Journal of Computers Vol. 33 No. 4, August 2022, pp. 159-168
doi: 10.53106/199115992022083304013

Facial Expression Recognition Based on Improved Residual Block Network

Hong-Jie Zhang[1,2], Guo-Jun Lin[1,2*], Tian-Tian Chen[1,2], Shun-Yong Zhou[1,2], Hong-Rong Jing[1,2]

[1] Artificial Intelligence Key Laboratory of Sichuan Province, Sichuan University of Science and Engineering, Zigong 643000, China
{2369095868, 386988463} @qq.com
[2] School of Automation and Information Engineering, Sichuan University of Science and Engineering, Zigong 643000, China
{1450935593, 2366464126} @qq.com

Received 25 October 2021; Revised 14 January 2022; Accepted 14 February 2022

Abstract. Facial expression recognition is widely used, but there are some problems such as complex scenes, lack of data sets and low recognition rate. In this paper, we construct a new network model and name it RNFC. The RNFC network adopts 6 improved residual blocks to extract features. Features are passed into the fully connected layer by flattening the data, and Dropout techniques are introduced between the fully connected layers to prevent overfitting of the model. Based on the pytorch framework, we use a cross-entropy loss function to improve the training speed of the network. And perform denoising and enhancement pre-processing on the FER2013 dataset. The RNFC network is trained and tested on the pretreated FER2013. It has a higher recognition rate than classical networks such as VGGnet19 and ResNet18.

Keywords: residual network, facial expression recognition, Dropout technology, deep learning

1 Introduction

Facial expression recognition (FER), as an important part of face recognition technology, can determine a person's mental state and internal intentions by analysing facial expressions on the face. Facial expression recognition can be applied in many fields. For example, during human-computer interaction, the expression analyzer in the camera analyzes the user's expression and only transmits the result, which can achieve a good communication result. In the field of safety, expression recognition can be used in jobs with an emphasis on safety, such as the management of nuclear power plants and coach drivers. Once person shows signs of fatigue and drowsiness, the identification system will issue an alarm to avoid the occurrence of dangerous situations. In the medical field, changes in a patient's body can be detected in time based on changes in his or her facial expressions. In remote teaching, the teacher can tell how well the students have mastered the lesson by their expressions. This requires an expression recognizer that defines the students' expressions as the mastery of the course, and feeds them back to the teacher. Teacher can make corresponding teaching adjustments. Face expression recognition includes three main steps: face detection and localization, expression feature extraction, and face expression classification. Among them, feature extraction is to extract identifiable features from face images. At present, the two most common feature extraction methods are geometry-based feature extraction and epistemology-based feature extraction. The feature extraction method based on geometric structure needs to accurately locate the key points of the face, and construct structural feature vectors such as geometric distance and angle of the face based on the key points. Appearance-based feature extraction method mainly uses texture information of images for facial expression recognition.

Compared with face recognition, facial expression recognition is more sensitive to changes in the texture, shape and other details of the face, which places more stringent requirements on face feature extraction. However, traditional face expression recognition methods involve human factors. When the amount of data is large, higher semantic and deeper features cannot be extracted from the original image, and only shallow features can be extracted, leading to large classification errors.

Today's facial expression recognition is mainly based on deep learning, which is mainly data-driven for feature extraction. Learning from a large number of samples can lead to deep, dataset-specific feature representations, which are more efficient and accurate representations of the dataset, and the extracted abstract features are more robust and have better generalization capabilities. Traditional neural networks suffer from information loss during information transfer. There is also a risk of gradient disappearance or gradient explosion, making the

* Corresponding Author

deeper network unable to train. The residual network can solve this problem through a deeper network structure. The residual network is characterized by their easy of optimization. It is able to improve the performance of the network by increasing the depth of the network. Its residual block uses skip connections to alleviate the problem of gradient disappearance in deep neural networks. Due to the problems of long recognition time, insufficient feature extraction, low recognition rate and easy over-fitting of network models in face expression recognition. In this paper, a neural network architecture is customized to extract rich facial expression features with six improved residual units, enhancing the representational power of the network. The extracted features are passed to the fully connected layer, and the Dropout technique [1] is introduced between the fully connected layers, as shown in Fig. 1. The introduction of Dropout reduces the number of parameters and prevents over-fitting. In the design, based on the pytorch framework, we use the Cross Entropy Loss function [2] to improve the training speed of the model. The softmax layer normalizes the probabilities to solve multi-classification problems. At the same time, adjust the learning rate StepLR at equal intervals to help the algorithm converge and make the trained network model reach the best. The StepLR formula is as follows.

$$Ir = Ir * gamma. \quad \textbf{(1)}$$

In the formula, *Ir* is the learning rate and *gamma* is the adjustment coefficient. The formula is calculated once every artificially defined number of iterations.

Our contributions can be summarized as follows: 1) The custom neural network is more accurate on the FER2013 dataset. 2) The improved residual block enhances the representation ability of the network and effectively solves the disappearance of the gradient of face feature extraction. 3) Dropout technique is reasonably introduced in the fully connected layer. This reduces the number of parameters per training round, improves the training speed of the model, and prevents overfitting of the model.

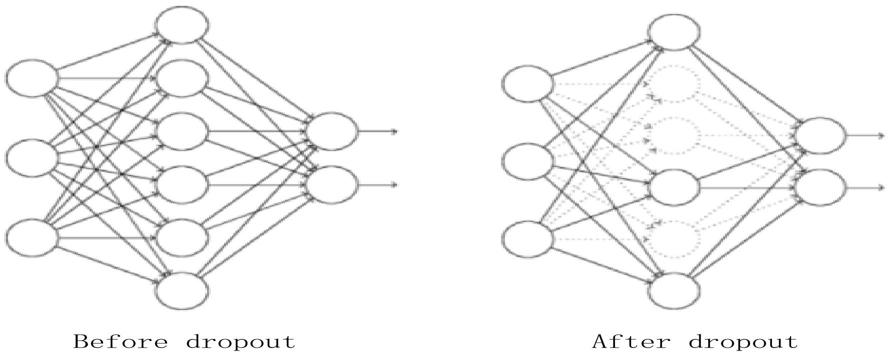

Fig. 1. Comparison of neural network dropout before and after

2 Related Work

Ekman and Friesen scientifically constructed a standard database of facial expressions [3]. A detailed representation of the changes in expression brought about by the movements of different facial muscles. Different expressions represented by changes in facial units including eyes, mouth, and eyelids. Then six basic human expressions are defined: Angry, Disgust, Fear, Happy, Sad and Surprise.

Cristinacce et al. [4] who combined two methods, the Feature Enhancement Algorithm (PRFR) and the Boundary Identification Algorithm (AAM), at the University of Manchester. This joint algorithm has excellent recognition of some edge features of the face, but classification is not ideal. In order to better implement expression classification, Yu et al. [5] investigated using the Gabor wavelet transform algorithm and the support vector machine (SVM) model. The former is used for feature extraction while the latter is used for feature classification. In particular, the use of linear and non-linear operators to characterize new features was also proposed. In order for the system to not only recognize a single facial expression, but also have a good recognition for complex mixed expressions. Jin H et al. [6] designed a classification system specifically for the recognition of complex expressions. The recognition system works in three broad steps. Firstly, the local features of each part of the human face are extracted and arranged into a feature matrix according to predefined rules. Then, the system automatically analyses the extracted feature information. Finally, a classification algorithm is used to accurately identify the category of complex expressions.

Due to the rapid development of deep learning technology, facial expression recognition has also entered the era of neural networks. Andre Teixeira Lopes et al. [7] proposed a simple solution for face expression recognition. Using convolutional neural networks and a specific image pre-processing step to obtain good recognition accuracy on a specified dataset. Although the method is rich in extracted features, it is susceptible to human interference. It does not work well in terms of stability and robustness. In order to reduce the complexity of feature extraction in traditional recognition methods, You-Mei Zhang [8] used a neural network structure to automatically extract facial expression features. The method of fusing image data features and marker point data features was proposed. Lue-Feng Cheng et al. [9] proposed a deep sparse self-coding network based on Softmax regression to recognize facial expressions in human-computer interaction, while dealing with the problem of too much information in the output during model training. In order to overcome the local extremes and gradient spreading problems during training, the overall weights of the network were fine-tuned so that the depth of the whole neural network was globally optimal. Later, Yan-Wei L et al. [10] used a combination of AAM, CNN and LBP features to classify expressions for recognition, and the results showed that the algorithm is an effective recognition method, but there is room for further improvement in computational efficiency. Octavio Arriaga [11] proposed a framework for building an expression recognition network using convolutional neural networks. And a real-time visual system is built to verify the accuracy of the model. However, his model has some limitations, the data is biased towards Western celebrities and Its generalization ability is not good. Kuang Liu [12] solves the FER problem using a CNN integration model, where he designs three CNN sub-networks with different structures. Finally, an overall structure is introduced to merge these sub-networks, and the model focuses on different CNNs to obtain better performance. All of these current network models have certain shortcomings, with some having low recognition rates and others being too complex in structure and slow to run. Compared with these models, we constructed a CNN network model. It has fewer number of parameters and faster running speed according to the size of the dataset, while enhancing the extraction of expression features. The problem of disappearing gradients of deep neural networks is also solved, and the recognition effect is good on the FER2013 face expression dataset.

3 RNFC Network Model

3.1 Selection of ResNet Residual Block

The trend in deep learning is towards larger and deeper networks. However, this has led to a degradation problem, where the deeper the network is, the poorer the performance of the network model. This is because the deeper network layer does not learn identity mapping, but introduces errors through non-identity mapping. The formula of identity mapping is

$$y = H(X) = x. \tag{2}$$

In the above formula, x represents input, y represents output, $H(x)$ represents the mapping of the network layer. This formula can be described as follows: a neural network with m layers, where the first n layers act ($n \ll m$) and the later m-n layers only need to complete the constant mapping, then the network has the ability to fit. However, experiments have shown that the later layers of the network learn non-constant mappings, introducing errors and leading to poor models.

To solve this problem, in 2015, Kai-Ming He [13] proposed the Deep Residual Network. The residual network is characterized by its ease of optimization and its ability to improve accuracy by adding considerable depth. Its internal residual blocks use jump connections, which mitigate the gradient disappearance associated with increasing depth in deep neural networks. Conventional neural networks suffer from information loss when information is passed through them. There is a risk of gradient disappearance or gradient explosion, making it impossible to train deeper networks. The residual network optimizes this problem by bypassing the input information directly to the output, protecting the integrity of the information. The residual cell is shown in Fig. 2.

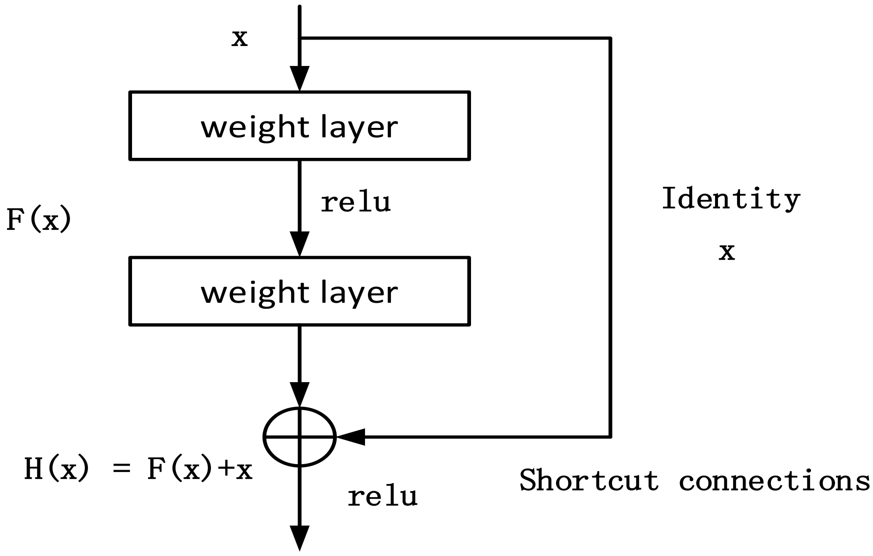

Fig. 2. Residual cell

The weight layer represents a convolutional layer. A convolutional layer is composed of several convolutional kernels. The local feature response is achieved mainly by convolution operations, and the whole image is scanned for feature extraction using the same convolution kernel. The convolution operation formula is as follows:

$$C(x, y) = \sum_{k.l} F(k, l)G(x - k, y - l). \quad \textbf{(3)}$$

where $C(x, y) = F_{k\times 1} \otimes G_{u\times n}$ is the convolution operation, $F_{k\times 1}$ is the convolution kernel of size $k \times 1$, $G_{u\times n}$ is the $u \times n$ input matrix, and $C(x, y)$ is the output matrix after convolution, $Vx \in [0, u]$, $V_y \in [0, n]$. A BN layer is designed after the convolution layer, which has three main purposes: 1) to speed up the training and convergence of the network, 2) to control the gradient explosion and disappearance, 3) to prevent overfitting. The BN layer is generally used after the linear and convolutional layers. Because the output of the fully connected and convolutional layers is generally a symmetric, non-sparse distribution, more akin to a Gaussian distribution. Normalizing them will produce a more stable distribution.

Relu is the activation function [14], and the common activation functions are sigmoid and tanh. Among them, sigmoid is to map neurons to (0, 1) intervals, Relu converges faster than the sigmoid and tanh. And is the simplest activation function, which solves the problem of gradient disappearance. Relu the function formula is as follows:

$$F(x) = M_{ax}(0, x). \quad \textbf{(4)}$$

Where: $F(x)$ is the activation result, when $x < 0$, $F(x) = 0$. when $x \in [0, x]$, $F(x) = x$.

As can be seen from Fig. 2, the input data is divided into two branches, one through two convolutional layers to obtain the output $F(x)$, and the other through the "shortcut connections" method, the direct output data x. The two data are added together to obtain $H(x)$: $H(x) = F(x) + x$, $H(x)$ is activated by the relu function to output. When $F(x)$ =0, $H(x)$ =0, which is the identity mapping. The residual block changes the learning objective. No longer is it to learn a complete output, but rather the difference between the target value $H(x)$ and x, which is $F(x)$: = $H(x) - x$. The ultimate goal of training is to approach the residual result to 0, to ensure the accuracy of the network model and solve the problem of network degradation. A residual network is a number of different or identical residual units connected in series to form a network. Commonly used residual network structures are 18-layer, 34-layer, 50-layer, 101-layer and 152-layer.

3.2 Selection of ResNet Residual Block

Compared with the classical residual block, the improved residual block in this paper is shown in Fig. 3.

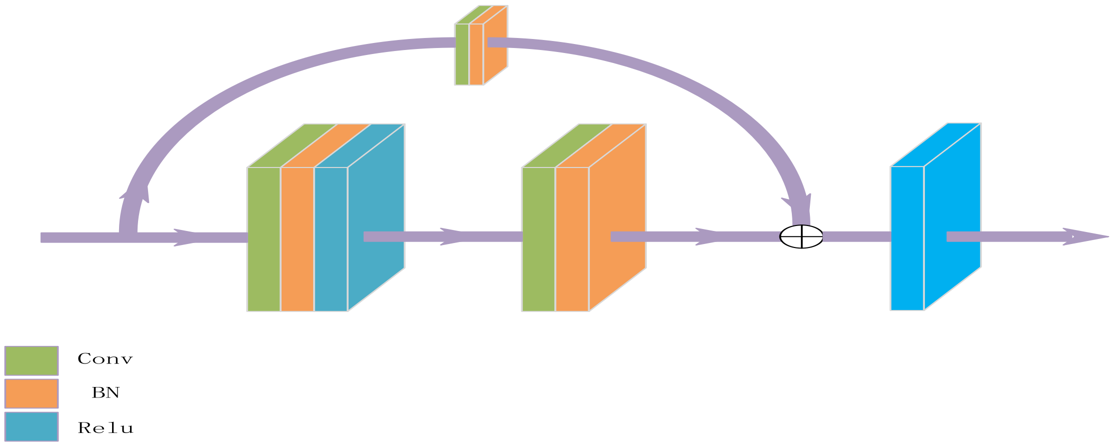

Fig. 3. Residual block network structure

In Fig. 3, the input data of the residual block is divided into two paths, one path passing through 2 convolutional layers of size 3×3 and 2 normalization layers (BN layers) before the output. The other shortcut connections path does not output x directly, but after a convolutional layer of size 1 × 1 and 1 BN layer. The convolution layer in the shortcut path is used to adjust the input *x* to a different dimension. Deepening the structure of the network and enhancing the feature extraction capability of the network. After the data of the two paths are merged, they are activated by the *Relu* function and passed to the next block.

3.3 Network Model

3.3.1 RNFC Network Structure

As can be seen in Fig. 4 and Table 1, we set the input image size to 48 × 48 and the number of channels to 3. After extracting features from six residual blocks, a feature map of 6 × 6 × 256 is obtained. The feature map is flattened and processed into a one-dimensional tensor of length 6 × 6 × 256 = 921. The data is fed into a hidden layer with 4096 neurons after dropout1 (dropout1=0.2), into a hidden layer with 1024 neurons after dropout2 (dropout2=0.5). Then passed through a hidden layer with 256 neurons, and finally passed through an output layer with 7 neurons.

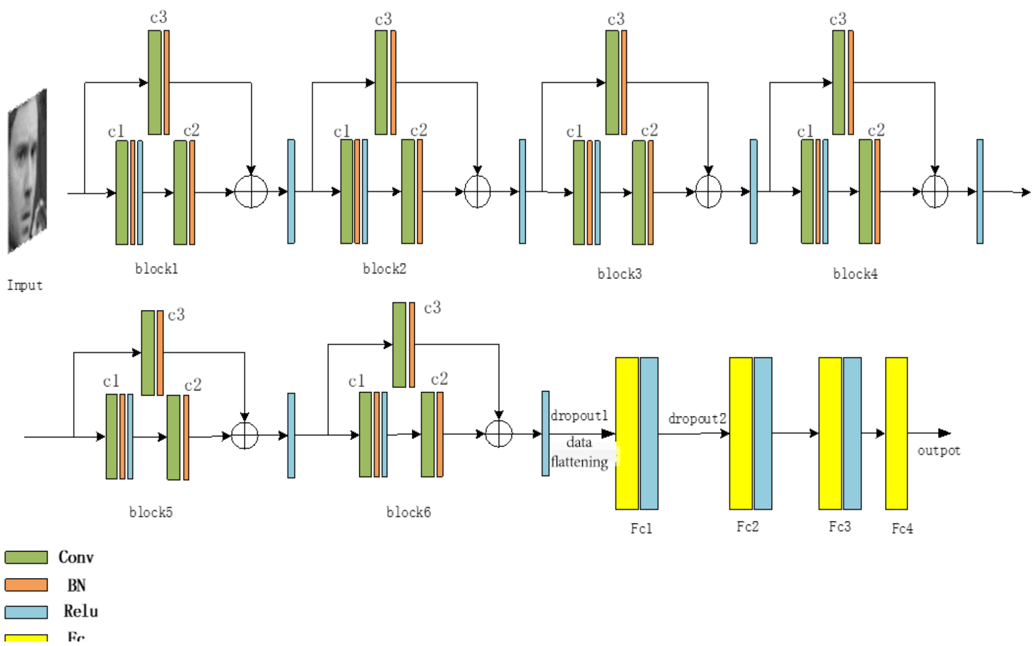

Fig. 4. RNFC network structure diagram

The input and output data for six of these residual blocks are shown in Table 1.

Table 1. The input and output data for six of these residual blocks

Conv layer	Block1	Block2	Block3	Block4	Block5	Block6
In-features	48 × 48 × 3	48 × 48 × 32	48 × 48 × 64	24 × 24 × 128	12 × 12 × 256	12 × 12 × 512
Out-features	48 × 48 × 32	48 × 48 × 64	24 × 24 × 128	12 × 12 × 256	12 × 12 × 512	6 × 6 × 256

3.3.2 Selection of Super Parameters

Through many comparative experiments and experience, the model parameters were set as shown in Table 2.

Table 2. Model parameter setting

Parameter	Epoch	Batch	I r	Step size	Gamma	Dropout1	Dropout1
Value	200	60	0.001	50	0.1	0.2	0.5

4 Experiments and Results

4.1 Experimental Environment

The experimental environment is Windows10, 64-bit operating system, python programming language, Pytorch deep learning framework, NVIDIA GeForce RTX 2060 GPU and 11th Gen Intel (R) Core (TM) i7-11700 @ 2.50 GHz CPU.

4.2 Data Set Selection

The commonly used face expression datasets are: JAFFE face expression dataset, FER2013 [15], CMU PIE face database [16], ORL face database [17], and Yale face expression database [18]. Among them, the FER2013 face expression dataset has a total of 35886 face expression images, including 28708 test images, 3589 public test images and 3589 private test images, and 7 types of expressions, including angry, disgusted, fearful, happy, sad, surprised and neutral. Each image is fixed as a 48 x 48 grayscale image. JAFFE has only 213 images and ORL has only 400 images. Their datasets are too small and the trained models do not have sufficient generalization capability. The CMU PIE face database has only four expressions, including neutral, smile, wink, and talk. The variety of expressions is too small and of little utility. Therefore, the FER2013 dataset was selected for training and prediction of the network model in this paper.

4.3 Pre-processing

This is because there are many cartoons, text and blank images in the FER2013 dataset. In order for the network to learn better feature information, image denoising was performed before the experiment. Filter out the images shown in Fig. 5.

Fig. 5. Noise picture

Then, face detection is performed. Face detection is basically a pre-processing process that is included in all face-related tasks. It extracts faces from complex images. Subsequently, you only need to focus on extracting relevant features of the face to enhance the relevant tasks. This paper uses OpenCV's own Deep Neural Network (DNN) classifier model, which is trained by the ResNet10 network and has good detection speed and accuracy on images with complex backgrounds.

The detected face images are then scale normalized and grayscale normalized. The purpose of scale normalization is to set the images to a uniform size. It ensures that the faces are all in the same position in the picture. It facilitates the extraction of expression features by the neural network. The purpose of greyscale normalization is to adjust the image contrast, reduce the effect of local shadows and light variations on the image on the results, and also to suppress the interference of noise. Local surface exposure contributes a large proportion of the image's texture intensity, so this compression process is effective in reducing local shadows and lighting variations in the image.

Finally, during the training process, we need to enhance the data by performing geometric operations such as random rotation, deflation and mirroring of the original images. It can expand the dataset and improve the robustness of the model. The enhanced images are shown in Fig. 6, from left to right, the original image, the rotated image, the zoomed image and the mirrored image.

Fig. 6. Enhanced image

4.4 Results

The learning rate was set to 0.01 at the beginning, and after 100 training sessions, the loss curve fluctuated back and forth in a toothed pattern and could not be converged. Then set the equal interval adjustment learning rate StepLR, the initial value was set to 0.001, and the learning rate decreased by 0.1 times after every 25 iterations. Setting the equal interval adjustment learning rate helped the algorithm to converge, and the final loss curve showed a smooth decrease, but did not converge to the level. Considering that the number of training sessions was probably too low, resulting in the model not converging completely, the number of training sessions was increased to 200, and the learning rate was adjusted every 50 iterations. On the other hand, when the dropout is set to 0.5, it is found that the loss curve is slow to drop and the dropout is excessive. Then, the dropout was changed to 0.2 and the model was found to have low loss rate and high accuracy on the test set, but poor results and overfitting on the test set. It was decided to set the dropout to 0.2 and 0.5 respectively and the desired result was achieved. The final model converged and the prediction accuracy acc is shown in Fig. 7.

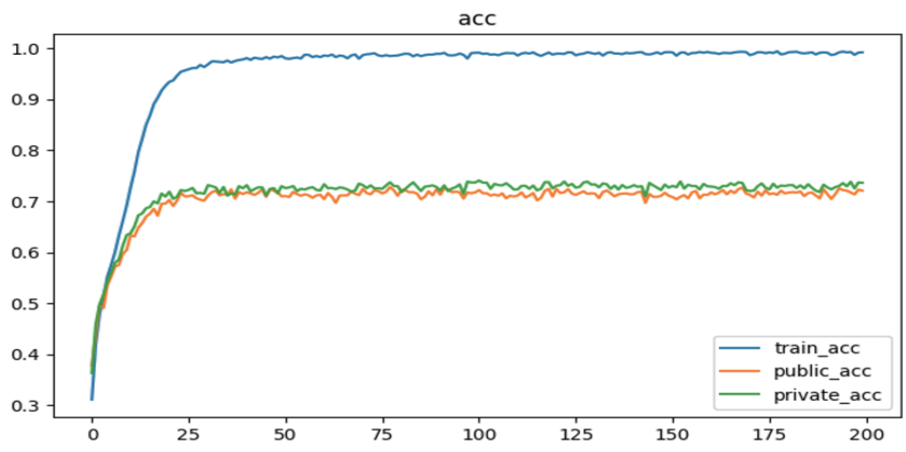

Fig. 7. Acc Curves of Training, Public Test, Private Test

The blue line represents the accuracy of the Training set, and the yellow and green represent the accuracy of the Public Test set and the Private Test set. After 200 iterations, the train acc was as high as 0.981. Accuracy rates of 0.728 and 0.743 were achieved for the two test sets respectively. There is a gap between the accuracy of the test set and that of the training set. The reason for this is that the gap between the test set and the training set images is too large. The network learns a large number of expression features from the training set during training. Due to the picture gap, when the model reaches optimization, the model's prediction on the test set is much less effective than that on the training set.

In order to observe the real prediction effect of each type of expression, there are two confusion matrices generated in this paper, the confusion matrix can represent the exact accuracy of each type of expression, the row labels of the confusion matrix represent the predicted expression results and the column labels represent the actual expression results. Fig. 8 shows the confusion matrix of the training model on the public test set. The average accuracy of the expressions reached 72.8%. Fig. 9 shows the confusion matrix for the private test set. The average accuracy of the expressions reached 74.3%. This can be seen from the confusion matrix of the FER2013 test set. The model best classifies the emotions "happy" and "surprise". On the other hand, it makes the most mistakes when it comes to distinguishing between "angry" and "sad". Secondly, the low classification accuracy of negative expressions can be attributed to two aspects: 1) they have a small number of samples in the original training set. 2) the facial features of negative expressions have some similarity and are confusingly difficult to distinguish, leading to inconspicuous spacing between classes and eventual classification errors.

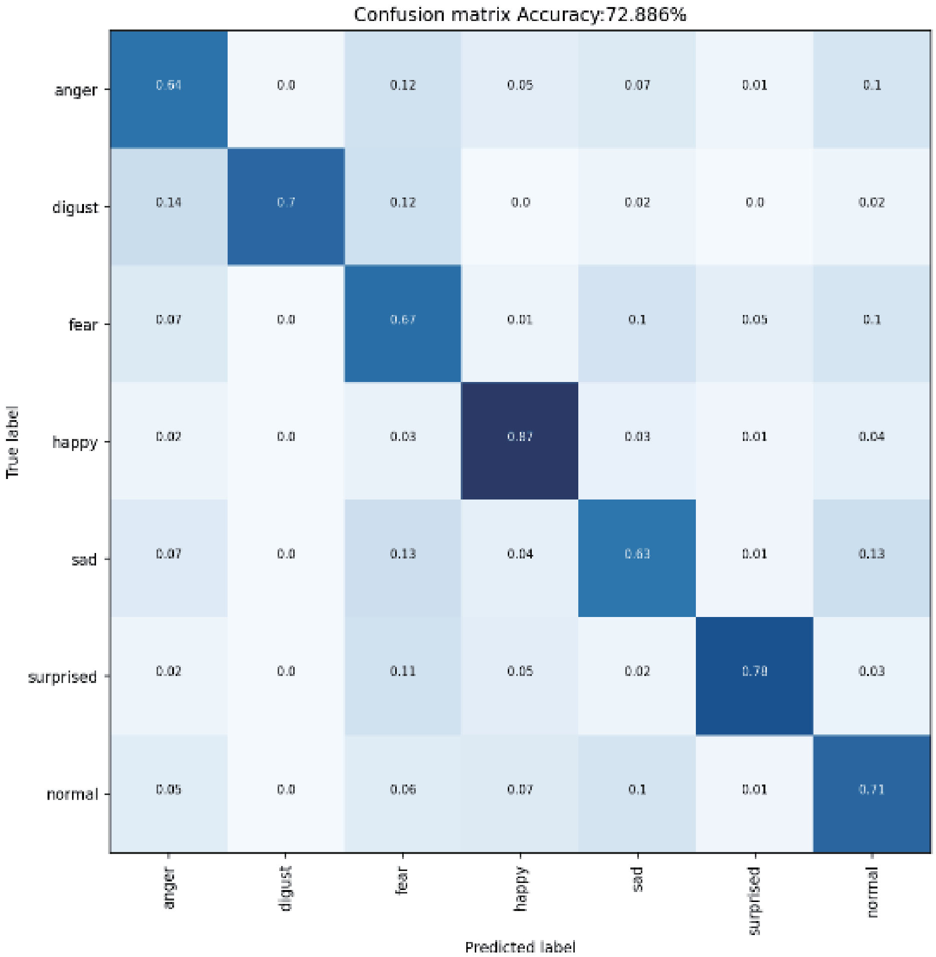

Fig. 8. Shows the confusion matrix of the training model on the Public Test

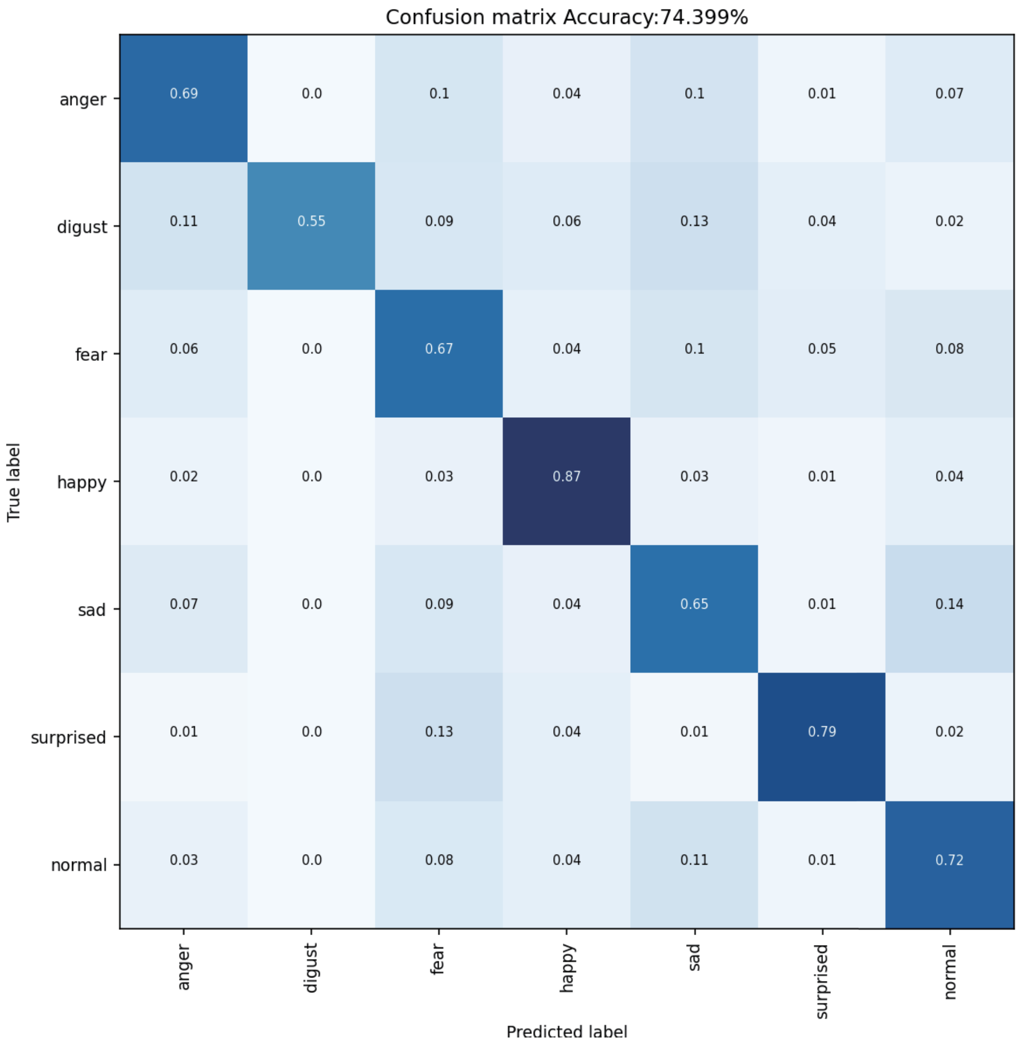

Fig. 9. Shows the confusion matrix of the training model on the Public Test

Finally, we compared the existing FER feature extraction methods. Table 3 shows the recognition accuracy of the different FERs, and our proposed algorithm achieves competitive results on the FER2013 dataset.

Table 3. Comparison of accuracy between this algorithm and other algorithms

Algorithm	Liu	Incerption	CPC [19]	Yan [20]	ResNet18	VGG19	Qu [21]	RNFC
Accuracy	63.6%	67.1%	71.4%	72.0%	73.0%	73.1%	73.00%	74.3%

5 Conclusion

In this paper, a self-defined neural network is proposed, which extracts facial features through improved residual units, and the features extracted are richer. The convolution layer of each residual unit adopts different step size, which plays a role in dimensionality reduction and replaces the dimensionality reduction function of pooling layer. Dropout technology is introduced between different hidden layers, which reduces the computational cost of the network and prevents the occurrence of over-fitting phenomenon. The network does a good job of classifying expressions on the FER2013 dataset. But the network model does not work well for classifying negative expressions. The class spacing of negative expression features is not obvious, and the commonly used Softmax does not require intra-class compactness and inter-class separation, leading to poor classification results. Future research efforts will focus on: 1) Improve the Softmax function. In addition to ensuring feature vector separability, it is important to achieve intra-class compactness and inter-class separation of feature vectors. 2) Optimize and improve the algorithm. Try to fuse with other networks to meet the high requirements of feature extraction for expression recognition. 3) We improved the recognition rate by denoising the dataset. In the future, it is necessary to find or create higher-level facial expression datasets, so that the trained model has better generalization.

Acknowledgements

This work was supported in part by the Sichuan Science and Technology Program of China (2020YFSY0027); in part by the Scientific Research Foundation of Sichuan University of Science and Engineering under Grant 2019RC11 and Grant 2019RC12.

References

[1] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, The journal of machine learning research 15(56)(2014) 1929-1958.
[2] Z. Zhang, M. Sabuncu, Generalized cross entropy loss for training deep neural networks with noisy labels, in: Proc. Advances in neural information processing systems, 2018.
[3] R.W. Levenson, L.L. Carstensen, W.V. Friesen, P. Ekman, Emotion, physiology, and expression in old age, Psychology and aging 6(1)(1991) 28-35.
[4] D. Cristinacce, Automatic detection of facial features in grey scale images, [dissertation] United Kingdom: The University of Manchester, 2004.
[5] J. Yu, B. Bhanu, Evolutionary feature synthesis for facial expression recognition, Pattern Recognition Letters 27(11)(2006) 1289-1298.
[6] H. Jin, W. Gao, The Human Facial Combined Expression Recognition System, Chinese Journal of Computers 23(6)(2000) 602-608.
[7] A.T. Lopes, E. De Aguiar, A.F. De Souza, T. Oliveira-Santos, Facial expression recognition with convolutional neural networks: coping with few data and the training sample order, Pattern recognition 61(2017) 610-628.
[8] Y.-M. Zhang, W. Zhang, Facial expression recognition based on data fusion, Journal of Sichuan University (Engineering Science Edition) 48(6)(2016) 160-164.
[9] L.-F. Chen, M. Zhou, W. Su, M. Wua, J. She, K. Hirota, Softmax regression based deep sparse autoencoder network for facial emotion recognition in human-robot interaction, Information Sciences 428 (2018) 49-61.
[10]Y.-W. Li, W.-Y. Zheng, N. Lin, Facial Expression Recognition Method Fusing AAM, CNN and LBP Features, Computer Engineering and Design 38(12)(2017) 3436-3440.
[11]O. Arriaga, M. Valdenegro-Toro, P. Plöger, Real-time convolutional neural networks for emotion and gender classification. <https://arxiv.org/abs/1710.07557>, 2017 (accessed 21.08.07).
[12]K. Liu, M. Zhang, Z. Pan, Facial expression recognition with CNN ensemble, in: Proc. 2016 international conference on cyberworlds (CW), 2016.
[13]K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proc. of the IEEE conference on computer vision and pattern recognition, 2016.
[14]M. Agarwal, S. Gupta, K.K. Biswas, A new conv2d model with modified relu activation function for identification of disease type and severity in cucumber plant, Sustainable Computing: Informatics and Systems 30(2021) 100473.
[15]I. Hatzilygeroudis, V. Palade (Eds.), Advances in Hybridization of Intelligent Methods. Smart Innovation, Systems and Technologies, Vol. 85, Springer, Cham, 2018 (pp. 1-16).
[16]H.K. Ekenel, R. Stiefelhagen, Local appearance based face recognition using discrete cosine transform, in: Proc. 2005 13th European Signal Processing Conference, 2005.
[17]M. Saraswathi, S. Sivakumari, Evaluation of PCA and LDA techniques for Face recognition using ORL face database, International Journal of Computer Science and Information Technologies 6(1)(2015) 810-813.
[18]S. Nigam, R. Singh, A.K. Misra, Efficient facial expression recognition using histogram of oriented gradients in wavelet domain, Multimedia tools and applications 77(21)(2018) 28725-28747.
[19]T. Chang, G. Wen, Y. Hu, J. Ma, Facial expression recognition based on complexity perception classification algorithm. <https://arxiv.org/abs/1803.00185>, 2018 (accessed 21.08.11).
[20]J. Yan, W. Zheng, Z. Cui, P. Song, A joint convolutional bidirectional lstm framework for facial expression recognition, IEICE TRANSACTIONS on Information and Systems E101.D(4)(2018) 1217-1220.
[21]J. Qu, R. Zhang, Z. Zhang, J. Pan, Facial expression recognition based on deep residual network, Journal of Computers 31(2)(2020) 12-19.