Proceedings of the 35th Chinese Control Conference
July 27-29, 2016, Chengdu, China

Deep Learning Neural Network for Power System Fault Diagnosis

WANG Yixing[1,2], LIU Meiqin[1,2] BAO Zhejing[2]
1. State Key Laboratory of Industrial Control Technology, Zhejiang University, Hangzhou 310027, P. R. China
2. College of Electrical Engineering, Zhejiang University, Hangzhou 310027, P. R. China
E-mail: {wangyixing,liumeiqin,zjbao}@zju.edu.cn

Abstract: This paper deals with application of deep learning neural network for power system fault diagnosis. Deep learning is a more effective approach than traditional neural network to solve problems including availability of data, better local optimum, and diffusion of gradients. In the paper, data is extracted from power system dispatching department and preprocessed before training in the deep learning network. Then, processed data is put into auto-encoders and the hidden features are observed in different dimensions so that we can preliminarily judge about the fault. Afterwards, trained stacked auto-encoders (SAE) is used to initialize and train a deep learning neural network (DLNN). The hidden features are observed in different dimensions so that the fault is preliminarily judged. The classifier is the last part of the network to reflect the types and possibility of diagnosis. The method of data availability, preprocess, and modeling is proposed in the paper. The result of simulation proves the feasibility of the approach and the influence factors are shown in the paper.

Key Words: Deep Learning, Stacked Auto-Encoders, Power System, Fault Diagnosis, Neural Network

1 Introduction

The safety and high quality operation of power system are faced with challenges as the load increases and the scale of grid expands. The fault of power system can cause large-scale power outage, which leads to a huge loss. Therefore, power system fault diagnosis is increasingly important and attracts high attention. Power system fault diagnosis is the process of analysing historical fault data to detect and predict current or future fault. It aims to provide assistance for staffs of power system to prevent fault or make a reasonable decision after fault, thereby reducing the negative effects. Research scholars have put forward some feasible methods in the field of power system fault diagnosis, which include diagnosis method based on expert system, Bayesian networks, Petri net, neural network, etc. Expert system is mainly used in power system equipment diagnosis. The automatic modeling and knowledge update ability of Bayesian networks is still immature. The modeling of Petri net is complex and difficult. Neural network easily runs into local optimum and turn up gradient diffusion problems. However, deep learning[1, 2], improvement of Neural network, has great advantages in fault diagnosis. The concept of deep learning is the result of artificial neural network research.

Deep learning is a machine learning method proposed by Hinton[3, 4], a prominent researcher in the field of machine learning and artificial intelligence in 2006, which imitates the mechanism of human brain to analyze data, such as images, sounds, and so on. In recent years, LeCun[5] and Bengio[6–9] also made a tremendous contribution to the development of deep learning. Deep learning has been applied to signal recognition, image identification, and target recognition[10–13], but it has just started in the field of power system fault diagnosis.

The training mechanism of deep learning is layer-wise pre-training and the parameters of each layer update using the gradient descent method. The mechanism bring strong ability of extracting features from samples and transforming features, which is advantageous to the visualization and classification of features. Deep learning also has great learning ability so it is a focus in the field of machine learning and artificial intelligence.

Unsupervised learning is the machine learning task of inferring a function to describe hidden structure from unlabeled data. There is a huge mass of unlabeled data which traditional artificial neural network has difficulty processing during operation period of power system. Therefore, sample acquirement is an important problem of artificial neural network. However, deep learning uses the unsupervised machine learning method in training, and makes use of a large amount of data to reflect hidden features. It is an effective method of unsupervised learning. Training a shallow network using traditional artificial neural network usually results in the parameters converging to reasonable values, but it is easy to run into local optimum when training a deep network, so that it fails to build a proper network for fault diagnosis. Moreover, since the training mechanism of traditional artificial neural network is updating each layer of the whole network from an iteration and each layers weight is restricted to the output layers error function, the traditional artificial neural network has gradient diffusion problems. Specifically, the gradients reduce sharply as the depth of the network increases, which are propagated backwards from the output layer to the earlier layers of the network. Then, the overall cost in relation to the weights in the earlier layers is very small. Therefore, the weights of the earlier layers update so slowly when using gradient descent that the earlier layers fail to learn effectively.

To solve the problems given above, deep learning is an efficient method because of its training mechanism that parameters iterate and update in each layer. Specifically, the weights of encoders are only restricted to the input of current layer and after training, the better parameters are propagated

This work is supported by the Zhejiang Provincial Natural Science Foundation of China under Grants LZ15F030001 and LZ14F030002, and the Specialized Research Fund for the Doctoral Program of Higher Education of China (SRFDP) under Grant 20120101110115.

to the next layer.

Researches about power system fault diagnosis based on traditional artificial neural network have been proposed in recent years. In order to solve the defect that artificial neural network is easy to fall into local minimum and sensitive to the initial values of weights and thresholds, researchers combine genetic algorithm, wavelet theory with neural network[14, 15]. They proposed a method that could reduce the uncertainty in the complex system fault diagnosis and improve the accuracy and efficiency of fault diagnosis. Wang[16] designed the learning, memory, and training of artificial neural network and design the online fault detection and diagnosis system and simulation environment based on neural network. However, they did not pay attention to the traditional problems of neural network mentioned above and fail to solve them in fundamentally.

An effective method is proposed in the paper to solve three key problems including availability of data, better local optimum, and diffusion of gradients. Data is extracted from power system dispatching department and preprocessed before training in the deep learning network. Then, processed data is put into auto-encoders and train a deep learning neural network(DLNN). The classifier is the last part of the network to reflect the types and possibility of diagnosis. The result of simulation proves the feasibility and accuracy of the method.

2 Preliminaries of Deep Learning

There are three main methods of deep learning including stacked auto-encoders(SAE), deep belief networks(DBN), and convolutional neural network(CNN). The paper mainly aims at research of SAE for power system fault diagnosis. The basic knowledge about the research is introduced in this section.

2.1 Fundamental of Auto-Encoder

An auto-encoder neural network is an unsupervised learning algorithm. It applies back-propagation, setting the output values to be equal to the input values. The model of auto-encoder was introduced in unsupervised feature learning and deep learning(UFLDL), course notes by Andrew[17].

An auto-encoder neural network can be considered a three-layer neural network. It is assumed that a set of unlabeled training samples $\{x^1, x^2, x^3, \ldots\}$, where $x^i \in R^n$. A typical example is shown in Fig. 1, where x_i' represents the output. The circles with +1 are called bias units, and correspond to the intercept term. Generally, the transformation process from the input layer to the hidden layer is called encoding, and the transformation process from the hidden layer to the output layer is called decoding. Therefore, the encoding and decoding process is presented by

$$\boldsymbol{d} = S(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1) \tag{1}$$

$$\boldsymbol{y} = S(\boldsymbol{W}_2\boldsymbol{x} + \boldsymbol{b}_2) \tag{2}$$

where S represents sigmoid function; $\boldsymbol{W}_1$ represents the weights matrix between the input layer and the hidden layer, $\boldsymbol{W}_2$ represents the weights matrix between the hidden layer and the output layer and $\boldsymbol{b}_i$ represents the bias.

According to the concept mentioned in the beginning of this section, the auto-encoder tries to learn a function $h_{\boldsymbol{w},\boldsymbol{b}}(\boldsymbol{x}) \approx \boldsymbol{x}$. In other words, it is training to learn an approximation to the identity function so that output is similar to input. By putting some constraints into auto-encoder, such as by the number of hidden units, meaningful structure about the data can be discovered. Actually, this auto-encoder can work out a low dimensional representation of data.

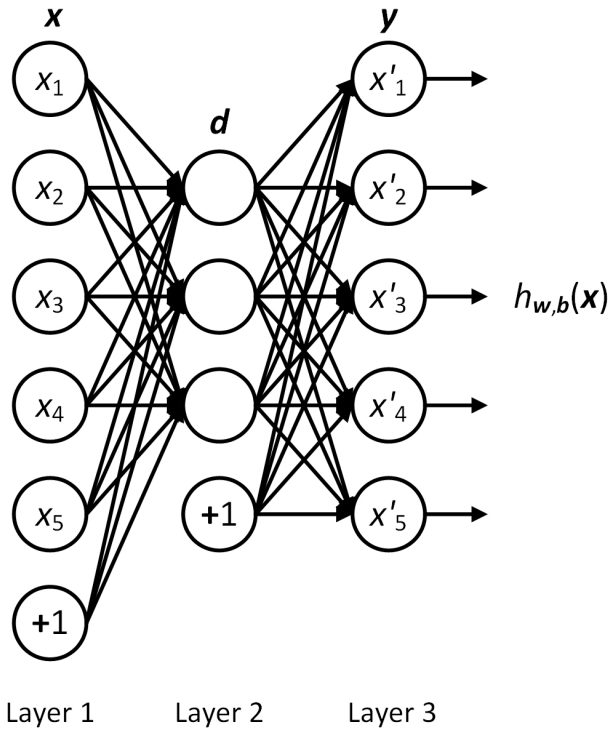

Fig. 1: An auto-encoder neural network

2.2 Back-propagation Algorithm

Back-propagation Algorithm is a common method of training artificial neural network used in conjunction with an optimization method such as gradient descent. It can also be applied to auto-encoders, training to make the output equal to the input.

Supposed that there is a fixed training set $\{(x^{(1)}, y^{(1)}), \ldots, (x^{(m)}, y^{(m)})\}$ of m training samples. The network can be train by using batch gradient descent. In particular, for one sample, the cost function is

$$J(W, b; x, y) = \frac{1}{2}\|h_{w,b}(x) - y\|^2 \tag{3}$$

For a training set of m samples, the cost function is

$$\begin{aligned} J(W, b) =& [\frac{1}{m}\sum_{i=1}^{m}(\frac{1}{2}\|h_{w,b}(x^{(i)}) - y^{(i)})\|^2] \\ &+ \frac{\lambda}{2}\sum_{l=1}^{n_l-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_l-1}(W_{ji}^{(l)})^2 \end{aligned} \tag{4}$$

where λ represents the weight decay parameter, which controls the relative importance of the two terms in formula (4), n_l represents the number of layers in the network, s_l represents the number of units of layer l. In formula (4), the first term is an average sum-of-squares error term and the second

term represents a regularization term, also called a weight decay term, which tends to decrease the magnitude of the weights and helps to prevent overfitting.

The weight W and bias b are updates with gradient descent by

$$W_{ij}^{\prime(l)} = W_{ij}^{(l)} - \alpha \frac{\partial}{\partial W_{ij}^{(l)}} J(W,b) \tag{5}$$

$$b_i^{\prime(l)} = b_i^{(l)} - \alpha \frac{\partial}{\partial b_i^{(l)}} J(W,b) \tag{6}$$

where α represents the learning rate. The backpropagation algorithm calculates the partial derivative by

$$\frac{\partial}{\partial W_{ij}^{(l)}} J(W,b) = [\frac{1}{m}\sum_{i=1}^{m} \frac{\partial}{\partial W_{ij}^{(l)}} J(W,b,x^{(i)},y^{(i)})] + \lambda W_{ij}^{(l)} \tag{7}$$

$$\frac{\partial}{\partial b_i^{(l)}} J(W,b) = \frac{1}{m}\sum_{i=1}^{m} \frac{\partial}{\partial b_i^{(l)}} J(W,b,x^{(i)},y^{(i)}) \tag{8}$$

where

$$\frac{\partial}{\partial W_{ij}^{(l)}} J(W,b;x,y) = a_j^{(l)} \delta_i^{(l+1)} \tag{9}$$

$$\frac{\partial}{\partial b_i^{(l)}} J(W,b;x,y) = \delta_i^{(l+1)} \tag{10}$$

where $a_j^{(l)}$ represents the activation of unit j of layer l and $\delta_i^{(l+1)}$ represents the error term of layer $l+1$, given by

$$\delta_i^{(l)} = (\sum_{j=1}^{s_l+1} W_{ji}^{(l)} \delta_i^{(l+1)}) f'(z_i^{(l)}) \tag{11}$$

and the error term of layer n_l is given by

$$\delta_i^{(n_l)} = -(y_i - a_i^{(n_l)}) f'(z_i^{(n_l)}) \tag{12}$$

where z represents the input weighted sum of a unit.

Repeating formula (3) to (12) can reduce the overall cost function to training the network including deep learning neural network.

2.3 Sparsity of Auto-Encoder

The number of hidden units is relatively smaller than that of input units, which is called the compact representation of data. When the number of hidden units is larger than the number of input units, meaningful features can also be discovered by imposing other constraints on the network. Specifically, imposing sparsity constraint on the network can help to find the useful features, which constrain the activation of hidden units to a small value.

Adding sparsity constraint makes the overall cost function to be

$$J_{\text{sparse}}(W,b) = J(W,b) + \beta \sum_{j=1}^{s} \text{KL}(\rho \| \hat{\rho}_j) \tag{13}$$

where

$$\text{KL}(\rho \| \hat{\rho}_j) = \rho \log \frac{\rho}{\hat{\rho}_j} + (1-\rho) \log \frac{1-\rho}{1-\hat{\rho}_j} \tag{14}$$

where $\sum_{j=1}^{s} \text{KL}(\rho \| \hat{\rho}_j)$ represents the sparsity penalty term, β controls the weight of the sparsity penalty term,$\hat{\rho}_j$ represents the average activation of hidden unit j,ρ represents a sparsity parameter, s represents the number of hidden units.

The sparsity penalty term has property that if $\hat{\rho}_j = \rho$,then $\text{KL}(\rho \| \hat{\rho}_j) = 0$, and it increases with the difference between $\hat{\rho}_j$ and ρ .Therefore, the activations of hidden units are small enough when ρ is set close to zero and the hidden units are considered to be inactive.

2.4 Stacked Auto-Encoders and Deep Learning Neural Network

Stacked auto-encoders are neural network consisting of multiple layers of sparse auto-encoders in which the outputs of each layer is wired to the inputs of the next layer. The schematic plot of an example is shown in Fig. 2. The features of data can be reflected in different dimension, so that the meaning structure of data is obtained to for further research. The algorithm of SAE is called Greedy layer-wise

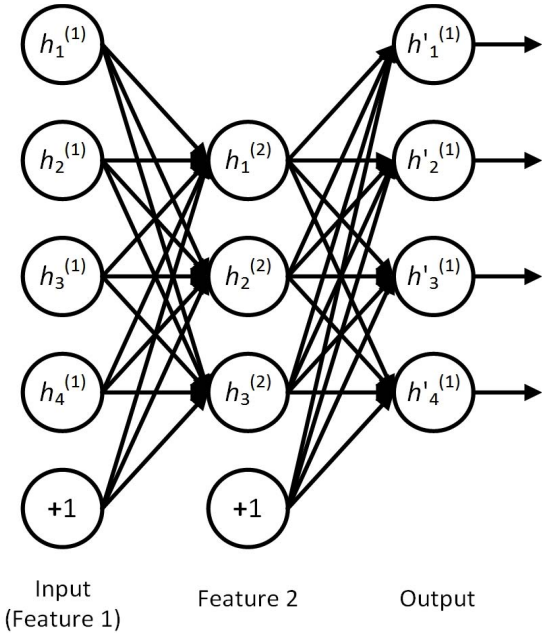

Fig. 2: Stacked Auto-Encoders

training algorithm. The advantages of the algorithm are that the better network can be built through using a huge number of unlabeled data and the weights of each layer after training would locate in better parameter space compared with their random initialization. Moreover, it does not share the problem of gradient diffusion.

In general, the trained SAE is connected with a classifier to complete the network, which includes Logistic classifier and Softmax classifier. Logistic classifier aims at the binary classification setting while Softmax classifier focuses on the multi-class classification setting. Therefore, the whole network combined SAE with classifier is complete deep learning neural network with SAE, shown in Fig. 3, where P represents the result of classifier.

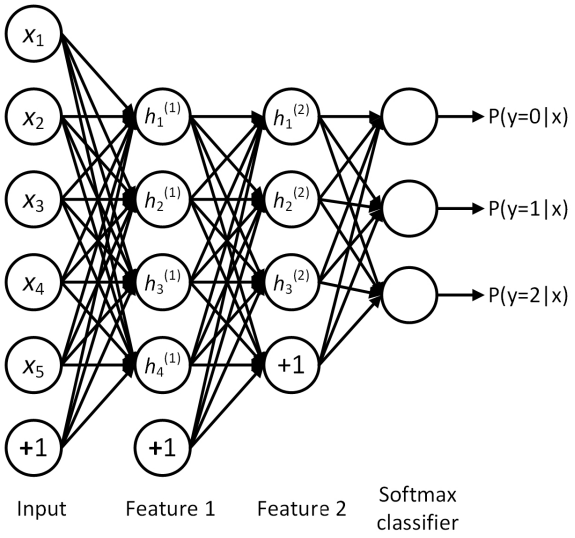

Fig. 3: Deep learning neural network with SAE

3 Application of DLNN in Power System Fault Diagnosis

It is discussed how to apply deep learning neural network to power system fault diagnosis in this section. The main idea is extracting useful data from a huge mass of unlabeled data and preprocessing with it so that the model can be built with the network mentioned before.

3.1 Availability of Data

Data of power system operation in the paper was obtained from administration of power supply in Dongguan. The data from SCADA includes data of measurement, equipment ledger, health of equipment, weather, and topology.

The paper focuses on the relation between fault and data of measurement. The measurement of data involves voltage, current, active power, and reactive power of every users. Data-collection interval is half an hour. The voltage of users reaches a stable level on 220V. The current of users, changing as the online load, can reflect the operating state of the node. Therefore, current change before fault happening must be meaningful and the hidden features can be discovered so that the fault can be diagnosed and forecasted.

3.2 Preprocessing

The value of current of every users is related to the load of users, so the magnitude is not in the same level. In order to reduce the influence of magnitude, the data should be preprocessed by

$$y_p = \frac{y_{\max} - y_{\min}}{x_{\max} - x_{\min}}(x - x_{\min}) + y_{\min} \qquad (15)$$

where y_p represents the preprocessed data, $y_{\max}$ and $y_{\min}$ represent the maximum and the minimum of the preprocessed data, $x_{\max}$ and $x_{\min}$ represent the maximum and the minimum of the unprocessed data. In the simulation experiment of this paper, $y_{\max}$ and $y_{\min}$ are given by 1 and 0 respectively so that the processed data can be constrained in interval [0,1] and constantly be nonnegative number. Therefore, the inner change rule of the current data is retained and the magnitude is transformed to the same level for training.

3.3 Modeling and Parameter Adjustment

The acquired data of current is divided into three modes including data of reclosing failure fault, data of reclosing success fault, and data of normal operation. In the simulation experiment, 500 points of current per 0.5 hour are recorded for a sample before fault or during normal operation. Then, 500 points of current is put into network as a sample so that the input layer contains 500 units. As the data is divided into three modes, the number of output units is given by 3. The network is shown in Fig. 4, where F represents the mode of fault.

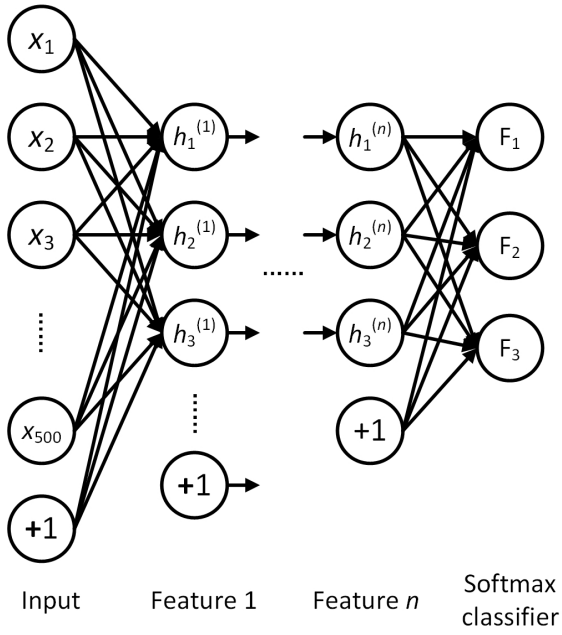

Fig. 4: Network of fault diagnosis in simulation

The number of hidden layers and hidden units of each layer should be adjusted according to size of samples, size of batch in training the network and so on. The ultimate purpose is making the performance better and result in accurate fault diagnosis.

4 Simulation Experiment

In this section, the result of simulation experiment is shown and the discussion of performance is indicated specifically.

4.1 Samples of Simulation Experiment

The data of simulation experiment was obtained from Wan Jiang substation in Dongguan where 500 points of current per 0.5 hour are recorded for a sample. The number of training samples and test samples among three modes are shown in Table 1. The performance of fault diagnosis is represented by accuracy rate of diagnosis, which is calculated by the ratio of proper diagnosis samples and all the test samples.

4.2 Discussion and Comparison of Results

After testing and adjusting experiment, the parameters was chosen as shown in Table 2. The features comparison of reclosing failure fault and normal operation in second dimension is shown in Fig. 5.

It is observed that the features of reclosing failure fault

Table 1: Samples of Simulation Experiment

Fault modes	Reclosing Failure Fault	Reclosing Success Fault	Normal Operation
Number of Training Samples	400	320	480
Number of Testing Samples	150	100	90

Table 2: Parameters in the Experiment

Parameters	Value
Input Units	500
Output Units	3
Size of Batch	10
Number of Epoch	30
Number of Hidden Layer	2
Number of Each Hidden Units	100, 100
Learning Rate	1

and normal operation are significantly different on positive plane. The difference of features is reflected among other samples in second dimension. The fault can be primarily diagnosed from the obvious difference. It proves that the meaningful structures can be discovered by SAE, but they may be reflected in other dimensions instead of the second dimension.

The relation between accuracy rate of diagnosis and number of epochs is shown in Fig. 6 where the other parameters are selected as Table 2. It is observed that the rate increases as the increase of the number of epochs. Moreover, the magnitude of the increase is large between epochs of 3 and 10 and tends to smooth when the number of epoch is more than 30. The result of simulation experiment is explicable because the deep learning neural network is more close to the ideal network with more epochs being trained according to the training samples. The relation between accuracy rate of diagnosis and size of batch is shown in Fig. 7 where the other parameters are selected as Table 2. It can be seen that the rate decreases as the increase of the size of batch. Batch in the method of batch gradient descent represents the number of input data in each training. It is also up to expectation because the error would be bigger if there are overmuch samples in each back-propagation of error.

The influence of hidden layer on accuracy rate of diagnosis is shown in Table 3. It can be concluded that number of epochs should be big enough to ensure the accuracy when the number units of hidden layer is lager. However, large number of units and epochs would cost more time to train, which influences the online fault diagnosis. Therefore, parameter selection should balance the training time and rate of accuracy. In general, it is can be observed that the accuracy rate of diagnosis is more than 80% if the parameters are chosen properly, which is relatively accurate for power system fault diagnosis. The simulation experiment is did in the same way based on back propagation neural network. However, it costs more time to train a back propagation neural network for these samples of power system because of diffusion of gradients. Moreover, it easily runs into better local

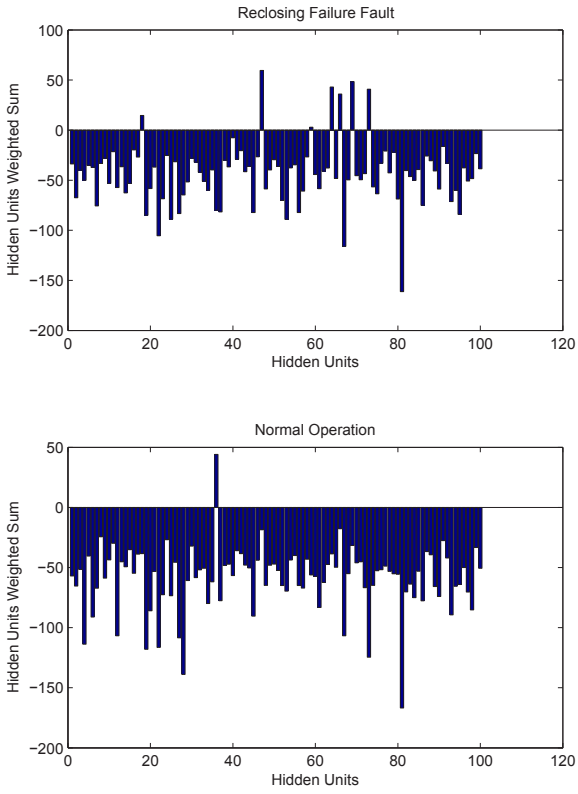

Fig. 5: Comparison of reclosing failure fault and normal operation

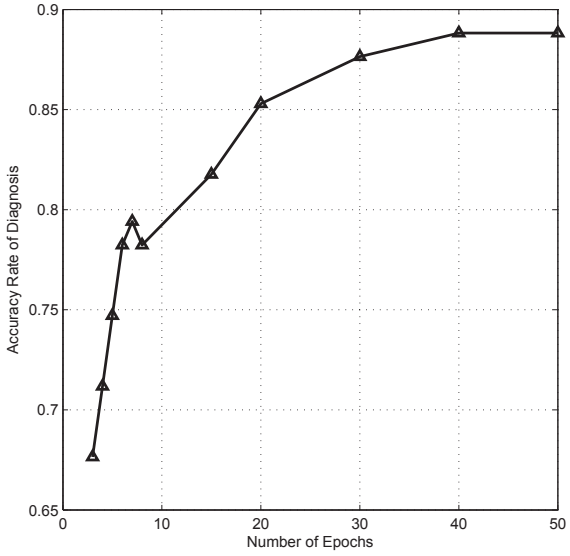

Fig. 6: Relation between accuracy rate of diagnosis and number of epochs

optimum and fails to build a useful network. The compared performance of successfully built network is shown in Table 4. It is concluded that the performance of deep learning neural network is generally better than that of back propagation neural network.

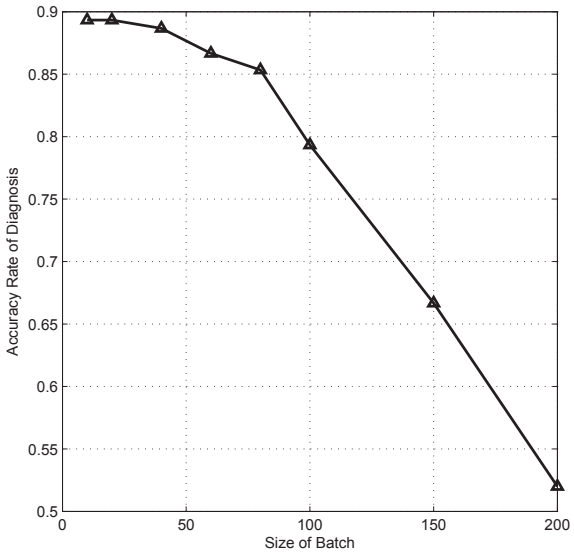

Fig. 7: Relation between accuracy rate of diagnosis and size of batch

Table 3: Samples of Simulation Experiment

Number of Hidden Layer	Number of Each Hidden Unit	Epochs	Accuracy Rate
3	500, 100, 100	30	88.8%
3	100, 100, 100	30	82.9%
3	100, 100, 100	10	70.0%
2	500, 100	10	82.9%
2	100, 100	10	80.6%
1	100	10	81.8%
1	700	10	85.3%

Table 4: Compared Performance

Average Accuracy Rate	Reclosing Failure Fault	Reclosing Success Fault	Normal Operation
Deep Learning Neural Network	65.3%	63.8%	71.3%
Back propagation Neural network	81.5%	78.8%	82.8%

5 Conclusion

This paper deals with application of deep learning neural network for power system fault diagnosis. An approach is proposed in the paper aiming to solve the three major problems of traditional neural network, including availability of data, better local optimum, and diffusion of gradients. First, data of power system operation in the paper is obtained from SCADA in administration of power supply. Then data should be extracted and preprocessed, and put into SAE to train the network. The hidden features are observed in different dimensions so that we can preliminarily judge about the fault. Afterwards, trained SAE is used to initialize and train a DLNN. The classifier of the network aims to reflect the types and possibility of diagnosis. The result proves the feasibility and accuracy of the method. There are three main methods of deep learning including SAE, DBN, and CNN. The application and comparison of DBN and CNN in power system fault diagnosis will be a significative future work.

References

[1] F. Q. Lauzon, An introduction to deep learning, in *11th International Conference on Information Science, Signal Processing and their Applications(ISSPA)*, 1438-1439, 2012.

[2] J. Schmidhuber, Deep learning in neural networks: An overview, in *Neural Networks*, Elsevier, 61, 85-117, 2015.

[3] G. E. Hinton, R. R. Salakhutdinov, Reducing the dimensionality of data with neural networks, in *Science*, sciencemag.org, 313(5786), 504-507, 2006.

[4] G. E. Hinton, S. Osindero, Y.W. Teh, A fast learning algorithm for deep belief nets, in *Neural computation*, Cambridge, USA: the MIT Press, 18, 1527-1554, 2006.

[5] Y. LeCun, Y. Bengio, G. E. Hinton, Deep learning, in *Nature*, Nature Publishing Group, 521(7553), 436-444, 2015.

[6] Y. Bengio, P. Lamblin, D. Popovici, etc, Greedy layer-wise training of deep networks, in *Advances in Neural Information Processing Systems*, Cambridge, USA: the MIT Press, 19, 153-160, 2007.

[7] Y. Bengio, Learning Deep architectures for AI, in *Foundations and trends in Machine Learning*, dl.acm.org, 2(1), 1-127, 2009.

[8] D. Erhan, Y. Bengio, A. Courville, etc, Why does unsupervised pre-training help deep learning?, in *The Journal of Machine Learning Research*, Cambridge, USA: the MIT Press, 11, 625-660, 2010.

[9] Y. Bengio, A. C. Courville, P. Vincent, Unsupervised feature learning and deep learning: A review and new perspectives, in *Department of Computer Science and Operations Research*, Montreal: University of Montreal, 2012.

[10] I. Arel, D. C. Rose, T. P. Karnowski, Deep machine learning-A new frontier in artificial intelligence research, in *IEEE Computational Intelligence Magazine*, Knoxville: The University of Tennessee, 5(4), 13-18, 2010.

[11] D Yu, L Deng, Deep learning and its applications to signal and information processing, in *Signal Processing Magazine*, IEEE, 28(1), 145-154, 2011.

[12] L Zhang, Y Liang, A fast method of face detection in video image, in *The 2nd IEEE International Conference on Advanced Computer Control(ICACC 2010)*, 490-494, 2010.

[13] R. Collobert, J. Weston, A unified architecture for natural language processing: Deep neural networks with multitask learning, in *Proceedings of the 25th international conference on Machine Learning*, 160-167, 2008.

[14] H. Xiang, D. Wang, The application of genetic BP neural network and D-S evidence theory in the complex system fault diagnosis, in *Springer Berlin Heidelberg*, Germany, 219-224, 2012.

[15] P. C. Fritzen, G. Cardoso, J. M. Zauk, etc, Alarm processing and fault diagnosis in power systems using artificial neural networks and genetic algorithms, in *International Conference on Industrial Technology*, Chile: Vi a del Mar, 891-896, 2010.

[16] F. Wang, Fault diagnosis for power systems based on neural networks, in *2nd International Conference on Software Engineering and Service Science*, 352-355, 2011.

[17] Andrew Ng, Unsupervised feature learning and deep learning tutorial, in *openclassroom.stanford.edu*.