Research and Application of Facial Expression Recognition Based on Attention Mechanism

2021 IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC) | 978-1-7281-9018-1/20/$31.00 ©2021 IEEE | DOI: 10.1109/IPEC51340.2021.9421081

Xin Zhang*
School of Software Engineering
Chongqing University of Posts and Telecommunications
Chongqing 400065, China
1724351590@qq.com

Zhuang Chen
School of Computer Science and Engineering
Chongqing University of Technology
Chongqing 400065, China

Qingjie Wei
School of Software Engineering
Chongqing University of Posts and Telecommunications
Chongqing 400065, China

***Abstract*—The scale of the existing convolutional neural network is getting larger and larger, resulting in too large amounts of parameters, and the structure is not light enough. Moreover, existing convolutional neural networks are difficult to recognize the subtle changes of facial expressions and cannot extract facial expression features accurately. Therefore, the performance of facial expression recognition needs to be improved. To solve the above problems, a new deep learning network model is proposed for facial expression recognition. Based on the deep residual network, the attention mechanism module (Convolutional Block Attention Module, CBAM) is added to the last layer of convolution and the first layer of convolution of the network. The spatial attention mechanism and channel attention mechanism are used to suppress the unimportant feature information and focus on the effective feature information. In the bottom layer, the influence of other factors is eliminated as much as possible, and more attention is paid to the extraction of facial expression features, which enriches the learning of facial expression features and improves the accuracy of facial expression recognition. The method proposed in this paper has been tested and verified on two public data sets FER2013 and CK+, and the results prove that the method has a high accuracy rate.**

Keywords—Expression recognition, attention mechanism, deep learning, deep residual network

I. Introduction

With the rapid development of computer vision technology, automatic facial expression analysis and recognition has important applications in areas such as fatigue driving, security video surveillance, automatic control, intelligent medical care, and criminal polygraph detection, which has attracted great attention.

Compared with traditional recognition methods, they have higher recognition accuracy and better stability. SE-Net attention network to obtain the channel dependence of features, which brings significant performance improvement to the convolutional neural network. Woo S et al. proposed the CBAM attention network, which combines space and channel attention, a good recognition effect is obtained, but its parameter amount is also larger. In addition, due to the small distance between expression classes, the performance of the Soft Max loss function is not ideal. Cai J. et al. [1] proposed the Island loss function to increase the distance between classes of different expressions, and combined with the Soft Max loss function to jointly monitor the expression features. Li S et al. [2] proposed the Local Proserving loss function to reduce the distance within the expression class and enhance the ability to discriminate the depth features of expression in an uncontrolled environment. However, the complex network model of convolutional neural network and the difficulty of real-time influence its further application in actual scenes, and the difficulty of expression recognition is that the difference between different expression classes is small, and the gap within the same expression class is large, and it is easily affected by the external environment. It becomes especially important to extract key features that change significantly between expressions.

This paper proposes a facial expression recognition method based on the attention mechanism of convolutional neural network. This method is based on the ResNet [3], and the attention mechanism module (CBAM) is embedded in the first and last layers of the convolutional layer, Increasing the complexity of the network connection, enhancing the feature learning of facial expressions, improving the performance of expression recognition, and the overall structure is lighter. Experiments on the datasets FER2013 and CK+ show that the method in this paper has a better effect on facial expression recognition tasks.

II. Related Work

A. Facial Expression Recognition Method

The main work can be divided into three areas: design a new network architecture, increase network width and depth to improve performance; study different optimization methods, such as introducing attention mechanism and other optimization Feature extraction improves the model's ability to learn effective features; explores different loss functions to supervise network training, etc.

Deep learning has developed rapidly, and many network architectures, such as VGG [4], AlexNet [5], ResNet [5], etc., have been widely used in facial expression recognition. Karen Simonyan [4] proposed the VGG network in ILSVRC in 2014. Unlike the AlexNet proposed by Alex Krizhevsky [6], the number of network layers is deepened, and all convolution kernels use 3×3 small convolutions. Core, thereby reducing the amount of parameters. Cheng et al. [7] optimized the network structure and its parameters on the basis of VGG19, and used transfer learning technology to overcome the lack of training samples. He et al. [3] proposed ResNet in 2015. Using the principle of identity path, the network can be further deepened without causing gradient explosion or increasing error. Wang et al. [8] used both global features and regional features in the proposed deep convolutional network. In

978-1-7281-9018-1/21/$31.00 ©2021 IEEE

addition, they also used facial action units to improve it, and established a Bayesian network model to analyze the probability of action units. Finally, the learned features are integrated for expression feature classification. Xu et al. [9] used the method of combining LBP (Local Binary Pattern) and convolutional neural network to extract features separately through two branches, then merged the two features and used PCA (Principal Components Analysis, PCA) to reduce dimensionality , To reduce the impact of image rotation on facial expression recognition.

B. Research on Attention Mechanism

In terms of the attention mechanism, most of them are formed in the form of masks. The principle is to give new weights to features, mark out the relevant features in the image, and let the neural network train. Recently, Wang et al. proposed a new type of regional attention network, which can better deal with facial expression recognition under occlusion and posture changes. The features of each region generated by the backbone convolutional network are aggregated and embedded in a compact fixed Length characteristics to improve its accuracy. Li [5] combines the features of the key areas of the face to weight each feature, and the gate unit obtains the facial occlusion area by calculating the weight according to the patency and importance. Focus on identifiable non-occluded areas. Gan et al. [6] proposed a multi-attention mechanism fusion network to deal with facial expression recognition under complex conditions. The network contains two modules: a regional perception module and an expression recognition module. The mask is learned through the regional perception module. It is used to locate important areas related to expressions, and then through an expression recognition module with multiple types of attention mechanisms, learn features with strong discrimination from these important areas.

More and more computer vision research tends to explore the application of attention mechanism. So based on the above analysis, this paper proposes a network model based on attention mechanism to improve the accuracy of facial expression recognition.

III. Methodology

The depth of the convolutional neural network determines whether it can extract deeper features, but as the network depth continues to deepen, it will cause network degradation problems. The deep residual network introduces a residual module into the network. The introduction of this module effectively alleviates the problem of the gradient disappearance of back propagation during network model training, thereby solving the difficulty of training and performance of the deep network the problem of degradation.

We embed the CBAM attention mechanism module in the first layer of convolution and the last layer of convolution. The network structure of this article is shown in Figure 1. We embed the CBAM attention mechanism module in the first layer of convolution and the last layer of convolution. Through the spatial attention mechanism and channel attention mechanism, the unimportant feature information is suppressed, and the effective feature information is focused on. Try to eliminate the influence of other factors at the bottom layer, and pay more attention to the extraction of facial features. On different classification and detection data sets, integrating CBAM into different models, the performance of the model has been consistently improved, demonstrating its wide applicability.

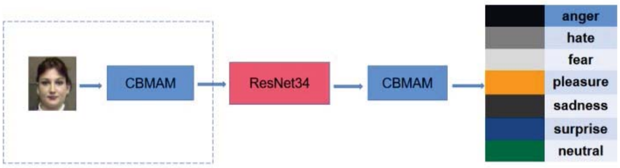

Fig. 1. Network structure diagram

A. Residual Learning Unit

Several residual learning units form the residual module, and the structure of the residual learning unit is shown in Figure 2. The unit is mainly composed of two branches, the first is the residual learning branch, and the other is the identity mapping branch of the input. Use x to represent the input and F(x) to represent the residual mapping, then the output of the residual learning unit is H(x)=F(x)+x. If the residual F(x)=0, then the accumulation layer is only used for identity mapping. Therefore, the following training goal is to approach the residual result to 0, so that as the network deepens, the accuracy will no longer decrease.

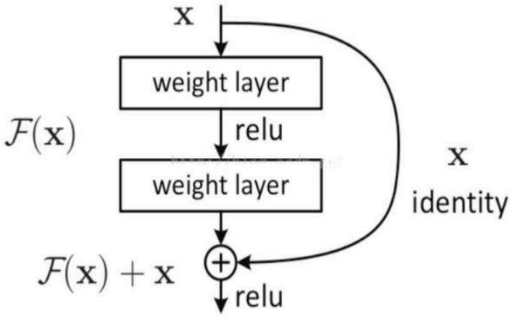
x
weight layer
relu
weight layer
$\mathcal{F}(\mathbf{x})$
x
identity
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu

Fig. 2. Residual learning unit structure diagram

B. Softmax Classifier

Softmax classifier is often used in multi-class recognition, and its input is a vector of arbitrary real numbers. The output is a vector, where the value of each element is between (0, 1), and the sum is 1. Assuming there is an array, its Softmax expression is as equation (1).

$$S_i = \frac{e^i}{\sum_j e^j} \tag{1}$$

Among them, S_i represents the value of Softmax, and e^i represents the i-th element vegetarian. $\sum_j e^j$ represents the sum of all elements.

C. CBAM Convolution Module

In this article, we use Convolutional Attention Block Module (CBAM) to implement the attention mechanism. CBAM stands for the attention mechanism module of the convolution module, which is a kind of convolutional neural. The simple and effective attention module designed by the network combines the attention module of space and channel. Compared with SENet, it has one more space attention, which can achieve better results.

Its structure is shown in Figure 3. For an input feature matrix $F \in R^{C\times H\times W}$ of the middle layer, CBAM undergoes a 1-dimensional channel compression operation and multiplies it with the input feature matrix to obtain F'. Then, after a 2-dimensional space compression operation, the space weight matrix of $F' \in R^{C\times H\times W}$ is calculated to obtain F''. Where $\odot$ represents the matrix elements are multiplied in sequence.

$$F' = M_C(F) \odot F \tag{2}$$

$$F'' = M_S(F') \odot F' \tag{3}$$

Among them, $F \in R^{C\times H\times W}$, $F' \in R^{C\times H\times W}$, and $F'' \in R^{C\times H\times W}$ respectively represent the input feature matrix, the feature map selected by channel attention, and the feature map selected by spatial attention. Among them, $M_C \in R^{C\times 1\times 1}$ and $M_S \in R^{1\times R\times H}$ respectively represent the channel compression weight matrix and the space compression weight matrix.

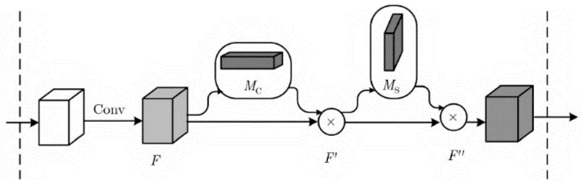

Fig. 3. CBAM structure diagram

IV. Experimental Process and Analysis

A. Data Set

FER2013, the database used in the International Conferen-CE on Machine Learning (ICML) 2013 competition, uses the Google facial recognition API (Application Programming Interface) to capture images, which are then transformed into 48×48 grayscale images through various bounds and cutouts. Some of the images are shown in Figure 4. FER2013 consists of three parts: the training set (28,708 images), the public test set (3,589 images) and the private test set (3,589 images). Each image in the data set has seven corresponding expression attributes (angry, disgusted, scared, happy, sad, surprised and neutral).

Fig. 4. FER2013

The number distribution of each category of training sets, validation sets, and test sets for the FER2013 dataset is shown in Table I.

TABLE I. Number of categories distribution for the FER2013 dataset

Category	Training set	Validation set	Test set
Angry	3995	467	491
Hate	436	56	55
Fear	4097	496	528
Happy	7215	895	879
Sad	4830	653	594
Surprised	3171	415	416
Normal	4965	607	626

CK + data set is Lucey, etc in Cohn Kanade development is put forward on the basis of data sets, the data set, a total of 123 objects, including 327 picture, label contains seven kinds of expression (anger, disgust, fear, happiness, sadness, surprise and contempt), each picture is taken in the experiment under specific conditions, so there is no other noise, as shown in figure 5. However, due to the small number of images in the database, cross-validation method is generally used for evaluation, and cross-validation is usually 5 times, 8 times and 10 times for verification.

Fig. 5. CK+

B. Experimental Parameters

The experimental environment is run on Windows system using NVIDIA GeForce GTX 1660 TI, using PyTorch 1.7.0 as the basic framework to write the program. In the process of training, stochastic gradient descent was used to optimize cross-entropy loss. The learning rate was set at 0.01 at the beginning, and the total batch was set at 500 times in FER2013. Decay began when 80 batches were iterated, total batches were set to 100 in CK+, and it began to decay when 20 batches were iterated and 10-fold cross-validation was used. At the same time, in order to prevent overfitting, the data increment strategy is adopted to enlarge the number of data sets. The original image (48×48) was randomly cropped into 10 images of 44×44 size, and the images were also fixed cropped, and the test data was increased by cropping in the upper left corner, lower left corner, upper right corner, lower right corner and central position respectively. Then the results obtained from these fixed cropped images are averaged as the final result to improve the classification accuracy.

C. Results and Analysis

1) Comparison of experimental results in FER2013 (Table II)

TABLE II. The proposed method is compared with some current methods in FER2013+

Method	Data set size	Image size	Accuracy
VGG16 [4]	256*256	256*256	0.628
LBCNN [10]	256*256	256*256	0.574
VGGface [6]	256*256	256*256	0.702
Microexpnet [8]	256*256	256*256	0.682
Our	256*256	256*256	0.749

2) Comparison of experimental results in CK+

TABLE III. THE PROPOSED METHOD IS COMPARED WITH SOME CURRENT METHODS IN CK+

Name	**Network structure**	**Accuracy**
Fei [9]	ResNet50	93.5
GPS [11]	Gabor filter	**95.1**
ROI [12]	Alex and GoogleNet	94.7
Our	ResNet34	**95.1**

The proposed model was compared with some newer classical methods on CK+ dataset, and the experimental results are shown in Table III. It can be seen from Table III that the accuracy of the method in this paper reaches 95.1%.

V. CONCLUSION

Aiming at facial expression recognition and classification, this paper designs a facial expression recognition model incorporating attention mechanism based on residual network. The attention mechanism can make the network pay more attention to the important feature information and suppress the background interference. The improved network in this paper has the advantages of simple structure, low complexity and few parameters, so it can train and forecast the model quickly and effectively. The experimental results show that the improved model has better performance than the existing facial expression recognition methods.

The next step will continue to improve the recognition accuracy of the model and further realize the intelligent and real-time expression recognition.

REFERENCES

[1] Y. Zhong, S. Qiu, X. Luo, et al., "Facial Expression Recognition Based on Optimized ResNet//2020 2nd World Symposium on Artificial Intelligence (WSAI)," *IEEE*, pp. 84-91, 2020.

[2] Y. Li, Y. Liu, W.G. Cui, et al., "Epileptic seizure detection in EEG signals using a unified temporal-spectral squeeze-and-excitation network," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28, no. 4, pp. 782-794, 2020.

[3] Y. Chen, H. Hu, "Facial Expression Recognition by Inter-Class Relational Learning," *IEEE Access*, no. 7, pp. 94106-94117, 2019.

[4] K. Wang, X. Peng, J. Yang, et al., "Region attention networks for pose and occlusion robust facial expression recognition," *IEEE Transactions on Image Processing*, no. 29, pp. 4057-4069, 2020.

[5] Y. Li, J. Zeng, S. Shan, et al., "Occlusion aware facial expression recognition using cnn with attention mechanism," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2439-2450, 2018.

[6] O.M. Parkhi, A. Vedaldi, A. Zisserman, "Deep face recognition," 2015.

[7] Y. Gan, J. Chen, Z. Yang, et al., "Multiple Attention Network for Facial Expression Recognition," *IEEE Access*, 2020, 8: 7383-7393.

[8] İ. Çuğu, E. Şener, E. Akbaş, "Microexpnet: An extremely small and fast model for expression recognition from frontal face images," *arXiv preprint arXiv: 1711.* 7011, 2017.

[9] Z. Fei, E. Yang, D. Li, et al., "Combining deep neural network with traditional classifier to recognize facial expressions," *2019 25th International Conference on Automation and Computing (ICAC)*, *IEEE*, pp. 1-6, 2019.

[10] F. Juefei-Xu, V. Naresh Boddeti, M. Savvides, "Local binary convolutional neural networks," *Proceedings of the IEEE conference on computer vision and pattern recognition,* pp. 19-28, 2017.

[11] D.G.R. Kola, S.K. Samayamantula, "A novel approach for facial expression recognition using local binary pattern with adaptive window," *Multimedia Tools and Applications*, vol. 12, pp. 1-20, 2020.

[12] X. Sun, P. Xia, L. Zhang, et al., "A ROI-guided deep architecture for robust facial expressions recognition," *Information Sciences*, vol. 522, pp. 35-48, 2020.