Research Article

Facial Expression Recognition Based on Attention Mechanism

Jiang Daihong , Hu yuanzheng, Dai Lei, and Peng Jin

Xuzhou University of Technology, College of Information Engineering, Xuzhou 221000, China

Correspondence should be addressed to Jiang Daihong; jdaihong@xzit.edu.cn

Received 3 January 2021; Revised 19 January 2021; Accepted 14 February 2021; Published 3 March 2021

Academic Editor: Qinhu Zhang

Copyright © 2021 Jiang Daihong et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

At present, traditional facial expression recognition methods of convolutional neural networks are based on local ideas for feature expression, which results in the model's low efficiency in capturing the dependence between long-range pixels, leading to poor performance for facial expression recognition. In order to solve the above problems, this paper combines a self-attention mechanism with a residual network and proposes a new facial expression recognition model based on the global operation idea. This paper first introduces the self-attention mechanism on the basis of the residual network and finds the relative importance of a location by calculating the weighted average of all location pixels, then introduces channel attention to learn different features on the channel domain, and generates channel attention to focus on the interactive features in different channels so that the robustness can be improved; finally, it merges the self-attention mechanism and the channel attention mechanism to increase the model's ability to extract globally important features. The accuracy of this paper on the CK+ and FER2013 datasets is 97.89% and 74.15%, respectively, which fully confirmed the effectiveness and superiority of the model in extracting global features.

1. Introduction

In human daily communication, the expression represents people's current emotional state, which can express more accurate information than language and plays an indispensable role in human emotional communication. In the 1970s, psychologists Ekman and Friesen defined six basic emotions, including happiness, anger, surprise, fear, disgust, and sadness, and later contempt was added. These seven emotions became the basis for people to study expression recognition.

As a research direction in the field of computer vision, facial expression recognition is closely related to face detection and recognition and has been gradually applied to daily life, such as driver fatigue driving detection, criminal investigation, and entertainment. At present, the research of facial expression recognition is mainly divided into two directions: artificial feature extraction based on traditional methods and depth based research. The expression was classified by combining LDA and PCA [1]. Feng et al. [2] proposed an expression recognition method based on the combination of LBP features and SVM classification and made corresponding improvements to the model in the case

of low resolution in practical application; Zhi et al. [3] proposed a novel graph-preserving nonnegative matrix factorization algorithm for facial expression recognition, which is robust to partial occlusions than other methods; Zhong et al. [4] proposed a multitask sparse learning method based on LBP features, which transformed the expression recognition problem into a multitask sparse learning problem, and good results were obtained on multiple data sets.

Since 2013, deep learning has been gradually applied to facial expression recognition. Matsugu et al. [5] used convolutional neural networks (CNN) to solve the translation, rotation, and scale invariance of expression images. Yao et al. [6] proposed HoloNet, a network model specially used for expression recognition, in which CReLU is used to replace ReLU, and the residual module and CReLU are combined to build an intermediate layer, which got great achievements. Zhao et al. [7] designed a feature extraction network by adding a feature selection mechanism in AlexNet. Zhang et al. [8] proposed an end-to-end deep learning model by exploiting different poses and expressions which is based on the generative adversarial network; Cai et al. [9] proposed a new loss function to optimize the distance between

expression classes and maximize the distance between classes so that the network can learn more discriminative features. He and Shen [10] used the improved deep residual network to deepen the depth of the network, and introduced migration learning to solve the problem that the current expression recognition data set is too small and achieved the accuracy rate of 91.33% on the CK + data set. In reference [11], the pairwise random forest method was used to process face pose change in order to solve the problem of face pose change in facial expression recognition. The above expression recognition methods based on deep learning show that the convolution neural network can achieve good recognition results, but the convolution operation is a partial operation in space. To capture the dependence between long-range features, it can only be achieved by repeatedly stacking the convolution layers. However, the efficiency of this method is too low, and due to a large number of network layers, it is necessary to design a reasonable module to keep the gradient from disappearing.

The main contributions of this paper are as follows:

- (i) A facial expression recognition model based on attention mechanism is proposed. A self-attention mechanism is introduced on the basis of the residual network to overcome the limitation of the local operation of convolution operation, which improves the ability of the model to capture long-range associated features. Considering the correlation between channels of feature graph, channel attention is introduced to learn the weight distribution among channels.
- (ii) Through a large number of experiments, the effects of different combinations of self-attention mechanism and channel attention mechanism on the performance of the model are analyzed, and the optimal model is selected.
- (iii) Experimental results on FER2013 and CK + datasets show that the proposed attention based facial expression recognition model is superior to other methods.

2. Relational Work

2.1. Expression Recognition Method Based on Deep Learning. With the improvement of data scale and computer computing power, deep convolutional neural network (CNN) has made important progress in the field of computer vision. In the task of facial expression recognition, the convolutional neural network has become the mainstream method.

Liu et al. [12] proposed a BDBN model based on the CNN network, using a deep belief network to complete feature extraction and expression classification to avoid complex artificial design features and obtaining better performance. Jung et al. [11] proposed the DTGAN model, which is composed of two CNN networks. The first network extracts temporal appearance features from image sequences, and the other extracts temporal geometric features from time-domain facial key points. The combination of the two networks improves the performance of expression

recognition. Zhao et al. [13] proposed a deep region and multilabel learning method (DRML), which uses the feed-forward function to obtain the face area, and forces the learning weight to capture the structural information of the key points of the face. It can automatically learn and has certain robustness to the fixed changes in the local region. In order to overcome the influence of face pose on expression recognition, Zhang et al. [8] used the counter automatic encoder to generate facial images with different expressions and different poses to expand the data set. However, because the generated facial images are not realistic and natural, the generalization performance of the model is poor. He and Chen [14] proposed an expression recognition model based on multiscale feature fusion. Two different networks were trained to extract features of expression images at different resolutions, and then the extracted features were fused to improve the recognition performance of the model and enhance the generalization ability of the model. Yao and Zhao [15] proposed a method of local feature fusion, which extracts the features of eyes, eyebrows, and mouth through CNN, and then uses SVM classifier for decision-making level weighted fusion. This method has a good recognition effect and strong real-time performance.

2.2. Attention Mechanism. Attention mechanism plays an important role in human perception. At present, attention mechanism has been successfully applied to machine translation [16], image generation [17], and other fields. In the field of expression recognition, there are also some researches on attention mechanism. Jaderberg et al. [18] proposed a STN module. The module learns an affine transformation matrix through an end-to-end manner, enabling it to perform spatial transformation such as translation and rotation of the feature map so as to focus on the key target area. Sun et al. [19] proposed a CNN model based on an attention mechanism to learn the attention weight of the feature graph to improve the performance of the model. Pei et al. [20] proposed an end-to-end attention network, where the attention weight on the feature map is automatically learned by CNN. In the process of long-term and long-term memory network processing information, it learned the weight of each time step and then weighed the input eigenvector to obtain the final output vector.

3. Attention Model of Residuals

3.1. Overall Network Framework. In this paper, a deep convolution neural network based on an attention mechanism is proposed and applied to facial expression recognition. In order to solve the limitation of local receptive field in CNN operation, a self-attention mechanism is adopted, which can make the network pay more attention to the regions related to expression; in addition, in order to learn the weight distribution between feature graphs, channel attention mechanism is used to make the network automatically learn the importance of each channel.

Figure 1 shows the overall framework of the attention mechanism model proposed in this paper. In the first part,

FIGURE 1: Facial expression recognition model.

we use downsampling to extract the expression feature map; then input the feature map into the attention residual module for feature transformation to improve the performance of the model; finally, the expression classification is realized through the full connection layer. The attention residual module includes the self-attention module and the channel attention module. The structure of each part is introduced in detail below.

3.2. Residual Network. In deep learning, the performance of the model is often improved by increasing the scale of the model. However, with the deepening of the network layers, the gradient vanishing problem will appear, which makes the model training difficult. To solve this problem, the residual network [21] adopts a short-circuit connection mode, which allows the information before the network to be directly transmitted to the module output layer, as shown in Figure 1.

As shown in Figure 2, the residual module establishes a connection between the input and the output by means of identity mapping so that the convolution layer can learn the residual between the input and output. If $F(x, \{W_i\})$ is used to represent the residual mapping, then the output of the residual module is as follows:

$$y = F(x, \{W_i\}) + x. \quad (1)$$

In which x and y represent the input and output information of the module, respectively.

3.3. Self-Attention Module. In a convolution neural network, the size of the convolution kernel is usually less than 7 due to the limitation of computing resources. Therefore, each convolution operation can only cover a small neighborhood around the pixel, and it is not easy to capture the features with a long distance, such as the association feature between two eyes. In order to capture the dependence between pixels in a long range, it is necessary to stack convolution operations repeatedly and get them by backpropagation, but this will easily lead to the problems of gradient vanishing and slow convergence. Because of the deep network, it is necessary to design a reasonable network structure without affecting gradient propagation.

Different from convolution local computation, the core idea of the nonlocal operation is that when calculating the output of each position in the feature map, it is no longer only calculated with the pixels in the local neighborhood, but also by focusing on all the signals related to the current representation in the image. The correlation weight is

FIGURE 2: Residual module.

obtained to represent the correlation between other positions and the current position to be calculated, which is defined as follows:

$$y_i = \frac{1}{C(x)} \sum_{\forall j} f(x_i, x_j) g(x_j) \quad (2)$$

where i is a position in the output feature map, j is the index of all possible positions in the feature map, x is the input feature map, y is the output feature map, the size is the same as the input feature map, f is the function to calculate the correlation between any two points, g is a function of one variable, the purpose is to transform information, and $C(x)$ is a normalization function. Since f and g are general expressions, the concrete form should be considered when combined with the neural network.

Firstly, since g is a univariate output, which could be replaced by 1×1 convolution, as follows:

$$g(x_j) = W_g x_j. \quad (3)$$

For the function f , which is used to calculate the location correlation, the similarity is calculated in the embedding space.

The mathematical expression is as follows:

$$f(x_i, x_j) = e^{\theta(x_i) \phi(x_j)}. \quad (4)$$

In $\theta(x_i) = W_\theta x_i$, $\phi(x_j) = W_\phi x_j$, the normalization parameters $C(x) = \sum f(x, x)$. For a given location i , $[1] f(x_i, x_j)$ is the softmax to calculate all the position j , the output of the Self-attention layer is as follows:

$$y = \text{soft max}(x^T W^T W_\phi x) g(x). \quad (5)$$

The structure of the self-attention network is shown in Figure 3. Suppose that the input $F^{H \times W \times C}$ of the network is transformed into two embedding spaces by two convolution weights W_θ and W_ϕ to obtain $F^{H \times W \times C^l}$ and $F^{H \times W \times C^l}$, usually $C^l < C$. The purpose of the step is to reduce the number of channels and reduce the amount of calculation. Secondly, reshape the output characteristic graph to $F^{HW \times HW}$, and then perform transposition operation on one of the matrices, and then multiply the matrix. The similarity matrix $F^{HW \times HW}$ is obtained by calculating the similarity, and then the softmax operation on the last dimension is equivalent to obtaining the normalized correlation between each pixel in the current feature map and pixels in other positions. Finally, the same operation is taken for g , which first reduces the dimension, then performs the reshape operation, and then multiplies with the matrix $F^{HW \times HW}$, and applies the attention mechanism to all channels of the feature map. After that, a 1×1 convolution recovery channel is used to ensure that the input and output sizes are exactly the same. From a mathematical perspective, assuming that the feature map of the previous layer is $x \in R^{C \times N}$, it is first mapped into two feature spaces f and g , and $f = W_f x$, $g = W_g x$

$$\beta_{j,i} = \frac{\exp(s_{ij})}{\sum_{i=1}^N \exp(s_{ij})}, s_{ij} = \theta(x_i^T) \phi(s_j), \quad (6)$$

where $\beta_{j,i}$ is the contribution of the position i to the region j of the feature map, where C is the number of channels in the previous layer of the feature map, and N is the number of pixels in the previous layer. This is the output of the attention layer $o = (o_1, o_2, \dots, o_j, \dots, o_N)$

$$o_j = v \left(\sum_{i=1}^N \beta_{j,i} g(x_i) \right), \quad (7)$$

where $W_\theta \in R^{C^l \times C}$, $W_\phi \in R^{C^l \times C}$, $W_g \in R^{C^l \times C}$, $W_v \in R^{C \times C^l}$ as the weight of convolution kernel, where C^l is super parameter, and $C^l < C$.

In addition, the residual connection is introduced in order to better carry out gradient backpropagation, so the final output of the attention module is as follows:

$$y_i = \gamma o_i + x_i, \quad (8)$$

where γ is a learning super parameter, which is initialized to 0 and gradually increases the weight in the process of training.

3.4. Channel Attention Module. Each channel of a feature graph plays a role as a feature detector [22]. Therefore, the channel of the feature graph focuses on what features are useful for the task. However, in the conventional convolution neural network, the importance of each channel is not distinguished. That is to say, each channel is treated equally, ignoring the fact that the contribution of each channel to the task is different. In view of this, this paper introduces channel attention to learn the weight distribution between

FIGURE 3: Self-attention module.

channels, strengthens the channels that are useful for expression recognition task, and weakens the irrelevant channels.

In order to calculate the channel attention more efficiently, for each channel of the intermediate feature map, the global average pooling and global maximum pooling operations based on height and width are, respectively, used to compress the feature graph into two different spaces, and then the two feature graphs are input into the fully connected network with shared parameters, and the output vectors of the full connection layer are processed according to the corresponding elements. Finally, the final channel weight is obtained by the sigmoid activation function. The detailed structure is shown in Figure 4.

Suppose the input feature graph is $F^{H \times W \times C}$, where H , W , C are the height, width, and channel number of the feature graph, respectively. After pooling, the maximum pooled feature graph $F_{\max} \in R^{1 \times 1 \times C}$ and the global average pooled feature graph $F_{\text{avg}} \in R^{1 \times 1 \times C}$ are obtained, respectively. Then these two feature maps are sent into the fully connected network with only one hidden layer. The calculation process is as follows:

$$M_c = \text{sigmoid}(\text{MLP}(\text{AvgPool}(F)) + \text{MLP}(\text{MaxPool}(F))). \quad (9)$$

further:

$$M_c = \text{sigmoid}(W_1((W_0 F_{\text{avg}})) + W_1(W_0(F_{\max}))), \quad (10)$$

where W_0 , W_1 are the shared weight of the full connection layer, and $W_0 \in R^{C/r \times C}$, $W_1 \in R^{C \times C/r}$.

3.5. Attention Fusion. In order to enhance the feature extraction ability of the network model and capture the dependence between long-range features, this paper adds a self-attention mechanism and channel attention to form the attention residual module based on the residual module so as

FIGURE 4: Channel attention module.

to improve the sensitivity of the model to useful information and suppress useless information. There are two kinds of adding methods: serial and parallel, in which serial is divided into self-attention before channel attention and channel attention before self-attention.

3.5.1. Self-Attention before Channel Attention. In serial mode, self-attention is performed first and then channel attention is performed, as shown in Figure 5. Feature graph F_{in} is obtained by convolution of the previous layer as input. Channel attention graph F_{mid} is obtained by channel attention M_c , and then fused with input feature map as input of self-attention M_a . Finally, feature maps obtained by M_a and F_{mid} are fused to obtain the output of the final attention module Out. The formal description of the whole process is shown in Figure 5.

$$F_{mid} = M_c(F_{in}) \otimes F_{in}, \quad (11)$$

$$F_{out} = M_a(F_{mid}) \otimes F_{mid}. \quad (12)$$

3.5.2. Channel Attention before Self-Attention. The serial mode of the first channel followed by self-attention is shown in Figure 6. The former layer is convoluted to obtain feature map F_{in} as input. First, the feature map is obtained by the action of self-attention M_a and channel attention M_c . Then, the self-attention diagram F_{amid} and channel attention map F_c are obtained by fusing with the input feature graph F_{in} . Finally, the corresponding elements of the two attention maps are added. The final attention map output F_{out} is obtained. The formal description of the whole process is as follows:

$$F_{mid} = M_c(F_{in}) \otimes F_{in}, \quad (13)$$

$$F_{out} = M_a(F_{mid}) \otimes F_{mid}, \quad (14)$$

where $\otimes$ represents the multiplication of the corresponding elements.

3.5.3. Parallel Mode. The parallel connection mode is shown in Figure 7. The feature graph F_{in} obtained by convolution of the previous layer is used as the input. Firstly, the respective feature graphs are obtained by the action of self-attention M_a and channel attention M_c .

Then, the corresponding elements of the self-attention graph F_{amid} and channel attention graph F_c are obtained by multiplying the corresponding elements of the obtained feature graph and the input feature graph, respectively. Finally, by adding the corresponding elements of the two

FIGURE 5: Self-attention before channel attention.

FIGURE 6: Channel attention before Self-attention.

FIGURE 7: Parallel mode.

attention graphs, the final output is obtained. The formal description of the whole process is as follows:

$$F_{mid}^c = M_c(F_{in}) \otimes F_{in}, \quad (15)$$

$$F_{mid}^a = M_a(F_{in}) \otimes F_{in}, \quad (16)$$

$$F_{out} = F_{mid}^c \oplus F_{mid}^a. \quad (17)$$

where $\oplus$ represents the addition of corresponding elements and $\otimes$ represents the multiplication of the corresponding elements.

3.6. Attention Residuals. In order to make better use of the channel self-attention designed above, this paper inserts it into the residual module. It can be divided into three kinds of structural design, including using self-attention mechanism alone, using channel attention mechanism alone, and using self-attention and channel attention mechanism simultaneously. The attention residual module adds an attention mechanism to the original residual module. The specific structure is shown in Figure 8.

4. Experiment

In order to verify the validity of the model, experiments are carried out on two datasets: FER2013 and CK+. The experiment is based on TensorFlow framework [23]. The experimental platform is Intel Core i7-6850 with six-core memory of 64GB, graphics card of GTX1080Ti, and system of Ubuntu16.04. All experiments were single card training.

FIGURE 8: Attention residuals module. (a) Residual channel attention. (b) Residual self-attention. (c) Residual mixed attention.

4.1. Data Set. FER2013 data set [24] is derived from Kaggle's expression recognition competition. There are 35888 facial expression images, including faces with different lighting and posture. Among them, 28709 images are in the training set, and 3589 images are in the public test set and private test set. The image size is 48×48 . There are seven categories of gray image, which are anger, disgust, fear, happiness, surprise, sadness, and neutrality. The sample image is shown in Figure 9.

FIGURE 9: Schematic diagram of the FER2013 data image.

FIGURE 10: Schematic diagram of CK + data image.

The CK + data set [25] is also a common data set for facial expression recognition. The modified data set contains 593 image sequences of 123 people, showing the change process of the expression of the test object from the natural state to the expression peak value. There were 327 expression tags, including 8 expressions of nature, disgust, contempt, fear, happiness, sadness, surprise, and anger. In this paper, 981 images of 7 kinds of expression images are selected for the experiment, and the image preprocessing is 48×48 , as shown in Figure 10.

Due to the small number of the two data sets, this paper uses data enhancement to expand the data sets, including random rotation, random brightness adjustment, random graying, etc. The number of CK + data sets was increased to over 29000, and that of FER2013 was increased to over 63000. Through the data enhancement operation, the accuracy of the model can be effectively improved and the overfitting phenomenon can be prevented at the same time.

4.2. The Results and Analysis of the Experiment

4.2.1. Ablation Experiment. In this section, we verify the effectiveness of the self-attention mechanism and channel attention mechanism through experiments. For the ablation experiment, we used FER2013 and CK + data sets and used the residual module as the basic module to build the benchmark model. In the experiment of FER2013 data set, 28709 images were used for training, 3589 images were used to verify the model, and 3589 images were used to test the

accuracy of the final model. For the CK + data set, we divided the training set, validation set, and test set according to the ratio of 7:2:1.

In the training process, Adam is selected as the optimizer, the learning rate is set to 0.0001, and the total training steps are 50 epochs, batch_size is set to 64. The experimental results are shown in Table 1.

From Table 1, we can draw the following conclusions: (1) on the two data sets of FER2013 and CK+, the performance of the benchmark model is obviously worse than that of the model with the attention mechanism. No matter what attention is added and what kind of adding method, this shows that the attention mechanism can improve the feature extraction ability of neural network, which is helpful to improve the performance of the expression recognition model. (2) For the attention machine added, the use of mixed attention mode is obviously better than that of single attention mode, which shows that the nonlinear mapping of the model is effective for the expression recognition task. (3) For the mixed attention model, on the FER2013 data set, the effect of channel attention before self-attention is the best. Compared with parallel mode and self-attention before channel attention, the accuracy rate is improved by 3.98% and 2.89%, respectively. On the CK + data set, the self-attention before channel attention mechanism has the best effect. Compared with the parallel mode and the self-attention before

TABLE 1: The result of the ablation experiment.

Model	FER2013 (%)	CK+(%)
Baseline	69.57	94.37
Baseline + self	72.32	94.89
Baseline + channel	69.63	95.28
Baseline + channel + self	74.15	97.89
Baseline + self + channel	71.20	98.54
Baseline + self and channel (in parallel)	72.01	97.48

channel attention mechanism, the accuracy rate is improved by 0.66% and 1.48%, respectively.

4.2.2. Scheme Selection. According to the analysis of the previous ablation experiment, the combination of channel attention before self-attention has the best comprehensive performance. It has high accuracy on both FER2013 and CK + data sets. Therefore, this model is selected as the final model. In order to verify the effectiveness of this model, the model is compared with other current methods. The experimental results are shown in Tables 2 and 3.

From the experimental data in Tables 2 and 3, we can draw the following conclusions: (1) Compared with the previous three traditional expression recognition methods, the deep learning method can significantly improve the accuracy of expression recognition. The results show that the features extracted by the convolutional neural network can better describe the expression than the artificial feature operator. (2) Compared with the current mainstream deep learning based methods, the self-attention mechanism model achieves higher accuracy on both data sets. (3) The accuracy rate of the FER2013 data set is significantly lower than that of CK + data set, which indicates that the quality of the data set has a certain influence on the experimental results. The scales of the FER2013 data set and CK + data set are relatively small, and there are also wrong tags and nonfacial expression tags in the FER2013 data set, which will bring interference to the training of the model, thus affecting the performance of the model.

Figures 11 and 12 show the training loss and accuracy curves of the model in this paper on the FER2013 and CK + data sets. It can be seen from the graphs that the training process of this model on the FER2013 data set is not stable on the CK + data set, which has a certain relationship with the two data sets. By checking the two data sets, we can find that the expression images of the FER2013 data set are quite different, the image resolution is low, and the image quality is different, which brings certain interference to the training process, and the accuracy rate is finally stable at about 75%. However, the image quality of the CK + data set is high, and the distribution is uniform, so the training model on this data set is relatively stable, and the final accuracy rate is also high. The accuracy rate of the training set and the verification set is about 98%.

TABLE 2: FER2013 comparative experiment.

Model	FER2013 (%)
CNN + SVM [26]	69.30
LBP + SVM [27]	70.86
CPC [28]	71.35
Inception [29]	66.40
Text model	74.15

TABLE 3: CK + comparative experiment.

Model	CK+(%)
Inception [29]	93.20
FRR-CNN [30]	92.06
Reference [11]	97.25
3DCNN [31]	85.90
Text model	97.89

Figure 13 is the confusion matrix obtained from experiments on the FER2013 data set, showing the classification accuracy of face images on seven expressions. The abscissa represents the prediction label, and the ordinate represents the real label. From the matrix, we can see that the accuracy rate of the model with the self-attention unit has been improved in each expression, among which the “sad” expression has the greatest improvement. With about 13% improvement, it shows that the addition of self-attention unit makes the expression classification more accurate. However, there is a certain gap between the accuracy of the seven expressions. For example, the highest accuracy rate of “happy” expression was 92%, while that of “sadness,” “fear,” and “anger” was 49%, 53%, and 64%, respectively. On the one hand, due to the small amount of data and unbalanced samples, these three kinds of expressions have certain negative effects on network training. On the other hand, these three kinds of expressions have a certain similarity, and their feature differences are not obvious, so they are not easy to distinguish.

Figure 14 shows the confusion matrix obtained on the CK + test. It can be seen that most of the expression recognition accuracy has been improved, which is the same as that of FER2013. Due to the relatively small amount of data of anger, sadness, and contempt, and the feature difference between expressions is not very obvious, and the recognition rate is slightly lower than those of nausea, fear, happiness, and surprise.

FIGURE 11: Training curve of FER2013.

FIGURE 12: Training curve of CK+.

FIGURE 13: FER2013 data set obfuscation matrix.

FIGURE 14: CK + Data set obfuscation matrix.

5. Conclusion

In this paper, a facial expression recognition model based on attention mechanism is proposed. The model consists of self-attention and channel attention. Self-attention gives different weights to different positions of the feature map to achieve different attention to different regions while the channel attention learns different weights for each channel of the feature map, strengthens the features useful for classification, and weakens the features irrelevant to the task. Experimental results show that the proposed model improves the accuracy of FER2013 and CK + data sets.

Data Availability

All data needed to evaluate the conclusions in the paper are present in the paper or the references cited here within.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The study was supported by the Major Project of Natural Science Research of the Jiangsu Higher Education Institutions of China (18KJA520012), the Xuzhou Science and Technology Plan Project (KC19197), and the One Stop Service Platform of Pocket Campus Based on Wechat Applet (XCX2020002).

References

- [1] S. Li and W. Deng, "Deep facial expression recognition: a survey," *IEEE Transactions on Affective Computing*, <https://arxiv.org/pdf/1804.08348.pdf>, 2020.
- [2] X. Feng, M. Pietikainen, and A. Hadid, "Facial expression recognition based on local binary patterns," *Pattern Recognition and Image Analysis*, vol. 17, no. 4, pp. 592–598, 2007.
- [3] R. Zhi, M. Flierl, Q. Ruan, and W. B. Kleijn, "Graph-preserving sparse nonnegative matrix factorization with application to facial expression recognition," *IEEE Transactions on Systems, Man, and Cybernetics Part B*, vol. 41, pp. 38–52, 2010.
- [4] L. Zhong, Q. Liu, P. Yang, and B. Liu, "Learning active facial patches for expression analysis," in *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2562–2569, IEEE, Providence, RI, USA, June 2012.
- [5] M. Matsugu, K. Mori, Y. Mitari, and Y. Kaneda, "Subject independent facial expression recognition with robust face detection using a convolutional neural network," *Neural Networks*, vol. 16, no. 5–6, pp. 555–559, 2003.
- [6] A. Yao, D. Cai, P. Hu et al., "Holonet: towards robust emotion recognition in the wild," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pp. 472–478, Tokyo, Japan, November 2016.
- [7] S. Zhao, H. Cai, H. Liu, J. Zhang, and S. Chen, "Feature selection mechanism in cnns for facial expression recognition," *BMVC*, p. 317, 2018.
- [8] F. Zhang, T. Zhang, Q. Mao, and C. Xu, "Joint pose and expression modeling for facial expression recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3359–3368, Salt City, UT, USA, June 2018.
- [9] J. Cai, Z. Meng, Z. Li et al., "Island loss for learning discriminative features in facial expression recognition," in *Proceedings of the 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pp. 302–309, IEEE, Xian, China, May 2018.
- [10] L. He and L. Shen, "Research on expression recognition based on improved deep residual network," *Journal of Applied Computing Research*, vol. 37, pp. 1578–1581, 2020.
- [11] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2983–2991, Santiago, CL, USA, December 2015.

- [12] P. Liu, S. Han, Z. Meng, and Y. Tong, "Facial expression recognition via a boosted deep belief network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1805–1812, Columbus, OH, USA, June 2014.
- [13] K. Zhao, W.-S. Chu, and H. Zhang, "Deep region and multi-label learning for facial action unit detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3391–3399, Las Vegas, NV, USA, June 2016.
- [14] Z. He and Chen, "Convolution neural network with multi-resolution feature fusion for facial expression recognition," *Laser Optoelectronics Program*, vol. 55, pp. 370–375, 2018.
- [15] X. Yao and Zhao, "Facial expression recognition based on convolutional neural network local feature fusion," *Program Laser Optoelectronics*, vol. 57, pp. 338–345, 2020.
- [16] V. Mnih, N. Heess, A. Graves et al., "Recurrent models of visual attention," in *Advances in Neural Information Processing Systems*, pp. 2204–2212, 2014.
- [17] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *Proceedings of the International Conference on Machine Learning*, pp. 7354–7363, PMLR, Long Beach, CA, USA, June 2019.
- [18] M. Jaderberg, K. Simonyan, A. Zisserman et al., "Spatial transformer networks," in *Advances in Neural Information Processing Systems*, pp. 2017–2025, 2015.
- [19] W. Sun, H. Zhao, and Z. Jin, "A visual attention based roi detection method for facial expression recognition," *Neuro-computing*, vol. 296, pp. 12–22, 2018.
- [20] W. Pei, H. Dibeklioglu, T. Baltrušaitis, and D. M. Tax, "Attended end-to-end architecture for age estimation from facial expression videos," *IEEE Transactions on Image Process*, vol. 29, pp. 1972–1984, 2019.
- [21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, Las Vegas, NV, USA, June 2016.
- [22] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proceedings of the European Conference on Computer Vision*, pp. 818–833, Springer, Zurich, Switzerland, September 2014.
- [23] M. Abadi, A. Agarwal, P. Barham et al., "Tensorflow: large-scale machine learning on heterogeneous distributed systems," 2016, <https://arxiv.org/pdf/1603.04467.pdf>.
- [24] I. J. Goodfellow, D. Erham, P. L. Carrier et al., "Challenges in representation learning: a report on three machine learning contests," in *Proceedings of the International Conference on Neural Information Processing*, vol. 117–124, Springer, Lake Tahoe, Nevada, December 2013.
- [25] P. Lucey, J. F. Cohn, T. Kanade et al., "The extended cohn-kanade dataset (ck+): a complete dataset for action unit and emotion-specified expression," in *Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision And Pattern Recognition-Workshops*, vol. 94–101, IEEE, San Francisco, CA, USA, June 2010.
- [26] Y. Tang, "Deep learning using linear support vector machines," 2013, <https://arxiv.org/pdf/1306.0239.pdf>.
- [27] W. Li, M. Li, Z. Su, and Z. Zhu, "A deep-learning approach to facial expression recognition with candid images," in *Proceedings of the 2015 14th IAPR International Conference on Machine Vision Applications (MVA)*, pp. 279–282, IEEE, Tokyo, Japan, May 2015.
- [28] T. Chang, G. Wen, Y. Hu, and J. Ma, "Facial expression recognition based on complexity perception classification algorithm," 2018, <https://arxiv.org/pdf/1803.00185.pdf>.
- [29] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in *Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1–10, IEEE, Lake Placid, NY, USA, March 2016.
- [30] S. Xie and H. Hu, "Facial expression recognition with FRR-CNN," *Electronics Letters*, vol. 53, no. 4, pp. 235–237, 2017.
- [31] M. Liu, S. Li, S. Shan, R. Wang, and X. Chen, "Deeply learning deformable facial action parts model for dynamic expression analysis," in *Proceedings of the Asian Conference on Computer Vision*, pp. 143–157, Springer, Singapore, November 2014.