Facial Expression Recognition Based on Convolutional Neural Network

Zhou Yue, Feng Yanyan
College of Electronic Engineering
Guangxi Normal University
Guilin, China
744427475@qq.com,1095115523@qq.com

Zeng Shangyou, Pan Bing
College of Electronic Engineering
Guangxi Normal University
Guilin, China
zsy@mailbox.gxnu.edu.com, 276696163@qq.com

Abstract—**Facial expression recognition is an important field of pattern recognition research. Traditional machine learning methods extract features manually. It has insufficient generalization ability and poor stability. Moreover, its accuracy is difficult to improve. In order to achieve better facial expression recognition, this paper designs a modular multi-channel deep convolutional neural network. To avoid overfitting, the network output uses a global average layer. Data enhancement on the dataset before training can improve the generalization ability of the model. Test the performance of network on the FER2013 emoticon dataset. The accuracy of expression recognition is 68.4%. It performs a prediction for about 0.12s. Compared to other recognition algorithms, network has certain advantages. Finally, a real-time facial expression recognition system is constructed by using the trained recognition model. The experimental results show that the system can effectively recognize facial expressions in real time.**

Keywords- convolutional neural network; network performance; real-time; expression recognition

I. INTRODUCTION

In 2006 Geoffrey Hinton proposed the concept of deep learning[1]. After that, the theory of deep learning gradually improved. Deep learning has become the AQhottest research field of artificial intelligence. Breakthrough progress has been made in the field of artificial intelligence. So far, deep learning algorithms have become the top algorithms in the field of artificial intelligence. In the fields of image recognition and natural language processing, deep learning algorithms perform much better than other algorithms.

The current top algorithm in the field of image recognition is the Convolutional Neural Network (CNN). The accuracy of image recognition on large-scale datasets has exceeded the human average. At present, CNN has been widely used in the field of image recognition, such as bank handwritten character recognition, face recognition[2], cell image recognition, facial expression recognition[3], etc.

The CNN model has been pursuing a high accuracy, with the more representative models being VGG[4], GoogLeNet[5], and ResNet[6].The parameters along with the improvement of accuracy are also increasing. This process is likely to cause the network to over-fit the training set. In addition, the large network structure will cause low efficiency of network training and the network requirements for computer hardware are too high to be portable. How to design a efficient and lightweight model has become an urgent problem to be solved.

Expression recognition technology is a means of judging emotions through human facial expressions. It has a wide range of applications in medical field, teaching field, intelligent transportation, human-computer interaction and so on. In this paper, the deep learning method is applied to facial expression recognition, and a convolutional neural network model suitable for expression recognition is constructed.

In order to obtain a high accuracy and lightweight facial expression recognition model, we mainly made the following contributions:

- Designed by ExpressionNet, a lightweight convolutional neural network architecture.
- Using the deep learning framework Caffe[7] to train ExpressionNet on the facial expression dataset. Then, it obtain a facial expression recognition network model. This model has a better advantage than the published expression recognition algorithm.
- Based on the trained ExpressionNet expression recognition model, a real-time facial expression recognition system is designed.

II. FACIAL EXPRESSION RECOGNITION MODEL

The ExpressionNet network is modular in design and the core component is the reduceV2[8] module. The structure of the ReduceV2 is shown in Fig. 1. The reduceV2 module divides a convolutional layer into two parts, a dimension reduction layer and a sampling layer, each with a ReLU active layer. The dimension reduction layer uses S 1×1 convolution kernels to reduce the input feature map to S. The sampling layer uses two sets of convolutions (G1 and G2) to extract features. G1 uses S 3×3 convolution kernels. G2 uses two layers of S 3×3 convolutions. At the same time, add shortcut connection between the input and output of the module. It allows the network to better pass the gradient to a shallower level during backpropagation. The ReduceV2 module greatly reduces the size of the network model while ensuring network accuracy. As shown in Fig. 2, ExpressionNet structure consists

978-1-7281-0945-9/19/$31.00©2019 IEEE

of alternately stacking the Reduce module and the downsampling layer.

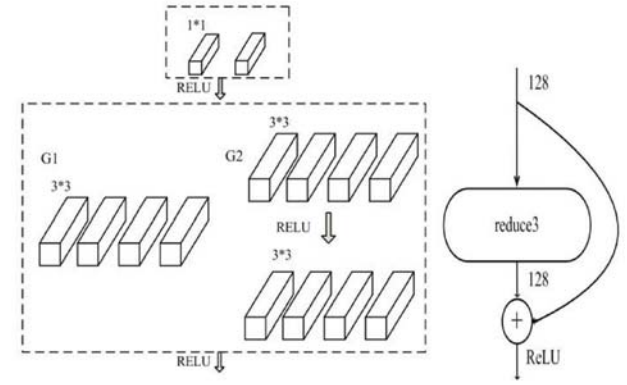

Figure 1 Reduce and reduceV2 module

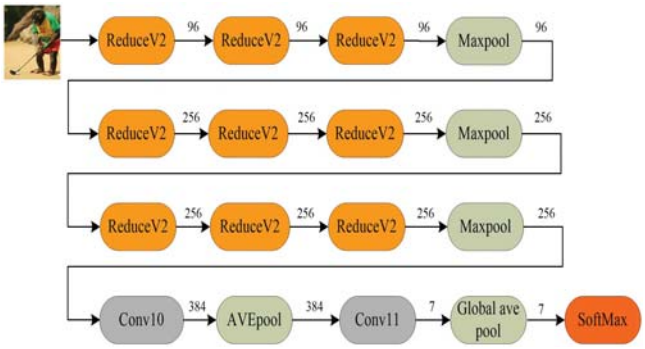

Figure 2 The ExpressionNet structure

The ExpressionNet model has 9 layers of ReduceV2 modules. Each module is followed by a BN layer and a ReLU activation function. Four downsampling layers are interspersed between the modules. The last layer uses the global average pooling to connect all the feature maps. In order to ensure that the size of the input and output can be mastered, ExpressionNet retains the original top and bottom convolution layers .

Overlapping pooling can alleviate overfitting. The non-overlapping pooling layer ignores the effects of neighboring pixels on feature map, which leads to a decrease in network accuracy. However, overlapping structures may introduce noise. This noise will be amplified when Max Pooling is constantly overlapping. Therefore, ExpressionNet uses a combination of Max+Max+Max+Avg to overlap the pooling layers. The sparsity of Avg Pooling can reduce the introduced noise . The fully connected layer acts as a connection layer for the feature map, and its parameters are huge. In general, the entire network parameters are mostly concentrated in the fully connected layer. Excessive parameters can easily lead to over-fitting of network training. It is also difficult to achieve lightweight. The global average pooling layer is a sparse connection. It can reduce the introduced noise and the parameters of the network. Therefore, ExpressionNet uses the global average pooling layer instead of the fully connected layer[9].

III. MODEL TRAINING

The framework that the experiment depends on is the caffe deep learning framework. The computer is configured with i7-6700K quad-core CPU, Ubuntu 14.04 operating system, 32GB memory, and NVIDIA-GTx 1070 GPU.

The facial expression dataset is the FER2013 dataset in the experiment. The FER2013 dataset has a total of 35,887 images. The dataset is divided into three parts: training set (28,709 pictures), test set (3,589 pictures), and verification set (3,589 pictures). All images are uniform 48×48 grayscale images. The dataset is divided into 7 categories, namely, anger, disgust, fear, happiness, sadness, surprise, and neutrality. Because the dataset has some noise (all black pictures, cartoon pictures, non-expression images and non-expression pictures), it is difficult to identify. The eye recognition accuracy of the FER2013 dataset is 65%. A sample of the FER2013 dataset is shown in Fig. 3.

Figure 3 Sample of the FER2013 dataset

The images are pre-processed prior to training. Using data amplification techniques, 48*48 pixel image blocks are taken from the upper left, lower left, upper right, lower right, and middle of each image respectively, and then flip it horizontally for a total of 10 new images. After data amplification, the numbers of the entire training set is expanded to 10 times of the original size, and then all images are subtracted from the mean value. In the network test, the test picture is input into the pre-trained network model to obtain the probability of all the expression categories of the network SoftMax layer. The output of the SoftMax layer is the expression class corresponding to the maximum probability. Data amplification effectively mitigates overfitting and improves network performance.

The experimental parameters are set as follows. The initial learning rate is 0.005, the learning rate varies by multistep, the gamma is 0.1, the stepvalue is 15000, 30000, and 45000, and the maximum number of iterations is set to 60000. The network performance is tested every 500 times.

Test ExpressionNet recognition performance on the FER2013 dataset. Fig. 4 shows the variation of the test accuracy of ExpressionNet during FER2013 training. When the training reaches 15000 and 30,000 times, the learning rate is reduced and the recognition accuracy is significantly improved. When the training reaches 40,000 times, the expressionNe tends to be stable, and the recognition accuracy changes little. When the maximum number of iterations is 60000, the recognition accuracy can reach 68.4%.

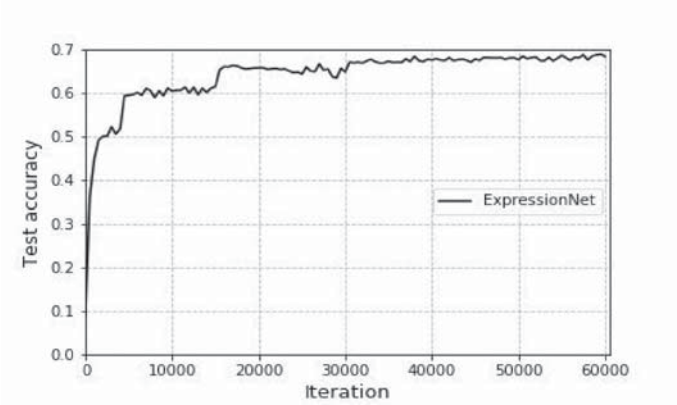

Figure 4 Test accuracy on the FER2013 dataset

Enter 35,89 images into ExpressionNet to test the recognition time. In order to accurately judge the speed of model recognition, the model was tested 5 times. The test time is shown in TABLE I.

TABLE I. ACCURACY AND TIME CORRESPONDING TO THE NUMBER OF RECOGNITIONS

numbers	Accuracy(%)	Time(s)
1	68.6	429
2	67.6	426
3	68.4	430
4	67.7	428
5	67.6	425

It takes about 426 seconds to identify all the emoticons, and the average time to recognize an image is 0.12 seconds. It has a faster recognition speed and a higher recognition accuracy. TABLE II gives the accuracy and parameters of ExpressionNet and other methods in the FER2013 dataset. The experimental results show that ExpressionNet has better recognition performance than other methods on the FER2013 dataset. ExpressionNet has the highest accuracy and fewer parameters. Its model is only 15 megabytes, which is basically lightweight. Although the model size is a bit larger than ShallowNet and SN (D & BN), its accuracy is higher.

TABLE II. THE ACCURACY AND PARAMETERS OF EXPRESSIONNET AND OTHER METHODS IN THE FER2013 DATASET

methods	Accuracy(%)	Model size(MB)
AlexNet[10]	67.50	208.2
ExpressionNet	68.40	15.0
CNN[11]	57.10	--
CNN-based+Softmax[12]	65.03	--
Net EXP_DAL_MSE[13]	61.59	--
ShallowNet[14]	63.49	11.4
SN(D&BN) [14]	64.78	11.4
VGG-11[14]	67.52	22.3
Paper[15]	65.60	--
Paper[16]	66.00	--
MVFE-LightNet[17]	68.00	--

IV. REAL-TIME FACIAL EXPRESSION RECOGNITION USING EXPRESSIONNET MODEL

Based on the trained ExpressionNet facial expression recognition model, a real-time facial expression recognition system is designed. This system is the practical application of personal face information recognition. It locates the face from the video stream captured by the camera. It then identifies the information of the positioned face. In this system, the front face image of the target person is pre-stored, and after the face region of the image is located, the feature of the face region is converted into a 128D face code. The system block diagram is shown in Fig. 5. The programming language of the facial expression recognition system is python. The Python IDE PyCharm is used as the development software of the program, and libraries such as opencv, dlib, and face_recognition are used.

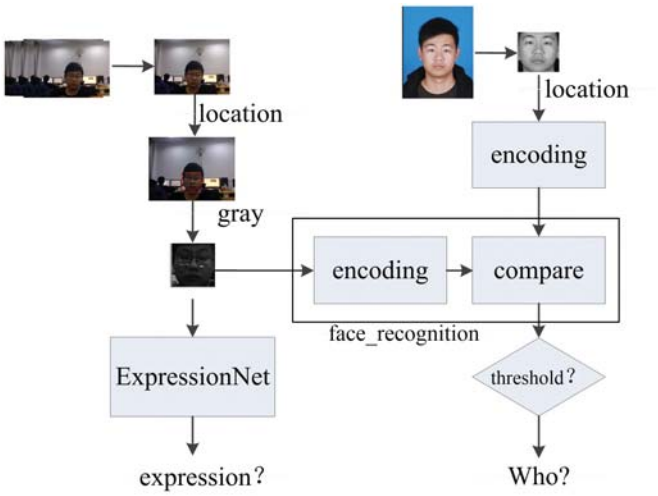

Figure 5 The system block diagram

The system is divided into four steps: image acquisition, Face positioning, face recognition, and expression recognition. The specific process is as follows.

- Image acquisition. Face and expression recognition images come from laptop cameras. The system reads the video stream in the camera through OpenCV library, and obtains one of the frames in the video stream.
- Face positioning. Face positioning is achieved through the face_recognition library. The face area is extracted by the data processing method of the slice in python.
- Face recognition. The system performs face recognition on the detected face area. It converts the face area into a 128D face code. Then, calculate the distance between them and the pre-stored face code. If the encoding distance is less than the threshold, it is the same face.

- Expression recognition. The detected face area is resized to 46 × 46 and converted into a grayscale image, loaded into the trained ExpressionNet model, forward computing. Finally, the output of the Softmax layer is extracted as the expression recognition result.

The recognition effect is shown in Fig. 6.

Figure 6 System recognition effect

V. CONCLUSION

This paper designs a cascading module for feature extraction. The module uses three different channels, increasing the diversity of extract feature. Before entering the convolution operation, the 1*1 convolution kernel is used to reduce the feature maps, which is beneficial to the reduction of parameters. Then, the feature maps of the three channel outputs are cascaded together. Finally, use this module to build ExpressionNet. Compare ExpressionNet to the classic AlexNet and others algorithm in the FER2013 dataset. The results show that the accuracy of ExpressionNet is significantly improved, and the parameters are greatly reduced. It proves that the module can effectively improve network performance and control model size. The next work is to continue to improve on this module to further improve network performance and reduce the size of the model. At the same time, look for larger datasets for experimentation.

REFERENCES

[1] Hinton G E, Salakhutdinov R R. Reducing the Dimensionality of Data with Neural Networks[J]. Science, 2006, 313(5786):504.

[2] Zhang K, Zhang Z, Li Z, et al. Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks[J]. IEEE Journal of Solid-State Circuits, 2016, 23(99):1161-1173.

[3] Hasani B, Mahoor M H. Spatio-Temporal Facial Expression Recognition Using Convolutional Neural Networks and Conditional Random Fields[C]// IEEE International Conference on Automatic Face & Gesture Recognition. IEEE, 2017:790-795.

[4] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition[J].ar Xiv preprint ar Xiv:1409.1556, 2015.

[5] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]// IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2015:1-9.

[6] He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition[C]// IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society,2016:770-778.

[7] Jia Y, Shelhamer E, Donahue J, et al. Caffe: Convolutional Architecture for Fast Feature Embedding[C]// ACM International Conference on Multimedia. ACM, 2014:675-678.

[8] Zhou, Yue & Feng, Yanyan & Zeng, Shangyou & Pan, Bing. (2019). Design of Lightweight Convolutional Neural Network Based on Dimensionality Reduction Module. IOP Conference Series: Materials Science and Engineering. 533. 012045. 10.1088/1757-899X/533/1/012045.

[9] Lin M, Chen Q, Yan S. Network In Network[J]. Computer Science, 2013.

[10] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[C]// International Conference on Neural Information Processing Systems.Curran Associates Inc.2012:1097-1105.

[11] Vedat Tümen, Ömer Faruk Söylemez, Ergen B . Facial emotion recognition on a dataset using convolutional neural network[C]// 2017 International Artificial Intelligence and Data Processing Symposium (IDAP). IEEE, 2017.

[12] Liu K, Zhang M, Pan Z. Facial Expression Recognition with CNN Ensemble[C]//International Conference on Cyberworlds. IEEE, 2016:163-166.

[13] ZHAI YI-kui,LIU Jian. Facial Expression Recognition based on Transfer ring Convolutional Neural Network[J]. Journal of Signal Processing, 201 8,34(6): 729-738.

[14] Yuan Fang.Research of Facial Expression Recognition Based on Convolutional Neural Network[D].XIDIAN UNIVERSITY,2017.

[15] Xu L L, Zhang S M, Zhao J L. Expression recognition algorithm for parallel convolutional neural networks [J]. Journal of Image and Graphics, 2019, 24(02) : 0227-0236.

[16] Arriaga O , Valdenegro-Toro M , Plöger, Paul. Real-time Convolutional Neural Networks for Emotion and Gender Classification[J]. 2017.

[17] QIAN Yongsheng, SHAO Jie, JI Xinxin, et al. Multi-view facial expression recognition based on improved convolutional neural network. Computer Engineering and Applications, 2018, 54（24）: 12-19.