A Robust Improved Network for Facial Expression Recognition

Hao Gao, Bo Ma*

College of Electrical & Information Engineering, Southwest Minzu University, Chengdu 610041, China
Email: mabo@swun.edu.cn, 973356927@qq.com

Abstract. With the development of deep learning, even important progress has been made in the field of image classification and recognition. But facial expression recognition still faces many problems. This article is an experiment on the FER2013 dataset, the purpose is to get the facial expression attributes from the facial image. Because the pictures in this dataset have low resolution, and some pictures have no faces at all. This reduces the accuracy of facial expression recognition. In this paper, we propose a robust improved model. In this model, we introduce attention mechanism and separable convolution to improve the extraction of image features, and use data argumentation techniques to enhance the generalization ability of the model. The model obtained 65.2% test set accuracy on the FER2013 dataset.

Keywords: attention mechanism, separable convolution, FER2013

1 Introduction

Facial expression recognition is a hot topic in computer vision. As a direct expression of human emotions, facial expressions are a form of nonverbal communication. The main application fields of facial expression recognition technology include human-computer interaction, security, robot manufacturing, medical treatment, communications, and automobiles. In emerging applications such as human-computer interaction, online distance education, interactive games, and intelligent transportation, an automatic facial expression recognition system is necessary. The key point of facial expression recognition is the extraction of facial expression features. For the extraction of facial expressions, two types of feature extraction methods have emerged. One is based on traditional artificially designed expression feature extraction methods, such as local binary patterns, oriented gradient histograms, scale-invariant feature transformation, etc. These methods are not only difficult to design, but also difficult to extract high-order statistical features of images. The other is an expression feature extraction method based on deep learning. At present, deep neural networks have been widely used in various fields such as images, speech, and natural language processing. In order to adapt to different application scenarios, more and more deep neural network models have been proposed, such as AlexNet, VGG, GoogleNet and ResNet. These network models are widely used in various fields, and have achieved good results in facial expression feature extraction and classification.

Yu and Zhang [1] pre-trained the model on the FER2013 dataset and fine-tuned the model on the Wild 2.0 [2] dataset. They used three combined face detectors to detect and extract faces in the Wild 2.0 dataset. Then they proposed a method of data perturbation and voting to increase the generalization ability of the model. The model achieved an accuracy of 0.612 on the FER2013 dataset. Kahou et al. used a model based on CNN-RNN architecture and achieved an accuracy of 0.528 [3]. In FER2013 challenge of the ICML 2013 Representation Learning [5], Tang introduced linear support vector machine (SVM) in CNN for facial expression recognition [6]. They used simple CNN and SVM instead of softmax classifier, and the model won the first place in the challenge at that time. In 2016, Zhou et al. [4] proposed a multi-scale CNN model composed of three networks of different scales, and obtained the final classification results using later fusion techniques.

In this paper, we propose an improved network model based on the attention mechanism, which enhances the ability to extract image features by adding attention mechanisms. At the same time, data argumentation technology is used to enhance the generalization ability of the model. And introduce separable convolution to reduce the amount of training parameters. This model has achieved considerable accuracy on the FER2013 dataset.

The structure of the paper is as follows. Section 2 introduces the dataset used in the experiment and

Copyright © 2020 Isaac Scientific Publishing

the division of the dataset. Section 3 describes the improved network model. Section 4 analyzes the experiment and results. Section 5 is a summary.

2 Dataset

In this experiment, we used the FER2013 dataset. The FER2013 dataset has a total of 35,887 images, which are composed of 28709 training images, 3589 public test images and 3589 private test images. Each picture is a grayscale picture with 48*48 pixels. The distribution of its categories is shown in Figure 1. There are 7 expressions in the FER2013 database: anger, disgust, fear, happiness, sadness, surprise and neutrality. We selected 9 pictures from it, as shown in Figure 2.

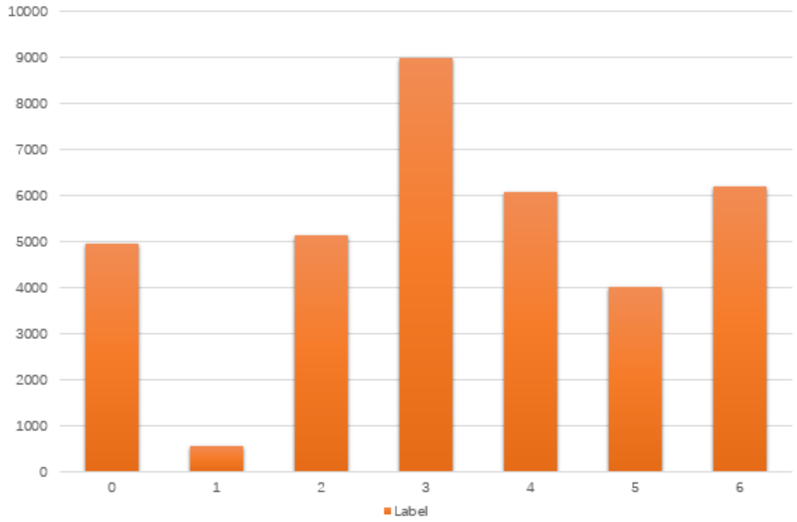


Figure 1. Distribution of FER2013 dataset categories

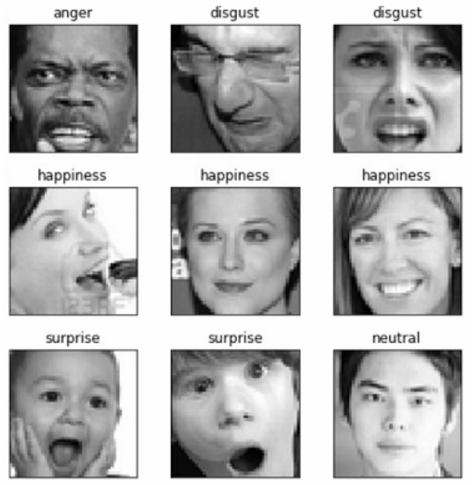


Figure 2. Images in FER2013 dataset

The database is the data of the 2013 Kaggle competition. Since this database is mostly downloaded from web crawlers, there are various pictures in it, face pictures, anime pictures and error pictures, so there is a certain error in itself, this error will cause The algorithm has poor effect on the classification and recognition of pictures in the FER2013 dataset. The error picture is shown in Figure 3. In the experiment, we added these images with errors to the model training and testing to enhance the robustness of the model. Before training the model, we first divide the dataset into a training set, a

Copyright © 2020 Isaac Scientific Publishing

validation set, and a test set. The role of the training set is a sample set used for training, which is mainly used to train the parameters in the neural network, the verification set is a sample set used to verify the performance of the model, and the test set is used to objectively evaluate the performance of the neural network. We select 80% of the total dataset for model training, and the rest as validation set and test set.

Figure 3. Pictures with errors

3 The Improved Model

In order to better perform feature extraction on data, we introduce attention mechanism and short connection in the network. To deal with the problem of too large parameters, we use separable convolution instead of ordinary convolution. At the same time, dropout [10] and batch normalization [11] are introduced to prevent overfitting.

3.1 Attention Mechanism

The basic idea of the attention mechanism [9] is to ignore irrelevant information and focus on key information. In the network design, we use the Squeeze-and-Excitation attention mechanism module in the SeNet [12] network, which mainly increases the attention on the channel. The attention mechanism module is shown in Figure 4.

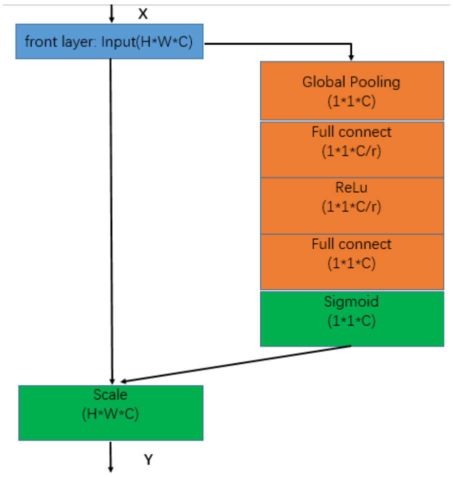


Figure 4. Attention mechanism module

We connect the output of the upper layer as the input to the attention mechanism module. First, through the global average pooling operation, the input H*W*C is compressed to 1*1*C, which is equivalent to compressing H*W into one dimension. After H*W is compressed into one dimension, it is equivalent to this dimension parameter to obtain the previous H*W global vision, and the perception area is wider. After obtaining the representation of 1*1*C, the first fully connected layer compresses 1*1*C to 1*1*C/r, reduces the amount of calculation, and is followed by Relu to perform nonlinear activation. The second fully connected layer is restored to the original C channels, followed by the sigmoid function, and the importance of each channel is obtained through sigmoid. Finally, the channel-by-channel weighting is applied to the previous features to complete the re-calibration of the original features in the channel dimension.

3.2 Separable Convolution

Separable convolution [8] is a type of convolution, which solves the standard convolution integral into deep convolution and 1x1 point convolution. Deep convolution applies a single convolution kernel to each single input channel for filtering, and then applies a 1x1 convolution operation to the point-by-point convolution to merge the output of all depth convolutions. This decomposition mode of separable convolution can effectively reduce the amount of parameters.

If the input channel is C1, the size of the convolution kernel is K*K, and the output channel is C_2. After standard convolution, its parameter is C_1*K*K*C_2. In the separable convolution, the depth convolution parameter is C_1*K*K, the point convolution parameter is 1*1*C_1*C_2, and the total separable convolution parameter is C_1*K*K+1*1*C_1*C_2. The parameters of standard convolution and separable convolution are shown in Equation 1. It can be seen from Equation 1 that the use of separable convolution can reduce the amount of parameters by K^2.

$$\frac{C_1 * K * K + 1 * 1 * C_1 * C_2}{C_1 * K * K * C_2} = \frac{1}{C_2} + \frac{1}{K^2} \tag{1}$$

3.3 Network Structure

The improved network model is shown in Figure 5, where the attention module is the attention mechanism module in Figure 4 and the block is shown in Figure 6. The network uses a standard convolutional neural network as the backbone network. In order to prevent the degradation of the network, which leads to the continuous decline in the accuracy of model training, we add short connections to the network. At the same time, due to the low image quality of the FER2013 dataset, the common network structure has weak feature extraction. In the network structure, we add the attention mechanism module to improve the feature extraction ability by assigning weights to the channels. At the same time, in order to improve the generalization ability of the model, we use data argumentation technology.

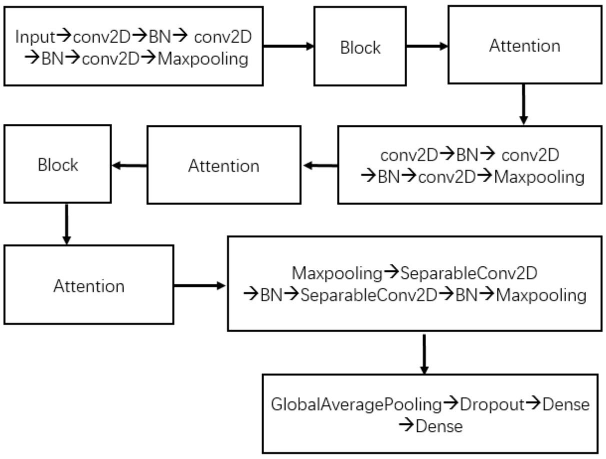

Input→conv2D→BN→ conv2D →BN→conv2D→Maxpooling
Block
Attention
conv2D→BN→ conv2D →BN→conv2D→Maxpooling
Attention
Block
Attention
Maxpooling→SeparableConv2D →BN→SeparableConv2D→BN→Maxpooling
GlobalAveragePooling→Dropout→Dense →Dense

Figure 5. The improved model structure

Copyright © 2020 Isaac Scientific Publishing

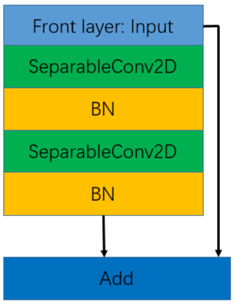


Figure 6. The block structure

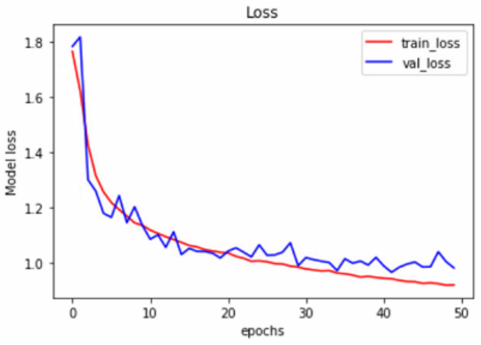
4 Experiments and Results

4.1 Experimental Environment

During the experiment, the training was conducted on a PC with 3.6 GHz Intel Core processors. Python was adopted as the programming language. The 1.14.1 version of tensorflow [7] is selected as the framework for neural network learning.

4.2 Experimental Results for FER2013 Dataset

In the experiment, we divided the FER2013 dataset into 3 datasets, in which the training set was used to fit the model for training. The purpose of the validation set is to find the best model. The test set is used to measure the performance and classification ability of the model. We choose 80% of the dataset for the training set, and the rest as the test set and validation set. During the training process, we use Adam as the optimizer and use its default parameters. And use the callback function EarlyStopping in keras, its function is to intercept the parameter model with the best saving result in the whole process of model training to prevent overfitting. Feature extraction was performed on the data through the model. After training 50 times, we obtained the best accuracy rate of 65.2% on the FER2013 test set. The accuracy and loss curves of the validation set on the FER2013 dataset are shown in Figures 7 and 8. As shown in table 1, it can be seen that the improved model in this paper has a significant improvement in test accuracy.



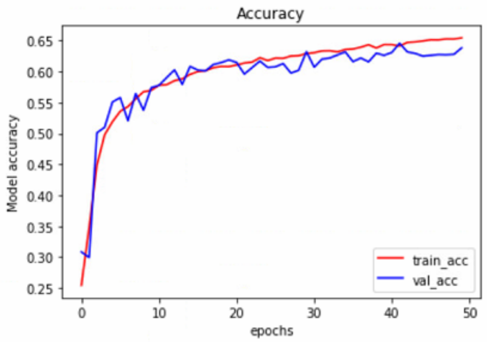


Figure 7. Model loss curve Figure 8. Model accuracy curve

Table 1. Compare recognition accuracy on FER2013 dataset

Algorithm	Test accuracy
Zhai Y et al [13]	59.10%
Tumen V et al [14]	57.10%
Liu K et al [15]	65.03%
Gan Y et al [16]	64.24%
Ours	65.20%

Copyright © 2020 Isaac Scientific Publishing

We use the model to generate a confusion matrix for the test set data. Through the confusion matrix, we can clearly see that the predicted value matches the true value. The confusion matrix is shown in the Figure 9. We use the trained model to perform attribute recognition on the test set pictures. We select 6 pictures from the recognition results, as shown in Figure 10.

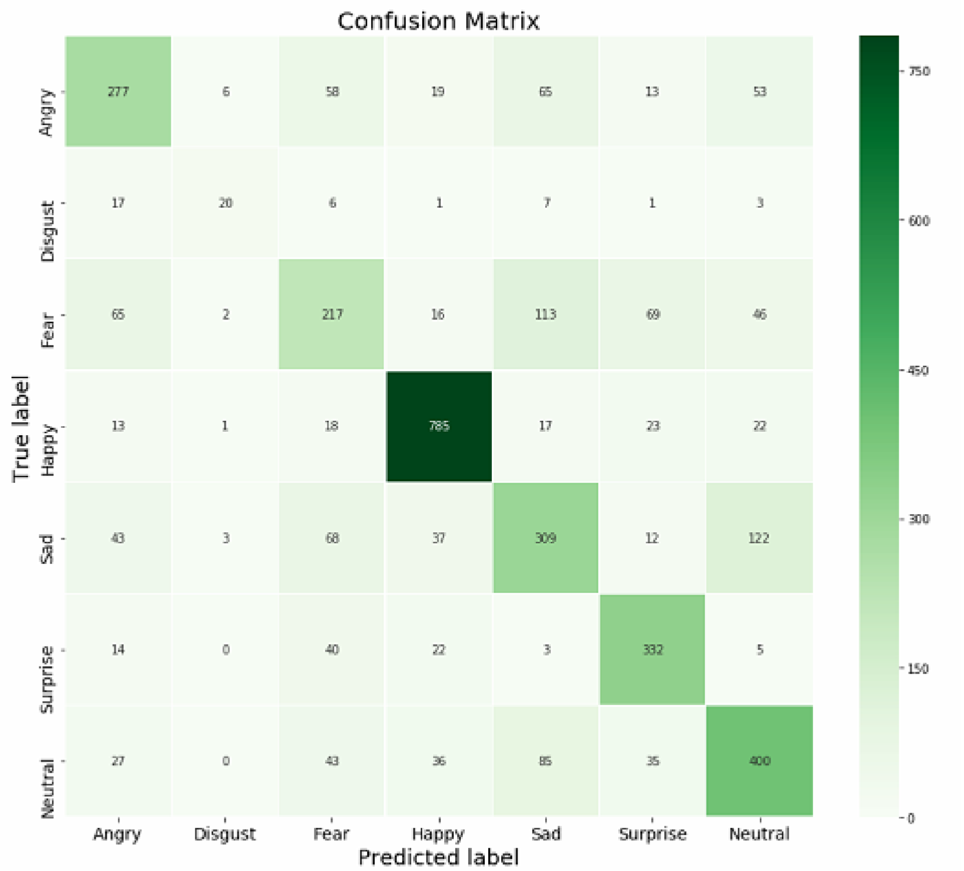


Figure 9. Confusion matrix of model on FER2013 testing set

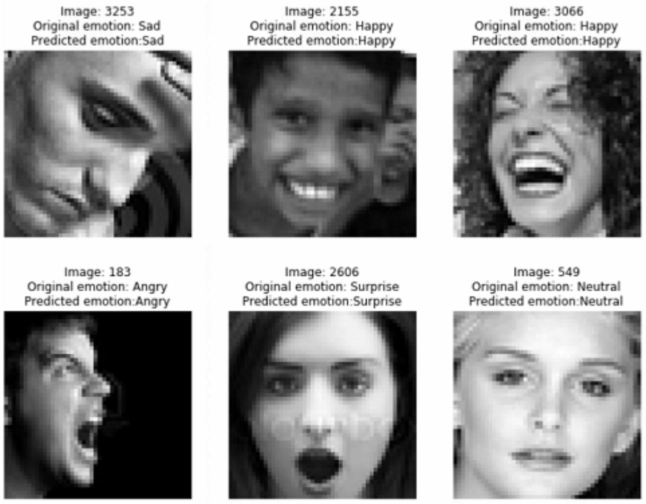


Figure 10. The results of test set pictures

Copyright © 2020 Isaac Scientific Publishing

5 Conclusion

In this article, we conducted a classification and recognition study on facial expressions. We propose a network structure that combines the attention mechanism and the short connection module through separable convolution. The attention mechanism is used in the network to enhance the feature extraction ability of the picture. At the same time, the separable convolution is introduced to reduce the amount of parameters, and the data is used to enhance Technology improves the generalization ability of the model. In the experiment, we achieved 65.2% accuracy on the FER2013 test set.

Because the combination of high-complexity networks and small data does not perform satisfactorily. Therefore, our next step is to do more work on dataset preprocessing and deep network construction to improve the recognition rate of facial expressions.

References

1. Zhiding Yu & Cha Zhang. (2015). Image based Static Facial Expression Recognition with Multiple Deep Network Learning. Proceedings of the 2015 ACM on International Conference on Multimodal Interaction. 435-442.
2. Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. (2011).Static Facial Expressions in Tough Conditions: Data, Evaluation Protocol And Benchmark, First IEEE International Workshop on Benchmarking Facial Image Analysis Technologies BeFIT, IEEE International Confer-ence on Computer Vision ICCV2011, Barcelona, Spain, 6-13 November 2011.
3. Ebrahimi Kahou, S., Michalski, V., Konda, K., Memisevic, R., and Pal,C. (2015). Recurrent neural networks for emotion recognition in video.Proceedings of the 2015 ACM on International Conference on Multimodal Interaction. 467-474. ACM.
4. S. Zhou, Y. Liang, J. Wan, and S. Z. Li, "Facial expression recognition based on multi-scale cnns," in Biometric Recognition, pp. 503–510,Springer, 2016.
5. P. Carrier and A. Courville, "Challenges in representation learning:Facial expression recognition challenge." https://goo.gl/kVzT48, 2013.
6. Y. Tang, "Deep learning using linear support vector machines," in Workshop on Challenges in Representation Learning, International Conference on Machine Learning (ICML), 2013.
7. Yuan Tang. 2016. TF.Learn: TensorFlow's High-level Module for Distributed Machine Learning. ArXiv preprint arXiv:1612.04251 (2016).
8. F. Chollet. Xception: Deep learning with depthwise separable convolutions. arXiv preprint arXiv:1610.02357v2, 2016.
9. Zhao M, Zhong S, Fu X, et al. Deep Residual Shrinkage Networks for Fault Diagnosis[J]. IEEE Transactions on Industrial Informatics, 2020, PP(99):1-1.
10. Wan L, Zeiler M, Zhang S, et al.: "Regularization of neural networks using dropconnect," Proceedings of the 30th International Conference on Machine Learning (ICML-13), 2013, pp.1058–1066.
11. S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network raining by reducing internal covariate shift. In ICML, 2015.
12. Hu J, Shen L, Albanie S, et al. Squeeze-and-Excitation Networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, PP(99).
13. Zhai Y, Liu J, Zeng J, et al. Deep convolutional neural network for facial expression recognition[J]. 2017.
14. Tumen V, Soylemez O F, Ergen B. Facial emotion recognition on a dataset using convolutional neural network[C]// 2017 International Artificial Intelligence and Data Processing Symposium (IDAP). IEEE, 2017.
15. Liu K, Zhang M, Pan Z. Facial Expression Recognition with CNN Ensemble[C]// International Conference on Cyberworlds. IEEE Computer Society, 2016.
16. Gan Y. Facial Expression Recognition Using Convolutional Neural Network[C]// the 2nd International Conference. 2018.