the differences between misclassified and other samples demand to be further explored. In addition, it is intractable for DCA-BiGRU to cope with the extremely imbalanced data set. In the future, machine learning such as meta learning, active sensitive cost learning, integrated learning or domain adaptation and generalization in transfer learning, will be combined with attention mechanism or other structures to address more complicated fault diagnosis situation with small sample and imbalanced data, which is worth further studying.

CRediT authorship contribution statement

Xin Zhang: Analysis, Funding acquisition. **Chao He:** Writing – original draft, Methodology, Software, Validation, Visualization, Investigation. **Yanping Lu:** Experiment. **Biao Chen:** Experiment. **Le Zhu:** Conceptualization, Software. **Li Zhang:** Supervision, Proofreading, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The authors are grateful for the supports of the National Key R&D Program of China (2018YFB1308700).

Appendix

Algorithm 1 AMSGradP

Input: learning rate, $\eta > 1$; momentum, $\beta_1, \beta_2 \subseteq (0,1)$; critical value, $\delta, \varepsilon > 0$; time step, t; step size, α;
Output: Resulting parameter, w_t;

for w_t not converged **do**
  $g_t \leftarrow \nabla_w f_t(w_t)$
  $m_t \leftarrow \beta_1 m_{t-1} + (1-\beta_1) g_t$
  $v_t \leftarrow \beta_2 v_{t-1} + (1-\beta_2) {g_t}^2$
  $\widehat{v}_t \leftarrow \max(\widehat{v}_{t-1}, v_t)$ *and* $\widehat{V}_t \leftarrow diag(\widehat{v}_t)$
  $p_t \leftarrow m_t / (\sqrt{\widehat{v}_t} + \varepsilon)$
  if $\cos(w_t, g_t) < \delta / \sqrt{\dim(w)}$ **then**
    $q_t = \prod_{w_t}(p_t)$
  else
    $q_t = p_t$
  end if
  $w_t \leftarrow w_{t-1} - \alpha q_t$
end for

Algorithm 2 1D-Grad-CAM++

Input: signal, x; category weight, y^c_{att}; feature map, A^k_{att};
Output: heatmap, h;

$grad \leftarrow \frac{y^c_{att}}{A^k_{att}}$
$a_i^{kc} \leftarrow \frac{grad^2}{2grad^2 + \sum_i A^k_{att} * grad^3}$
if $grad > 0$ **then**
  $weight \leftarrow grad \times a_i^{kc}$
else
  $weight \leftarrow 0$
end if
$weight$.size $\leftarrow$ x.size by linear interpolation
$h \leftarrow$ MinMaxScaler($weight$)

References

[1] J. Jiao, M. Zhao, J. Lin, K. Liang, A comprehensive review on convolutional neural network in machine fault diagnosis, Neurocomputing 417 (2020) 36–63.
[2] S. Zhang, S. Zhang, B. Wang, T.G. Habetler, Deep learning algorithms for bearing fault Diagnosticsx— A comprehensive review, IEEE Access 8 (2020) 29857–29881.
[3] H. Luo, C. He, J. Zhou, L. Zhang, Rolling bearing sub-health recognition via extreme learning machine based on deep belief network optimized by improved fireworks, IEEE Access 9 (2021) 42013–42026.
[4] Y. Ke, C. Yao, E. Song, Q. Dong, L. Yang, An early fault diagnosis method of common-rail injector based on improved CYCBD and hierarchical fluctuation dispersion entropy, Digit. Signal Process. 114 (2021) 103049.
[5] X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, X. Huang, Pre-trained models for natural language processing: A survey, Sci. China Technol. Sci. (2020) 1–26.
[6] S. Hao, Y. Zhou, Y. Guo, A brief survey on semantic segmentation with deep learning, Neurocomputing 406 (2020) 302–321.
[7] K. Tong, Y. Wu, F. Zhou, Recent advances in small object detection based on deep learning: A review, Image Vis. Comput. 97 (2020) 103910.
[8] G. Algan, I. Ulusoy, Image classification with deep learning in the presence of noisy labels: A survey, Knowl. Based. Syst. 215 (2021) 106771.
[9] Z. Zhao, T. Li, J. Wu, C. Sun, S. Wang, R. Yan, X. Chen, Deep learning algorithms for rotating machinery intelligent diagnosis: An open source benchmark study, ISA Trans. 107 (2020) 224–255.
[10] J. Li, X. Li, D. He, Y. Qu, Unsupervised rotating machinery fault diagnosis method based on integrated SAE–DBN and a binary processor, J. Intell. Manuf. 31 (8) (2020) 1899–1916.
[11] Y. Wang, G. Sun, Q. Jin, Imbalanced sample fault diagnosis of rotating machinery using conditional variational auto-encoder generative adversarial network, Appl. Soft Comput. 92 (2020) 106333.
[12] Z. Wang, Y. Dong, W. Liu, Z. Ma, A novel fault diagnosis approach for chillers based on 1-D convolutional neural network and gated recurrent unit, Sensors 20 (9) (2020) 2458.
[13] X. Wang, D. Mao, X. Li, Bearing fault diagnosis based on vibro-acoustic data fusion and 1D-CNN network, Measurement 173 (2021) 108518.
[14] X. Chen, B. Zhang, D. Gao, Bearing fault diagnosis base on multi-scale CNN and LSTM model, J. Intell. Manuf. 32 (4) (2021) 971–987.
[15] D. Huang, Y. Fu, N. Qin, S. Gao, Fault diagnosis of high-speed train bogie based on LSTM neural network, Sci. China Inf. Sci. 64 (1) (2021) 119203.
[16] X. Li, X. Kong, J. Zhang, Z. Hu, C. Shi, A study on fault diagnosis of bearing pitting under different speed condition based on an improved inception capsule network, Measurement 181 (2021) 109656.
[17] F. Zhou, S. Yang, H. Fujita, D. Chen, C. Wen, Deep learning fault diagnosis method based on global optimization GAN for unbalanced data, Knowl. Based Syst. 187 (2020) 104837.
[18] P. Kumar, A.S. Hati, Deep convolutional neural network based on adaptive gradient optimizer for fault detection in SCIM, ISA Trans. 111 (2021) 350–359.
[19] A. Zhang, S. Li, Y. Cui, W. Yang, R. Dong, J. Hu, Limited data rolling bearing fault diagnosis with few-shot learning, IEEE Access 7 (2019) 110895–110904.
[20] C. Wang, Z. Xu, An intelligent fault diagnosis model based on deep neural network for few-shot fault diagnosis, Neurocomputing (2021) http://dx.doi.org/10.1016/j.neucom.2020.11.070.
[21] J. Wu, Z. Zhao, C. Sun, R. Yan, X. Chen, Few-shot transfer learning for intelligent fault diagnosis of machine, Measurement 166 (2020) 108202.
[22] S.R. Saufi, Z.A.B. Ahmad, M.S. Leong, M.H. Lim, Gearbox fault diagnosis using a deep learning model with limited data sample, IEEE Trans. Ind. Inform. 16 (10) (2020) 6263–6271.
[23] T. Han, R. Ma, J. Zheng, Combination bidirectional long short-term memory and capsule network for rotating machinery fault diagnosis, Measurement 176 (2021) 109208.
[24] C. Li, K. Yang, H. Tang, P. Wang, J. Li, Q. He, Fault diagnosis for rolling bearings of a freight train under limited fault data: Few-shot learning method, J. Transp. Eng. A Syst. 147 (8) (2021) 04021041.
[25] W. Zhang, G. Peng, C. Li, Y. Chen, Z. Zhang, A new deep learning model for fault diagnosis with good anti-noise and domain adaptation ability on raw vibration signals, Sensors 17 (2) (2017) 425.
[26] K. Zhao, H. Jiang, Z. Wu, T. Lu, A novel transfer learning fault diagnosis method based on Manifold Embedded Distribution Alignment with a little labeled data, J. Intell. Manuf. (2020) 1–15.
[27] Z. Yang, J. Zhang, Z. Zhao, Z. Zhai, X. Chen, Interpreting network knowledge with attention mechanism for bearing fault diagnosis, Appl. Soft Comput. 97 (2020) 106829.
[28] T. Zhang, J. Chen, F. Li, K. Zhang, H. Lv, S. He, E. Xu, Intelligent fault diagnosis of machines with small & imbalanced data: A state-of-the-art review and possible extensions, ISA Trans. (2021) http://dx.doi.org/10.1016/j.isatra.2021.02.042.
[29] J. Gu, V. Tresp, H. Hu, Capsule network is not more robust than convolutional network, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognit, CVPR, 2021, pp. 14309–14317.
[30] T. Huang, Q. Zhang, X. Tang, S. Zhao, X. Lu, A novel fault diagnosis method based on CNN and LSTM and its application in fault diagnosis for complex systems, Artif. Intell. Rev. (2021) 1–27.
[31] M. Jalayer, C. Orsenigo, C. Vercellis, Fault detection and diagnosis for rotating machinery: A model based on convolutional LSTM, fast Fourier and continuous wavelet transforms, Comput. Ind. 125 (2021) 103378.
[32] Y. Li, N. Wang, J. Shi, X. Hou, J. Liu, Adaptive batch normalization for practical domain adaptation, Pattern Recognit. 80 (2018) 109–117.
[33] N. Ma, X. Zhang, M. Liu, J. Sun, Activate or not: Learning customized activation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognit, CVPR, 2021, pp. 8032–8042.
[34] B. Heo, S. Chun, S.J. Oh, D. Han, S. Yun, G. Kim, Y. Uh, J.-W. Ha, AdamP: Slowing down the slowdown for momentum optimizers on scale-invariant weights, in: International Conference on Learning Representations, ICLR, 2021.
[35] A. Shewalkar, D. Nyavanandi, S.A. Ludwig, Performance evaluation of deep neural networks applied to speech recognition: RNN, LSTM and GRU, J. Artif. Intell. Soft Comput. Res. 9 (4) (2019) 235–245.
[36] Y. Dong, Y. Li, H. Zheng, R. Wang, M. Xu, A new dynamic model and transfer learning based intelligent fault diagnosis framework for rolling element bearings race faults: Solving the small sample problem, ISA Trans. (2021) http://dx.doi.org/10.1016/j.isatra.2021.03.042.
[37] X. Li, Y. Hu, M. Li, J. Zheng, Fault diagnostics between different type of components: A transfer learning approach, Appl. Soft Comput. 86 (2020) 105950.
[38] K.A. Loparo, Bearing data center, Case Western Reserve University.
[39] P. Cao, S. Zhang, J. Tang, Preprocessing-free gear fault diagnosis using small datasets with deep convolutional neural network-based transfer learning, IEEE Access 6 (2018) 26241–26253.
[40] P. Cao, S. Zhang, J. Tang, Gear fault data. figshare. Dataset, http://dx.doi.org/10.6084/m9.figshare.6127874.v1.