Facial Expression Recognition Based on Pruning Optimization Technology

Chengyu Jia[1, a, †], Xinyi Li[2, *, †], Rui Qian[3, b, †], Huayu Sun[4, c, †]

[1]Department of Computer Science, Singapore Institute of Management (SIM) Singapore

[2]College of Media Engineering, Communication University of Zhejiang Hangzhou, China

[3]School of oversea education, Changzhou University Changzhou, China

[4]Materials Science and Engineering, Northeastern University Shenyang, China

*Corresponding author: 190807810@stu.cuz.edu.cn, [a]cjia004@mymail.sim.edu.sg, [b]20435515@smail.cczu.edu.cn, [c]1597920988@qq.com

[†]These authors contributed equally.

Abstract. Facial expression recognition is based on deep learning, which has become one of the main topics in artificial intelligence. Nevertheless, the complex network structure causes numerous parameter redundancy, which leads to low recognition efficiency. Therefore, researching how to reduce parameter redundancy makes sense. In addition, today's machine expression recognition has some limitations and constraints, while the generalization ability is poor. Aiming at addressing the problems that excessive parameter redundancy brings, for example, low efficiency and massive limitations. This paper will put forward a facial expression recognition technology that is based on pruning optimization to improve existing models. Based on VGGNet, ResNet and SENet, these three convolutional neural networks modify the convolution layer of the network, then add the attention module and pruning optimization. Through face detection, face area feature representation, expression recognition, and other steps. Realize the recognition of multiple facial expressions. The CK+ dataset and JAFFE dataset are independently divided into 4/5 for training and 1/5 for testing. Experimental results show that after inserting CBAM, except ResNet18, the accuracy of the other two networks all improved, SENet18 improved the most significantly. After pruning, the accuracy and speed of VGG16 and ResNet18 improved, but the accuracy of SENet18 decreased. In summary, the optimized model has good applicability and generalization ability.

Keywords: Network Sliming; image recognition; VGGNet; ResNet; SENet; CBAM.

1. Introduction

Deep Learning also called Neural Networks, brings a great leap to Artificial Intelligence in recent years, particularly in Computer Vision, for example, Image Recognition, Object Detection, Image Super-Resolution and so on. As technology has been developing fast for a long time, the accuracy of our eyes has gradually lost its advantage over machines [1-3].

Nowadays, due to advanced hardware and software foundations and large market demands, facial expression recognition (FER) has become embraced in many research directions. Through the function of a bridge between people and machines, FER enables machines to learn more about what people think so that they can better serve humans. Therefore, FER research has huge potential meaning. Neural networks such as AlexNet [4], VGG [5], GoogleNet [6], ResNet [7] and so on, are constantly emerging to help to measure a lot of hard work. However, as neural networks become more complex, they still cause some problems, such as more network redundancy and lower efficiency.

In addition, larger neural networks require more memory, which brings a significant burden on emergency consumption and hash rate to mobile devices. Furthermore, FER research is primarily in short distances and concentrates on identifying positive photos, so they have obvious limitations. Particularly when it comes to pictures of occluded face or chin, the identification accuracy may undergo a tremendous slide. Thus, how to find a more efficient way to extract pictures' features and

build a more efficient neural network model has become the key point of upgrading identify precision and is also the main direction of FER currently.

Neural networks can be put on mobile devices more easily if people can find an effective way to compress the structure. Therefore, under the premise of ensuring its performance, it can be given certain optimizations to reduce the cost of operation, such as the use of low-rank filters, NSA (Non-Standalone), and knowledge distillation [8-9] and other methods. Many studies have proven that pruning is the simplest and most effective method of cutting based on neural network models. Only a small part of the main calculation is involved, and the removal of redundant parts is the theoretical basis for the pruning technology to achieve. Therefore, on the premise of ensuring its performance, cutting off some unnecessary connections, and using pruning technology to improve the recognition efficiency and rate of accuracy.

Based on the above description, because of the problems of excessive redundancy, low efficiency, and large recognition limitations. The pruning optimization technology based on VGGNet, ResNet and SENet are proposed. Then the attention module is added to realize the recognition of expressions such as joy, anger, surprise, sadness, fear, disgust and other expressions of the face through face detection, face area feature representation, expression recognition and so on. The CK+ dataset and JAFFE dataset can be randomly split into 80% training data and 20% test data. Its results are validated after training. Improve the accuracy of feature map extraction, reduce the parameters and improve the recognition efficiency.

2. Model and method

Here are three representative convolution neural networks in this section, VGG16, ResNet18 and SENet18. This article will illuminate them respectively in different dimensions so that it will be easier to help learn these three neural networks and then make a comparison between them.

2.1. VGGNet Model and Its Optimization

Nowadays, VGG16 is widely used in image recognition. It is a member of the VGG, which was proposed by Karen et al. in 2014 [5]. Comprising 13 convolution layers and 3 fully connected layers, VGG16 adopts 3*3 convolution filters although its architecture substitutes large convolution filters by superposing several 3*3 convolution filters, as shown in Figure 1. In this way, we can cut quantities of parameters in the case of improving quantities of layers, thus increasing the linear expression of VGG16. Therefore, VGG16 has a high operating speed and excellent accuracy performance. Here is its architecture in Figure 2.

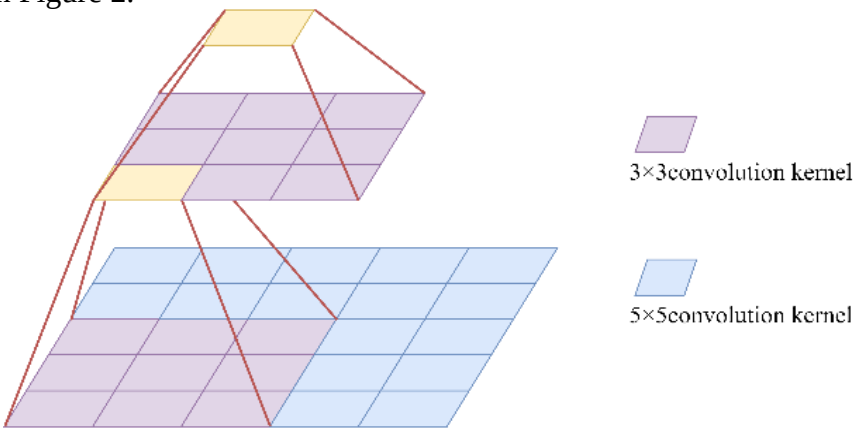

Figure 1. Legend of the convolution kernel

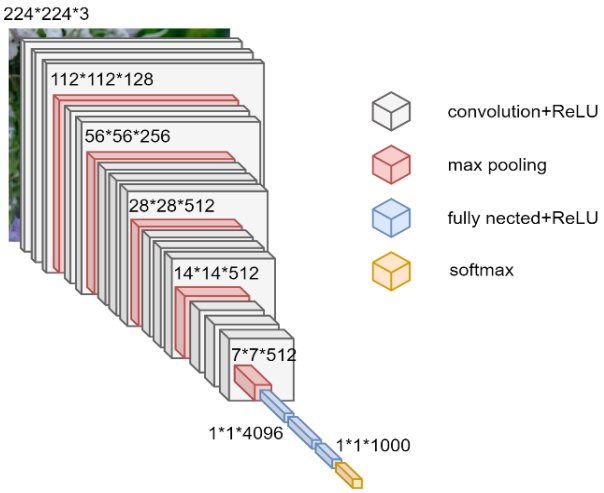

Figure 2. The structure of VGG

However, the three fully connected layers in VGG16 also caused the parameters that needed to be calculated to increase greatly and lost the advantage brought by small convolution filters. Given this situation, this article makes optimization to VGG16, replacing the first convolution layer between the second and the third pooling layer with a 3*3 convolution filter and a 1*1 convolution filter. In addition, cut a complete connected layer. Proven by the calculation, the parameter quantities undergo a great slide. Finally, add several different pruning methods to reduce unnecessary parameters and improve the calculation speed.

2.2. ResNet Model and Its Optimization

With the expansion of the network, features are more abundant and people can get more information. However, as the network gets deeper, it increases to a certain extent. It brings some problems, such as gradient explosion and gradient disappearance. Instead, the accuracy is reduced when testing the trained model. To get more information and features while training a better model, the ResNet network structure is proposed [7]. ResNet is a residual network. It can be regarded as a subnetwork of the network structure. Many networks can make up a deep network. The gradient explosion and gradient disappearance are solved by adding the jump connection of the residual unit. After each convolution, a ReLU activation function [10] is added. The neural network with nonlinear factors can be applied to the nonlinear model. Improving the expression ability of neural networks to models. Figure 3 provides a graphic view of how basic residual block structure works in ResNet. To improve the basic residual block. Three convolution kernels of 1*1, 3*3, and 1*1 are used to replace two convolution kernels of 3*3. Not only to ensure accuracy but also to reduce the parameter number, the improved residual block structure [11] is shown in Figure 4.

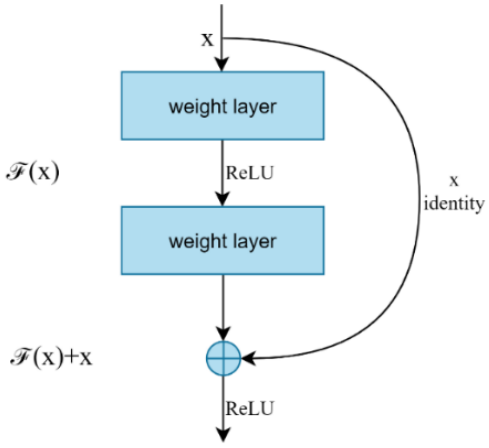

Figure 3. The basic structure of a residual block of ResNet

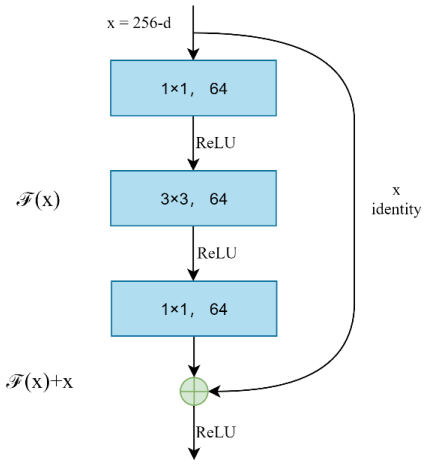

Figure 4. The optimized residual block structure of ResNet

The error term is the gap between the observed and estimated values. Therefore, the problem can be transformed into solving the residual mapping function f (x) of the network, can get the formula

$$H(x)=F(x)+x \tag{1}$$

Where H (x) is the observed value, and x is the estimated value(Feature mapping output from the previous layer). General convolutional neural network requires H (x) =f (x). Exist at a certain moment, when the network can reach the optimal solution, and has the lowest error rate. However, if the network continues to expand, the error rate will increase and the accuracy will decrease. It is necessary to find a method. At the same time, the network is deeper to achieve the same optimal state. ResNet can solve this problem very well. In order to ensure that the network state of the next layer is still the optimal solution, let f (x) =0.

This article uses the ResNet18 model. It consists of 17 convolutional layers and 1 fully connected layer. The network uses zero padding when it needs to add dimensions. Therefore, no other parameters have been added. However, a neural network model based on ResNet18 still has some problems, such as excessive redundancy and low efficiency. Therefore, this paper embeds CBAM in the network to improve the ability of expression feature extraction. Then add channel pruning optimization technology to optimize ResNet18.

2.3. SENet Model and Its Optimization

SENet [12] automatically acquired the importance of each feature channel through learning. Then promoted the useful features according to this importance, and suppressed the features that were not useful for the current task. Given an input X with a characteristic channel number of C1, after a series of convolutions, the channel number becomes C2, turning each two-dimensional channel into a real number, which has a global feeling to some extent, so that the layer close to the input can also obtain all the feel fields.

The weight is generated for each feature channel through parameter W to show the correlation between channels. The weighted output of extension is regarded as the importance of each feature channel after feature selection, and then it is added to the previous feature by multiplication to complete the re-calibration of the original feature in the channel dimension. The structure of SENet is shown in Figure 5.

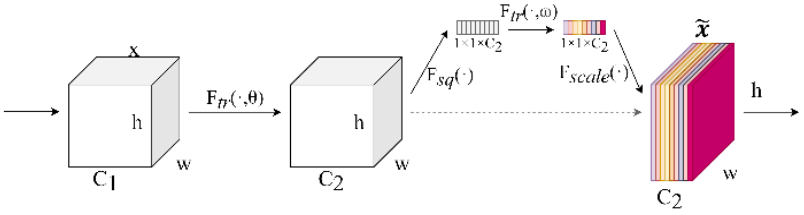

Figure 5. The structure of SENet

First, perform the convolution operation, and the input and output are defined as:

$$\mathbf{F}_{tr}: \mathbf{X} \rightarrow \boldsymbol{U}, \mathbf{X} \in \mathbb{R}^{H' \times W' \times C'}, \mathbf{U} \in \mathbb{R}^{H \times H \times C}. \tag{2}$$

After convolution, we can get U. It is a feature map of size H*W. The Squeeze operation is then performed to convert the input of H*W*C into an output of 1*1*C.

$$z_c = \mathbf{F}_{sq}(\mathbf{u}_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i,j). \tag{3}$$

The result of Squeeze is Z. The dimension of W1 is c/r*c. r is a scaling parameter. The dimension of z is 1*1*C. Therefore, the result of W1 is 1*1*C/r. After passing a RELU layer, the output dimension has not changed. Then multiply with W2, because the dimension of W2 is C*C/r. So the output dimension is 1*1*C. Finally, pass the sigmoid function,

$$\mathbf{s} = \mathbf{F}_{ex}(\mathbf{z}, \mathbf{W}) = \sigma\big(g(\mathbf{z}, \mathbf{W})\big) = \sigma\big(\mathbf{W}_2 \delta(\mathbf{W}_1 \mathbf{z})\big). \tag{4}$$

Then it can be concluded that

$$\tilde{x}_c = \mathbf{F}_{\mathbf{scale}}(\mathbf{u}_c, s_c) = s_c \cdot \mathbf{u}_c. \tag{5}$$

The structure of SENet is very simple and easy to use. There is no need to introduce new functions or layers. However, due to excessive parameters, the training speed is reduced. Prone overfitting, then insert pruning optimization technology to reduce the parameter quantity and redundancy.

2.4. NetWork Sliming

Network Sliming [13] proved a notion of channel-level pruning, it aims at decreasing the parameter quantities in the model by cutting down channels with small weights in way of comparing the scaling factors (γ) with the model's global threshold, so we can reduce the time when doing MAC and then increasing the induct rate of the model. The principle of channel-level pruning based on the BN layer is shown in Figure 6.

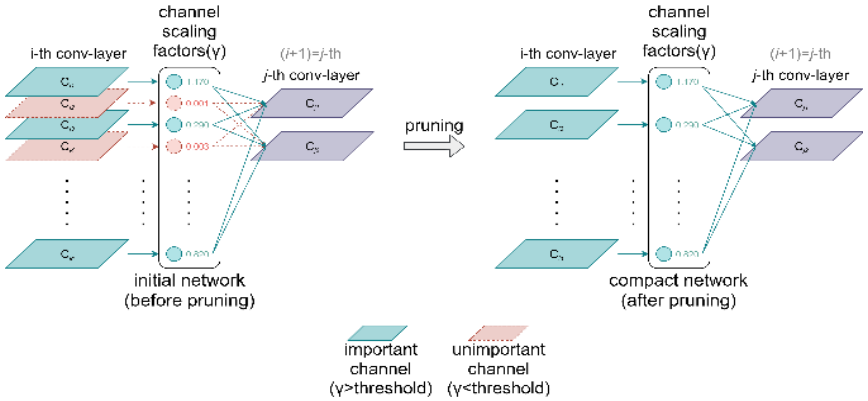

Figure 6. Schematic diagram of BN layer channel clipping principle

The program is as follows. Firstly, applying regularization to γ when we are training convolution neural networks, so that it will be easier to distinguish useless channels as the indicators of data become more comparable. Then, cut down channels that have small γ after setting a threshold and compare it with γ layer by layer. Finally, fine-tune the neural networks to recover the decline in accuracy caused by pruning.

2.5. CBAM

CBAM (Convolutional Block Attention Module) [14] is a lightweight attention module proposed in 2018 that can perform attention operations on both spatial and channel dimensions. The attention mechanism is a mechanism that focuses on local information. On a certain image, the area of attention tends to change from task to task.The attention mechanism is to find the most useful information. CBAM contains two submodules, CAM (Channel Attention Module) and SAM (Spatial Attention Module), which respective perform channel and spatial attention.

CAM passes the input map through two parallel MaxPool layers and AvgPool layers. Change the feature map from size C*H*W to size C*1*1. Then passes through the Share MLP module, in which it first compresses the number of channels to the original 1/r (Reduction, reduction rate) times. After that expands to the original number of channels, and passes through the ReLU activation function to get the result of two activations. The two outputs are added, and then the output of Channel Attention is obtained by a sigmoid activation function. Finally, the output result is multiplied by the original map to change back to the size of C*H*W. Its structure diagram is shown in Figure 7.

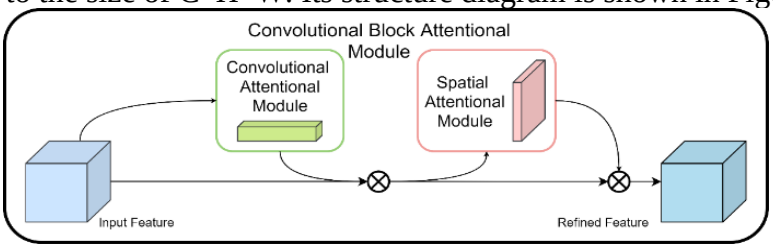

Figure 7. The structure of the CAM module

SAM will channel Attention output results through maximum pooling and average pooling to obtain two 1*H*W feature maps, and then through the Concat operation to join the two feature maps, through the 7*7 convolution into a 1 channel feature map (experimentally proved that the 7*7 effect is better than 3*3). Then through a sigmoid to get the feature map of Spatial Attention, and finally, the output result is multiplied by the original map back to C*H*W size. Its structure diagram is shown in Figure 8.

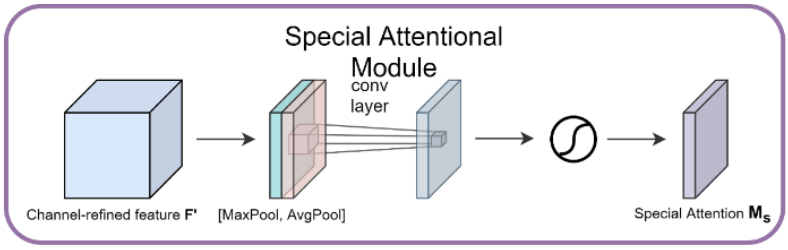

Figure 8. The structure of the SAM module

The two modules of channel attention and spatial attention can be combined in parallel or serial connection, and experimental comparisons have been conducted on the serial order and parallelism on the Channel Attention Module and Spatial Attention Module, and the results of the first Channel Attention Module and then Spatial Attention Module are slightly better. In summary, this article uses the method of combining two aspect modules in the order of Channel Attention and Spatial Attention.

3. Experiment and discussion

Aiming at evaluating the effect after pruning, this article selects an ideal dataset to help conduct the verification. This section will exhibit the feature of 5 datasets and then compare their different points. Besides, how did the dataset processed and what conclusions were cast after verification will also be published in this section.

3.1. Data Set

The emergence of the CK+ dataset is based on the CK dataset, compared with the CK dataset, the number of CK+ dataset increased by 22%, containing 593 face pictures, correcting and verifying the emotional labels in the CK dataset, including anger, disgust, fear, happiness, sadness, surprise, contempt, neutral 8 types of expressions. More advanced algorithms are used to support the architecture of the database.

The JAFFE dataset contains a total of 312 images, containing seven expressions: anger, disgust, fear, happiness, sadness, surprise, and neutrality. Table 1 makes comparisons between several classic datasets.

Table 1. Comparison of data sets

dataset	Sample parameters	Expression
FER2013	It contains 26190 48*48 grayscale maps, and the resolution of the pictures is relatively low	Angry, disgusted, fearful, happy, sad, surprised, neutral
CK+	593 expression sequences	Neutral, angry, contemptuous, disgusted, fearful, happy, sad, surprised
JAFFE	213 images in total	Angry, disgusted, fearful, happy, sad, surprised, neutral
FAR	Contains a total of 29672 images	7 basic expressions and 12 compound expressions
GENKI	GENKI -R2009A contains 11159 images, and GENKI-4K contains 4000 images	Laugh, don't laugh

In summary, the CK+ dataset uses the Caucasian and Yellow face expression sets compared to the JAFFE dataset, which has a better and meets the test requirements of the project, the test accuracy is higher, the sample size is larger, and the robustness of the network can be well improved. Therefore, this paper uses the CK+ and JAFFE datasets and divides them into 80% training and 20% test sets for training and detection. The CK+ dataset is shown in Figure 9 and the JAFFE dataset is shown in Figure 10.

Figure 9. CK+ dataset

Figure 10. JAFFE dataset

3.2. Data Processing

First, the data is preprocessed, the part of the specified picture is cropped, the image is normalized, transformed into tensor data type, and the data set is deeply learned and verified by using VGG16, ResNet18 and SENet18 convolutional neural network models, which are divided into 80% training set and 20% test set for training and detection, and each network is trained with 200epoches because of the rapid network convergence. After adding the CBAM attention module to the convolutional neural network models VGG16, ResNet18 and SENet18, repeat the above learning and verification. Based on adding the CBAM attention module, the pruning optimization technique is added to the three networks to compare the accuracy after the experiment.

3.3. Result Analysis

After the insertion of the attention module, the accuracy of the other two networks has improved except for ResNet18, and the improvement of SeNet is the most significant.

After pruning, the accuracy of VGG16 and ResNet18 is improved, and the computing speed is also greatly improved. However, SENet18 has decreased accuracy (it may be singular data).

Figure 11 shows the loss function of LR = 0.01 and epochs = 200. Tables 1, 2, 3 and 4 show the test accuracy of VGG16, ResNet18 and SENet18 respectively.

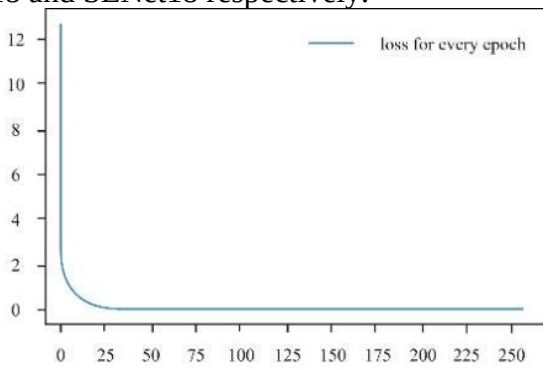

Figure 11. Loss function

Table 2. The Test Accuracy of VGG16

VGG16			
	1	2	3
Test accuracy(%)	91.9	92.9	94.9
Average accuracy(%)	93.23		

Table 3. The Test Accuracy of ResNet

ResNet18			
	1	2	3
Test accuracy(%)	93.9	87.9	91.9
Average accuracy(%)	91.23		

Table 4. The Test Accuracy of SENet

SENet18			
	1	2	3
Test accuracy(%)	80.8	90.9	71.7
Average accuracy(%)	81.13		

Table 5. The Test Accuracy of VGG16 and ResNet After Pruning

	accuracy after pruning	Parameter quantity after pruning
VGG16	93.94%	14726727
ResNet18	89.90%	11172423

4. Conclusion

This research goal is to find an effective lightweight network model that adds an attention module to a different convolutional neural network and then compare the consequences to them after pruning. We selected three representative convolutional neural networks for the experiment, VGG16, ResNet18 and SENet18 so that the results of this research will be more universal.

VGG made optimizations based on AlexNet, which proposed replacing large convolution filters with small convolution filters, and this remarkably improved the performance of the convolutional

neural network. In the VGG family, VGG16 is one of the members for it adopts 3 × 3 convolution filters although the architecture, and so that VGG16 becomes the representative of classical deep neural network.

The appearance of ResNet is only one year behind VGG, it raised a milestone structure -- residual network. Residual networks improved the accuracy of recognition by solving the problem that the gradient gradually decreases or even disappears as the number of network layers cleverly increases. In short, ResNet has opened a new era, after which the residual network has been adopted by many new modules, YOLOV3 is one of them.

SENet won the crown of the last ImageNet with an overwhelming advantage and brought Squeeze-and-Excitation Module (SE Module) to the public. This architecture encouraged scientists to conduct further research. Furthermore, SE Module inheres powerful portability, it can help improve the performance of the different neural networks by porting SE Module into them, for example, the deployment of SE Module on ResNet has achieved ideal results.

Based on the above considerations, we chose these three classic convolutional neural networks as the carriers of this experiment. In addition, we compared many different pruning methods. From classic BN layer pruning algorithms, Alpha-Beta pruning algorithms to Taylor series pruning algorithms. At the same time, we consulted a lot of high-level literature. Finally, the BN layer channel pruning was selected to optimize these three models.

At present, expression recognition has many uses. For example, judge the psychological activities of criminals through micro-expressions, use mobile terminals for psychological monitoring, observe users' expressions in the mall in real time to judge their satisfaction and so on. Due to limited time and effort, we only select some classical convolutional neural networks and common pruning algorithms for improvement and experiments. In the following study, we will focus on adopting a more diversified network lightweight approach to find more efficient network models in the field of facial expression recognition. In addition, we will also be committed to deploying the lightweight network model to mobile devices to achieve the actual application requirements.

References

[1] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, “The Extended Cohn-Kanade Dataset (CK+)_ A complete dataset for action unit and emotion- specified expression,” in Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on, 2010, pp. 94-101.

[2] LYONS M J, AKAMATSU s, KAMACHI M G, et al. Coding facial expressions with Gabor wavelets [C]//Third IEEE International Conference on Automatic Face and Gesture Recognition, 1998:200-205.

[3] Ian Goodfellow, Yoshua Bengio, Aaron Courville Deep Learning [M] Boston: MIT Press http://www.deeplearningbook.org 2016.

[4] Krizhevsky A, Sutskever I, Hinton G. lmageNet classification with deep convolutional neural networks [C]//Proceedings of the 25th International Conference on Neural Information Processing Systems, 2012:1097 - 1105.

[5] A. Zisserman, 2014 “Very Deep Convolutional Networks for Large-Scale Image Recognition” [Submitted on 4 Sep 2014 (v1), last revised 10 Apr 2015 (this version, v6)]

[6] Szegedy C, Liu W, Jia Y Q, et al. Going deeper with convolutions [C]//Proceedings of IEEE Conference on ComputerVision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2015: 1-9.

[7] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[c]//Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition,2016:770-778.

[8] LIN Jing-Dong, WU Xin-Yi, CHAI Yi, et al. Structure Optimization of Convolutional Neural Networks: A Survey. ACTA AUTOMATICA SINICA, 2020, 46(1): 24-37. doi: 10.16383/j.aas.c180275.

[9] LI Jiang-Yun, ZHAO Yi-Kai, XUE Zhuo-Er, et al.. A survey of model compression for deep neural networks [J]. Chinese Journal of Engineering, 2019, 41(10): 1229-1239.doi: 10.13374/j.issn2095-9389.2019.03.27.002.

[10] WANG Hong-xia, ZHOU Jia-qi, GU Cheng-hao, LIN Hong, Design of activation function in CNN for image classification [J] Journal of Zhejiang University (Engineering Science), 2019, 53(7): 1363-1373.

[11] ZHOU lie, MA Ming-dong (School of Telecommunications & Information Engineering. Nanjing University of Posts and Telecommunications.Nanjing210003.ChinaSchool of Geographical and Biological Information, Nanjing University of Posts and Telecommunications. Nanjing 210003. China)

[12] Hu J, Shen L, SunG. Squeeze-and-excitation networks [C]. IEEE/CVF Conference on ComputerVision and Pattern Recognition. Salt Lake City, 2018:7132-7141.

[13] Zhuang. L, Jian. G. L, Zhi. Q. S, Gao. H, Shou. M. Y, Chang. S. Z, 2017, "Learning Efficient Networks through Network Slimming" IEEE 22-29 DOI: 10.1109/ICCV.2017.298

[14] WOO, PARK J, LEE J Y, et al. CBAM: convolutional block attention module [C]//European Conference on Computer Vision, 2018: 3-19.