Face Expression Recognition Based on Improved Convolutional Neural Network

Quanming Liu
School of Computer and Information Technology, Shanxi University
Shanxi, China
+8618635185151
liuqm@sxu.edu.cn

Jing Zhang
School of Computer and Information Technology, Shanxi University
Shanxi, China
+8615536831209
854224610@qq.com

Yangyang Xin
School of Computer and Information Technology, Shanxi University
Shanxi, China
+8618435122778
1355640087@qq.com

ABSTRACT
Aiming at the problems of huge parameters and network degradation caused by simple linear stacked convolution layers or continuous full connection layers in traditional expression recognition methods, two convolution neural network models are designed through depth separation convolution and residual module respectively to widen and deepen the network. Firstly, model A adopts depth separation convolution instead of regular convolution layer, and the global average pooling layer replaces the final full connection layer, utilizes the methods of dropout, batch normalization, activation function of PReLU and image augmentation to avoid over-fitting effectively. Model B adopts pre-trained ResNet50 model to extract facial features, magnifies the images twice by the SRGAN method. Using ensemble method to fuse model A and B, the accuracy is further improved. To verify the feasibility of the method, the model was tested on the FER2013 facial expression dataset, and the performance was compared with the other facial expression recognition algorithms. The final results showed the improved convolutional neural network (CNN) reached the advanced precision of 73.244% in FER2013 dataset, and the experiment data and the number of model parameters all proved the effectiveness of this method.

CCS Concepts
• **Computing methodologies → Image processing**

Keywords
Facial expression recognition; CNN; Depth separable convolution; Residual module; Center loss

1. INTRODUCTION
In 1971, the psychologist Mehrabian [1] found that about 55% of the total contact information is conveyed by body language such as facial expressions, while less than 10% is conveyed by conversation, which proves capture and accurate recognition facial expression features is a very meaningful and challenging task. According to the contraction of facial muscles and the movement of facial organs, facial expressions can be classified as neutral, angry, disgusting, fearful, happy, surprised and sad.

AIPR 2019, August 16–18, 2019, Beijing, China

ACM ISBN 978-1-4503-7229-9/19/08…$15.00
DOI: https://doi.org/10.1145/ 3357254.3357275

Traditional facial expression recognition methods can be divided into face detection, facial landmark localization, feature extraction and facial expression recognition. However, the artificial design filters such as Gabor and LBP have limited recognition ability and poor robustness. The recognition rate is slightly lower in complex situations such as illumination change and face angle in real situations, and it takes a lot of time and effort.

The CNN model is an end-to-end training process. The network is trained layer by layer by gradient optimization, back propagation and other methods, the low-level features are continuously merged into deeper features, which is superior to the abstract features of artificial design and can accurately complete classification and recognition tasks. In the ILSVRC Challenge in 2014, the GoogLeNet [2] won the championship and the vggNet [3] won the second place. VGG model continues the classic network model AlexNet's [4] style of 5 convolution layers and 3 full connection layers. The small convolution kernel of 3×3 is used to replace the large size filter, and the maxpooling layer of 2×2 is added. However, the last three full-connection layers account for 89.36% of the total parameters, and the model size is 528 MB. The structure of Inception in GoogLeNet uses convolution cores of different sizes and pointwise convolutions of 1×1, which can combine non-linear features with stronger recognition ability. Using pointwise convolutions before conventional can reduce the number of features of the input, and the global average pooling layer replaces the last fully connected layer to effectively compress the model, so compared to the 16-layers of vggNet, the GoogLeNet model has a depth of 22 layers but reduces the number of parameters. Inception V2 [5] mainly adds batch normalization operation to adjust the intermediate neurons to the standard normal distribution, which solves the vanishing gradient problem caused by the change of the input probability distribution of each layers. Using two filters of 3×3 instead of the convolution core of 5×5 in Inception can reduce the computational cost. Inception V3 uses two asymmetric convolution cores to reduce the large convolution, which reduces the training time and increases the depth of the network. Inception V4 model has more uniform network structure and more Inception modules, which greatly improves the training speed of the model by adding residual connection structure. The Google team proposed the Xception [6] model, which is a linear stack of deep separable convolution structures with residual connections. Depth separable convolution layer decomposes the traditional convolution layer. First, the spatial convolution is performed independently on each channel by depthwise convolutions, and the output channel can be mapped to a new channel space by pointwise convolutions.

After the design and optimization of a series of CNN models, scholars have found that when the number of layers of the network

becomes deeper, further deepening of the network will slow down the convergence speed of the model and reduce the accuracy of classification because of network degradation. ResNet [7] model reduces the training difficulty of the network by adding residual learning module. It successfully reaches 152 layers of network depth and gets the first place in ILSVRC 2015 classification task competition. In the ResNet, the latter layer can directly learn the residual of the former network layer, and ingeniously solve the network degradation problem. Moreover, the identity shortcut connection skips over more layers to perform identical mapping, no additional parameters or computational complexity are added. In view of the fact that Inception module and residual module can widen and deepen the network model and improve the accuracy of expression classification. The expression recognition model designed in this paper uses Inception module and residual module respectively, and uses ensemble learning method to fuse the feature vectors derived from the network models, which further improves the performance of the model.

2. Two CNN MODELS

2.1 Depthwise Separable Convolution

The model A used in the facial expression recognition method is a lightweight CNN based on depthwise separable convolution. By using the depthwise separable convolution layer to increase the non-linear modules in the network, the parameters of the network are sufficiently reduced and the convergence speed of the model is accelerated. Taking convolution core of 3×3 as an example, the traditional convolution method jointly maps the cross-channel and spatial correlations, without dividing the input channels; The depthwise separable convolution first extracts features through depthwise convolution of 3×3, then uses pointwise convolution across the channel to map the input data to different spaces for fusion features. By using the Inception module to widen the network, more features can be extracted. The size of the convolution core is 3×3, N is the number of convolution kernels, M is the size of the input channels, K is the size of feature map.

The calculation amount of standard convolution lists as follows.

$$3 \times 3 \times M \times N \times K \times K \quad (1)$$

The calculation amount of depthwise separable convolution is

$$3 \times 3 \times M \times K \times K + M \times N \times K \times K \quad (2)$$

The result is $\frac{3\times3\times M\times N\times K\times K}{3\times3\times M\times K\times K+M\times N\times K\times K} = \frac{1}{N} + \frac{1}{9}$ (3)

The formula (2) shows that the cross-channel correlation and spatial correlation in the feature graph of CNN are fully decoupled by this method. By decomposing the traditional convolution into two steps such as formula (3), the computational complexity is reduced to 1/9 of the original parameters, which solves the problem of model convergence caused by network deepening and speeds up the training speed.

2.2 The Model Architecture

The architecture of the model is shown in Figure 1. The network can be divided into three modules: the entry flow, the middle flow and the exit flow. The depth of the network is 83 and 25 convolution layers are used. The parameters of the whole network architecture are 2720439. It consists of three convolution layers and 22 depthwise separable convolution layers. The middle flow uses four repetitive deep separable convolution layers. In order to improve the utilization of the whole parameters, the global average pooling layer is used to replace the full connection layer before the final layer to complete the classification and prediction of expression recognition. The model consists of depthwise separable convolution layers, PReLU activation function, batch normalization, maxpooling layers and global average pooling layers.

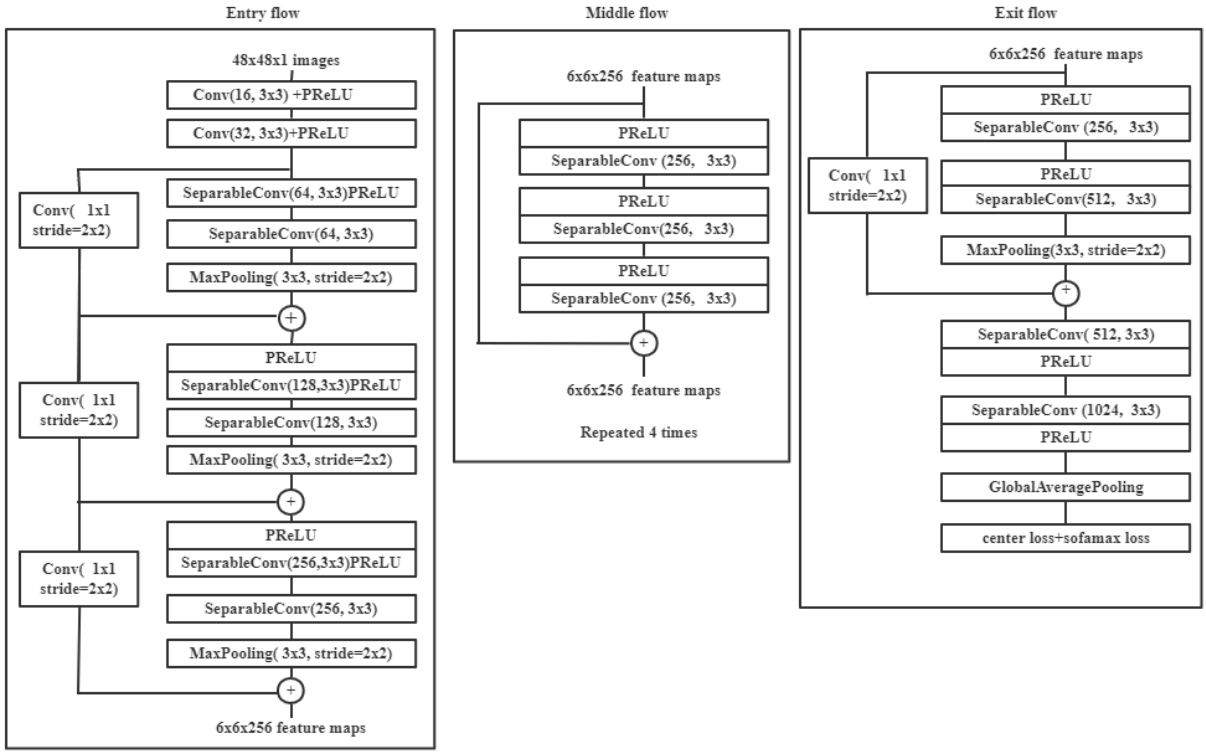

Figure 1. The model architecture.

The input data of the model is the original gray scale image of $48 \times 48 \times 1$. Two convolution layers of 3×3 and six deep separable convolution layers are used to extract features in the entry flow. Using batch normalization method to dynamically adjust network parameters and PReLU activation function to improve the ability of extracting non-linear features of the model. Two consecutive convolution kernels of 3×3 are used to replace the large convolution kernels of 5×5, which adds extra layers of convolution layer and activation layer to make the non-linear structure stronger. The size of the feature map is 6×6×256 in the entry flow. The middle entry consists of four repetitive depth-separable convolution modules, each module contains three depth-separable convolution modules, and the output of the feature map is 6×6×256. Two deep separable convolution cores and one spatial convolution layer are used in the exit entry, and the global average pooling layer is used instead of the traditional full connection layer. Finally, the multi-classification of facial expressions is realized by using the softmax activation function. The number of parameters is greatly reduced, which effectively reduces the computational complexity and the capacity of the model. Finally, the AdaMax algorithm is used to dynamically adjust the learning rate of the parameters, which further improves the stability of the model during the training process and the convergence speed of the network.

2.3 PReLU Function

In this paper, PReLU (Parametric Rectified Linear Unit) is selected as activation function. When α_i=0 PReLU will degenerate into ReLU function; when α_i is a small fixed value (such as α_i=0.01), PReLU will degenerate into LeakyReLU. The positive partial derivative of ReLU function is 1, which makes the weights of some inputs constant. This leads to the neglect of some effective features of the network, and the output offset of the function also affects the convergence of the network. The PReLU activation function improves the fitting ability of the model near the zero value by using the same calculation cost, effectively reduces the risk of over-fitting and the vulnerability of the nerve unit in the training process, and improves the recognition accuracy by using adaptive learning parameters.

2.4 Loss Function

The paper uses the center loss and softmax loss joint supervision method to classify the facial expressions. Softmax loss is usually used in the classification process of CNN, which is fast in calculation, but has a large dispersion distance within the class. Center loss increases the compactness of the class by reducing the sum of the squares of all samples from the center of the sample. The joint supervision of center loss and sofamax loss not only increases the distance between classes, but also reduces the compactness between classes and enhances the ability to distinguish expression classification. The formulation is given in Equation 4.

$$\mathcal{L} = \mathcal{L}_S + \mathcal{L}_C = -\sum_{i=1}^{m} log \frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T x_i + b_j}} + \frac{\lambda}{2} \sum_{i=1}^{m} \left\| x_i - c_{y_i} \right\|_2^2 \quad (4)$$

In Equation 4, x_i denotes the ith deep feature which belonging to the y_ith class. w_j denotes the jth column of the weights in the fully connected layer and b is bias, m is the size of mini-batch and n is the number of class. The c_{yi} denotes the y_ith class center of deep features and λ is used for balancing the two loss functions.

3. RESNET50

The model B used in the facial expression recognition method in this paper is a pre-trained ResNet50. In order to improve the generalization of face recognition in different sizes of the model, the super-resolution method is used to enlarge the images by twice for training.

3.1 Transfer Learning

When training a network model for a specific task, it will be restricted or affected by the quantity, quality and even distribution of training data, and it will waste a lot of computing resources and time to adjust the network weight. At this point, it becomes necessary to transfer the useful knowledge learned from the source domain to the target domain [8]. Therefore, the paper transfers and fine-tunes the ResNet50 model with pre-training weights for facial expression recognition. The pre-training weights are trained by the ImageNet datasets, which can fully represent the image features and information that are ubiquitous in image classification tasks, thus realizing fast learning and saving enormous resource overhead. The ResNet50 model has the size of 99 MB and a parameter of 25636712. It contains 49 convolution layers and one full connection layer. The number of ResNet50 models is 99 MB and the parameters are 25636712, which includes 49 convolutional layers and one fully connected layer. ResNet can be divided into 5 segments, the number of residual blocks in each module is different, and each residual block contains 3 convolution layers. Therefore, there are 1+3×(3+4+6+3)=49 convolutional layers.

3.2 SRGAN

All the pictures in the Fer2013 dataset are grayscale images of size 48×48 with low resolution, and the details of facial expressions are not fully expressed. Experiments show that the classification effect in ResNet network is better when the image is magnified twice. In recent years, most of the super-resolution algorithms are based on deep learning. SRCNN is the first deep learning algorithm successfully applied in super-resolution, which has been greatly improved compared with previous methods, but still has the problem of texture blurring. In 2017, Christian Ledig [9] used the method of generative adversarial networks. Even if the image magnification is more than 4 times, the image will remain realistic and natural, retaining more detailmore details. This paper uses SRGAN method and ResNet as generator for training. SRGAN uses perceptual loss function to enhance the authenticity of images, which is composed of content loss and adversarial loss. The enlarged image is shown in Figure 2, which shows that more details and textures are preserved.

Figure 2 Super-resolution results

4. EXPERIMENT RESULTS AND ANALYSIS

To verify the recognition effect of the designed network model on facial expression, this paper chooses the FER2013 public database to carry out the experiment and test of facial expression

recognition. The operating system chooses the Ubuntu 18.04 system, builds the deep learning framework of TensorFlow-Gpu==1.5 and the keras==2.1.5 under the Python 3.6 environment, installs the acceleration libraries cuDNN V7 and CUDA 9.0 for GPU computing, and the computer hardware configuration is Intel Xeon E5-2637 and 64 GB memory, and NVIDIA Tesla K40C with 12GB memory.

4.1 Training Data

The experimental dataset in this paper is the FER2013, which is the public dataset of facial expressions in kaggle competition. FER2013 is sorted out by Google image search tool and divided into seven basic expressions: angry, disgust, fear, happy, sad, surprise and neutral. It contains 28,709 training images, 3589 PublicTest images and 3589 PrivateTest images, all of them are 48×48 pixel grayscale images and the face is basically centered. Because the pictures in FER2013 are collected from the web and not deliberately shot, the features of the same kind of expressions is not standard. There are a large number of side faces or hand occlusion of facial organs and other behaviors in the dataset, and there are interference from unrelated pictures such as text, pure chromatograms or cartoons. Currently, the average recognition rate of humans on this database is 65%±5% [10].

4.2 Implementation Details

28709 pictures in FER2013 are used for model training, but the classification of these facial expression data is not balanced. There are only 436 disgusting expressions, which make the network lack enough sample data for deep feature extraction and training. Therefore, this paper uses image data generator methods such as random horizontal flip, random rotation, horizontal and vertical offset and random scaling. Expanding the number of pictures in the training set to specify the number of epochs while retaining the relevant facial features effectively. The robustness of the model and the recognition ability of the network are improved.

In training, the data augmentation method mentioned above can still lead to over-fitting phenomenon. In this paper, the image is randomly clipped to the size of 44×44, and the data can be expanded to 16 times by cropping the image and mirroring, which effectively reduces the over-fitting and improves the recognition accuracy. The picture obtained by random cropping is shown in Figure 3.

After many experiments, the initial learning rate of the model is set to 0.01, the size of mini-batch is 64, and the number of iterations is 250. The AdamMax optimization algorithm is used to dynamically adjust the learning rate to avoid falling into the local optimal solution.

Figure 3 Random crops

Figure 4 shows the feature map obtained by the guided backpropagation. In the process of training, features change from simple to complex and finally to more abstract and comprehensive facial organs. It can be found that the final feature map generated by the model is very sensitive to the changes of eyebrows, mouth and facial muscles, and can fully capture facial expressions.

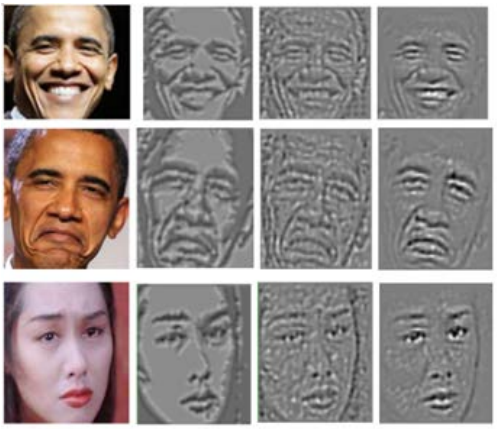

Figure 4 Guided back-propagation visualization

4.3 Experimental Results

The recognition results of FER2013 dataset are shown in Table 1. The recognition rate of this method in FER2013 database is the highest at 73.244% and the network depth is 83, but the number of parameters is only 18.5% of VGG16 model and 13.6% of VGG19 model. Compared with the champion of the Kaggle competition, the RBM team has 69.768% in the Public Test and 71.161% in the Private Test. The method of this paper has increased by 1.84% in PublicTest and by 2.083% in the Private Test. This method can effectively find subtle changes between different expressions and reach advanced levels.

Tabel 1 Facial expression recognition rate

Model	Public Test	Private Test	Parameters	Depth
VGG16	0.65589	0.660072	14714688	22
VGG19	0.6644079	0.6777821	20024384	25
Improved model	0.71608	0.73244	2720439	83

The confusion matrix of predicted expression and real expression in the Public Test and Private Test is shown in Figure 5. The research found that the highest recognition rate for happy expressions was 91%, followed by the accuracy rate of 83% for surprised, and the lowest recognition rate for fear was 56%. Fear is easily confused with sad and anger, with 18 percent being mistaken for sad and another 10 percent for anger. Because angry people scratch their cheeks with their hands and open their mouths to shout. When they are sad, they cover their faces with their hands and open their mouths to cry. These are very similar to fear expressions that cover their faces and close their eyes at the same time. In addition, some pictures with blank eyes and expressionless sad expressions are easily mistaken for neutral expressions. The expression images in FER2013 dataset are gray scale images of 48×48, it is impossible to clearly describe the organs and facial muscles. Therefore, the overall recognition rate of 73.244% has achieved good recognition.

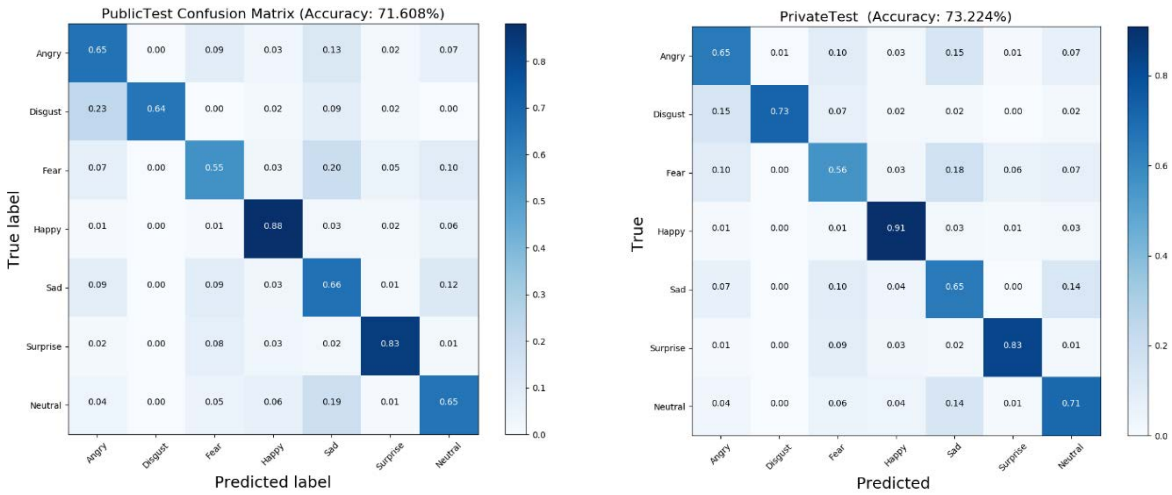

Figure 5 Normalized confusion matrix.

5. CONCLUSION

In this paper, a convolution neural network model is designed, which combines the depth separable convolution and residual module. It fully reduces the redundant parameters and enhances the effectiveness of features. Experimental results show that the accuracy of the facial expression recognition method in FER2013 reaches 73.244% and reaches the advanced level. The high recognition rate is guaranteed, and the validity of the model method is proved. However, the model is still inadequate for subtle facial muscle changes and facial occlusion. Therefore, the classification of different subtle expressions representing the rich emotional changes in human emotions is still a new challenge, which is the focus of the follow-up research in this paper.

6. ACKNOWLEDGMENTS

This paper is supported by National Natural Science Fund (No. 61673295，U1805263)，Scientific research of returnees from Shanxi Province (No. 2016-004).

7. REFERENCES

[1] Mehrabian A. Communication without words. Psychology Today,1968,2(4):53-56

[2] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ:IEEE Press, 2015:1-9.

[3] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition// 3rd International Conference on Learning Representations (ICLR). Hilton San Diego: Computer Science, 2015:1150-1210.

[4] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks //Proceedings of the 25th International Conference on Neural Information Processing Systems. New York: ACM Press, 2012,1:1097-1105

[5] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the Inception Architecture for Computer Vision // Computer Vision and Pattern Recognition. IEEE, 2016:2818-2826.

[6] Chollet F. Xception: deep learning with depthwise separable convolutions / /Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA:IEEE, 2017:1800-1807.

[7] He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition //IEEE International Conference on Computer Vision. 2015:770-778.

[8] Long M, Wang J, Cao Y, et al. Deep Learning of Transferable Representation for Scalable Domain Adaptation. IEEE Transactions on Knowledge & Data Engineering, 2016,28(8):2027-2040.

[9] Ledig C, Theis L, Huszar F, et al. Photo-realistic single image super-resolution using a generative adversarial network // CVPR 2017: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Washington, DC: IEEE Computer Society, 2017:105-114.

[10] Goodfellow I, Erhan D, Carrier P L, et al. Challenges in representation learning: a report on three machine learning contests. Neural Network, 2015,64(1):59-63.