2020 IEEE 9th Data Driven Control and Learning Systems Conference
November 20-22, 2020, Liuzhou, China

A Deep Learning Model with Adaptive Learning Rate for Fault Diagnosis

Xiaodong Zhai[1], Fei Qiao[1]

1. School of Electronics and Information Engineering, Tongji University, Shanghai 201804
E-mail: 1710332@tongji.edu.cn, fqiao@tongji.edu.cn

Abstract: With the increasing amount of data in the field of equipment fault diagnosis, deep learning is playing an increasingly important role in the process of fault diagnosis, during which the timeliness requirement is high and the fault diagnosis results need to be obtained accurately and timely. However, with the increase of network layers, the training time of deep learning model becomes longer. Learning rate in the deep learning model plays an important role in the process of model training, and a well-designed learning rate adjustment strategy can effectively reduce the training time and satisfy the requirements of fault diagnosis. At present, some deep learning models usually adopt a globally uniform learning rate strategy, which is unreasonable for different parameters. This paper has designed an adaptive learning rate strategy for the parameters of weight and bias respectively in deep learning model. Specifically, the strategy contains a learning rate strategy based on stochastic gradient descent method for weight, and a power exponential learning rate strategy for bias. Experiments are carried out to validate the effectiveness of proposed learning rate strategy. Results suggest that the strategy can reduce the training time and reconstruction error rate of deep learning model, and improve the classification accuracy of fault diagnosis.

Key Words: Deep learning, Learning rate, Adaptive, Fault diagnosis

1 Introduction

With the development of modern industrial technology, the safety, stability, reliability and operation efficiency of equipment have become the core competitiveness of manufacturing enterprises [1], and equipment management has become an important field in enterprise management. In the process of production, the performance of equipment deteriorates with the increase of service time, and various faults will occur in the process of equipment operation. When the equipment fails, the production efficiency will be reduced. More seriously, the equipment will be shut down, and malignant accidents such as machine damage and human death will occur. Therefore, it is particularly important to find and identify the types and locations of faults in time. With the development of computer technology, many artificial intelligence algorithms have been applied in the field of equipment fault diagnosis. It is predicted that the growing Internet of Things will connect 30 billion devices by 2020 [2], and the huge amount of data will also promote the innovation of the monitoring process of the physical network system of industrial 4.0. With the increasing amount of data, the advantages of deep learning using in dealing with large-scale data are highlighted.

The motivation of deep learning is to build and simulate the neural network of human brain for analysis and learning. It imitates the mechanism of human brain to interpret data, such as images, sounds and texts [3-5]. Deep learning is a multi-layer neural network model essentially. By combining low-level features, we can get a higher-level and more abstract feature representation to discover the distributed feature representation of data. At the same time, it weakens the adverse effects of unrelated factors and improves the accuracy of classification and prediction [6]. Meanwhile, the excellent performance of deep learning is mainly based on a large number of training data and deep-level network structure, as a result, the training time of deep learning model is longer than other machine learning algorithms generally [7]. Therefore, how to speed up the training time of deep learning model is a problem which is worth of intensive study, especially when it is applied in engineering practice.

2 Related Work

Traditional fault diagnosis methods include model driven methods, knowledge driven methods, and data driven methods. However, the first two methods are often limited by professional technology, expert experience and other knowledge. In addition, with the continuous development of equipment status monitoring technology, more and more equipment status data can be utilized. As a result, data driven methods based on machine learning and artificial intelligence have attracted people's attention in recent years [8-9]. Data driven methods can discover the intrinsic law of equipment status trends and estimate the fault types of equipment by advanced methods based on equipment status data. With the increasing amount of equipment status data, more and more attention has been paid to the deep learning method in machine learning.

There are two significant parameters in deep learning model, which are weight and bias. However, traditional deep learning models often use a global uniform constant parameter for these two parameters, and the setting of this constant parameter requires previous experience. Meanwhile, it should be noted that there are a large number of weight and bias parameters in deep learning model, and they are two different types of parameters. Different parameters play different roles. With this in mind, it is unreasonable to provide the same learning rate strategy for different parameters. A global uniform learning rate is not necessarily suitable for all parameters, and it will reduce the iteration efficiency and increase the model training time of deep learning model.

At present, there have been some studies on the adjustment strategies of the learning rate in the deep

*This work is supported by National Natural Science Foundation, China(No. 71690234, 61873191), National Science and Technology Major Project (2017-V-0011-0063) and the National Key R&D Program, China (No. 2017YFE0101400).

978-1-7281-5922-5/20/$31.00 ©2020 IEEE DDCLS'20

learning model. A downward trend learning rate strategy can significantly improve the convergence speed and reduce the training time of the model [10], and the learning rate adjusts according to the characteristics of the function. In many cases, a downward trend learning rate strategy is still a simple and effective learning rate strategy relatively. In 2010, Duchi *et al.* proposed an adaptive all-parameter learning rate strategy AadGrad [11]. This method designs a learning rate for each parameter in the process of deep learning model training, and uses the sum of gradients to ensure the downward trend of learning rate. This method is the first to propose an all-parameter learning rate strategy, which is an effective way to accelerate the convergence of deep learning model. In 2013, Senior *et al.* proposed an improved learning rate strategy AadDec [12], which is based on the learning rate strategy of AadGrad. Each learning rate is simplified from the sum of squares of all previous round gradients to the sum of squares of the current gradient and the last round gradient in this strategy. The convergence speed of this method is further improved compared with that of AadGrad, and this strategy has achieved good results in practical application.

The above literatures provide some feasible methods to increase the iteration efficiency and reduce the model training time of deep learning model. However, these methods do not distinguish the weight and bias in the deep learning model, and use a uniform adaptive learning rate strategy, which will have some limitations. Because these existing studies can not solve the above problems well, the original intention of the work in this paper is proposed, and specific improvement programs are put forward. This paper proposed an deep learning model with adaptive learning rate for fault diagnosis. In the process of model training, the learning rate is adaptively adjusted according to the current gradient value of the objective loss function in each iteration which is based on stochastic gradient descent (SGD), and independent learning rate is designed for the weight and bias separately. This method speeds up the iteration process of the model and weakens the dependence of the model on the initial value of the learning rate. In this paper, the method is applied to the gear fault diagnosis process, and comparison experiments are carried out to demonstrate the improvement of convergence efficiency and the classification accuracy of the proposed methodology.

3 Methodology

3.1 Deep learning Model

The essence of deep learning model is a kind of multi-layer neural network, and the general neural network only has several layers of network, but deep learning model contains a large number of hidden layers, so it has strong ability of feature learning. Through multi-layer non-linear conversion, we can learn the deep abstract features from complex training data and describe the intrinsic information of data. In order to avoid falling into the problem of local optimum, it usually adopts layer-by-layer training algorithm to realize parameter training of multi-layer neural network in deep learning.

At present, there are many mature deep learning models, and this paper focuses on Deep Belief Network (DBN). The basic unit of DBN is the Restricted Boltzmann Machine (RBM) [13]. DBN is formed by layers of unsupervised RBM which are trained and stacked. In the model proposed in this paper, the structure of Softmax regression, which is often used in the process of multi-classification, is added to the top layer. In order to achieve multi-classification, it maps the output of multiple neurons into the interval of (0, 1), which can be regarded as probability.

A basic model structure of DBN is shown in Fig. 1. Firstly, layer-by-layer training algorithm is adopted to complete pre-training, and through supervised reverse fine-tuning training, the whole deep neural network is trained to realize feature learning and classification.

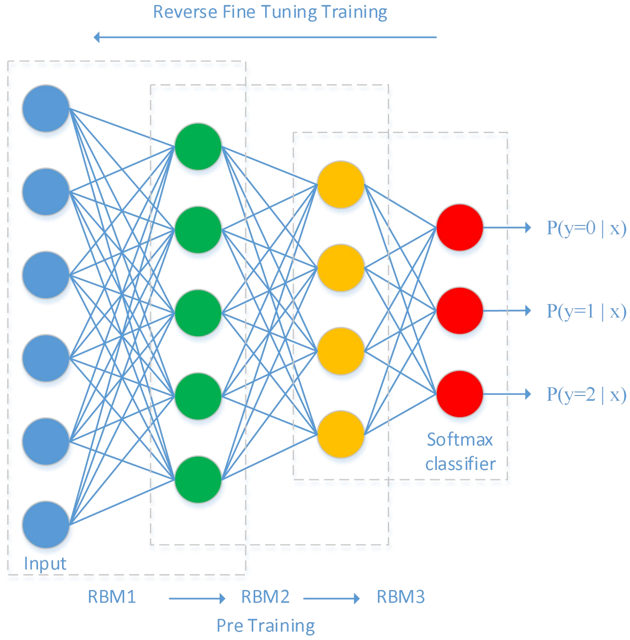

Fig. 1: Basic structure of DBN

3.2 Weight and Bias

The basic unit of the deep learning model is neuron, and its structure is shown in Fig. 2. Where v_i is the input neuron, x_i is the state of the input neuron, w_{ij} is the connection weight between input neuron and output neuron, b_j is the bias of the output neuron, F(.) is the activation function, and y_j is the state of the output neuron. The mathematical expressions are as follows:

$$y_j = f(u_j) \tag{1}$$

$$u_j = \sum_{i=1}^{I} \omega_{ij} x_i + b_j \tag{2}$$

In the model of deep learning, data are expressed by connecting weights, and they are distinguished by sharing weights and biases. Therefore, weights are important for feature extraction and layer-by-layer abstraction of deep learning model. According to formula (2), it can be seen that the bias b_j can be regarded as a neuron with the state of b_j and weight of 1. It can be seen as adding a dimension to the original data which is beneficial to data differentiation, especially when the dimension of input data is low. However, when the dimension of input data is high, which is enough to distinguish the data, the role of bias will be relatively weakened. Therefore, for the fault diagnosis model based on deep learning, when the dimension of input data is high, the calculation amount of bias could be reduced appropriately.

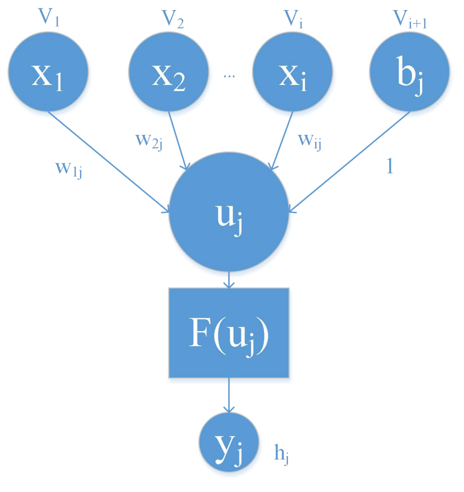

Fig. 2: The network graph of an artificial neuron

At present, the adjustment of weight and bias generally adopts a globally uniform way. However, for the parameter of weight, designing a learning rate for each weight parameter which could adjust increment adaptively according to its own state can accelerate the stable expression of input data and improve the convergence speed of the model. Meanwhile, in the process of dealing with high-dimensional data, although the role of bias could be weakened, it will also slow down the convergence rate of the model in the condition that learning rate adjustment strategy is improper. With this in mind, we should select a function model with less computation on the basis of ensuring the downward trend for the parameter of bias.

3.3 Definition of Learning Rate

For a RBM model with parameters $\theta=\{\omega_{ij}, b_{1i}, b_{2j}\}$, as shown in Fig. 3, the upper layer is the hidden unit layer, and the lower layer is the visible unit layer. The connection between visible unit and hidden unit is bi-directional, but the neurons in the same layer are not connected with each other. According to the probability theory, the hidden elements are mutually independent when the visible element states are given, and the visible elements are mutually independent when hidden element states are given. In the process of model training and calculation, the updating criteria of model parameters are as follows [14]:

$$\Delta\omega_{ij} = \alpha(E_{\text{data}}(v_i h_j) - E_{\text{model}}(v_i h_j)) \qquad (3)$$

$$\Delta b_{1i} = \beta(E_{\text{data}}(v_i v_i^{\mathrm{T}}) - E_{\text{model}}(h_i h_i^{\mathrm{T}})) \qquad (4)$$

$$\Delta b_{2j} = \gamma(E_{\text{data}}(v_j v_j^{\mathrm{T}}) - E_{\text{model}}(h_j h_j^{\mathrm{T}})) \qquad (5)$$

Where α is the learning rate of the weight between visible unit layer v_i and hidden unit layer h_j, $\Delta\omega_{ij}$ is the weight increment, β is the learning rate of the bias in visible unit layer, Δb_{1i} is the bias increment, γ is the learning rate of the bias in hidden unit layer, Δb_{2j} is the bias increment, E_{data} is the expectation obtained by input data label, E_{model} is the expectation obtained by model. $\eta=\{\alpha, \beta, \gamma\}$ is called the learning rate of the model. $\Delta\omega_{ij}$, Δb_{1i}, Δb_{2j} obtained by the above formulas are utilized to update the corresponding weights and biases in formula (2). With the constant updating of parameters, the training process of the model is completed until the iteration termination condition is reached.

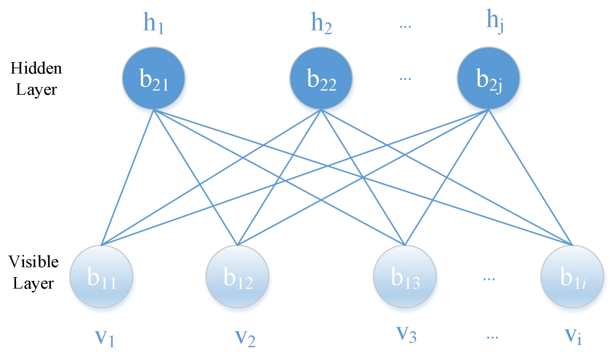

Fig. 3: The network graph of an RBM

3.4 Reverse Fine-tuning Training

The training of RBM is unsupervised, and according to the distribution of training data, the initial values of DBN model parameters can be obtained. Reverse fine-tuning training stage is a process of supervised learning, and it fine-tunes DBN layer parameters according to the known label from top to bottom. RBM is a typical energy model, and the loss cost function of the model can be obtained on the basis of the defined energy function. In the process of model training, the aim of reverse fine-tuning training is to minimize the value of loss cost function by adjusting model parameters. The gradient descent method is utilized widely [15-16] in order to obtain the appropriate model parameters and minimize the loss cost function. Its general mathematical expression is:

$$\theta(t+1) = \theta(t) - \eta(t)\nabla \mathrm{L}(\theta(t)) \qquad (6)$$

Where $\mathrm{L}(\theta)$ is defined as the loss cost function on the data set, $\nabla \mathrm{L}(\theta)$ is the gradient value of the loss cost function, $\theta(t+1)$ is the model parameter at the iteration time of $t+1$, $\theta(t)$ is the model parameter at the iteration time of t, i.e. the weight or bias parameter, $\eta(t)$ is the learning rate, i.e. the step size, and it is generally using a small positive number. Gradient descent method can solve most optimization problems quickly. However, because deep learning is usually based on large amount of training data, the calculation of $\nabla \mathrm{L}(\theta)$ will be huge, and even can not be calculated when using gradient descent method to optimize model parameters.

Therefore, stochastic gradient descent (SGD) [17] is adopted to optimize the parameters of deep learning model in this paper. Essentially, SGD is a deformation of gradient descent method. Unlike gradient descent method, SGD calculates the gradient of loss cost function by randomly selecting some samples from training data. Its mathematical expressions are as follows:

$$\begin{gathered}\theta(t+1) = \theta(t) - \eta(t)\nabla \mathrm{L}_m(\theta(t)), \\ m \in (1,2,3,\cdots,M)\end{gathered} \qquad (7)$$

$$\nabla L_m(\theta)=\sum_{n=1}^{N} l_n(\theta) \tag{8}$$

Where $\nabla L_m(\theta)$ is the gradient value of loss function calculated from the m-th batch data, N is the number of samples in the m-th batch data set. As we can see in these two formulas, the computational complexity of SGD is greatly reduced compared with gradient descent method.

3.5 Learning Rate Scheduling

For $\eta(t)$ in Formula(7), traditional DBN models usually set a global uniform constant parameter based on experience. However, with the number of iterations increasing, more precise step size of iteration is needed. Constant learning rate will slow down convergence speed of the model because it keeps the step size of each iteration unchanged during the iteration process. A good learning rate strategy can significantly improve the convergence speed and operation efficiency of deep learning model. In terms of mechanism, the full-parameter learning rate will further reduce the training time of the model and the final classification error of the model. Based on AdaGrad and AdaDec, and combined with SGD method, the mathematical expression of learning rate strategy is designed according to the different characteristics and functions of weight and bias, which are formulated as follows:

$$\alpha_{ij}(t)=\frac{\alpha_{ij}(t-1)}{\sqrt{K+g(t)^2}} \tag{9}$$

$$\beta_i(t)=\beta_i(0)(1+\frac{t}{R})^{-q} \tag{10}$$

$$\gamma_j(t)=\gamma_j(0)(1+\frac{t}{R})^{-q} \tag{11}$$

Where $\alpha_{ij}(t)$ is the learning rate of weight in next round, $\alpha_{ij}(t-1)$ is the learning rate in current round, $g(t)^2$ is the sum of squares of the gradients of the loss function in the current round, K is a constant term, and it equals one in general which mainly ensures that the learning rate is bounded and in a downward trend. $\beta_i(t)$ and $\lambda_j(t)$ are the learning rates of bias terms for visible and hidden units respectively, which use power exponential functions with a downward trend where R is the number of iterations, and the value of q is 0.75 generally.

According to the above formulas, the main idea of the learning rate adjustment strategy is that a larger learning rate can make the value of the target loss function decrease rapidly in the initial stage of iteration process. In the process of iteration, the learning rate decreases gradually, which can accelerate the stable expression of data samples and help the model to find the convergence point of data samples more quickly and steadily. In this paper, the learning rate of weight is adaptively adjusted by using the current gradient value based on the learning rate of previous round, which is adaptive. As a result, the learning rate can describe the current running state of model more accurately, and reduce the amount of computation of historical gradient data compared with other adaptive methods [11]. For the model dealing with fault diagnosis, sometimes the original data is high-dimensional, which can weaken the role of bias relatively. Therefore, a simple power exponential function is chosen as the learning rate strategy for bias, which simply ensures that the learning rate is in a downward trend, so as to further reduce the amount of calculation and improve the final classification accuracy.

3.6 Evaluating Indicator

In this paper, Reconstruction Error Rate (RER) of test data in the reverse fine-tuning stage is used as the quantitative evaluation index, which can well describe the convergence state of the model. For a test data set with N samples, the mathematical expressions of the reconstruction error rate are as follows:

$$RER=\frac{1}{N}\sum_{n=1}^{N} MSE(data(n)) \tag{12}$$

$$MSE(data)=\frac{1}{D}\sum_{d=1}^{D}(In(data(d)-Out(data(d)))^{2} \tag{13}$$

Formula(13) is the calculation formula of Mean Squared Error(MSE), where $In(data)$ is the input data of the model, $Out(data)$ is the data generated by the model, and D is the number of samples. Under the condition of same number of iterations, when the reconstruction error rate is high, the convergence of the model is bad. On the contrary, when the reconstruction error rate is low, the convergence of the model is good.

4 Case Study

In order to verify the performance of the learning rate strategy proposed in this paper, a constant learning rate(Cons) is introduced. Experiment in Section 4.1 was designed to compare the convergence and computational complexity of Cons, AdaGrad, AdaDec and the learning rate strategy proposed in this paper. On this basis, the classification accuracy of each method is compared and analyzed in Section 4.2. Finally, in order to verify the reliability of setting the learning rates of weight and bias respectively, experiment in Section 4.3 was designed to elaborate the relationship between weight and bias.

In this paper, the data set of rolling bearings are used [18], including bearings in good condition, bearings with peeling off in outer ring, bearings with peeling off in inner ring, bearings with peeling off in ball and bearings with broken cage. The neural network model used in the experiment has a five-layer structure. The number of neurons in the input layer is 1000, and the numbers of neurons in the three hidden layers are 1000, 500, 250 respectively. The number of neurons in the output layer is 5. The initial connection weights between layers obey the Gauss distribution with the mean value of 0 and the variance of 0.001. The initial bias of the first layer is determined by the training data, and the initial biases of the other layers are set to 0. All the methods mentioned in the experiment adopt the same initial value of learning rate, the initial value of learning rate in the pretreatment stage is 0.1, and the initial value of learning rate in the reverse fine-tuning stage is 0.001. Reconstruction error rate of the model is calculated by formula (12) and formula (13).

4.1 Reconstruction Error Rate Comparison

In this section, the influence of four methods on the convergence of the deep learning model is compared. Meanwhile, the running time of the model with 100 iterations is counted. The experimental results are shown in Fig. 4.

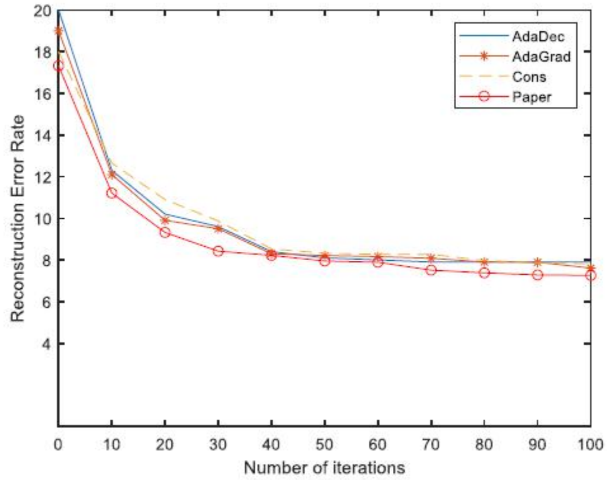

Fig. 4: Comparison of the convergence performance of four methods

According to Fig. 4, we can see that the Cons, AdaGrad, AdaDec and the learning rate strategy proposed in this paper make the reconstruction error rate of the model decrease with the increase of iterations and stabilize eventually. In the whole process of iteration, the reconstruction error rate curves of four strategies are close, but the reconstruction error curves of the method proposed in this paper are obviously lower than those of other three strategies. When the number of iteration reaches 100, the reconstruction error rate of constant learning rate is 7.81, that of the AdaGrad is 7.61 and that of the AdaDec is 7.90. The reconstruction error rate of the model proposed in this paper is 7.26, which shows that the convergence of model with proposed learning rate strategy is better.

At the same time, the training time of four strategies was counted during the experiment, among which the time of constant learning rate was the shortest, followed by the learning rate strategy proposed in this paper, AdaGrad and AdaDec. Although the training time of the learning rate strategy proposed in this paper is longer than that of constant learning rate, the disparity is not significant. However, if we want to achieve the same convergence effect, the constant learning rate needs more iterations and more training time. Considering the reconstruction error rate and training time, the learning rate strategy proposed in this paper is obviously better than other three strategies.

4.2 Classification Accuracy Comparison

The classification accuracies of the four strategies vary with the process of iteration are showed in Fig. 5. From the figure, we can see that with the increasing of iterations, the classification accuracies of four learning rate strategies increase as well. When the number of iterations reaches 100, the classification accuracy of the learning rate strategy proposed in this paper can reach 99.2%, while the classification accuracies of the other three learning rates are 98.7%, 98.1% and 98.3%, respectively. In addition, the curve of the learning rate strategy proposed in this paper is higher than those of other three strategies in the whole iteration process, that is, its comprehensive performance is better.

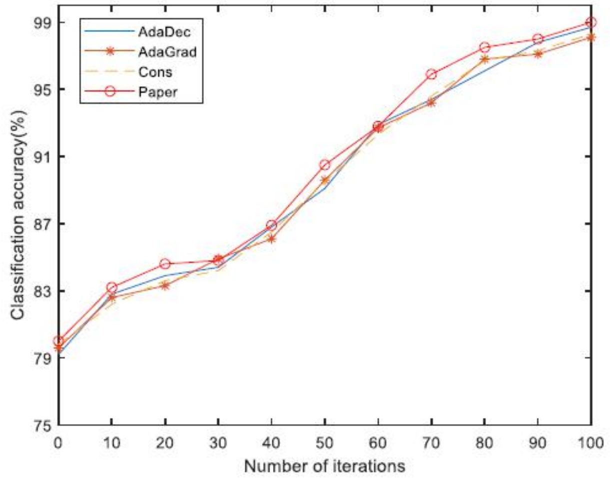

Fig. 5: Classification accuracy comparison

4.3 Function Contrast of Weights and Bias Function

In this paper, we set different learning rate strategies for weight and bias, and the learning rate strategy for bias is only set as the power exponents. In order to verify the effectiveness of this strategy, three learning rate strategies are compared in this section. They are learning rate strategy in which weight and bias are constant (Cons+Cons), learning rate strategy in which weight is constant, bias is 0 (Cons+Zero), and learning rate strategy in which weight is 0, bias is constant(Zero+Cons). The main purpose is to compare and analyze the influence of weight and bias on the convergence of deep learning model. The experimental results are shown in Fig. 6.

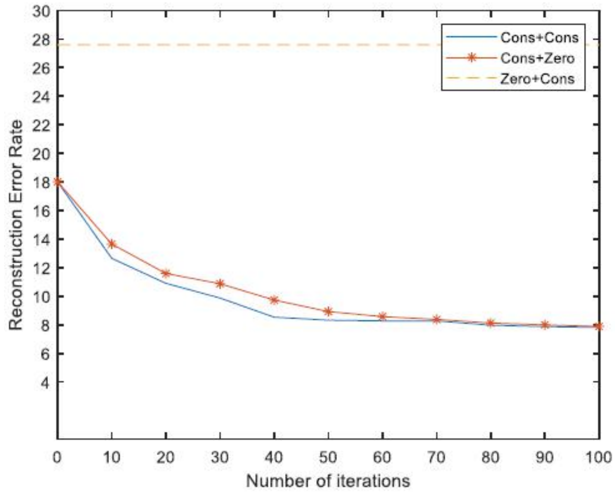

Fig. 6: The influence of weight and bias on the convergence of deep learning model

According to Fig. 6, it can be seen that the two learning rate strategies, the weight and bias are constant and the weight is constant with bias being zero, make the reconstruction error rate of model decrease gradually in the process of iterations. When the number of iterations reaches

100, the reconstruction errors of the two models are 7.81 and 7.88 respectively. By increasing the number of iterations, there is little difference between the two results. However, the learning rate strategy in which weight is zero and bias is constant does not reduce the reconstruction error rate in the process of iterations, which keeps a high reconstruction error rate. Therefore, we can get a conclusion that the weight plays a decisive role in the process of model convergence, and the bias term does not play an important role in the same process.

5 Conclusions

In this paper, a deep learning model with adaptive learning rate for fault diagnosis is proposed. According to the different roles of weight and bias in the deep learning model, SGD method is used to design a suitable learning rate strategy for the parameter of weight, and a power exponential function is chosen as the learning rate strategy for the parameter of bias. Experiments show that the strategy proposed in this paper can extract the characteristics of data samples better, reduce the reconstruction error rate of data samples, improve the training efficiency and classification accuracy of the model and has better performance than some existing learning rate strategies.

However, there are still some areas for further improvement for the work proposed in this paper. For example, when the dimensional of data set is low, the learning rate strategy proposed in this paper needs to be adjusted appropriately, and further study is needed when the deep learning model with adaptive learning rate applied to more practical problems.

References

[1] E. Pan, W. Z. Liao, and M. l. Zhuo, Periodic Preventive Maintenance Policy with Infinite Time and Limit of Reliability Based on Health Index, *Journal of Shanghai Jiaotong University*, 15(2):231-235, 2010.

[2] C. MacGillivary, V. Turner, and D. Lund, Worldwide internet of things (IOT) 2013-2020 forecast: Billions of things, trillions of dollars, IDC Q1 Doc, 243661(3):1-22, Oct. 2013.

[3] L. Li, and S. S. Yu, Image quality assessment based on deep learning model, *Journal of Huazhong University of Science and technology (Nature Science edition)*, 44(12): 70-75, 2016.

[4] S. K. Kim, Y. J. Park, and S. Lee, Voice activity detection based on deep belief networks using likelihood ratio, *Journal of Central South University*, 23: 145-149, 2016.

[5] K. Liu, and W. Y. Yuan, Short Texts Feature Extraction and Clustering Based on Auto-Encoder, *Acta Scientiarum Naturalium Universitatis Pekinensis*, 51 (2):282-288, 2015.

[6] Y. Lecun, Y. Bengio, and G. E. Hinton, Deep learning, *Nature*, 521: 436-444, 2015.

[7] K. Liu, L. M. Zhang, X. L. Fan, New image deep feature extraction based on improved CRBM, *Journal of Harbin Institute of Technology*, 48(5):155-159, 2016.

[8] Y. Peng, and M. Dong, A prognosis method using age-dependent hidden semi-Markov model for equipment health prediction, *Mechanical Systems & Signal Processing*, 25(1): 237-252, 2011.

[9] T. P. Hong, and B. S. Yang, Estimation and forecasting of machine health condition using ARMA/GARCH model, *Mechanical Systems & Signal Processing*, 24(2):546-558, 2010.

[10] G. E. Hinton, and R. R. Salakhutdinov, Reducing the dimensional-ity of data with neural networks, *Science*, 313(5786): 504-507, 2006.

[11] J. Duchi, E. Hazan, and Y. Singer, Adaptive subgradient methods for online learning and stochastic optimization, *The Journal of Machine Learning Research*, 12: 2121-2159, 2011.

[12] A. Senior, G. Heigold, M. A. Ranzato, and K. Yang, An empirical study of learning rates in deep neural networks for speech recognition, in *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2013: 6724-6728.

[13] J. W. Liu, Y. Liu, and X. L. Luo, Research and Development on Boltzmann Machine, *Journal of Computer Research and Development*, 51(1): 1-16, 2014.

[14] R. Salakhutdinov, and G. Hinton, An efficient learning procedure for deep Boltzmann machines, *Neural Computation*, 24(8): 1967-2006, 2012.

[15] H. Robbins, and S. Monro, A stochastic approximation method, *The Annals of Mathematical Statistics*, 22(3): 400-407,1951.

[16] Z. You, X. R. Wang, and B. Xu, Exploring one pass learning for deep neural network training with averaged stochastic gradient descent, in *Proceedings of the 2014 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2014: 6854-6858.

[17] S. Klein, J. P. W. Pluim, M. Staring, and M. A. Viergever, Adaptive stochastic gradient descent optimisation for image registration, *International Journal of Computer Vision*, 81(3): 227-239, 2009.

[18] https://download.csdn.net/download/qq_34133884/11017362