PAPER • OPEN ACCESS

Research on Expression Recognition Algorithm Based on Attention Mechanism

To cite this article: Lei Sun 2021 *J. Phys.: Conf. Ser.* **1952** 042069

View the article online for updates and enhancements.

Research on Expression Recognition Algorithm Based on Attention Mechanism

Lei Sun[*]

Faculty of Information Technology, Beijing University of Technology, Beijing, China

[*]Corresponding author: leisun2020@bjut.edu.cn

Abstract. Facial expressions carry the emotional state of human beings and are one of the most powerful signals for humans to express themselves. Therefore, the research of facial expression recognition has very important significance and broad application prospects. However, in complex scenes, the blurring of facial images, the interference of redundant information, and the similarity of expressions will cause unsatisfactory expression recognition. In response to these shortcomings, this article adds a convolutional attention module to the basic Resnet network structure, infers the attention map in turn along two independent dimensions (channel and space), performs adaptive feature optimization, captures expression feature information, and reduces human faces. The interference of redundant information; and the introduction of island loss ISlandloss classification to optimize the distance between classes, so that the network can learn more discriminative features, and improve the discrimination of expressions. Compared with the existing mainstream facial expression classification algorithms, this method has certain advantages in objective evaluation indicators. The experimental results on FER2013 show that the accuracy of this algorithm is 73.85%, which is higher than 71.23% of the Resnet network.

1. Introduction

In daily social situations, facial expressions can express information that cannot be contained in language, and play a very important role in human communication. According to research scholar Mehrabian A [1], there are three main types in human daily communication. Expression methods: facial expressions, voices, and language. Among the information conveyed by the experimenter, facial expressions accounted for 55%, voice information accounted for 38%, and language information accounted for only 7% of the total information. Therefore, it can be said that facial expressions are The most expressive way of expressing human emotions. The rapid development of computer science technology and deep learning technology has made the analysis and recognition of facial expressions widely used in education [2], human &computer interaction [3], treatment [4], transportation [5] and other fields. It has increasingly become a hot spot in the field of scientific research and application, and is attracting the attention and input of countless research scholars and application technicians.

With the maturity of computer science technology and computer vision technology, the performance of deep learning technology in the field of information technology has become more prominent. With the further research of many scientific researches, deep learning algorithms are more accurate than previous traditional algorithms in the field of target recognition and classification. Perform well in terms of rate and speed. Compared with traditional algorithms, the deep learning algorithm integrates the two processes of feature extraction and classification, reduces the operation process, and has powerful feature extraction capabilities. It is used in various competitions related to Computer Vision (CV) [6]. The performance in China is very good. However, in real life, the collected facial images are very uncontrollable. These uncontrollable problems will increase the difficulty of facial expression recognition. Therefore, what method is used and how to design the network structure to achieve high efficiency and accuracy Recognizing facial expressions still requires constant exploration.

In facial expression recognition technology, the general process is divided into three parts: face detection, expression feature extraction and expression classification (as shown in Figure 1). Among them, facial expression feature extraction and classification algorithms are the research focus of facial expression recognition, and also the main research focus of this article.

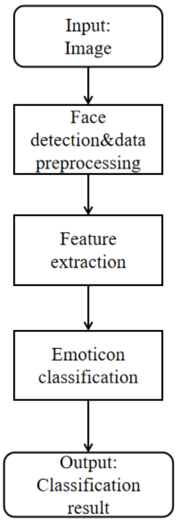

Fig 1. Facial expression recognition process

Its face detection algorithms include sliding window method [7], Faceness-Net [8], Cascade CNN [9], MTCNN [10], etc., feature learning and classification algorithms include FaceNet2ExpNet [11], DTAGN [12], DERL [13], etc.

In the process of facial expression recognition technology research, the effect of facial expression recognition is often interfered by redundant features such as different lighting, posture, occlusion, and facial identity information, and different expressions made by different people are different. The similarity between them will also affect the discrimination effect. In response to the above problems, this paper designs an expression recognition algorithm based on the attention mechanism. In the basic structure of the network, a Convolutional Block Attention Module (CBAM) is added. The CBAM module combines spatial and channel attention [14]. The mechanism module infers the attention map in turn along two independent dimensions (channel and space), and then multiplies the attention map with the input feature map for adaptive feature optimization. Introduce ISlandloss that can aggregate expression features of the same category and separate expression features of different categories, and use the fusion loss function for expression classification.

2. Methods

2.1. Model structure

In the process of facial expression recognition technology research, the effect of facial expression recognition is often interfered by redundant features such as different lighting, posture, occlusion, and facial identity information, and different expressions made by different people are different. The similarity between them will also affect the discrimination effect. In response to the above problems, this paper designs an expression recognition algorithm based on attention mechanism (The Schematic diagram of algorithm structure is shown in Figure 2). In the basic structure of the network, a Convolutional Block Attention Module (CBAM) is added. The CBAM module combines spatial and channel attention mechanism modules., Infer the attention map in turn along two independent dimensions (channel and space), and then multiply the attention map with the input feature map for adaptive feature optimization. Introduce ISlandloss that can aggregate expression features of the same category and separate expression features of different categories, and use the fusion loss function for expression classification.

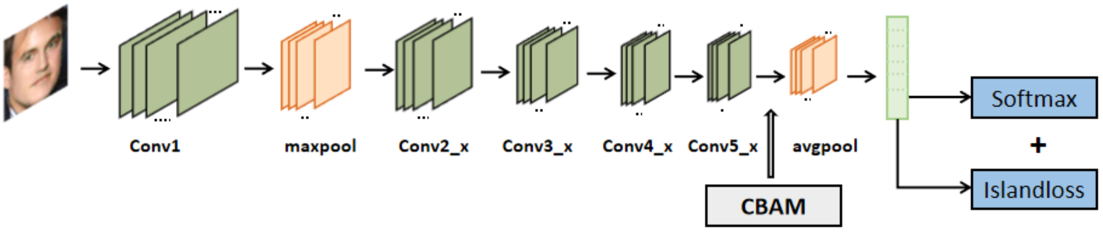

Fig 2. Schematic diagram of algorithm structure

2.2. *Attention mechanism*

An attention module is designed to integrate channel attention and spatial attention. The network can infer attention maps along two independent dimensions (channel and space), optimize adaptive features, capture expression feature information, and reduce facial redundancy. The interference of remaining information. The attention module is shown in Figure 3.

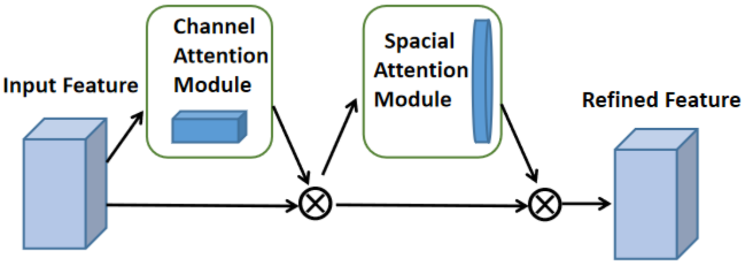

Fig 3. Schematic diagram of CBAM structure

Among them, the working process of the Channel Attention module is: the input feature map F (H×W×C) is respectively subjected to global maximum pooling and global average pooling based on width and height to obtain two 1×1×C feature maps, and then send them to a two-layer neural network. The number of neurons in the first layer is C/r (r is the reduction rate), the activation function is Relu, and the number of neurons in the second layer is C. This The two-layer neural network is shared. Then the output features are added and operated, and then the sigmoid activation operation is performed to generate the final Channel Attention Feature, namely M_c. Finally, multiply M_c and the input feature map F to generate the input features required by the Spatial Attention module.

The workflow of the Spatial Attention module is: take the feature map F output by the Channel Attention module as the input feature map of this module. First do a channel-based global maximum pooling and global average pooling to obtain two H×W×1 feature maps, and then combine these two feature maps as channels, and then go through a 7×7 convolution operation to reduce The dimension is

1 channel, namely H×W×1. After sigmoid, Spatial Attention Feature is generated, namely M_s. Finally, the feature and the input feature of the module are multiplied to obtain the final generated feature.

The CBAM module, which integrates channel attention and spatial attention, is adaptively optimized by learning to capture the importance of each area, emphasizing facial expressions, so that the model can focus on local features that have important contributions to expression recognition, thereby improving Discrimination of fusion features.

2.3. Loss function

The Softmax loss function is currently the most commonly used classification loss function. Among the loss functions used in the traditional classification network model, the SoftmaxLoss loss function mainly requires the features to be correctly classified in the last layer of classifier, that is, the SoftmaxLoss loss function can extract different categories The features are restricted, but the features extracted from the same category are not specifically restricted. It maps the output of multiple neurons to the (0,1) interval, which can be understood as a probability to perform multi-classification.See formula (1) for softmax loss calculation.

$$L_S = -\frac{1}{N}\sum_{j=1}^{N} y_i \log \frac{\exp(z_{y_i})}{\sum_{k=1}^{c}\exp(z_k)} \tag{1}$$

Its characteristic is that it is great to optimize the distance between classes, but it is weak when optimizing the distance between classes. Softmax loss is good at learning information between classes, because it uses an inter-class competition mechanism. It only cares about the accuracy of the prediction probability of the correct label, ignoring the difference of other incorrect labels, which leads to the scattered characteristics of the learned [15].

In order to solve this problem, we define a center for each label data. Everyone should approach the center. Cyi represents the feature center of the yi category. In this way, a new loss is added on the basis of softmaxloss:

$$L_C = \frac{1}{2}\sum_{j=1}^{n}\left\|x_i - c_{yi}\right\| \tag{2}$$

The overall loss is:

$$Lcs = L_S + \lambda L_C \tag{3}$$

The island loss function introduced in this article, on this basis, adds another loss

$$L_{\text{island}} = L_C + \lambda \sum_{c_j \in N} \sum_{c_k \in N, c_k \neq c_j} \frac{c_k \cdot c_j}{\|c_k\|_2 \|c_j\|_2} + 1 \tag{4}$$

That is, the cosine distance of each cluster center is calculated, and +1 makes the range 0-2. The closer to 0, the greater the category difference, so optimizing Loss will make the distance between clusters larger.

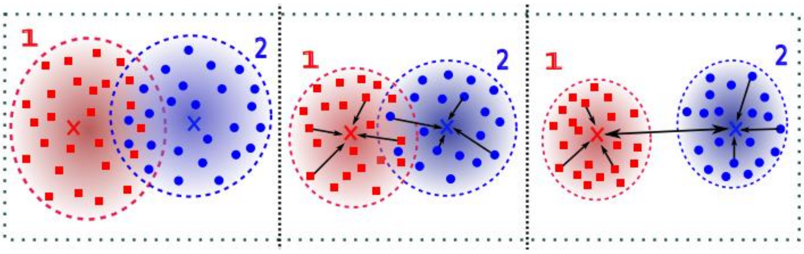

Fig 4. Visualization of three kinds of loss learning

As shown in the Figure 4, red represents expression category 1, blue represents expression category 2, (a) is softmaxloss feature icon, (b) centerloss feature icon, (c) is Islandloss feature icon, which can be seen from the figure , Centerloss optimizes the intra-class distance on the basis of softmaxloss, which makes the intra-class aggregation and the spatial distribution of similar features more compact, but the inter-class distance is not optimized, and the Islandloss introduced in this article optimizes the intra-class distance at the same time The distance between classes is also optimized to increase the distance between classes to separate the classes and improve the discrimination of expressions.

The entire loss function of the network training in this paper is: $\mathrm{L_S} + \lambda \mathrm{L_{island}}$

Among them, input represents the proportional weight of the Softmax loss function and the island loss function Lisland, and the specific parameter settings are explored through experiments.

3. Results

3.1. Experimental environment and platform

The current mainstream deep learning frameworks include PyTorch, Tensorflow, Caffe, PaddlePaddle, Keras, Theano, etc. Each framework has its own advantages, and there is no absolute distinction between good and bad. This article mainly uses the Pytorch framework. The computer's CPU is Intel(R)Core (TM) i7-8700K, the main frequency is 3.70GHz, the memory is 16G, and the GPU is GTX1080ti 11G. In the training process, the optimization algorithm uses the stochastic gradient descent algorithm (SGD), the number of training samples in a batch (batch_size) is 128, weight_decay is set to 0.05, a total of 200 epochs are iterated, and the return value of the loss function stabilizes, Save the parameters after training. In this paper, a comparative experiment of attention mechanism and loss function is carried out, and the trained model is tested on three public data sets of FER2013, CK+, and RAF-DB, and the objective evaluation index—accuracy is evaluated. And give experimental data and analysis.

3.2. Experiments and analysis of attention mechanism

This section sets up the relevant experiments of the attention mechanism. The channel attention module and the CBAM module are added to the Resnet network respectively, and the experiments are performed on FER2013[16], CK+ [17], RAF-DB [18], and the accuracy obtained by different methods The rate is shown in the Table 1.

Table 1. Experimental results related to attention mechanism

	FER2013	CK+	RAF-DB
Res18	68.46%	93.52%	80.77%
Res18+channel(SE)	69.94%	93.71%	82.43%
Res18+spacial+channel(CBAM)	**71.25%**	**94.57%**	**83.69%**

The experimental results show that the CBAM module is added to Resnet, which is a module that combines channel attention and spatial attention. The test accuracy rates on the three data sets of FER2013, CK+, and RAF-DB are 71.25%, 94.57%, and 83.69, respectively. %, the effect is better than the effect of only adding the channel attention module, which is 1.31%, 0.86%, 1.26% higher than that of the baseline network Resnet18, 2.79%, 1.05%, 2.92% higher than the baseline network Resnet18, indicating that the expression recognition algorithm in this article is added The effectiveness of the CBAM module.

3.3. Loss function related experiments and analysis

3.3.1. Parameter setting. In this section, experiments are designed for the proportional weights of softmaxloss and islanding loss. In order to determine the optimal weight value between LS and Lisland, this section sets up different weight values and conducts experiments on the CK+ data set. The accuracy of different weight values is shown in the Table 2.

Table 2. Loss function weight comparison experimental results

	λ_1	λ_2	λ_3	λ_4
Weight	0.1	0.3	0.4	0.5
Accuracy	0.8647	0.8702	0.8813	**0.8924**
	λ_5	λ_6	λ_7	λ_8
Weight	0.6	0.7	0.8	0.9
Accuracy	0.8867	0.8742	0.8721	0.8679

Experiments show that when λ=0.5, the model generated by training performs best on the test set and achieves the highest accuracy value. Therefore, the value of the island loss ratio introduced in this section is 0.5. The weight values λ in the experiments in this section are all set to 0.5.

3.3.2. Confusion matrix. Confusion matrix, also known as error matrix, is a standard format for precision evaluation, expressed in the form of a matrix with n rows and n columns. In the image accuracy evaluation, it is mainly used to compare the classification result with the actual measured value, and the accuracy of the classification result can be displayed in a confusion matrix. The confusion matrix is calculated by comparing the position and classification of each measured pixel with the corresponding position and classification in the classified image. Each column of the confusion matrix represents the predicted category, and the total number of each column represents the number of data predicted to be that category; each row represents the true attribution category of the data, and the total number of data in each row represents the number of data instances of that category. The value in each column represents the number of real data predicted to be of that class.

As shown in Table 3, it is the confusion matrix of FER2013 in the different loss function test experiments in the algorithm of this chapter. Table (a) is the confusion matrix of Resnet+L_S on the FER2013 data set, and table (b) is Resnet+ The confusion matrix of Resnet+L ($L = L_S + \lambda L_{island}$)on the FER2013 data set. You can clearly see the accuracy and improvement effects of various expressions (Angry, Disgust, Fear, Happy, Sad, Surprise, Normal).

Table 3. FER2013 confusion matrix(a)

	Ang	Dis	Fear	Happy	Sad	Sur	Nor
Ang	**64.35**	4.84	8.21	3.71	6.72	2.43	5.31
Dis	4.43	**73.12**	2.41	2.52	3.14	2.14	7.28
Fear	6.52	2.14	**50.33**	3.15	8.19	8.05	5.53
Happy	2.97	3.53	1.24	**91.13**	2.36	4.58	2.79
Sad	6.73	3.58	10.13	2.89	**59.35**	2.43	2.43
Sur	3.51	3.47	6.62	3.16	2.97	**85.14**	7.31
Nor	7.24	0.91	3.63	2.45	8.30	5.41	**69.12**

Table 4. FER2013 confusion matrix(b)

	Ang	Dis	Fear	Happy	Sad	Sur	Nor
Ang	**66.57**	2.44	6.53	2.39	3.58	1.35	2.57
Dis	3.57	**75.23**	2.12	1.25	2.45	0.91	6.26
Fear	4.77	0.81	**52.38**	1.89	7.33	4.82	5.46
Happy	1.87	2.34	0	**92.41**	1.66	2.65	1.98
Sad	5.73	2.63	6.32	2.11	**60.87**	1.37	2.35
Sur	2.53	1.81	4.78	2.34	2.54	**86.52**	5.32
Nor	6.37	0	2.83	1.77	7.38	4.57	**71.24**

As shown in Table 3-3, Table (a) is the confusion matrix of Resnet+LS on the FER2013 data set. It can be seen that the accuracy of the seven types of expressions: Angry, Disgust, Fear, Happy, Sad, Surprise, and Normal are 64.35. %, 73.12%, 49.33%, 90.13%, 59.35%, 85.14%, 69.12%, the misjudgment rate between various expressions is very high.

Table (b) is the confusion matrix of Resnet+L ($L = L_S + \lambda L_{island}$)on the FER2013 data set. The accuracy of the seven types of expressions: Angry, Disgust, fear, happy, Sad, Surprise, and Normal are 66.57%, 75.23%, 52.38%, 92.41%, respectively., 60.87%, 86.52%, 71.24%, the accuracy of the seven types of expressions has increased by 2.22%, 2.11%, 2.05%, 1.28%, 1.52%, 1.38%, 2.12%. The fusion loss function used in this article realizes the inter-class Separate the targets that are aggregated within the class, the misjudgment rate is reduced, and the accuracy of the seven types of expressions has been improved.

3.3.3. Comparative experiment. This section sets up the comparison experiment of the algorithm in this chapter using different loss functions for expression classification, which are the softmax loss function LS, the central loss function L_C and the function L of the fusion island loss proposed in this article, which are in FER2013, CK+, RAF-DB The test results on the data set are shown in the Table 4.

Table 5. Loss function comparison experiment

	FER2013	CK+	RAF-DB
Res18+CBAM	71.25%	93.57%	83.69%
Res18+CBAM+L_S	71.82%	95.26%	84.16%
Res18+CBAM+L_C	72.79%	96.43%	85.34%
Res18+CBAM+L	**73.65%**	**97.37%**	**86.55%**

Among $L = L_S + \lambda L_{island}$,λtake 0.5

The accuracy rates of the loss function using the integrated island loss on the FER2013, CK+, and RAF-DB data sets are 73.65%, 97.37%, and 86.55%, respectively, which are 0.86%, 0.94%, 1.21% higher than the accuracy of the central loss function. Compared with the traditional softmax loss function, the accuracy rate is 1.83%, 2.11%, and 2.39% higher. The experimental results show that the loss function of the fusion island loss proposed in this paper has the best effect, which illustrates the effectiveness of the improved method of loss function in this chapter.

4. Conclusions

This chapter mainly introduces the expression recognition algorithm based on the attention mechanism. First of all, it introduces the problem of low expression recognition rate caused by the interference of redundant feature information and expression similarity in the current expression recognition. A brief overview is given. In the backbone network part, this algorithm uses the Resnet18 network. The input is the entire facial expression image, and the output is a feature map containing deep expression features; an attention mechanism is added to capture the expression feature information and reduce the interference of redundant features., To obtain output features with stronger expressive ability; finally introduce the Islandloss loss function to optimize the distance between classes, improve the discrimination of expressions, and improve the recognition accuracy. In order to objectively evaluate the effect of the model, the accuracy rate is used as the evaluation index of the model, and it is verified on the public data set to prove the effectiveness of the algorithm.

References

[1] Mehrabian A.Communication without words[J].University of East London,1968, 24(4):1084-5.
[2] Bo Sun, Yongna Liu, Jiubing Chen, et al. Emotion analysis based on facial expression in intelligent learning environment [J]. Research on Modern Distance Education, 2015, 000(002):96-103.
[3] Y, Yang, S, et al. Facial expression recognition and tracking for intelligent human-robot

interaction[J]. Intelligent Service Robotics, 2008, 1(2):143-157.

[4] Guanming Lu, Xiaonan Li, Haibo Li. Facial expression recognition of neonatal pain [J]. Acta optica Sinica, 2008(11):75-80.

[5] Lal S , Craig A . DRIVER FATIGUE: PSYCHOPHYSIOLOGICAL EFFECTS[C]// International Conference on Fatigue & Transportation. 2000.

[6] Kai Yu , Lei Jia, Yuqiang Chen, et al. Yesterday, Today and Tomorrow of Deep Learning [J]. Computer Research and Development,2013, 050(009):1799-1804.

[7] Vaillant R , Monrocq C , Cun Y L . Original approach for the localisation of objects in images[J]. Vision, Image and Signal Processing, IEE Proceedings -, 1994, 141(4):245 - 250.

[8] Yang S , Luo P , Loy C C , et al. Faceness-Net: Face Detection through Deep Facial Part Responses[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017:1-1.

[9] Li H , Zhe L , Shen X , et al. A convolutional neural network cascade for face detection[C]// Computer Vision & Pattern Recognition. IEEE, 2015.

[10] Zhang K , Zhang Z , Li Z , et al. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks[J]. IEEE Signal Processing Letters, 2016, 23(10):1499-1503.

[11] Hui D , Zhou S K , Chellappa R . FaceNet2ExpNet: Regularizing a Deep Face Recognition Net for Expression Recognition[C]// 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017). IEEE, 2017.

[12] Jung H , Lee S , Yim J , et al. Joint Fine-Tuning in Deep Neural Networks for Facial Expression Recognition[C]// 2015 IEEE International Conference on Computer Vision (ICCV). IEEE, 2015.

[13] Yang H, Ciftci U, Yin L . Facial Expression Recognition by De-Expression Residue Learning[J]. International Journal on Computer Science & Engineering, 2018, 2(5):2220-2224 vol.3.

[14] Jie H , Li S , Gang S , et al. Squeeze-and-Excitation Networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, PP(99).

[15] Wen Y , Zhang K , Li Z , et al. A Discriminative Feature Learning Approach for Deep Face Recognition[C]// European Conference on Computer Vision. Springer, Cham, 2016.

[16] Challenges in representation learning: A report on three machine learning contests[J]. Neural Networks: The Official Journal of the International Neural Network Society, 2015.

[17] Lucey P , Cohn J F , Kanade T , et al. The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression[C]// Computer Vision & Pattern Recognition Workshops. IEEE, 2010.

[18] Expression Recognition in the Wild[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017.