Facial Expression Recognition Based on Convolutional Neural Network

Ruhua Lu1, Yalan Li2*, Pan Yang2, Wenfen Zhang1

1 College of Software and Communication Engineering, Xiangnan University, Chenzhou, Hunan 423000, China

2College of Software and Communication Engineering, Xiangnan University, Chenzhou, Hunan 423000, China

*liyalan@xnu.edu.cn

Abstract. At present, facial expression recognition technology is widely used in artificial intelligence, transportation, medical and other aspects, so it has important research value. Traditional facial expression recognition uses manual feature extraction method with low accuracy and weak generalization ability, which is difficult to be applied in real life. With the development of deep learning, convolution neural network appears in people's vision. Different from traditional manual feature extraction, convolution neural network can learn image features independently, and learn more features. In addition, it has the advantage of sharing the weight with the neural network. Although convolution neural network has multiple advantages, it also has some disadvantages, especially over fitting. In this paper, the model of convolution network is improved based on the classical VGGNet according to the working principle of convolution neural network. In this new model, the number of convolution kernels is reduced in parameters, and the global average pool layer is used to replace the full connection layer in the structure, so as to reduce the degree of over fitting and decrease the operation parameters. Finally, experiments show that the accuracy, generalization and consumption of resource are enhanced in the new model. It is proposed that the new method is better than the traditional convolution network VGGNet.

Keywords: Convolution Neural Network; Facial Expression Recognition; Over Fitting

1. Introduction
Facial expression is the main way of non-verbal communication between people. The change of facial expression can directly reflect the emotional change, and even reflect the personality and psychological state of people. Facial expression recognition is a technology to extract the features of human face in the image, then analyze and process the extracted features, so as to obtain the emotions such as happiness, anger, surprise and so on[1]. In the year of 1978, facial expression recognition is carried out in an animation video, then facial expression recognition was first applied to computer image processing by Suwa[2]. This technical core is to realize automatic recognition of facial expression through image sequence. In the early 1990s, K. mase and A. Pentland proposed to use the streamer method to determine the main movement direction of muscles, and extract the streamer value as the facial expression feature to realize the expression recognition function[3]. After that, automatic facial expression recognition entered a new era. In 2009, Waseda University of Japan developed the Kobian robot, which can use seven kinds of expressions to interact with people [4]. In 2000, MIT Media Laboratory of Massachusetts Institute of Technology developed a robot Kismet, which can recognize facial expressions, and then imitate the recognized facial expressions [5]. In 2015, Italian designers Simone rebaudengo and Paul Adams designed the unmask anti haze mask, which can recognize facial expressions and then convey people's expressions through the external LED screen [6]. With the development of facial expression technology, more and more scholars begin to study how to use more efficient and accurate methods to realize the function of facial expression recognition[7-8]. At present, it is widely used in intelligent monitoring, online teaching, medical services, safe driving and other daily life fields.

2. The Main Structure of Convolution Neural Network
The main structure of convolution neural network is composed of convolution layer for extracting image features, nonlinear activation function, pool layer for compressed data and full connection layer for outputting classification information.

2.1. Convolution Layer
The function of convolution layer in convolution neural network is to extract features for classification in image. The specific operation is that all the pixels in the local window centered on the pixels in the image are inner product with the convolution kernel. The operation of inner product is to multiply the data in the convolution kernel with the pixel values at the corresponding position in the local window, and then add them to produce a new pixel instead of the central pixel. At last, the local window convolutes the whole feature map with a specified step size to get a new feature image, which is send to the next layer.

When the convolution kernel is convoluted with the pixels on the edge of the image, because a part of the convolution kernel is beyond the image range, no pixels are multiplied with it. The solution to this situation is to artificially fill a number of 0 in the periphery of the input image to multiply it with the convolution kernel, so that the size of the new image will not change.

2.2. Activation Function
In real life, most of the problems are nonlinear, and the neurons in neural network can only carry out the weighted operation of data, so its essence is linear operation. The function of activation function is
to nonlinear the neural network and simulate the stimulated response of human brain nerve. When the output of neurons reaches a certain threshold, the behind neurons can be just activated. The traditional activation functions include Sigmoid function, Tanh function and ReLU function. ReLU function is one of the most commonly used activation function[9], which is an unsaturated piecewise function. Its formula is as follows.

\[ f(x) = \max (0, x) \]  

In above formula, the input x is compared with 0. x is output when x is greater than 0, and 0 is output when x is less than 0. ReLU function only compares with numerical value, therefore, it is simpler and faster than traditional activation function.

2.3. Pool Layer
The role of pool layer is to reduce the size of input feature map and at the same time keep the depth dimension of feature map unchanged. If the size of the feature map is not limited, the computational complexity of convolution layer will always be very large. If the size of the feature map is too large, the training parameters of the full connection layer will be too much, which will lead to the aggravation of over fitting phenomenon. Too large size of the feature map also lead to that the calculation amount will increase and the training time will be longer.

The operation mode of pool layer is to scan the whole feature map as a specified size window with step size greater than 1, and fuse the data in the window, so as to obtain a new feature map with reduced size. There are two ways of data fusion in the window: maximum pool and average pool. In this paper, we use the maximum pool, because when the input data of ReLU is negative, the output of 0 is an invalid neuron, and when the output is positive, the negative value can be avoided as far as possible. Moreover, the larger the activation value of the neuron, the more intense the response of the neuron to the window data, and the better the extracted features. Therefore, replacing all the data in the window with the maximum value can be regarded as a purification operation.

2.4. Full Connection Layer
The full connection layer is the last layer in the convolution neural network. With the movement of the network, the feature matrix of the image will be obtained after the feature extraction of the convolution layer and the feature purification of the pool layer. The function of the full connected layer is to synthesize the final extracted feature matrix to get the score vector of classification, which can be a single layer or multi-layer superposition. The full connection layer does not use local connection, but uses full connection to connect each node with all nodes of the previous layer, which is used to synthesize the features extracted from the front edge and output the score vector. There are two ways to realize the full connection layer. One is to stretch the input multi-dimensional data into one-dimensional data, and then carry out full connection calculation. The other is to use the convolution layer to realize the full connection layer. Assuming that the size of the input feature map is 3×3×128, and the output score vector of seven categories namely the output size is 1×1×7, then the convolution kernel of size can be 3×3×7, which used to convolution with the input feature map. Without filling and scanning, the output size of feature map 1×1×7 is obtained.

3. The Model of Facial Expression Recognition System
3.1. Facial Expression Data

3.1.1. Data Sources
The data sources of this paper are mainly from the latter three common facial expression recognition databases. The first is FER2013 database, it is composed of 35886 facial expression images, including 28708 in training set, 3589 in test set and 3589 in verification set. The size of each picture is 48×48, all of which are gray scale images. There are seven types of expression in the database, namely anger, disgust, fear, happy, sad, surprised, normal. But there are many wrong labels, this paper uses the relabeled FERPlus tag. The second is JAFFE database, which was released in 1998. There are 213 facial images in the database, including seven kinds of expressions, which is the same as the first database. The size of each picture is 256×256. The third is CK+Database, which was released in 2010 and can be obtained free of charge. This database includes 123 subjects and 593 image sequences. The last frame of each image sequence has the label of action units. Among the 593 image sequences, 327 sequences have emotion labels.

3.1.2. Data Processing
After comparing the data obtained from the above databases, it is found that the forms of data in different databases are different, mainly in the size of the pictures and the format of saving. It is not feasible to use these data directly for convolution network training. The pictures must be processed in order to unify the size and format of the pictures. Since there are most data in FER2013 database, the data size and format in FER2013 database are taken as the unified standard, that is, the size of all pictures is zoomed to 48×48, and the saved format is PNG.

In addition to the size and format of the image, it also needs to be preprocessed. The two most commonly used preprocessing methods are centralization and normalization. These two methods can eliminate the errors caused by different dimensions, self variation or large difference in values. The specific operation of image centralization is that the value of each element subtracts the average value of the entire image. The operation of centralization is essentially a translation process, so that the center of all data is (0,0). The specific operation of normalization is divided by standard deviation, which is attribute by attribute operation liking centralization. Normalization limits the data range to [-1,1]. The combination of data centralization and normalization is the standardization of data. The standardized operation makes the mean value of data is 0 and the variance is 1. The mathematical formula is formula 2.

\[ \hat{x} = (x - \mu)/\sigma \]  

(2)

In formula 2, \( \mu \) is the mean value and \( \sigma \) is the standard deviation.

3.2. Convolution Neural Network Model

3.2.1. Classical Model VGGNet
VGGNet is a deep convolution neural network researched by researchers from visual geometry group and Google deepmind company. VGGNet explores the relationship between the depth and performance of convolution neural networks. By stacking small convolution kernels with size 3×3 and
the maximum pool layer with size 2×2, VGGNet successfully constructs deep convolution neural networks with 16-19 layers.

VGGNet model has five convolutions, each of which is composed of 2-3 convolution layers. In these five convolutions, the number of convolution kernels is 64, 128, 256, 512, 512. The size of all convolution kernels is 3×3, and each convolution is followed by a maximum pool layer. The last three layers are full connection layers, at the same time, the neuron formats are 4096, 4096 and 1000. Dropout is used between the full connection layers and the softmax classifier is used for classification.

The biggest characteristic of VGGNet is that the convolution kernel of VGGNet with the size 3×3 replaces the convolution kernel with the size 5×5 and 7×7. The substitution rule is that the superposition of two convolution layers with the size 3×3 can be equivalent to one convolution layer with the size 5×5, and the superposition of three convolution layers with the size 3×3 can be equivalent to one convolution layer with the size 7×7. Since each convolution layer is nonlinearized by using activation function, multiple convolution layers with the size 3×3 have more nonlinear transformations than single convolution layers with the size 5×5 or 7×7, which makes the model more capable of learning features and has a receptive field.

3.2.2. Improved Model of VGGNet

The improved model in this paper bases on the VGGNet model of the class D, the input of which is color picture with the size 224×224, and the output is classified into 1000 categories. In this paper, we need to modify the parameters of the VGGNet model, and retain the overall structure of the model, which not only retains the five convolution segments and the last three full connection layers. Firstly, the size of input data is modified as 48×48×1, of course, the size of convolution kernel is not modified, but the number of convolution kernels is modified. The increasing rule of convolution kernels is that the number of convolution kernels is changed from 64, 128, 256, 512, 512 to 16, 32, 64, 128, 128, and the size of neurons in the whole connection layer is modified in turn to 256, 256, 7. The modified structure is shown in Fig.1.
Because there are many parameters when the convolution layer is transformed into the full connection layer, it is easy to increase the over fitting phenomenon, which leads to the decrease of the generalization ability of the model. It has been known that the main function between convolution layer and full connection layer is to transform the output characteristic graph of convolution layer with size $3 \times 3 \times 128$ into 7-dimensional score vector. In order to achieve this effect, we can first convert the feature map of $3 \times 3 \times 128$ to the feature map of $1 \times 1 \times 128$, and then output the classification information from the full connection layer of seven neurons. In order to complete the transformation from feature graph of $3 \times 3 \times 128$ to feature graph of $1 \times 1 \times 128$, each feature graph with size $3 \times 3$ can be processed to a value, and finally the feature map of $1 \times 1 \times 128$ can be output. Taking the average value of each feature graph is called global average pool [10]. At the last convolution layer, the two full connection layers are replaced with a global average pool. In Adam optimization algorithm in the improved model, the learning rate is set to 0.001, $\mu$ is set to 0.9, $\mu_2$ is set to 0.999, $\text{esp}$ is set to $10^{-8}$, and the regularization strength of L2 regularization is set to 0.01.

4. Experimental Simulation

In this paper, we mainly modify the number of convolution kernels in VGGNet model and keep the whole structure of the model unchanged. Then we improve the structure of the model on the basis of the modification, using the global average pool layer to replace part of the full connection layer. The performance of the two models can be compared through the results of training and testing.

4.1. Comparison of Loss Curves

By observing and comparing the loss curves, we can know the over fitting phenomenon of the model and the convergence speed of the model. The loss curves of the two models in this paper are shown in Fig.2.

![Fig. 2 Loss Curves](image)

The left side is the loss curve of the classical model, and the right side is the loss curve of the improved model. By observing the loss curves of the two models, it can be found that the loss curves of the two models are very similar. There is no "U" shape in the loss curves of the two models, but it tends to be flat when it drops to about 0.6. However, there is any gap between the loss of the training
set and that of the test set, which indicates that the two models have over fitting phenomenon but some inhibition. After 200 times of training, the loss of the classical model training set is 0.4362, and the loss of the test set is 0.6276. The loss of training set and test set is 0.4010 and 0.6098 respectively. By comparing the two models, it can be found that the loss of training set and test set of the improved model has decreased, and the observation curve can also find that the improved model converges faster.

4.2. Accuracy Comparison
The performance of the model can be directly reflected by comparing the accuracy rate of the model in the data set. Tab.1 shows the accuracy rate of the two models in different data sets.

| Models      | Training Set | testing set | CK+ set |
|-------------|--------------|-------------|---------|
| Classical Model | 84.31%       | 79%         | 82.23%  |
| Improve Model    | 84.74%       | 79.94%      | 83.57%  |

**Tab. 1 Accuracy**

Compared with the accuracy of training set and test set, it can be found that the improved model accuracy is higher than the classical model, which indicates that the performance of the improved model has been improved. Compared with the accuracy in ck+ data set, the improved model accuracy has also been improved, which indicates that the generalization ability of the improved model has also been improved.

The training parameters of the classical model are up to 1283255, while the improved training parameters are 921399, and the improved model parameters are 361856 less, which reduces a large number of training parameters, which means that the training time of the model will be shorter and the utilization of computer resources will be less.

5. Conclusion
This paper makes reference to the classical convolution network VGGNet, and improves the network structure. Firstly, the original network model structure is kept unchanged and its parameters are modified. Then, the performance of the model is improved by improving the structure of the model and replacing the full connection layer with global average pool. According to the comparison results of the two models, the improved model can be strengthen in the accuracy, generalization and resource consumption. This proves this new model has been enhanced.

**Acknowledgements**
This work has been supported by Hunan Education Department Scientific Research Key Project(17A201) and the Scientific Research Project of Hunan Provincial Education Department(19C1716,19C1723).

**References**
[1] Lewis A K , Porter M A , Williams T A, Bzishvili S and Payne J M 2017 Facial emotion recognition, face scan paths, and face perception in children with neurofibromatosis type Neuropsychology 31 361
[2] Yunxin H, Fei C, Shaohe L and Xiaodong W. 2019 Facial Expression Recognition: A Survey
Symmetry 11 1

[3] Mliki H, Fendri E and Hammami M 2015 Face Recognition Through Different Facial Expressions J. Signal Process. Sys. 81 433

[4] Goldman A I and Sripada C S 2005 Simulationist models of face-based emotion recognition Cognition 94 193

[5] Yuan L, Caiming W and Yi Z 2013 Facial expression recognition based on fusion feature of PCA and LBP with SVM Optik-International Journal for Light and Electron optics 124 2767

[6] Turk M and Pentland A 1991 Eigenfaces for Recognition J. Cognitive Neurosci. 3 71

[7] Jinpeng L, Shuang Q, Yuanyuan S, Chenglin L and Huiguang H 2020 Multisource Transfer Learning for Cross-Subject EEG Emotion Recognition IEEE trans. Cybern. 50 3281

[8] Grosefifer J, Lobel M, diFilipo D and Gordon J 2020 Low Spatial Frequency Sensitivity and Emotional Face Processing in Adolescents: An Event-related Potential Study Dev. Neuropsychology 45 1

[9] Shojacilangari S, Wei-Yun Y, Nandakumar K, Jun L and Eam Khwang T 2015 Robust representation and recognition of facial emotions using extreme sparse learning IEEE trans. image process. 24 2140

[10] Cruz A C, Bhanu B and Thakoor N S 2014 Vision and Attention Theory Based Sampling for Continuous Facial Emotion Recognition. IEEE Trans. Affect. Comput. 5 418