Signal, Image and Video Processing (2021) 15:1889–1898
https://doi.org/10.1007/s11760-021-01939-w

ORIGINAL PAPER

A sparse denoising deep neural network for improving fault diagnosis performance

Funa Zhou[1] · Tong Sun[1] · Xiong Hu[1] · Tianzhen Wang[1] · Chenglin Wen[2]

Received: 13 February 2021 / Revised: 8 April 2021 / Accepted: 17 May 2021 / Published online: 3 June 2021

Abstract

Deep neural network (DNN) has been recently used in the field of fault diagnosis, but still their applicability is restricted to high computational complexity. In addition, useless information transformation between adjacent layers of the network could have a negative influence on the diagnosis accuracy. In this paper, a new DNN structure with sparse gate is designed to highlight the role of neurons contributed more by making it directly transfer through layers rather than transfer via an activation function. So it can reduce the computational complexity of network training since only those contributed less are required to be transferred via a nonlinear transformation. The proposed sparse denoising DNN (SD-DNN)-based fault diagnosis method can achieve more accurate diagnosis result with less computational complexity. It shows significant superiority to other-related methods in the case when only small size of training samples polluted by strong noise is available, which is very common for the engineering field of fault diagnosis. The experimental testing of fault diagnosis for rolling bearings verifies the effectiveness of the proposed method.

Keywords Fault diagnosis · Deep learning · SD-DNN · Sparse gate · Contribution

1 Introduction

As one of the key components, the healthy operation of rolling bearing is critical to the safety of intelligent manufacturing process to avoid some economic losses or even disastrous phenomenon. Accurate real-time fault diagnosis is an important means to secure healthy operation of the components of the intelligent manufacturing process [1–3]. Rolling bearing fault diagnosis methods can be mainly divided into the following three categories: model-based methods, knowledge-based methods and data-driven methods. Data-based methods are increasingly favored by experts in the engineering field since it can get rid of too much dependence on physical model and experience [4–6].

Deep learning is an efficient tool to extract feature involved in data. Fault diagnosis method using deep learning has received extensive attention from scholars [2, 3, 7, 8]. Existing methods can be divided into four categories: convolutional neural network (CNN)-based methods, long short-term memory neural network (LSTM)-based methods, deep belief network (DBN)-based methods and deep neural network (DNN)-based on methods constructed by stacking autoencoders [9–12]. CNN-based fault diagnosis method can extract features in the image by designing multiple convolution layers and pooling layers with an additional fully connected layer [13]. But it is difficult to achieve a real-time fault diagnosis result since 1-D signal is reshaped as 2-D matrix before it is fed into CNN. DBN-based fault diagnosis method can eliminate some uncertainty in the faulty data since RBM rather than AE is stacked to construct DBN, but the initialization process of DBN is complex and calculation burden is large [14]. LSTM extracts features from the data by using gate structure. The forget gate is used to discard useless information, the transferring gate determines which information needs to be transferred, and the output gate determines the output of the LSTM [15]. Compared with the above three methods, DNN constructed by stacking

✉ Funa Zhou
zhoufn2002@163.com

✉ Tong Sun
441438682@qq.com

[1] School of Logistic Engineering, Shanghai Maritime University, Shanghai 201306, China

[2] Institute of Automation, Guangdong University of Petrochemical Technology, Maoming, China

multiple AEs shows its advantage when 1-D sequence is processed [2, 3, 7, 12].

Due to the complexity of operation environment of mechanical equipment, the collected monitoring data are usually polluted by noise, which will affect the accuracy of DNN-based fault diagnosis. In addition, fully connected network structure of DNN may transfer unnecessary information to the next layer. Therefore, unsatisfying fault diagnosis result may be resulted. Methods to solve this problem can be classified into two classes: methods use filtering as a preprocessing technique of DNN and methods use sparse learning mechanism by adding a penalty term.

To improve the accuracy of deep learning-based fault diagnosis model, Fourier transform, median filtering, wavelet transform, etc., are used for pre-processing of denoising [16–20]. In order to extract more accurate feature involved in non-stationary vibration signal sampled from rotating machinery, digital wavelet frame is used to extract the features of fault signal with DNN stacked by multiple autoencoders [19, 20]. Some experts use intelligent filtering methods to process noisy information [21, 22]. However, the efficiency of pre-processing step will have strong influence on the final diagnosis result of DNN.

On the other hands, most processing of noisy data is based on designing new learning mechanism or optimization principles [3, 9, 23, 24]. Stacked denoising autoencoders and stacked sparse autoencoders (SSAE) are two representative methods of this class [23–27]. Lu et al. studied deep learning method for fault diagnosis of rotating machinery by stacking denoising autoencoder (DAE) [24]. Comparing to AE, DAE aimed to make the network capable of restoring the unpolluted data by using the noisy polluted data as the training samples such that SDAE is more robust. Wang et al. combine SDAE and CNN to improve the accuracy of fault classification [25]. But SDAE-based methods and their variants have complex computational burden. Sparse autoencoder (SSAE) is designed to suppress some hidden neurons by adding a penalty factor to the loss function when optimization of the training is considered. Sun et al. use SSAE to diagnose fault of induction motor with a high accuracy [28]. In order to solve the problem of shaft speed fluctuation, Sohaib et al. used SSAE to well extract the fault features involved in the training samples [29]. Some variations of normalization were designed to further improve the performance of normalization technique [30, 31]. Zhang et al. used batch normalization for each layer of the DNN to reduce the difficulty of training [26]. Qi et al. used the integrated empirical model and autoregressive model to process non-stationary signals to design a stacked sparse denoising autoencoder (SSDAE) to mine more advanced features [30]. Zhang et al. proposed a stacked marginalized SDAE to improve the noise reduction ability to achieve accurate fault diagnosis result [31].

As an application field of deep learning, the accuracy of deep learning-based fault diagnosis method depends on the size of the training samples, the quality of the training samples, the network structure and learning mechanism. The above-mentioned methods tried to perform some preprocessing analysis to improve the quality of the training samples or tried to design an efficient learning mechanism. They all failed to design a new network structure to highlight the role of neurons with large contributions by directly transferring them to the next layer. How to design a new network structure rather than new learning mechanism to accurately extract feature from noisy data with low computational burden is significant. In this paper, a new deep neural network structure with sparse gate is designed to highlight the neurons that contribute more by directly transferring them to the next layer without additional nonlinear transform. The designed sparse denoising DNN (SD-DNN) structure can achieve the purpose of network sparsity as well as noise reduction at the same time. Thus, more accurate deep learning-based fault diagnosis method with low training computational complexity is developed.

Remark 1: Transferring directly without nonlinear transformation via an activation function means that the computational burden required by the nonlinear transformation can be saved. In this sense, sparsity means that the information involved in the sparse neuron does not need to be transferred across the activation function.

The main contributions of SD-DNN-based fault diagnosis method proposed in this paper are as follow:

1. A new DNN structure with sparse gate is designed to process noisy data as well as make those neurons contributed much transfer to the next layer directly.
2. The proposed SD-DNN-based fault diagnosis method can achieve more accurate diagnosis with less computational complexity.
3. SD-DNN is significantly superior to other-related methods in the case when only small size of training samples polluted by strong noise is available, which is very common for the field of fault diagnosis.

The remaining sections of this article are organized as follows: Sect. 2 introduces deep neural networks. Section 3 addresses the fault diagnosis algorithm based on SD-DNN. Section 4 provides the experimental results and comparative analysis. The paper is concluded in Sect. 5.

2 Preliminary of DNN

AE is an unsupervised learning neural network with one hidden layer, which includes two stages: encoding and decoding. The goal of AE training is to restore the input data of

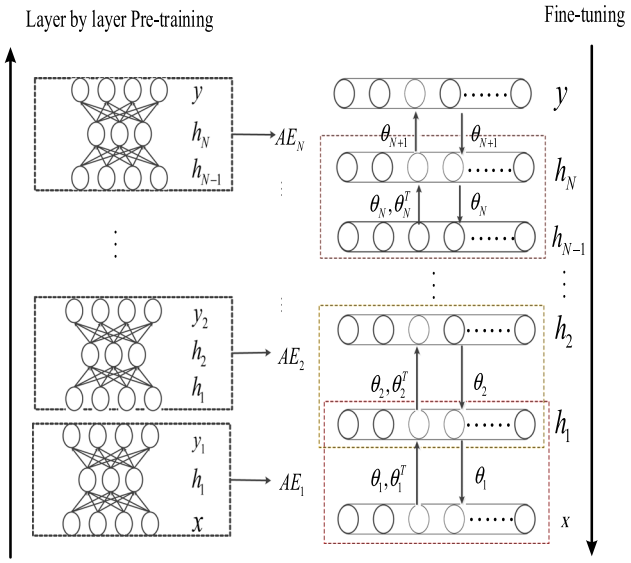

Fig. 1 Structure of DNN stacked with multiple AEs

AE, such that the trained AE has good data feature representation capability without any label information of the training samples. As shown in Fig. 1, DNN can be constructed by stacking multiple AEs to extract data's potential abstract features layer by layer in the means of bottom-up unsupervised learning with top-down supervised fine-tuning. The output of the previous AE's encoding is fed into the next AE's encoding [32, 33].

3 Fault diagnosis algorithm based on SD-DNN

3.1 Design of Sparse Gate

In the process of information transferring between neurons on adjacent layers, there will be some information correlated less with fault features. If the neurons transferring information correlated much with fault feature are highlighted before it is transferred to the next layer, those less contributed "noise" can be weakened. For this goal, a new DNN structure with sparse gate is designed in this paper to design a new transferring mechanism between layers by adjusting the weight of the sparse gate. The structure of the designed sparse gate between two layers can be shown in Fig. 2, where T is the switching gate, C is the carrying gate. Gate C means that information related to a specific neuron in the previous layer is directly transferred to the next layer without additional nonlinear transformation. While gate T means that information related to a specific neuron in the previous layer is transferred to the next layer via an activation function. In Fig. 2, the information

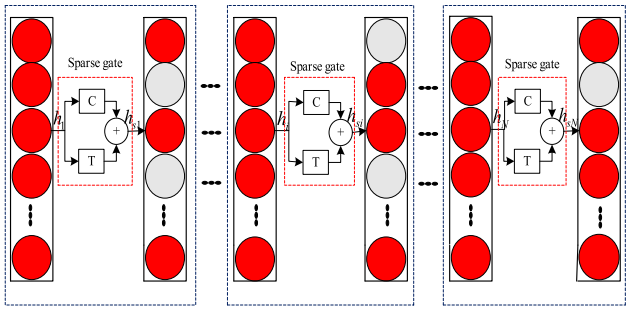

Fig. 2 The structure diagram of SD-DNN

related to red neurons can be directly transferred to the next layer, and the gray neurons are suppressed. SD-DNN is designed by adding a sparse gate on the basis of the DNN shown in Fig. 2.

The working mechanism of DNN with sparse gate shown in Fig. 2 is as follows.

The output of the sparse gate demonstrated in Eq. (1) is fed to the next hidden layer. If $h_i \in R^{n_i \times 1}$ is the output of a neuron on the previous layer, linear transformation without activation is required in forward propagation to get $H(h_i)$ as the input of the sparse gate. It is only used as a part of the sparse process

$$h_{si} = T(h_i) \circ H(h_i) + C(h_i) \circ h_i \tag{1}$$

$$T(h_i) = \sigma(W_s h_i + b_s) \tag{2}$$

$$C(h_i) = 1 - T(h_i) \tag{3}$$

$$H(h_i) = W_{hi} h_i + b_{hi} \tag{4}$$

where $\circ$ denotes the dot product operation. $h_{si} \in R^{n_i \times 1}$ is the output of the sparse gate. $H(h_i) \in R^{n_i \times 1}$ is the input of sparse gate.$T(h_i) \in R^{n_i \times 1}$ is the result of switching gate which is the forward propagation result via an activation function used in regular DNN. $T(h_i) \circ H(h_i)$ means that h_i should be forward transferred via an activation function. When the output of activation function $T(h_i) \in R^{n_i \times 1}$ is near 0, $C(h_i)$ is near 1,$C(h_i) \circ h_i$ means that h_i can be directly transferred without an activation function.

The function of the sparse gate is to establish a highway between two adjacent layers to make those information much correlated with fault features directly transferred through, while other correlated less needs to be transformed via an activation function in the forward propagation. In Fig. 3, gray neurons on the hidden layer are suppressed by sparse gate in the sense that they cannot be directly be transferred to the next layer. On the other hands, yellow neurons can be directly transferred to the next layer, just like transferring it through a highway, which will reduce the computational

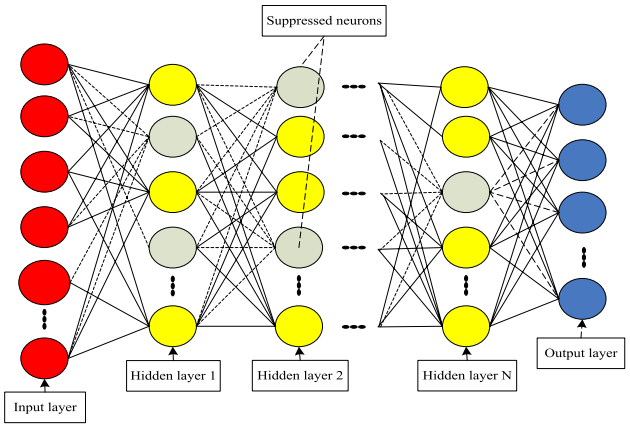

Fig. 3 The model of SD-DNN

complexity. By this means the computational complexity of network training can be saved once sparse gate is used.

Remark 2: The sparsity ability of the sparse gate means that some information contributed much can be directly transferred, while nonlinear transformation is required to the other information to extract more abstract feature.

Remark 3: In the case, when the training sample size is small, it is prone to suffer from overfitting problem since small number of training samples tries to learn a large number of connection weights. So SD-DNN with sparse gate is more significant in the field of fault diagnosis since small sample size of faulty data is common.

The flowchart of SD-DNN-based fault diagnosis algorithm is shown in Fig. 4. The detail algorithm includes the following steps:

3.1.1 Offline training

Build network model $\mathrm{NET}_{\mathrm{SD-DNN}}$:

$$\mathrm{NET}_{\mathrm{SD-DNN}} = \mathrm{Feedforward}(\theta_N, \theta_s, \theta_H) \tag{5}$$

where Feedforward is the function to construct neural network, and $\theta_H = \{W_{h1}, b_{h1}, W_{h2}, b_{h2}, \cdots, W_{hP}, b_{hP}\}$ is the parameters that needs to be transformed in the hidden layer.$\theta_{\mathrm{N}} = \{W_1, b_1, \cdots, W_N, b_N\}$ is the weight and bias of each layer, and $\theta_s = \{W_{s1}, b_{s1}, \cdots, W_{sP}, b_{sP}\}$ is the weight and bias of the sparse gate in each hidden layer. N is the number of network layers, $P{=}N-1$.

Forward propagation of SD-DNN is just similar to that of traditional DNN, as shown in Eq. (6)

$$h_1 = \sigma(W_1 X + b_1) \tag{6}$$

where X is the training sample, W_1 and b_1 are the weight and bias of the first hidden layer, and $\sigma(.)$ is the activation function.

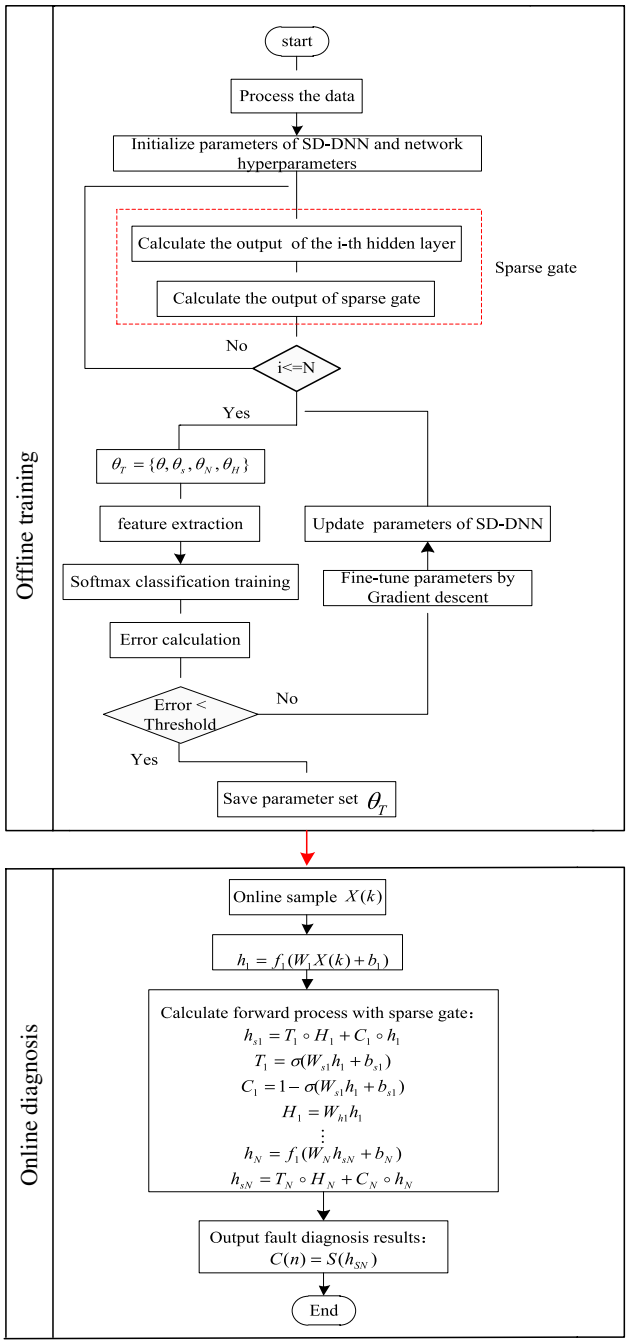

Fig. 4 Flowchart of the fault diagnosis algorithm based on SD-DNN

1 Sparsity of the first hidden layer. The sparse algorithm can be illustrated in Eq. (7)-(10):

$$h_{s1} = T_{s1} \circ H_1(h_1) + C_{s1} \circ h_1 \tag{7}$$

$$T_{s1} = \sigma(W_{s1} h_1 + b_{s1}) \tag{8}$$

$$C_{s1} = 1 - T_{s1} = 1 - \sigma(W_{s1} h_1 + b_{s1}) \tag{9}$$

$$H_1(h_1) = W_{h1}h_1 + b_{h1} \tag{10}$$

where W_{s1} and b_{s1} are weight and bias of the sparse gate. Equation (7) shows that the sparse gate T_{s1},C_{s1} are required to be learned to determine whether h_1 can be transferred directly or be transferred after nonlinear transformation.

2 Sparsity of the second hidden layer. The sparse algorithm is shown in Eq. (11)–(12):

$$h_2 = f_1(W_2h_{s1} + b_2) \tag{11}$$

$$h_{s2} = T_{s2} \circ H_2(h_2) + C_{s2} \circ h_2 \tag{12}$$

3 Sparsity of the Nth hidden layer. The sparse algorithm is shown in Eq. (13):

$$h_{sN} = T_{sN} \circ H_N(h_N) + C_{sN} \circ h_N \tag{13}$$

where h_N is the output of the Nth hidden layer.

4 Backpropagation of SD-DNN is similar to that of traditional DNN. Fed the output of SD-DNN h_{sN} into a classifier model to get the error of forward propagation. Then, BP algorithm is used to optimize the loss function, such that well-trained parameters of SD-DNN θ_T can be obtained, as is shown in Eq. (14).

$$\theta_T = \left\{ \theta_{\text{SD - DNN}}, \theta_C \right\} \tag{14}$$

Where $\theta_c = [W_c, b_c]$ is the trained model parameters of the classifier, $\theta_{\text{SD - DNN}} = \left\{ \theta, \theta_s, \theta_H \right\}$ is the trained model parameters of SD-DNN.

Remark 4: For the stage of offline training, the main differences between SD-DNN and DNN are pointed out as follows: (1) As shown in Eq. (7)–(10), in the forward process, additional operation corresponding to sparse gate between adjacent layers is required to highlight the role of neurons contributed much by forcing it a relatively large weight coefficient. (2) As shown in Eq. (4), the parameters of DNN as well as sparse gate are fine-tuned by backpropagation algorithm.

3.1.2 Online diagnosis

1. Feed the online sample at time k into the well-trained network to extract features $h_{sN,\text{online}}(k)$

$$h_{sN,\text{online}}(k) = G_{\text{SD - DNN}}(\text{NET}_{\text{SD - DNN}}, \theta_{\text{SD - DNN}}, X_{\text{online}}(k)) \tag{15}$$

where $G_{\text{SD - DNN}}$ is a function to describe the relation between the input and output of the well-trained SD-DNN.

2. Fed online features into the well-trained Softmax classifier to realize online diagnosis.

4 Experiment analysis

4.1 Experiment data and experiment design

The rolling bearing data set of Case Western Reserve University is used to verify the effectiveness of the proposed method [34]. The bearing data with sampling frequency of 12 kHz and fault diameter of 0.007 inches are used. Six categories of fault are included: normal, inner race fault, ball fault, outer race fault 1, outer race fault 2 and outer race fault 3. The experiment result of SD-DNN is compared with SAE, SDAE, SSAE,stacked sparse denoising autoencoder (SSDAE),CNN and LSTM.

The parameters of the network model are shown in Table 1. Table 2 shows the specific experimental design.

Remark 5: The second row of Table 2 means that there are 4 layers included in DNN constructed by stacking multiple AEs. The number of neurons on the input layer is 400, the number of neurons on the second layer is 200, the number

Table 1 Parameters of network

Model	Model parameters
SAE	Number of layers: 4, neurons in each layer:400/200/50/6, learning rate: 0.2
SDAE	Number of layers: 4, neurons in each layer:400/200/50/6, learning rate: 0.2
SSAE	Number of layers: 4, neurons in each layer: 400/200/50/6, learning rate: 0.2
SSDAE	Number of layers: 4, neurons in each layer:400/200/50/6, learning rate: 0.2
CNN	Convolutional layer:K_{size}:5*5 K_C:16 K_{step}:1, Pooling layer: P_{size}:2*2 P_{step}:2, fully connected layer: Number of neurons: 100, learning rate: 0.003
CNN with sparse gate	Convolutional layer:K_{size}:5*5 K_C: 16 K_{step}:1, Pooling layer: P_{size}:2*2 P_{step} :2,fully connected layer: Number of neurons: 100, learning rate: 0.003
LSTM	Cell number: 4, number of hidden neurons in the cell: 100, learning rate: 0.2
LSTM with sparse gate	Cell number: 4, number of hidden neurons in the cell: 100, learning rate: 0.2
SD-DNN	Number of layers: 4, neurons in each layer:400/200/50/6, learning rate: 0.2
Numbers of sparse gate: 2	

Table 2 Experimental design

Experiment	Number of fault type	Training sample size	Testing sample size
Experiment 1	6	300	300
Experiment 2	6	600	600

of neurons on the third layer is 50, and the number ot the output layer is 6.

4.2 Analysis of experimental results

Tables 3, 4, 5, 6 show the fault diagnosis accuracy of the corresponding five models. In order to reduce the influence of randomness, 10 times of average are conducted.

Table 3 Fault diagnosis result with training sample size 600

	Normal (%)	Inner race (%)	Ball (%)	Outer race1 (%)	Outer race2 (%)	Outer race3 (%)	Average accuracy (%)
SAE	100	63.89	94.4	91.00	73.60	45.30	78.27
SDAE	100	63.50	94.50	94.50	75.60	44.50	78.79
SSAE	100	55.6	92.20	87.70	67.70	44.20	74.66
SSDAE	100	70.40	93.30	91.10	71.70	47.30	79.16
SD-DNN	100	70.60	97.40	98.5	81.90	60.40	84.78
CNN	97.11	73.88	79.33	74.11	70.78	65.33	76.76
SD-CNN	98.57	80.00	80.71	81.71	81.29	75.29	82.981
LSTM	100	85.29	84.71	99.29	97.00	82.14	91.41
SD-LSTM	100	88.00	88.29	99.86	96.86	82.43	92.35

Table 4 Fault diagnosis result with training sample size 300

	Normal (%)	Inner race (%)	Ball (%)	Outer race1 (%)	Outer race2 (%)	Outer race3 (%)	Average accuracy (%)
SAE	100	20.60	44.60	72.60	25.2	29.80	48.77
SDAE	100	21.20	55.20	73.40	21.30	28.60	49.70
SSAE	100	16.29	35.43	67.71	25.43	24.26	44.86
SSDAE	100	22.20	47.20	73.40	21.30	28.60	49.73
SD-DNN	100	33.00	78.60	80.60	37.40	31.40	59.77
CNN	96.40	51.20	70.80	58.8	65.60	57.20	66.67
SD-CNN	97.00	65.00	87.00	72.00	72.00	64.87	76.28
LSTM	100	48.68	52.57	94.00	86.29	68.57	75.05
SD-LSTM	100	50.86	57.71	96.86	83.14	68.29	76.14

Table 5 Fault diagnosis result with different sizes of noise for sample size 600

Additional noisy variance	Normal (%)	Inner race (%)	Ball (%)	Outer race1 (%)	Outer race2 (%)	Outer race3 (%)	Average accuracy (%)
0	100	70.60	97.40	98.5	81.90	60.40	84.78
0.05	100	66.70	97.00	97.80	84.80	57.30	83.87
0.1	96.50	62.30	93.60	97.70	78.60	52.80	80.25

Table 6 Comparison of training time with different sample sizes (unit: second)

Training sample size	SAE	SDAE	SSAE	SSDAE	SD-DNN	CNN	SD-CNN	LSTM	SD-LSTM
300	43.91	50.47	46.62	54.83	20.51	218.85	163.88	123.78	67.88
600	76.38	78.01	79.44	85.15	31.56	1069.20	637.51	224.78	167.33

Fault diagnosis result with training sample size 600 is shown in Table 3. Column 2–Column 7 of Table 3 show specific diagnosis accuracy of each type of fault. The 8th column is the average diagnosis accuracy of different categories. The 9th column of is the increment to the traditional fault diagnosis method using DNN stacked with autoencoders.

From the second row that corresponding to the traditional fault diagnosis method using DNN stacked with AEs, it can be seen that when the sample size is small, traditional DNN-based method can well distinguish normal data but fail to diagnose the outer race 3 fault. Row 2 of Table 3 indicates that DNN model constructed by stacking DAEs can improve a little since the training data are collected from the experimental platform rather than the actual engineering field. If is difficult for SSAE to achieve a satisfying diagnosis accuracy when the specific learning algorithm is inappropriate, just as the diagnosis result Row 4 confirms. The 5th row indicates that by combining SDAE and SAE still cannot achieve a satisfying diagnosis result. The reason is that the above 3-mentioned methods all try to solve the problem using a revised learning algorithm of DNN without modifying the network structure. The 6th row of Table 3 shows that SD-DNN can achieve a higher fault diagnosis accuracy since it focuses on developing a spare gate to modify the structure of traditional DNN. This shows that SD-DNN inhibits neurons with small contributions and highlights those with large contributions during the propagation of fault features. For Ball and outer race2 fault, the diagnosis accuracy of SD-DNN is significantly higher than other models, which shows that SD-DNN can inhibit some neurons very well. The average fault diagnosis accuracy of SD-DNN is 6.51% higher than SAE. Comparing Row 6 with Row 7 and Row 9, it can be seen that the diagnosis accuracy of SD-DNN is lower than CNN and LSTM. But Table 3 indicates that both CNN and LSTM require more heavy computational burden. On the other hands, the comparison of Row 7 with Row 8 and comparison of Row 9 with Row 10 show that the designed sparse gate can improve the fault diagnosis accuracy at lower computational cost is suitable for DNN as well as CNN and LSTM.

Remark 6: The innovation of this paper is to design an improved network structure with sparse gate to achieve high accuracy with low computation complexity. The proposed method is developed for DNN. When network structure is changed to CNN and LSTM, the same conclusion can be achieved.

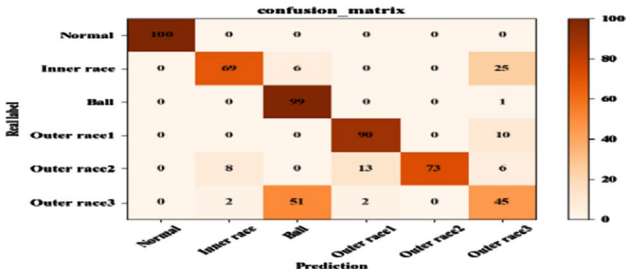

Fig. 5 Confusion matrix of SAE-based fault diagnosis

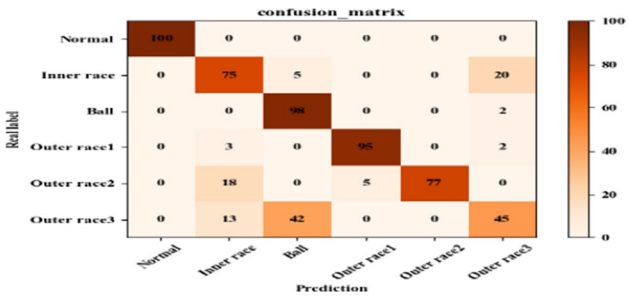

Fig. 6 Confusion matrix of SDAE-based fault diagnosis

Remark 7: In the experiment, all deep learning models use SDG as the optimizer for BP algorithm, when other optimizer is used, the same conclusion can also be achieved.

For the case, when only smaller sample size is available, in addition to affection by noise, smaller sample size usually makes DNN model suffered from overfitting problem. Sparse gate makes it possible for learning much less number of weighting coefficients between neurons on adjacent layers since it makes some information directly transferred. So it can partially alleviate the problem of overfitting problem. Fault diagnosis result for training sample size 300 is shown in Table 4. Comparing the 6th row of Table 4 with that of Table 3, it can be concluded that the proposed SD-DNN-based method is significantly superior to other methods since it can achieve an diagnosis accuracy increment of 11%.

Figure 5, 6, 7, 8, 9, 10, 11, 12, 13 are the confusion matrix of the corresponding nine methods when the training sample size is 600. The horizontal axis of the confusion matrix is the predicted number of correct classifications of each model. Other locations are misclassified. The darker the color, the high diagnosis accuracy.

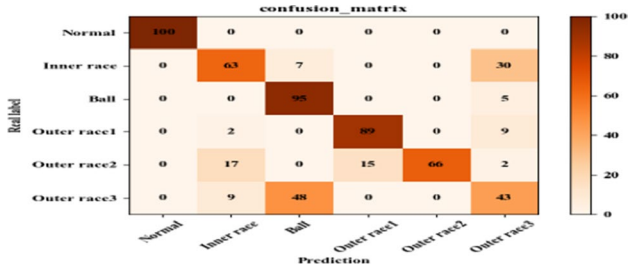

Fig. 7 Confusion matrix of SSAE-based fault diagnosis

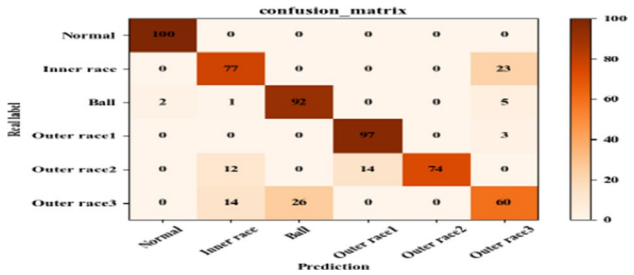

Fig. 8 Confusion matrix of SSDAE-based fault diagnosis

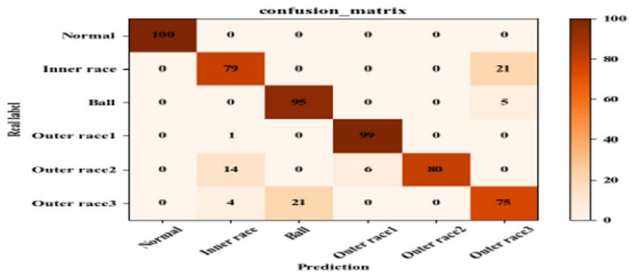

Fig. 9 Confusion matrix of SD-DNN-based fault diagnosis

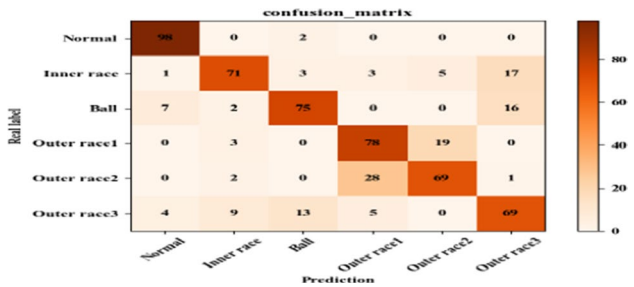

Fig. 10 Confusion matrix of CNN-based fault diagnosis

Rolling bearing data are collected from a simulated industrial platform, which is an ideal experimental platform in some sense. While data collected in actual industrial platform are usually polluted by strong noise. To test the diagnosis ability of our method in the scenario of actual

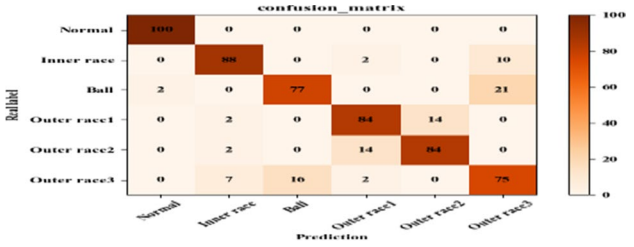

Fig. 11 Confusion matrix of SD-CNN-based fault diagnosis

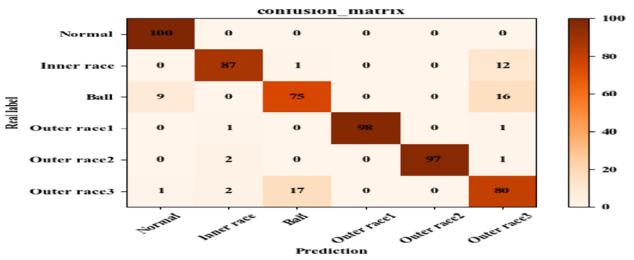

Fig. 12 Confusion matrix of LSTM-based fault diagnosis

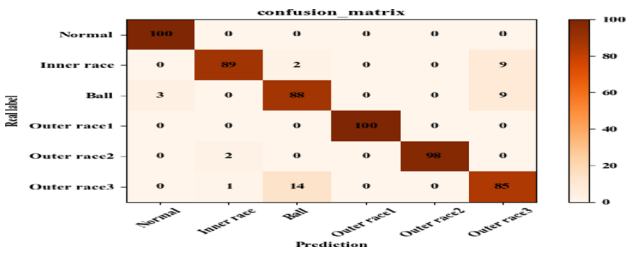

Fig. 13 Confusion matrix of SD-LSTM-based fault diagnosis

engineering field, a kind of normally distributed noise with variance 0.05 and 0.1 is added to the experiment scenarios, respectively. The experiment result is listed in Table 5. From Table 5, it can be seen that fault diagnosis of all corresponding methods decreases. Table 5 also indicates that once data polluted with strong noise are processed, it can achieve much high increment, which shows that the sparse gate can suppress noise better and SD-DNN based fault diagnosis model has good denoising performance.

4.3 Analysis of computational complexity

The computation complexity can be tested by the training time of each method. Table 6 shows the training time of SAE, SDAE, SSAE, SSDAE and SD-DNN in different experiment scenarios with different sample sizes since many neurons in SD-DNN are inhibited. In Table 6, the fault diagnosis capabilities of different structures of neural networks for different samples are compared for the fault diagnosis capability. LSTM showed better results of fault diagnosis.

SD_DNN has higher fault diagnosis accuracy than CNN when the samples are 1800 and 600.

It can be seen from Table 6 that SD-DNN can save more computational complexity. While SSDAE spends more computational complexity than SD-DNN since denoising and sparsity are implemented separately.

Remark 8: All training time listed in Table is just for the scenery when the training epochs come to 5000, which is the maximum number of epochs.

Combining the experiment result of Tables 3, 4 and 6, it is obvious to see that SD-DNN-based fault diagnosis method proposed in this paper can achieve much more accurate fault diagnosis result with much lower computational burden, especially in the case, when small size of training sample polluted by strong noise, which is common in the engineering field of fault diagnosis.

5 Conclusions and future work

Deep learning is a promising tool for fault diagnosis of rolling bearing. But existed structure of DNN may make that information correlated less with the fault feature transfer through layers. This will be destined to get inaccurate fault diagnosis with large computation burden. This paper focuses on developing a deep learning fault diagnosis algorithm by designing a sparse gate to make it possible for achieve the goal of sparsity and denoising simultaneously. SD-DNN is capable of achieving an accurate fault diagnosis result with less computational complexity.

Future research of our work will focus on designing a mechanism to combine limited size of training data with the available rough physical model to achieve more satisfying diagnosis accuracy.

Acknowledgements This research was supported in part by the Natural Science Fund of China (Grant No.62073213, U1604158, U1804163, 61751304, 61673160).

References:

1. Zhang, D., Chen, Y., Guo, F., Karimi, H.R., Dong, H., Xuan, Q.: A New Interpretable Learning Method for Fault Diagnosis of Rolling Bearings. IEEE Trans.Instrum. Meas. **70**, 1–10 (2021)
2. Hoanga, D., Kang, H.: Survey on Deep Learning based bearing fault diagnosis. Neurocomput. **335**, 327–335 (2019)
3. Sun, M., Wang, H., Liu, P., Huang, S., Fan, P.: A sparse stacked denoising autoencoder with optimized transfer learning applied to the fault diagnosis of rolling bearings. Meas. **146**, 305–314 (2019)
4. Cerrada, M., Sánchez, R.-V., Li, C., et al.: A review on data-driven fault severity assessment in rolling bearings. Mechl. Syst. Signal Process. **99**, 169–196 (2018)
5. Li, Y., Xu, M., Wei, Y., Huang, W.: A new rolling bearing fault diagnosis method based on multiscale permutation entropy and improved support vector machine based binary tree. Meas. **77**, 80–94 (2016)
6. Dhamande, L.S., Chaudhari, M.B.: Bearing fault diagnosis based on statistical feature extraction in time and frequency domain and neural network. Int. J. Veh. Struct. Syst. **8**(4), 229–240 (2016)
7. Jia, F., Lei, Y., Lin, J., et al.: Deep neural networks: a promising tool for fault characteristic mining and intelligent diagnosis of rotating machinery with massive data. Mech. Syst. Signal Process. **72–73**, 303–315 (2016)
8. Duan, L., Xie, M., Wang, J., Bai, T.: Deep learning enabled intelligent fault diagnosis: Overview and applications. J. Intell. Fuzzy Syst. **35**, 5771–5784 (2018)
9. Janssens, O., Slavkovikj, V., Vervisch, B., Stockman, K., Loccufier, M., Verstockt, S., Walle, R.V., Hoecke, S.V.: Convolutional neural network based fault detection for rotating machinery. J. Sound Vib. **377**(1), 331–345 (2016)
10. Shao, H., Jiang, H., Wang, F., Wang, Y.: Rolling bearing fault diagnosis using adaptive deep belief network with dual-tree complex wavelet packet. ISA Trans. **69**, 187–201 (2017)
11. Cao L., Zhang J., Wang J. and Qian Z., (2019) Intelligent fault diagnosis of wind turbine gearbox based on Long short-term memory networks. Proceeding of 2019 IEEE 28th International Symposium on Industrial Electronics (ISIE), Vancouver, B.C., Canada, 890-895
12. Yu, J.: Evolutionary manifold regularized stacked denoising autoencoders for gearbox fault diagnosis. Knowl.-Based Syst. **178**, 111–122 (2019)
13. Wen, L., Li, X., Gao, L., Zhang, Y.: A New convolutional neural network-nased data-driven fault diagnosis method. IEEE Trans. Industr. Electron. **65**(7), 5990–5998 (2018)
14. Shao, H., Jiang, H., Zhang, X., et al.: Rolling bearing fault diagnosis using an optimization deep belief network. Meas. Sci. Technol. **26**(11), 115–123 (2015)
15. Yu, L., Qu, J., Gao, F., Tian, Y., Mucchi, E.: A Novel Hierarchical Algorithm for Bearing Fault Diagnosis Based on Stacked LSTM. Shock Vib. **2019**, 1–10 (2019)
16. Lei, Y., Karimi, H.R., Cen, L., Chen, X., Xie, Y.: Processes soft modeling based on stacked autoencoders and wavelet extreme learning machine for aluminum plant-wide application. Control Eng. Prac. **108**, 104706 (2021)
17. Zhao, M., Kang, M., Tang, B., Pecht, M.: Deep Residual Networks With Dynamically Weighted Wavelet Coeffificients for Fault Diagnosis of Planetary Gearboxes. IEEE Trans. Industr. Electron. **65**(5), 4290–4300 (2018)
18. Wang, J., Mo, Z., Zhang, H., Miao, Q.: A Deep Learning Method for Bearing Fault Diagnosis Based on Time-Frequency Image. IEEE Access **7**, 42373–42383 (2019)
19. Tang J., Lu W., An J. and Wan X., (2015) Fault diagnosis method study in roller bearing based on wavelet transform and stacked auto-encoder. Proceeding of 27th Chinese Control and Decision Conference, 4608–4613.
20. M.Heydarzadeh, S. H. Kia, M. Nourani, H. Henao and G. Capolino, (2016) Gear fault diagnosis using discrete wavelet transform and deep neural networks. Proceeding of 42nd Annual Conference of the IEEE Industrial Electronics Society, 1494–1500.
21. Zarei, J., Tajeddini, M.A., Karimi, H.R.: Vibration analysis for bearing fault detection and classification using an intelligent filter. Mechatron. **24**(2), 151–157 (2014)
22. Qian, W., Li, S., Wang, J., Wu, Q.: A novel supervised sparse feature extraction method and its application on rotating machine fault diagnosis. Neurocomput. **320**, 129–140 (2018)
23. Meng, Z., Zhan, X., Li, J., Pan, Z.: An enhancement denoising autoencoder for rolling bearing fault diagnosis. Meas. **130**, 448–454 (2018)
24. Lu, C., Wang, Z., Qin, W., Ma, J.: Fault diagnosis of rotary machinery components using a stacked denoising

autoencoder-based health state identification. Signal Process. **130**, 377–388 (2017)
25. Wang Y., Han M. and Liu W., (2019) Rolling Bearing Fault Diagnosis Method Based on Stacked Denoising Autoencoder and Convolutional Neural Network, Proceeding of 2019 International Conference on Quality, Reliability, Risk, Maintenance, and Safety Engineering, 833–838.
26. Zhang, J., Chen, Z., Du, X., Yu, M.: Application of stack marginalised sparse denoising auto-encoder in fault diagnosis of rolling bearing. J. Eng. **16**, 1772–1777 (2018)
27. Li, Y., Lei, Y., Wang, P., Jiang, M., Liu, Y.: Embedded stacked group sparse autoencoder ensemble with L1 regularization and manifold reduction. Appl. Soft Comput. J. **101**, 107003 (2021)
28. Sun, W., Shao, S., Zhao, R., Yan, R., Zhang, X., Chen, X.: A sparse auto-encoder-based deep neural network approach for induction motor faults classification. Meas. **89**, 171–178 (2016)
29. Sohaib, M., Kim, J.-M.: Reliable Fault Diagnosis of Rotary Machine Bearings Using a Stacked Sparse Autoencoder-Based Deep Neural Network. Shock Vib. **2018**, 1–11 (2018)
30. Qi, Y., Shen, C., Wang, D.: Stacked sparse autoencoder-based deep network for fault diagnosis of rotating machinery. IEEE Access **5**, 15066–15079 (2017)
31. Zhang, J., Chen, Z., Du, X., Xu, X., Yu, M.: Application of stack marginalised sparse denoising auto-encoder in fault diagnosis of rolling bearing. J. Eng. **2018**(16), 1772–1777 (2018)
32. Hinton, G.E., Salakhutdinov, R.R.: Reducing the dimensionality of data with neural networks. Sci. **313**(5786), 504–507 (2006)
33. Bengio Y, Lamblin P, Popovici D, et al., (2007) Greedy layer-wise training of deep networks.Advances in Neural Information Processing Systems 19, Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems:153–160.
34. Bearing data Centre, Case Western Reserve University, Available: http://csegroups.case.edu/bearingdatacenter/home