DOI: 10.1007/978-3-319-71607-7_19

Deep Convolutional Neural Network for Facial Expression Recognition

Yikui Zhai[1], Jian Liu[1], Junying Zeng[1(✉)], Vincenzo Piuri[2], Fabio Scotti[2], Zilu Ying[1], Ying Xu[1], and Junying Gan[1]

[1] School of Electronic and Information on Engineer, Wuyi University, Jiangmen 529020, China
yikuizhai@163.com, iamjianliu@163.com, zengjunying@126.com, ziluy@163.com, xuyingll7@163.com, junyinggan@163.com

[2] Department of Computer Science, Università degli Studi di Milano, 26013 Crema, Italy
{vincenzo.piuri,fabio.scotti}@unimi.it

Abstract. In this paper, a deep convolutional neural network model and the method of transfer learning are used to solve the problems of facial expression recognition (FER). Firstly, the method of transfer learning was adopted and face recognition net was transferred into facial expression recognition net. And then, in order to enhance the classification ability of our proposed model, a modified Softmax loss function (Softmax-MSE) and a double activation layer (DAL) are proposed. We performed our experiment on enhanced SFEW2.0 dataset and FER2013 dataset. The experiments have achieved overall classification accuracy of 48.5% and 59.1% respectively, which achieved the state-of-art performance.

Keywords: Facial expression recognition (FER)
Deep convolutional neural network · Transfer learning

1 Introduction

Deep Convolutional Neural Networks (DCNN) are playing a more and more important role in most artificial intelligence domains such as face recognition [1], facial expression recognition [2], etc. Based on their large-scale databases, deep convolutional neural networks have achieved good performance on these subjects.

However, there are few datasets on facial expression recognition tasks and it is unrealistic to build a large-scale FER dataset. To resolve this problem, many researchers turned to using the method of transfer learning for answers. Because in related domains (such as face recognition and facial expression recognition), the pre-learned features from the first layers have almost the same generality. Most of researchers utilize transfer learning by fine-tuning a pre-trained model from other areas and have achieved great performance [3, 4].

Although the great progress has been made, the problems on facial expression recognition are still severe. Because usually facial expression recognition task has 6 or 7 classes, features generated from the first layers of pre-trained model are often too big and have lots of redundant information. The problem of over-fitting still remains.

In this paper, we use deep convolutional neural network model to classify facial expressions. Like the mostly used method, we firstly use transfer learning method to fine-tune a face recognition model into facial expression model on FER datasets. To restrain the over-fitting problem, we first did data augmentation on SFEW2.0 dataset and then we propose double activation layer (DAL) and Softmax-MSE loss function in our proposed network. We performed our experiments on both SFEW2.0 and FER 2013 datasets and achieved state-of-art performance.

The rest of this paper is organized as follows. Section 2 introduces the related works, the main algorithms are detailed in Sect. 3, experiments and conclusions are described in Sects. 4 and 5 respectively.

2 Related Works

In [5], Ekman and Friesen developed Facial Action Coding System (FACS) which can divide human faces into several independent and interrelated action units. They classified human facial expressions into 6 basic ones: “happy”, “sad”, “surprise”, “fear”, “anger” and “disgust” and they built a facial expression dataset based on these 6 facial expressions. Ekman and Friesen’s work gave a basic description of facial expressions. Motivated by their work, Lien and Kanade [6] developed a system that can automatically recognizes uncorrelated action units using Hidden Markov Models (HMMs). Other researchers did facial expression recognition by applying Gaussian Mixture Model (GMM) [7–9].

Recently, the rapid development of machine learning especially the development of DCNN brought a new direction to facial expression recognition. Lecun et al. [10] proposed Convolutional Neural Networks which is the first successful learning algorithm for multi-layer networks. As a kind of depth learning model, DCNN improves the training performance of the back propagation by reducing the number of parameters to be trained through local spatial mapping. Many researchers adapted DCNN framework in their facial expression recognition research and achieved good performance. Burkert et al. [11] proposed a DCNN framework for facial expression recognition which is irrelevant to any manual feature extraction and it outperforms earlier DCNN based methods. Yu and Zhang [12] put forward a DCNN facial expression recognition framework which gathers three state-of-art neural networks. The structure of this DCNN framework is made up of three state-of-art face detectors and a classification model that ensembles multiple DCNN. The framework was tested on SFEW 2.0 dataset and achieved good performance.

DCNN networks are suitable for tasks that have large databases in that deep convolutional neural networks have large amount of parameters and large database can fit them well. While in facial expression recognition task, it is impractical to have large database on FER tasks. Transfer learning solved this problem for its usage of other related knowledge. In [13] Zhang settled the problems of facial feature fusion and the relations of multiple action units (AUs) in building robust expression recognition systems by utilizing two transfer learning algorithms – multi-kernel learning and multi-task learning. In [14], Chen et al. used the method of transfer learning to train a person-specific FER model by utilizing the informative knowledge from other people.

Learning a FER model using transfer learning only consume a small amount of data. Because there is a great relevance between face recognition tasks and facial expression recognition tasks, many researchers used transfer learning method to train facial expression recognition based on the knowledge of face recognition tasks like in [3, 4].

3 Proposed Approach

3.1 Convolutional Neural Networks

Convolutional neural networks have been used for classification tasks for many years. Recently the rapid development of DCNN have greatly promoted the classification performance. DCNN mainly applied on two methods: Local Receptive Fields (LRF) and Sharing Weights.

General neural networks adopted fully-connected design between input layers and hidden layers. It is feasible to calculate the feature of the whole image when the image is in small size. But when the image becomes bigger, the calculation will be more time-consuming. Convolutional layer is the way to solve this problem. It restricts the connections between the input units and hidden units to make each hidden units only connect to a small area of input units.

Sharing weights is another strategy to save training expenditure. DCNN combines several convolutional layers and pooling layers, and then implement mapping between input matrix and output matrix in fully-connected layers. Each convolutional layer and pooling layer contains several feature mapping, and each feature maps a "plane" which consists of multiple neurons. Each "plane" extracts a kind of feature from input using convolutional filter. In our expression net, the size of input image is $144 * 144 * 3$, the size of filter in this net is $5 * 5 * 3$, the pad of convolutional operation is 2.

3.2 Transfer Face Net to Exp Net

In this paper, a representative transfer learning method in [15] was adopted. We transfer face recognition network in [16] to facial expression recognition network. There are two networks in [16] and we adopted Net B.

To apply the knowledge of face recognition network to facial expression network, we modify the architecture of face net into expression recognition net. First we changed the number of outcomes in fully-connected layers of the original face net and then kept the other parameters in face net invariable. For there are only 7 classes in facial expression recognition task while there are far more than 7 classes in face net, so we changed the number of outcomes in the last fully-connected layer into 7.

After modifying face net into expression recognition net, we append double activation layer on the last convolutional layer and input the outputs of the double activation layer to the loss layer.

3.3 Modified Softmax Loss Function (Softmax-MSE)

There are mainly two issues in transferring face net into facial expression recognition net. The first one is that the transferred face net may still contains information which is still helpful for face identification because of the big quantitative difference between face dataset and facial expression dataset. The second one is that the fine-tuned face net is so huge for FER tasks, so the over-fitting problem couldn't be solved appropriately.

In order to overcome these problems, we proposed a modified Softmax loss function (Softmax-MSE) which is a cost-sensitive learning method that takes classification error into consideration. At present, the existing Softmax loss function is a non-cost-sensitive function. Experimental results indicate that Softmax loss function would achieve better performance in some tasks such as face recognition which do not care the correlations of outcomes, but cost-sensitive function like Euclidean loss function would achieve better regressive performance on tasks that require high correlation of outcomes. Furthermore, our proposed Softmax-MSE loss function absorb their advantages and weaken their shortcomings. The function is detailed as follows.

Suppose that input neurons and output neurons have the same number of m, and the input of loss layer is m, and the input of loss layer is $X = \{x_0, x_1, \ldots, x_{m-1}\}$, then the Softmax function output of the kth neuron of the layer is defined as:

$$p_k = \frac{e^{x_k - \max(X)}}{\sum_{i=0}^{m-1} e^{x_i - \max(X)}} \tag{1}$$

where $k \in [0, m-1]$. If the batch size of the net is n, $p_k = \max([p_0, p_1, \ldots, p_{m-1}])$, the regressive prediction value is:

$$\hat{y}_j = \sum_{k=0}^{m-1} k p_k \tag{2}$$

Softmax-MSE loss value of the nth image is:

$$L = \frac{1}{n} \sum_{j=0}^{n-1} (\hat{y}_j - y_j)^2 \tag{3}$$

where y_j is the expectation value of $\hat{y}_j$, and it is also the label value of the jth image.

In order to maintain the advantage of Softmax loss function, we used the gradient of Softmax loss function in Softmax-MSE layer:

$$\frac{\partial L_j}{\partial x_i} = \begin{cases} p_i - 1, & i = y_j \\ p_i, & i \neq y_j \end{cases} \tag{4}$$

To illustrate the proposed Softmax-MSE function more clearly, we made two comparative experiments, one of which is based on implementing proposed Net B_DAL network on enhanced SFEW 2.0 dataset and the other is based on

implementing proposed Net B_DAL_MSE network on enhanced SFEW 2.0 dataset. The experiments results will be detailed in part 5.

3.4 Double Activation Layer

In [16], both Net A and Net B employ Maxout activation method in all their layers and achieved relatively good results. Motivated by this, we modified Net B by appending Double Activation Layer (DAL) on fully-connected layer to enhance the performance of the network.

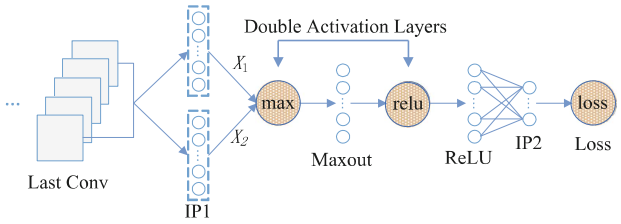

Fig. 1. The architecture of DAL layers

As shown in Fig. 1, DAL adapt Maxout + ReLU architecture by utilizing the nonlinearity of Maxout and the sparseness of ReLU function to make the network obtain its global optimal result. Specifically, there are two groups of inner product layer which are connected to the last convolutional layers of the network. Then the outputs of those two inner product layers are pumped into max-out layers which find the bigger output. During the last layer in double activation layers, "relu" layer get the maxout as input. In Fig. 1, "IP1" and "IP2" are inner product layers.

Suppose that the input of activation layer is X, then the output of ReLU layer is:

$$Y = \max(0, X) \tag{5}$$

In proposed network, there are two sets of inner product layers in IP1 layer and one set of inner product layer in IP2 layer. As depicted in Fig. 1, the outputs of last convolutional layers are the input of the two sets of inner product layers respectively.

Assume that the outputs of two sets of inner product layers in IP1 layer are X1 and X2, then the output of Maxout is:

$$Y = \max(X_1, X_2) \tag{6}$$

After the max function, then the output of relu layer is input to IP2 layer. Lastly, the output of IP2 layer is input to loss layer.

4 Experiments

4.1 Preprocessing on SFEW 2.0 New and FER2013

We did our experiments on SFEW 2.0 dataset [17] and FER2013 dataset [18]. Both the two datasets have 7 classes: “Angry”, “Disgust”, “Fear”, “Happy”, “Sad”, “Surprise” and “Neutral”.

The original SFEW dataset has 958 face images in training set and it is insufficient for fine-tuning a face net to expression net. So we did data augmentation on the training set of SFEW dataset by adding random information to three channels of the colored face image and remain the test set of SFEW dataset unchanged. To make the performance of expression net better, we did selective measures on the training set of SFEW dataset by excluding noisy image. These images either contain no human face or the face in the image could not be detected by face detector. In the next paper, we refer the processed SFEW dataset as enhanced dataset.

The detailed information about these two datasets is described in Table 1. “SFEW2.0 new” refers to data-augmented SFEW2.0 dataset. The numbers of images in each class is showing in Table 1.

Table 1. The information about FER2013 and SFEW2.0 new datasets

	Ang	Dis	Fear	Happy	Neural	Sad	Surprise
SFEW2.0 new	356	180	327	773	580	593	385
FER2013	3995	436	4097	7215	4965	4830	3171

Before training expression net, as shown in Fig. 2, we first applied Viola and Jones [19] face detector for face detection. Then we detected 5 basic facial key points in the image and crop the human face in the image into size of 144 * 144 based on these facial key points and set images gray.

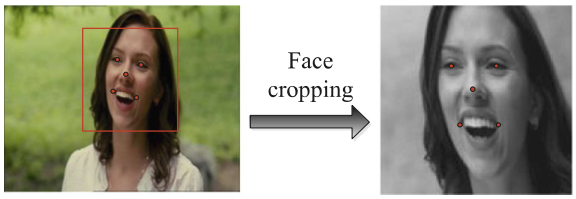

Fig. 2. The preprocessing on SFEW2.0 dataset

Because our fine-tuning is based on a pre-trained face net, it gave a good starting point of training stage. According to the experimental results, we set the base learning rate as 10e−5 and the learning policy as “fix” mode, besides we set the momentum as 0.9, gamma as 5 * 10e−6 and power as 0.75. The other parameters in the solver file are set based on the true conditions of expression dataset. The training is implemented on

deep learning framework Caffe [20]. In the training stage on FER2013 dataset, we also adopted the same solving parameters of training on SFEW new dataset. We didn't do data augmentation on FER2013 dataset because it has enough training images.

4.2 Experiments and Results on SFEW2.0 and FER2013

The experiments are implemented on SFEW 2.0 new dataset and FER 2013 dataset respectively. Firstly, we adopted the network described in Sect. 3.4 and set it as Net B, then we employed DAL to Net B by appending DAL to the first fully-connected layer and called the net as Net B_DAL. At last, we changed the Softmax layer in Net B_DAL into Softmax-MSE loss layer and call the changed net as Net B_DAL_MSE. Secondly, we trained these three networks on both SFEW dataset and FER 2013 dataset using fine-tuning method. The overall average accuracy is shown in Table 2.

Table 2. The overall average accuracy

Methods	Average accuracy	Datasets
AUDN [21]	26.14%	SFEW 2.0
STM-ExpLet [22]	31.73%	
Inception [23]	47.70%	
Mapped LBP [24]	41.92%	
Train from scratch	39.55%	
VGG fine-tune	41.23%	
Transfer learning [25]	48.50%	SFEW2.0 + FER2013
Multiple deep network [12]	52.29%	
Net B	46.52%	SFEW2.0 new
Net B_DAL	45.64%	
Net B_DAL_MSE	48.51%	
Net B	60.91%	FER2013
Net B_DAL	58.33%	
Net B_DAL_MSE	59.15%	

From Figs. 3, 4 and 5 report the confusion matrix of experiments on fine-tuning Net B, Net B_DAL and Net B_DAL_MSE on SFEW 2.0 new dataset respectively. Figures 6, 7 and 8 report the confusion matrix of experiments of fine-tuning Net B, Net B_DAL and Net B_DAL_MSE on FER 2013 dataset respectively. The diagonals in the pictures show the rates of correct classification rate of 7 facial expressions. The boxes in the figures are covered gray, the more dark the box, the more higher the correct classification rate.

In the three figures above, comparing Figs. 4 and 5 with Fig. 3, we can see that although the overall average accuracy of Net B_DAL and Net B_DAL_MSE are lower than Net B, some expression's accuracy in Net B_DAL and Net N_DAL_MSE are higher than its counterparts in Net B. For example, the accuracy of Neutral expression in Net B_DAL_MSE is 3% higher than Net B. In FER tasks, the expression of Disgust

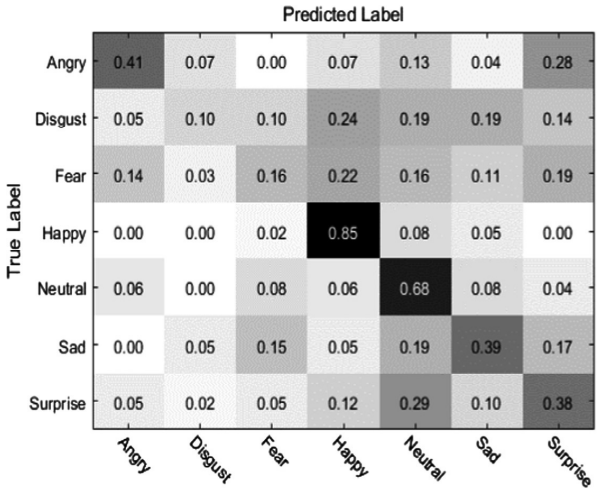

Fig. 3. The confusion matrix of Net_B on SFEW2.0 new

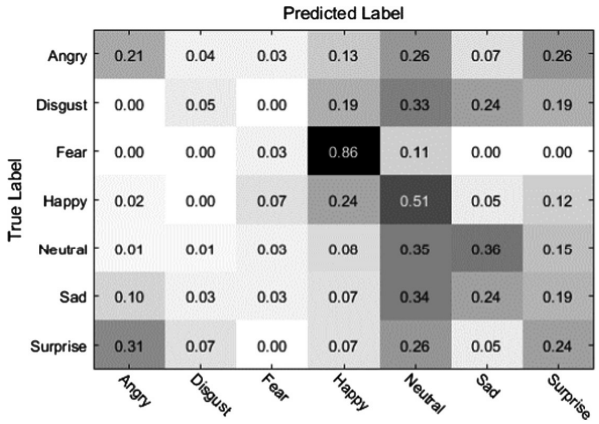

Fig. 4. The confusion matrix of Net B_DAL on SFEW2.0 new

and Neutral are the most two difficult ones to classify. Besides, SFEW dataset is built in unconstrained conditions, experiments on SFEW dataset are harder than the experiments on other datasets.

Figures 6, 7 and 8 below report the confusion matrix of experiments Net B, Net B_DAL and Net B_DAL_MSE fine-tuning on FER 2013 dataset respectively. For FER 2013 dataset contains more pictures of 7 facial expression classes than in SFEW 2.0 dataset, the experiments on FER 2013 dataset are more comprehensive. The experimental results draw the same conclusions as in experiments on SFEW2.0 new dataset.

The curve of train accuracy and test accuracy with iterations is shown in Figs. 9 and 10. Here test accuracy means validation accuracy. Both the two experiments are

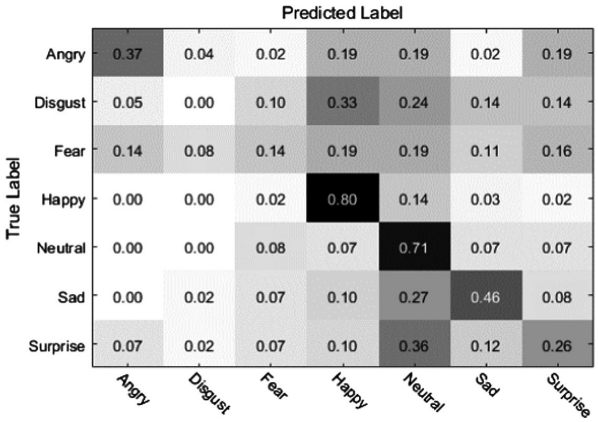

Fig. 5. The confusion matrix of Net B_DAL_MSE on SFEW2.0 new

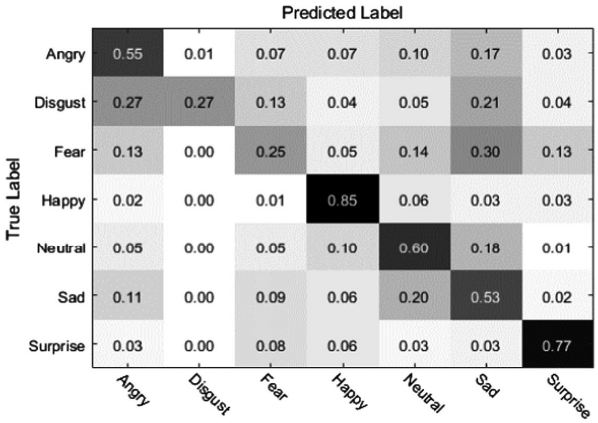

Fig. 6. The confusion matrix of Net_B on FER2013

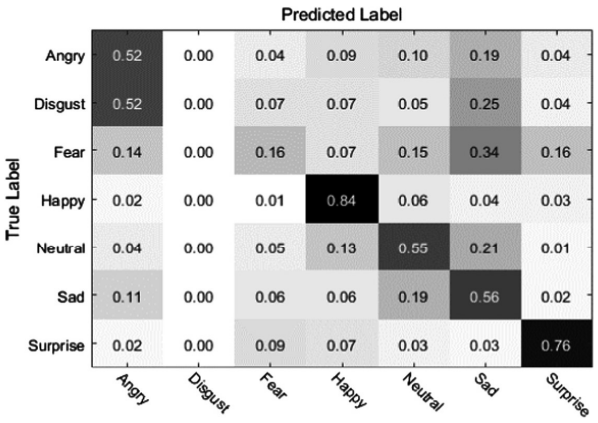

Fig. 7. The confusion matrix of Net B_DAL on FER2013

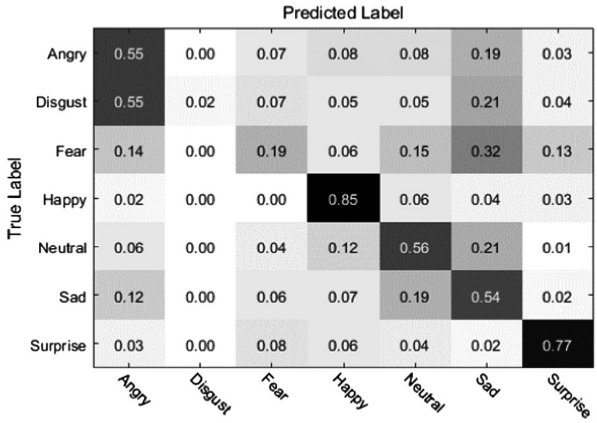

Fig. 8. The confusion matrix of Net B_DAL_MSE on FER2013

experimented on the SFEW 2.0 new dataset. In Fig. 9, the train and test process are conducted using Net B_DAL network. In Fig. 10, the train and test process are conducted using Net B_DAL_MSE network. As shown in these two figures, the test accuracy in Fig. 9 tends to be stabilized earlier than in Fig. 10.

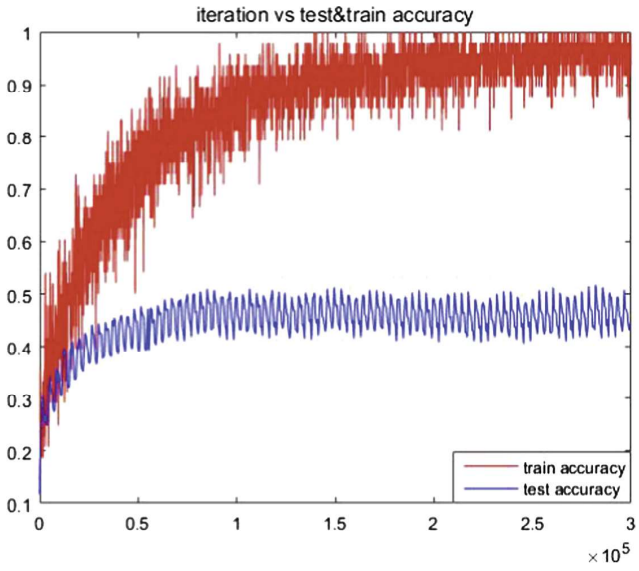

Fig. 9. The curve of test and train accuracy of Net B_DAL on SFEW2.0 new

Another statement about the shape the curves goes has to be made. As can be seen in Figs. 9 and 10, the test accuracy curves and train accuracy curves are jagged. The reason for this phenomenon is that the SFEW 2.0 new dataset contains few pictures.

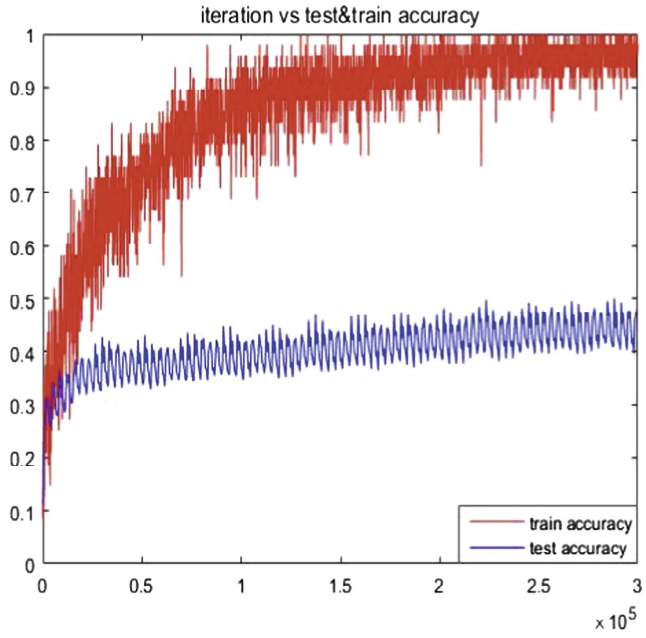

Fig. 10. The curve of test and train of Net B_DAL_MSE on SFEW2.0 new

5 Conclusions and Future Works

In this paper, we proposed double activation layer (DAL) and Softmax-MSE loss function to overcome the over-fitting problem. Based on Net B, we proposed Net B_DAL and Net B_DAL_MSE and did experiments of fine-tuning Net B, Net B_DAL and Net B_DAL_MSE on SFEW 2.0 new dataset and FER 2013 dataset respectively. The experiments achieved state-of-art performance. But the accuracies on Disgust and Neutral are not as good as expected. In future, we plan to work on improving the accuracy on these two expressions.

Acknowledgment. This work is supported by NNSF (No. 61372193), Guangdong Higher Education Outstanding Young Teachers Training Program Grant (No. SYQ2014001), Characteristic Innovation Project of Guangdong Province (Nos. 2015KTSCX 143, 2015KTSCX145, 2015KTSCX148), Youth Innovation Talent Project of Guangdong Province (Nos. 2015K QNCX172, 2016KQNCX171), Science and Technology Project of Jiangmen City (Nos. 20150 1003001556, 201601003002191), and China National Oversea Study Scholarship Foundation.

References

1. Sun, Y., Wang, X., Tang, X.: Deep convolutional network cascade for facial point detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3476–3483 (2013)

2. Kahou, S.E., Pal, C., Bouthillier, X., et al.: Combining modality specific deep neural networks for emotion recognition in video. In: Proceedings of the 15th ACM on International Conference on Multimodal Interaction, pp. 543–550. ACM (2013)
3. Ding, H., Zhou, S.K., Chellappa, R.: FaceNet2ExpNet: regularizing a deep face recognition net for expression recognition. arXiv preprint arXiv:1609.06591 (2016)
4. Wang, F., Xiang, X., Liu, C., et al.: Transferring face verification nets to pain and expression regression. arXiv preprint arXiv:1702.06925 (2017)
5. Ekman, P., Friesen, W.V.: Facial action coding system (1977)
6. Lien, J.J., Kanade, T., Cohn, J.F., et al.: Automated facial expression recognition based on FACS action units. In: International Conference on Face & Gesture Recognition, p. 390. IEEE Computer Society (1998)
7. Yang, J.: Maximum margin GMM learning for facial expression recognition. In: IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, pp. 1–6. IEEE (2013)
8. Tariq, U., Yang, J., Huang, T.S.: Maximum margin GMM learning for facial expression recognition (2013)
9. Rong, L.I., Wang, H.J., Yan-Hua, X.U., et al.: A face expression recognition method based on fusion of supervised super-vector encoding and adaptive GMM model. Comput. Mod. **02**, 15–20 (2016)
10. Lecun, Y., Boser, B., Denker, J.S., et al.: Backpropagation applied to handwritten zip code recognition. Neural Comput. **1**(4), 541–551 (2008)
11. Burkert, P., Trier, F., Afzal, M.Z., et al.: DeXpression: deep convolutional neural network for expression recognition. Comput. Sci. **22**(10), 217–222 (2015)
12. Yu, Z., Zhang, C.: Image based static facial expression recognition with multiple deep network learning. In: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, pp. 435–442. ACM (2015)
13. Zhang, X.: Facial expression analysis via transfer learning. Dissertations & Theses - Gradworks (2015)
14. Chen, J., Liu, X., Tu, P., et al.: Person-specific expression recognition with transfer learning. In: IEEE International Conference on Image Processing, pp. 2621–2624. IEEE (2012)
15. Girshick, R., Donahue, J., Darrell, T., et al.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587 (2014)
16. Wu, X., He, R., Sun, Z.: A lightened CNN for deep face representation. Comput. Sci. (2015)
17. Dhall, A., Goecke, R., Lucey, S., et al.: Static facial expressions in the wild: data and experiment protocol. CVHCI
18. Goodfellow, I.J., Erhan, D., Luc, C.P., et al.: Challenges in representation learning: a report on three machine learning contests. Neural Netw. **64**, 59–63 (2015)
19. Viola, P., Jones, M.: Robust real-time face detection. Int. J. Comput. Vis. **57**(2), 137–154 (2004)
20. Jia, Y., Shelhamer, E., et al.: Caffe: convolutional architecture for fast feature embedding. Eprint ArXiv, pp. 675–678 (2014)
21. Liu, M., Li, S., Shan, S., et al.: AU-inspired deep networks for facial expression feature learning. Neurocomputing **159**(C), 126–136 (2015)
22. Liu, M., Shan, S., Wang, R., et al.: Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1749–1756. IEEE (2014)

23. Mollahosseini, A., Chan, D., Mahoor, M.H.: Going deeper in facial expression recognition using deep neural networks. Comput. Sci., 1–10 (2015)
24. Levi, G., Hassner, T.: Emotion recognition in the wild via convolutional neural networks and mapped binary patterns. In: ACM on International Conference on Multimodal Interaction, vol. 2015, pp. 503–510. ACM (2015)
25. Ng, H.W., Nguyen, V.D., Vonikakis, V., et al.: Deep learning for emotion recognition on small datasets using transfer learning. In: ACM International Conference on Multimodal Interaction, vol. 2015, pp. 443–449. ACM (2015)