Research on Facial Expression Recognition Based on Improved Deep Residual Network Model

Ji Chen¹, Xiaoshu Luo*, Zhiming Meng, Wanting Li

College of Electronic Engineering, Guangxi Normal University, Guiling, Guangxi, 541000, China

¹email: 1047986497@qq.com,
*Corresponding author’s e-mail: lxs@mailbox.gxnu.edu.cn

Abstract. Facial expressions are the main external manifestations of human emotions. Facial expression recognition technology can be used in medical, investigation, education and other application scenarios. Aiming at the disadvantages of slow convergence speed and low recognition accuracy of traditional neural networks, in order to optimize the network more efficiently and improve the recognition rate of facial expressions, a facial expression recognition method based on improved convolutional neural networks is proposed. First, the deep residual network ResNet18 is improved, and then the improved ResNet18 network is used to extract global expression features of face images, and then the self-attention weighting module is introduced to calculate the expression features of each face image and output one. The corresponding weights are used to weight the loss function. Finally, the two public expression data sets of CK+ and RAF-DB are used to experimentally verify the method in the article. The accuracy of facial expression recognition reaches 98.89% and respectively 87.13%, compared with the original deep residual network ResNet18 model and other types of network models, the facial expression recognition rate has been significantly improved, which proves that this method can effectively improve the accuracy of facial expression recognition and has certain application value.

1. Introduction

Facial expression recognition (FER) is an important means of using image processing and analysis technology to effectively read facial expressions to convey information. At present, facial expression recognition technology has been applied in many fields¹, involving all aspects of our work and life. For example, by recognizing the facial expressions of a teacher in class, the effect of a teacher's teaching can be reflected from one side².

In recent years, there has been a period of rapid development of scientific research on facial expression recognition. Many scholars have applied deep learning methods to facial expression recognition and achieved good results. In 2018, Kaur et al. ³ used facial expression recognition to predict the concentration of students. In 2019, Xie et al. ⁴ constructed an adaptive weighted loss function to fuse manual features and deep features to effectively improve facial expression classification. At present, facial expression recognition based on deep learning is still under development.

In order to further improve the accuracy of facial expression recognition and enhance the generalization ability of the network model, this paper has improved on the basis of the traditional ResNet18 network, adding a convolutional layer and a maximum pooling layer, and weighting the loss
function to improve. Experimental results show that the improved model and method have significantly improved facial expression recognition accuracy and generalization ability, and have certain application value.

2. Deep Residual Network Resnet18
The basic unit of the deep residual network is the residual structural unit, and its structure is shown in Figure 1. The residual structure unit is realized in the form of cross-layer identity connection. The output of each residual unit is obtained by adding the output and input elements of multiple convolutional layers cascaded, and finally activated by the nonlinear activation function ReLU. The entire depth The residual network is composed of multiple residual structural units stacked to form a deep neural network. Introducing the deep residual network of shortcut connection and identity mapping enables the data stream to flow across layers, so the input feature information can be well characterized by the output of the network. The residual structure unit also well controls the number of network parameters and computational complexity.

3. Improved facial expression recognition of residual network model
3.1. Improved deep residual network
This paper is based on the network model of the deep residual network ResNet18, using a 1×1 convolutional layer, passing through the BN layer, then passing through the Relu activation function, and finally using a Maxpool layer to replace the last average of the original network. The structure of the Avgpool layer is shown in Figure 3. We named the improved model C-ResNet18.
3.1. Deep residual network C-ResNet18

1×1 activation and Relu activation can fatally produce roles under the function of no loss \[^5\], making the network more characteristic of discriminating information. Maximum pooling (Maxpool) can reduce the deviation of the estimated mean value caused by the parameter error of the convolutional layer, retain more texture information, prevent over-fitting, and improve the generalization ability of the model.

3.2. Improved self-attention weighted deep residual network model

The improved self-attention weighted deep residual network model is shown in Figure 4. The face image is first input into C-ResNet18 for feature extraction, and the extracted features are put into the fully connected layer and the weighting module. The weighting module is from The sample weights are learned in the facial features, which are used for loss weighting, and finally the facial expressions are classified by softmax.

Figure 3. Improved deep residual network C-ResNet18

Figure 4. Improved self-attention weighted C-ResNet18
The self-attention weighting module is used to obtain the contribution of input samples to training, output a weight in the range of 0 to 1, and weight the loss. Use \( F = [x_1, x_2, \ldots, x_N] \) to represent the facial features of \( N \) images, take \( F \) as input, and output a weight for each feature. Specifically, the weighting module is composed of a linear fully connected (FC) layer and a sigmoid activation function. The extracted expression feature \( F_i \) passes through the fully connected layer and the Sigmoid activation function, and the weight corresponding to the expression picture can be obtained. The expression is:

\[
\alpha_i = \sigma(W^T_{\alpha}x_i) \tag{1}
\]

among them, \( \alpha_i \) represents the weight of the \( i \)-th sample, \( W_{\alpha} \) indicates the parameters of the FC layer of the weighting module, \( x_i \) indicates the extracted feature of the \( i \)-th sample, \( \sigma \) represents the sigmoid function.

The loss value is the goal of the overall network optimization, and participates in the network optimization operation, so as to train a neural network model with higher accuracy. In this article, the weighting module is used to weight the cross entropy loss \([6]\), and its expression is:

\[
L_{wce} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{e^{\alpha_i(W^T_{\alpha}x_i)}}{\sum_{j=1}^{C} e^{\alpha_j(W^T_{\alpha}x_i)}} \tag{2}
\]

among them, \( W_j \) represents the \( j \)-th classifier, \( C \) is the number of categories. It can be seen from the above formula that \( L_{wce} \) and \( \alpha \) are positively closed.

4. Experiment and result analysis

4.1. Data set and preprocessing

The experimental data sets are JAFFE and RAF-DB, which include seven types of expressions: fear, surprise, happiness, anger, disgust, neutral, and sadness. The RAF-DB facial expression data set has a total of 15,339 expression pictures, of which there are 10226 pictures in the training set and 5113 pictures in the test set, as shown in Figure 5.

![Figure 5. RAF-DB data set](image)

The CK+ data set contains 327 continuous expression frames of 123 users, with a total of 40,221 images, including 30,155 in the training set and 10066 in the test set, as shown in Figure 6.

![Figure 6. CK+ data set](image)
4.2. Experimental parameter setting

Experimental environment: This experiment is based on the PyTorch framework under the Windows platform to carry out experimental analysis of the proposed algorithm. The calculation and configuration are Intel i5 processor, 8G memory, NVIDIA GeForce GTX 1050 Ti graphics card, 4G video memory, and Python3.7 is installed.

Parameter setting: The debugging of network training parameters is very important to the performance of the network. The learning rate is an important parameter in network training. If the learning rate is too high or too low, it will have a certain impact on the network. In the experiment process, the improved network model uses different learning rate parameter settings. When training on the CK+ data set, the initial learning rate is set to 0.001, and the learning rate is attenuated by 10% during the 15th and 30th iterations of training, 1. When training on the RAF-DB data set, the initial learning rate is set to 0.005, and the learning rate is attenuated by one-tenth each during the 15th and 30th iterations of training.

5. Experimental results

This paper uses an improved self-attention weighted network model to conduct experiments on the CK+ data set and RAF-DB data sets. The experimental results are shown in Table 1-4.

Table 1 and Table 2 respectively show the facial expression recognition results of the improved network model and the original network model on the CK+ data set and RAF-DB data set. It can be seen from Table 1 and Table 2 that the accuracy of facial expression recognition of the improved network is significantly higher than that of the original network.

### Table 1. The recognition results of the original model and the improved model in this article on the CK+ data set

| Network model            | Accuracy/% |
|-------------------------|------------|
| ResNet18                | 97.38      |
| C-ResNet18              | 98.08      |
| Self-attention weighted C-ResNet18 | 98.89      |

### Table 2. Recognition results of the original model and the improved model in this paper on the RAF-DB data set

| Network model            | Accuracy/% |
|-------------------------|------------|
| ResNet18                | 86.02      |
| C-ResNet18              | 86.45      |
| Self-attention weighted C-ResNet18 | 87.13      |

Table 3 and Table 4 respectively show the comparison results of the face recognition rate of the improved model in this paper and other models on the CK+ data set and RAF-DB data set. It can be seen from the results in Table 3-4 that the improved model in this paper has a significant improvement in the accuracy of facial expression recognition than other models.

### Table 3. Comparison of the accuracy of facial expression recognition between the improved model in this article and other models on the CK+ data set

| Network model            | Accuracy/% |
|-------------------------|------------|
| C-LetNet5[3]            | 94.37      |
| EM-AlexNet[4]           | 93.02      |
| WGAN[5]                 | 96.00      |
| Self-attention weighted C-ResNet18 | 98.89      |
Table 4. Comparison of the accuracy of facial expression recognition between the improved model in this article and other models on the RAF-DB data set

| Network model                      | Accuracy/% |
|------------------------------------|------------|
| VGG16[5]                           | 80.96      |
| Wang et. Al[6]                     | 82.83      |
| Li et. Al[7]                       | 85.07      |
| Self-attention weighted C-ResNet18 | 87.13      |

6. Concluding remarks

Based on the deep residual network ResNet18, this paper proposes an improved network C-ResNet18. The improvement is to replace the last average pooling layer of the original network with a convolutional layer and a maximum pooling layer, so that the network extracts more discriminative information features, prevent over-fitting, weight the loss function, and repeatedly train to achieve a more optimized neural network model. Experiments on the CK+ data set and RAF-DB data and the experimental results show that the accuracy of facial expression recognition has reached 98.89% and 87.13%, respectively, which is better than the facial expression recognition of the original ResNet18 network and other types of network models. The rate has been significantly improved, which proves that this method can effectively improve the accuracy of facial expression recognition and has certain application value.

Acknowledgments

This article is one of the phased achievements of the general project of the National Social Science Fund "Scientific Research Project of Guangxi Humanities and Social Sciences Development Research Center-Innovation and Entrepreneurship Special Project (Major Commissioned Project)" (contract number: ZDCXCY01).

References

[1] CHEN G L. Comparison of common interpolation methods in image scaling algorithm[J]. Fujian computer, 2017, 33(9): 98-99.
[2] BRITO C, MACHADO A, SOUSA A. Electrocardiogram beat-classification based on a resnet network[J]. Studies in Health Technology and Informatics, 2019, 30(6): 264-265.
[3] Kaur A, Mustafa A, Mehta L, et al. Prediction and localization of student engagement in the wild[C]//2018 Digital Image Computing: Techniques and Applications (DICTA). IEEE, 2018: 1-8.
[4] Xie W, Shen L, Duan J. Adaptive Weighting of Handcrafted Feature Losses for Facial Expression Recognition[J]. IEEE Transactions on Cybernetics, 2019.
[5] Zhong L, Bai C, Li J, et al. A Graph-Structured Representation with BRNN for Static-based Facial Expression Recognition[C]//2019 14 th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019). IEEE, 2019: 1-5.
[6] Liang XL and Xu J. Biased ReLU neural networks[J]. Neurocomputing, 2021, 423: 71-79.
[7] Li Y, Zeng J B, Shan S G, et al. Occlusion aware facial expression recognition using cnn with attention mechanism[J]. IEEE Transactions on Image Processing, 2018, 28(5): 2439-2450.