Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 174 (2020) 463–477

Procedia Computer Science

www.elsevier.com/locate/procedia

2019 International Conference on Identification, Information and Knowledge in the Internet of Things (IIKI2019)

Facial Expression Recognition via a CBAM Embedded Network

Wenhao Cao[a], Zhuoyu Feng[a], Dongyao Zhang[a], Yisiyuan Huang[a]

Northeastern University, Shenyang, Liaoning, China
502415368@qq.com
Shandong University, Qingdao, Shandong, China
201700121022@mail.sdu.edu.cn
Xi' an Jiaotong University, Xi'an, Shaanxi, China
lnsyldy@stu.xjtu.edu.cn
Boston Trinity Academy, Boston, US
1357873156@qq.com

These authors are contributed equally to this work

Abstract

Facial expression recognition plays an important role in face recognition. Face recognition helps computer to figure out people faces from a simple scenery and to recognize who they are. But facial expression recognition assists computer to analyze the emotion state of one single person in order to improve the human-computer interaction experience. For facial expressions, there are certainly lots of conspicuous features for observers to look on, like shapes of eyes, mouth. When people smile, their lips turn up and their eyebrows bend downward. In this way, we get features of smile. And in same way, we can also do the same to anger, sadness, and surprise etc. Single neural network can make it. However, the accuracy is always low, nearly 70% up and down, which always companied with fluctuation, like electrocardiogram, though not that severe, according to experiments. Therefore, we choose to include Convolutional Block Attention Module (CBAM) into some layers of VGG network we used to improve the accuracy and also the stability. CBAM is an effective attention module for neural networks to concentrate on features. With CBAM, we simplify the scenery that networks need to analyze, and then we can finally make an improvement. This way, we can build better human-computer interfaces.

Keywords: expression recognition; attention mechanism; embedded CBAM; feature extraction

1. Introduction

Face recognition [7] [8], based on the features information of human faces to process the identity recognition, is one

10.1016/j.procs.2020.06.115

kind of biometric recognition technology. It takes advantage of camera to collect both image and video flow having human faces, to detect and trace human faces in it automatically. At the same time, Facial Expression Recognition (FER), do the similar job as face recognition. But FER doesn't aim at figuring out the identity information of one single body. It extracts the specific state of facial expression from static images or dynamic video sequences to recognize the state of somebody's emotion. Then, computer can achieve the deep understanding of human faces. And ultimately, people can make the relationship between themselves and computers better to gain better human-computer interaction.

1970s, Ekman *et al* proposed six fundamental types of facial expression (viz. happiness, anger, surprise, fear, sadness, and disgust), which defined the types of recognition object. And then they built the Facial Action Coding System (FACS) to partition a series of human face Action Unit (AU), the description of motion of human face. 1990s, A Pentland and K Mase used light flow to judge the main direction of how facial muscles moved to derive the value of light flow in a partial space. Finally, they gained expression specification vector, with which they built up their facial expression recognition system, recognition rate of which reached nearly 80%. But they can only recognize four different kind of expression (viz. happiness, anger, disgust, and surprise). So people are going to improve the model.

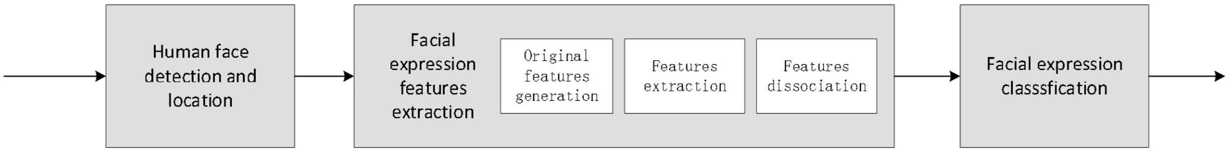

Human face detection and location
Facial expression features extraction
Original features generation
Features extraction
Features dissociation
Facial expression classsfication

Figure 1. Facial expression recognition flow

Fig.1 shows a complete FER system, which can be summarized into three steps (viz. preprocess, features extraction, and classification). The first step is detection and location, which means finding the position of human face in the input images. The fundamental idea is to use knowledge or statistic method to build human face model, and to compare the similarity between detection region and the model, in which way we can obtain the possible region containing human face. Because this step is the start of the entire system and also the basic part of the system, it has become one single independent study direction [1] [2] [4]. The second step is the extraction of facial expression features, which can be divided into two types according to the characters of images (viz. static image features extraction and sequential images features extraction). For static images, the extraction is to find the deformation characteristics of human faces, while for sequential images, we need to extract not only the deformation characteristics per frame, but also the motion characteristics of constant sequences. The studies in dynamic sequential textures aim at simplifying and extending the approached for recognizing facial expression in continuing facial images [3]. The last step is to classify the expression derived from the two forward steps. Classification means basing on the relationship among different features to put the facial expression into fitting class, like the six fundamental expression types [4].

There are four basic types of approaches for the last step:

- Module-based matching
- Using convolutional neural networks (CNN) to classify
- statistic methods
- Methods based on support vector machine [5-6]

Far from traditional approaches for extracting features and also classification, using CNN methods [9] [10], deep learning, can avoid big troubles for people to do the same job themselves, because CNN does well in features extraction fields, especially on images [10]. For the structure of networks, the common idea is to pick up the features through several layers of CNN, and then using full-connected layers to do non-linear classification. This way, the full-connected layer can work like Multi-Layer Perception (MLP) [12]. Although traditional machine learning methods such as the four basic beforementioned approaches have been successful in some situations such as controlled environments when human faces are fixed towards a fixed direction (Fig.2), they don't work as well as before when the applied datasets or databases change irregularly (Fig.3). Simultaneously, it's always difficult to obtain clear, accurate training data [13]. However, due to the growing of calculating ability and the increasing capacity of training databases, using neural networks has become more and more popular.

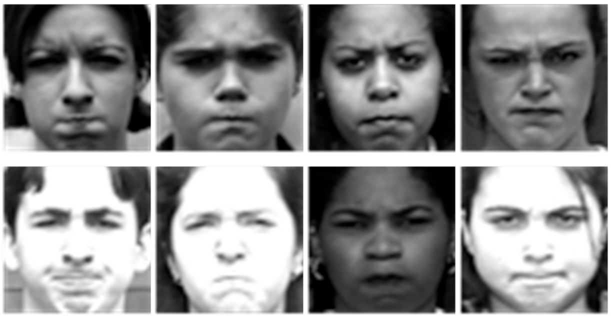


Figure 2. Fixed databases

Figure 3. Human faces toward different directions

Convolutional neural networks have dramatically improved the performance of recognition tasks [14] [15] due to the growing calculating power. Nowadays, there has been many networks proposed, like Residual-style Net (ResNet) [16] [18]s, VGG net etc. Besides these networks, scholars are also working on specific network layers in order to improve the performance by taking advantage of both hardware and concrete code. People have proposed many modules that help. For example, Shaoqing Ren *et al* had given out a faster R-CNN [17]. Sanghyun Woo *et al* also gave out a Convolutional Block Attention Module (CBAM) [19], with which we can finish our experiment in making FER a little bit better. Thanks to Ren's team, we insert CBAM into different layers so that we can compare the effect under different situations with different parameters.

In this paper, we will give out our experiment results using CBAM embedded VGG network in FER. Though CBAM has been put forward for several years, we still haven't seen many experiments using this module for FER. As follow parts of this paper will show, with the customized network and also FER2013 database, we can achieve better accuracy and stability in FER.

Oxford Visual Geometry Group (VGG) belonging to Robotics Research Group, firstly proposed VGG net. We chose VGG net as our base neural network because of its great power in extraction, which means it works really well in extracting features from both simple and complex images, although it has too many parameters we need to set. Our team insert CBAM into different layers of VGG net in order to obtain better recognition result. For instance, we embedded CBAM into every layer of the net, and next time we also did the same to max-pool layers, next time so did we to full-connected layers.

With our experiments, we can make following contributions:

- We take CBAM, an advanced attention mechanism module into VGG net so that we can raise the performance of FER from code layer.
- We set different situations and parameters, then we find out a better one by inserting channel attention function and special attention function into max-pool layers.

- We evaluate the performance (viz. accuracy and stability) of every approach that we took.

2. Related Work

There's a simple fact that currently most of researches on FER aim on how to make use of human face information automatically, stably, efficiently. These researches can be traced back until 1970s when Charles Darwin firstly proposed that consistency existed among different emotions of different sex, different race of people. Ekman and Frisen raised Facial Action Coding System (FACS) using 44 action units (AU) to descript the motion of human face, and defined six basic types of expressions, which gained broad agreement. That partition plays an essential role in today's researches. Pantic *et al* [20] summarized the techniques and approaches of FER in 2002. They provided readers with possible directions of FER. Bazzo *et al* [21] used Gabor in FER and succeeded in recognizing expressions from neutral faces. S. Lawrence *et al* [11] applied a convolutional neural network approach for FER in 1997. That way, many scholars began using CNNs to extract features and also to classify the expressions. S.K. Pal and S. Mitra [12] proposed a better method in classification step. They used a fuzzy neural network model based on the multilayer perceptron to compared with conventional MLP, and finally validated the effectiveness of their approach.

However, these previous works on FER were all based on an ideal situation, which assumed that each input images had a unique, fixed mode. This way, they can certainly tell the differences among different expressions. But it's not realistic. With both quality and quantity of study increasing, scholars start realizing the importance of real data. Deng *et al* [14] had thought about a big and complex database, which is not designed deliberately. That database is closer to reality, including faces towards different directions, images with different sizes, and also with hand gestures. It indeed brought great challenge into research, but made sense. Except this kind of database idea, people like Gauthier [22] attempted to generate ideal data with cGANs. Also, Radford *et al* [23] tried to enlarge the ability of GANs by using CNNs to generate data images. In their effort, they not only scaled up GANs, they also picked up their new structure of deep convolutional generative adversarial networks (DCGANs), which means people now can derive data by using vector arithmetic. These instances show people are always working on improve ability from simulating truth to reflecting it.

So far, there are lots of researchers proposing lots of neural networks for FER [9] [10] [13] [16] [17] [18]. Besides these researches on different kinds of whole networks, we also have many studies on specific content such as single layers of networks, different sequences of networks, and also embeddable module that can be inserted into networks. Sanghyun Woo *et al* [19] introduced such kind of embeddable module, Convolutional Block Attention Module (CBAM), including channel attention module and spatial attention module, with which our net can concentrate on concrete facial expression features so that we can simplify the work.

3. Proposed Method

The SE block [24] and CBAM [19] block are light-weight modules which are likely to be effectively integrated into standard structures like ResNet [16] and VGGNet [25] to improve the accuracy of convolutional neural networks. The flexibility of the blocks determines that they are light enough and feasible to be applied in many potential positions of the network. It is always recommended to insert the block after the non-linearity following convolution layers in VGGNet. To fulfill the facial expression recognition task, we combine CBAM block with standard VGG structure and apply several adjustments.

3.1 VGGNet

The ConvNet configurations of VGGNet, with width starting from 64 in the first layer and until 512 in the last layer, differ in depth which varies from 11 weight layers as the minimum to 19 weight layers as the maximum. The net with 19 weight layers achieves lower rate of test error, thus the configuration is utilized in our proposed structure. To avoid overfitting, dropout strategy is utilized in the model before Fully-Connected (FC) layers.

In standard VGGNet, three fully-connected layers follow a stack of convolutional layers to improve the ability of discrimination. In this case, the traditional non-linear layers of rectification are replaced by just a single layer with 7 channels representing 7 classes. The prediction of the network is determined by the index of the maximum value in the 7 channels.

3.2 Convolutional Block Attention Module

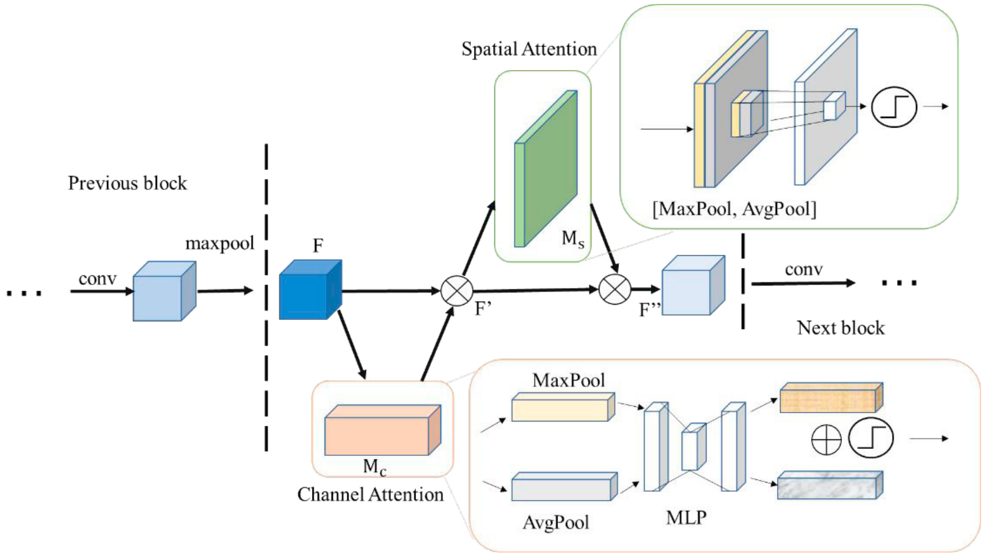


Figure 4. CBAM in VGGNet

Convolutional Block Attention Module (CBAM) applies attention mechanism. The attempts to incorporate channel attention processing and spatial attention processing help refine the feature maps and improve the performance. Compared to SENet, CBAM concerns about spatial features instead of merely channel features and utilizes average pooling method instead of merely max pooling. In that case, the attention generation process in CBAM performs better and is more robust.

CBAM sequentially produces a sub-module of channel attention as well as one of spatial attention. The channel sub-module applies the shared network of outputs of both max-pooling and average-pooling. The spatial sub-module arranges the two types of outputs along the channel axis. Finally, they are forwarded to a convolution layer.

Channel sub-module:

Average-pooling and max-pooling aggregates spatial statistic features and spatial information about more meaningful parts. Both descriptors of the average-pooling operation and max-pooling operation are forwarded to produce the channel attention map by a multi-layer perceptron (MLP). The multi-layer perceptron contains one hidden layer whose activation size is reduced by the reduction ratio 16. The output feature vectors are merged into the channel attention map.

The channel attention is computed as:

$$\begin{aligned} \mathrm{M_c}(\mathrm{F}) &= \sigma(MLP(AvgPool(\mathrm{F})) + MLP(MaxPool(\mathrm{F}))) \\ &= \sigma\big(\mathrm{W}_1\big(\mathrm{W}_0\big(\mathrm{F}^{\mathrm{c}}_{\mathrm{avg}}\big)\big) + \mathrm{W}_1(\mathrm{W}_0(\mathrm{F}^{\mathrm{c}}_{\mathrm{max}}))\big), \end{aligned}$$

where σ denotes the sigmoid function, $\mathrm{W}_0 \in \mathrm{R}^{C/r\times C}$, and $\mathrm{W}_1 \in \mathrm{R}^{C/r\times C}$. The weights of the multi-layer perceptron, W_0 and W_1, are shared for both inputs. The weight W_0 follows the ReLU activation function.

Spatial sub-module:

The efficient feature descriptor applies both average-pooling and max-pooling operations. Those treatments are put along the channel axis. A convolution layer is applied to generate a spatial attention map on the feature descriptor concatenated both operations.

The spatial attention is computed as:

$$\mathrm{M_S}(\mathrm{F}) = \sigma(f^{7\times7}([AvgPool(\mathrm{F});MaxPool(\mathrm{F})]))$$
$$= \sigma(f^{7\times7}([\mathrm{F}^{\mathrm{s}}_{\mathrm{avg}};\mathrm{F}^{\mathrm{s}}_{\mathrm{max}}])),$$

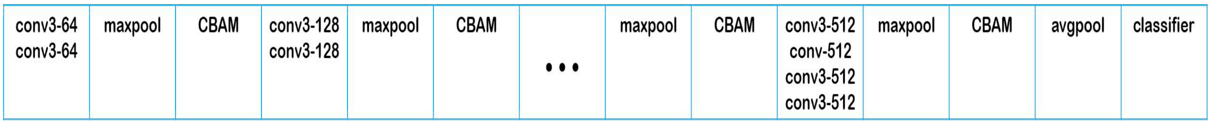
where σ denotes the sigmoid function. With the filter size of 7×7, $f^{7\times7}$ represents a convolution operation.

The whole attention is computed as:

$$\mathrm{F'} = \mathrm{M_C}(\mathrm{F})\circ\mathrm{F}$$
$$\mathrm{F''} = \mathrm{M_s}(\mathrm{F'})\circ\mathrm{F'},$$

where ○ denotes element-wise multiplication

3.3 CBAM in VGGNet



conv3-64 conv3-64	maxpool	CBAM	conv3-128 conv3-128	maxpool	CBAM	• • •	maxpool	CBAM	conv3-512 conv-512 conv3-512 conv3-512	maxpool	CBAM	avgpool	classifier

Figure 5. Configuration of CBAM in VGGNet

The inputs to our network are fixed-size 44×44 gray images which come from the cropping preprocess and flips of 48×48 images.

The CBAM block should be inserted after the non-linearity following convolution layers. In our VGGNet with 19 weight layers, 13 convolution layers and 5 max-pooling layers are utilized.

Inserting the CBAM block after each convolution layer will add to the complexity in the network architecture and training process. Although CBAM is a light-weight module, overutilization will definitely destroy its feature of light weight. The number of max-pooling layers is comparatively small, which will contribute less in complexity. Therefore, the CBAM block is inserted after each max-pooling layer in the refined network.

3.4 Loss function

Cross entropy loss function is utilized to indicate the rate of error in the multivariate classification.

The loss is calculated as:

$$loss =-\frac{1}{n}\sum_{i=1}^{n}(y_i \log(p_i) + (1-y_i)\log(1-p_i)),$$

where y_i denotes the labels of the classes and p_i denotes the possibilities of the classes

4. Experiments

We conduct experiments on two public facial expression databases, including CK+ [26] and FER-2013 [27]. And we use comparison to see the effects of embedding CBMA into the original network VGG19. The following sections report the performance of our designed methods on these two datasets.

4.1 Experiments on CK+

We first conduct experiment on the CK+ dataset. The last three frames from each sequence in the CK+ dataset are extracted, which contains a total of 981 images. Those 981 images are divided into 7 kinds of facial expressions, angry (135), disgust (177), fear (75), happy (207), sad (84), surprise (249) and contempt (54). We use 10-fold cross validation in the experiment, selecting 99 validation(public) images and 99 test(private) images according to the proportion of the 7 categories, the rest are training images.

In case the data is not enough so that then network overfits too quick, we use data augmentation to enlarge the data size in the database and make the network more robust. In the training phase, we randomly cut and flip the 44*44 images and then put them into the training set. While in the test phase, we do cutting and flipping operations on the top left corner, top right corner, left bottom, right bottom and center of every image. Those operations enlarge the database by ten times and effectively reduce the error rate of classification.

To test the training accuracy, we set 100 epochs, 128 as batch size, 0.01 as learning rate and then test a single fold. Fig. 6 shows the training accuracy curves of VGG19 network with and without CBAM embedded respectively.

The resulting accuracy performances are that with CBAM embedded, the accuracy converges to 96%, however, the accuracy number is only 91% without CBAM.

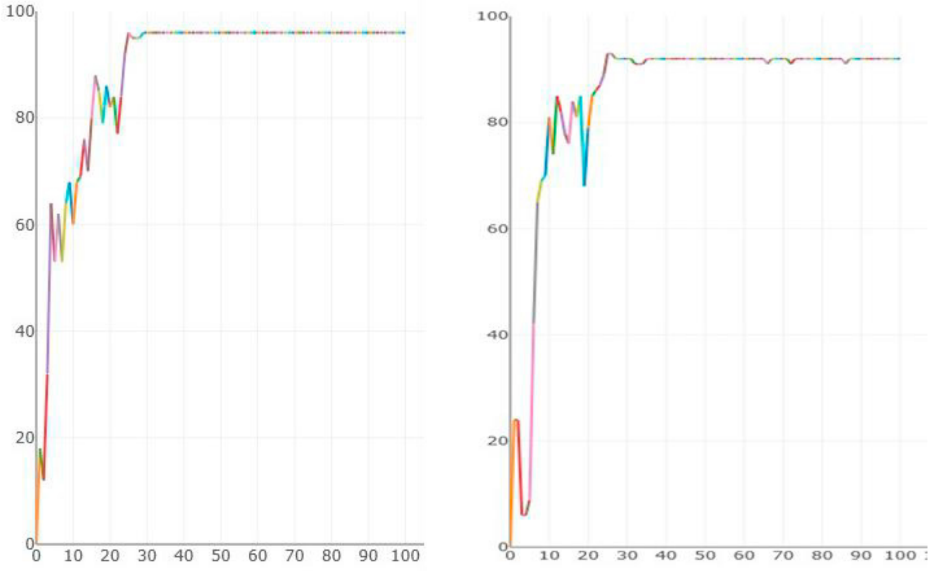


Figure 6. The validation(public) accuracy curves of VGG19 network with CBAM embedded(left) and without CBAM embedded(right) on CK+ dataset

We also compute and plot the loss functions in these two training situations, both using SGD (Stochastic Gradient Descent) as the optimizer. The performances are shown in Fig. 7. The loss value converges to only 0.067 with CBAM embedded while converges to 0.145 without CBAM embedded.

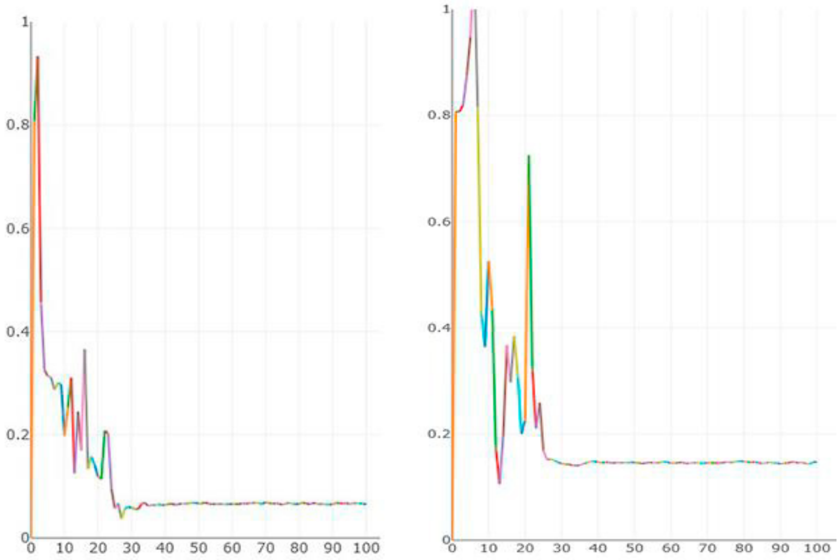


Figure 7. The loss function curves of VGG19 network with CBAM embedded(left) and without CBAM embedded(right) on CK+ dataset

Finally, we obtain 10 models after training and test the test(private) accuracy using test(private) images by calculating the average of these 10 models. Fig. 8 and Fig. 9 shows the resulting confusion matrixes.

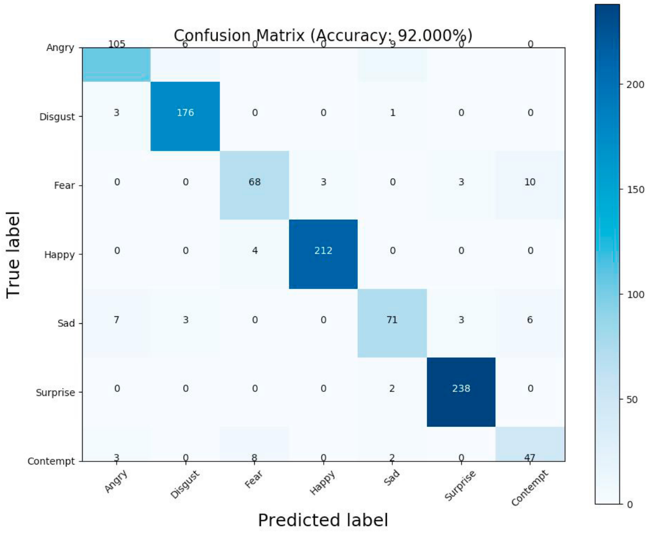


Figure 8. The confusion matrix of VGG19 network with CBAM embedded on CK+ dataset

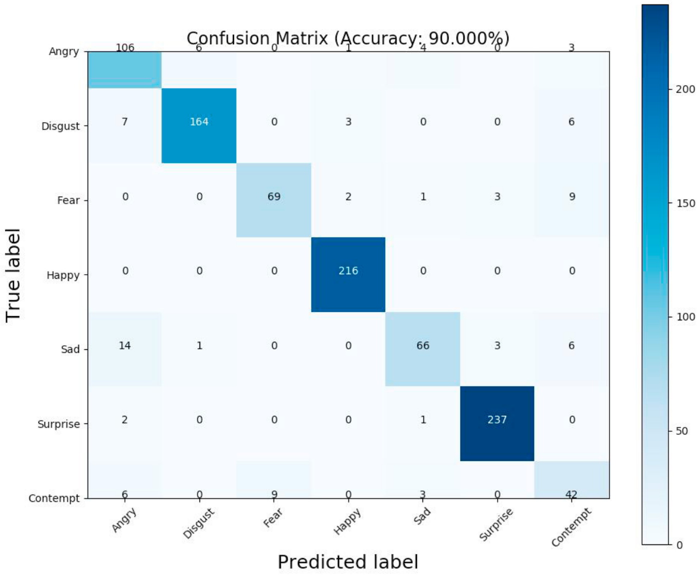


Figure 9. The confusion matrix of VGG19 network without CBAM embedded on CK+ dataset

And we put the all the comparative results into a table shown in Table 1.

Table 1. The comparative results on CK+ dataset

	VGG_CBAM	VGG

Validation (Public) Accuracy	96%	91%
Test (Private) Accuracy	92%	90%
Training Loss	0.067	0.145

4.2 Experiments on FER-2013

Next we conduct experiment on the FER-2013 [27] dataset, which is composed of 28709 training images, 3589 public images and 3589 private images. Every image is a 48*48 grey-scale image. The database contains 7 kinds of expressions: angry, disgust, fear, happy, sad, surprise and neutral.

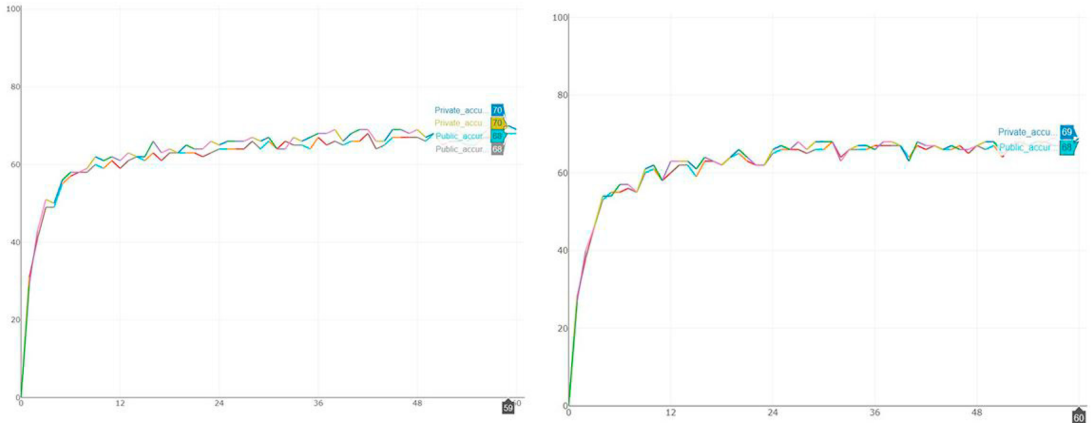


Figure 10. The validation(public) accuracy and test(private) accuracy curves of VGG19 network with CBAM embedded(left) and without CBAM embedded(right) on FER-2013 dataset

We use the same data augmentation method and the same optimizer SGD with these in the last experiments, setting 60 epochs, 128 as batch size, 0.01 as learning rate and then train the model. The difference is that we no longer use 10-fold cross validation in this experiment, so we only obtain two models, one has CBAM embedded and the other hasn't. We put both validation(public) and test(private) accuracies of a model together in a single figure, which are shown in Fig. 10.

As for the loss function, the performances are shown in Fig. 11. The loss values of these two models both converge to quite small numbers, the performance of VGG19 with CBAM embedded is a little better than VGG19 without CBAM embedded.

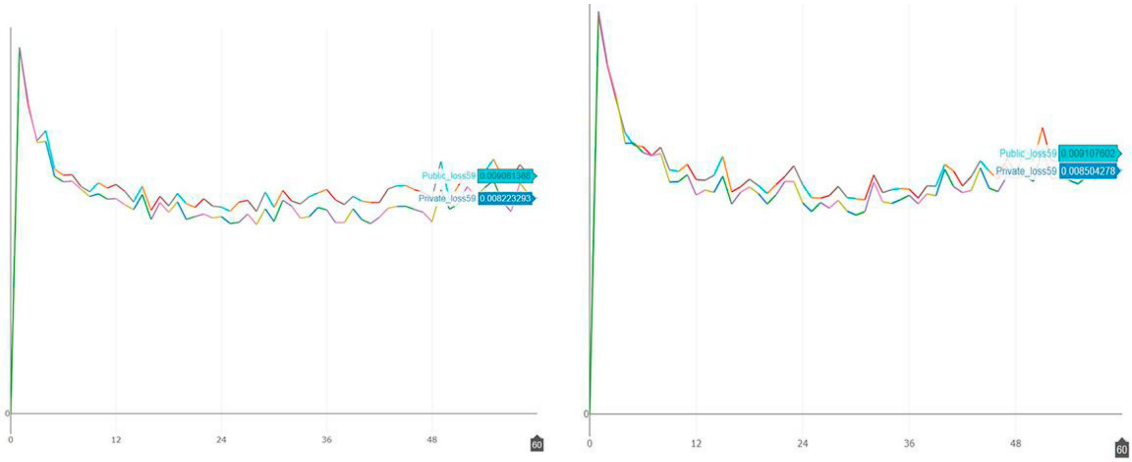


Figure 11. The loss function curves of VGG19 network with CBAM embedded(left) and without CBAM embedded(right) on FER2013 dataset

Finally, we plot the public test and private test confusion matrixes of two models.

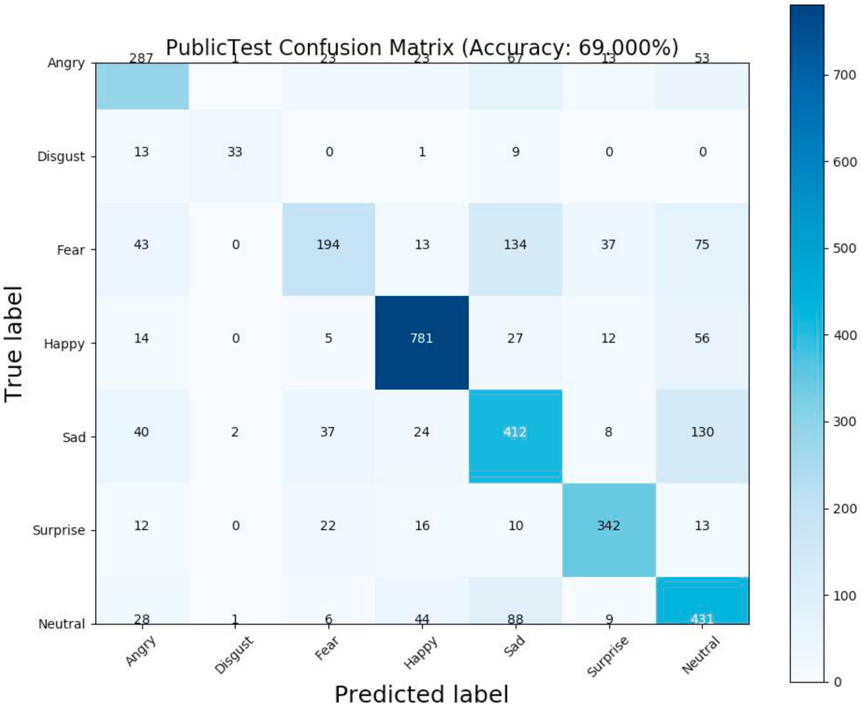


Figure 12. The public test confusion matrix of VGG19 network with CBAM embedded on FER-2013 dataset

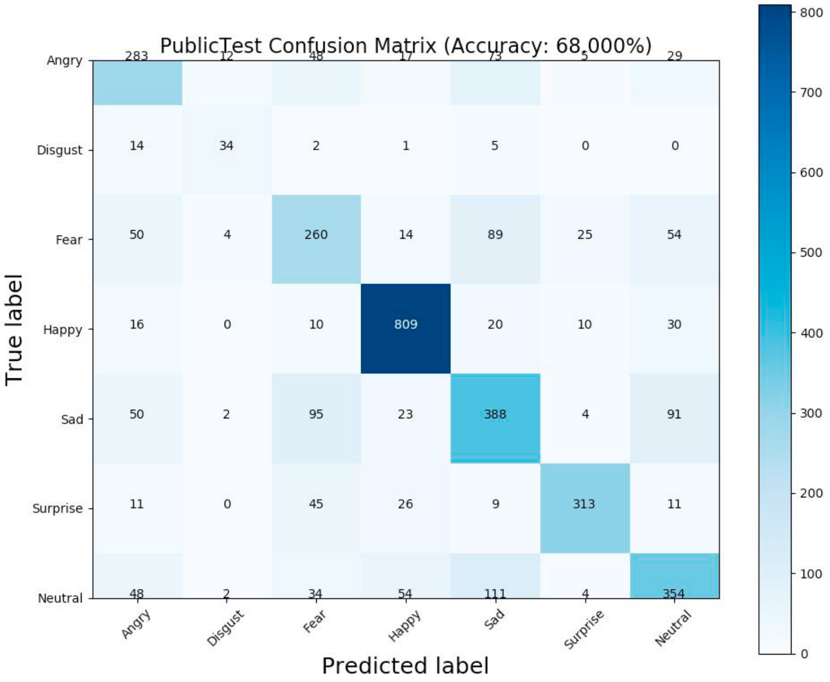


Figure 13. The public test confusion matrix of VGG19 network without CBAM embedded on FER-2013 dataset

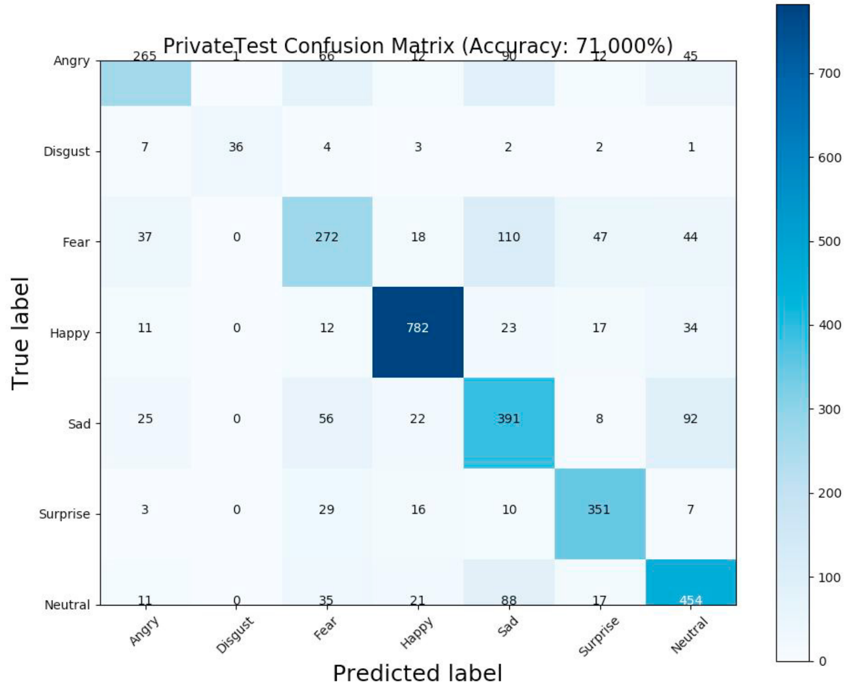


Figure 14. The private test confusion matrix of VGG19 network with CBAM embedded on FER-2013 dataset

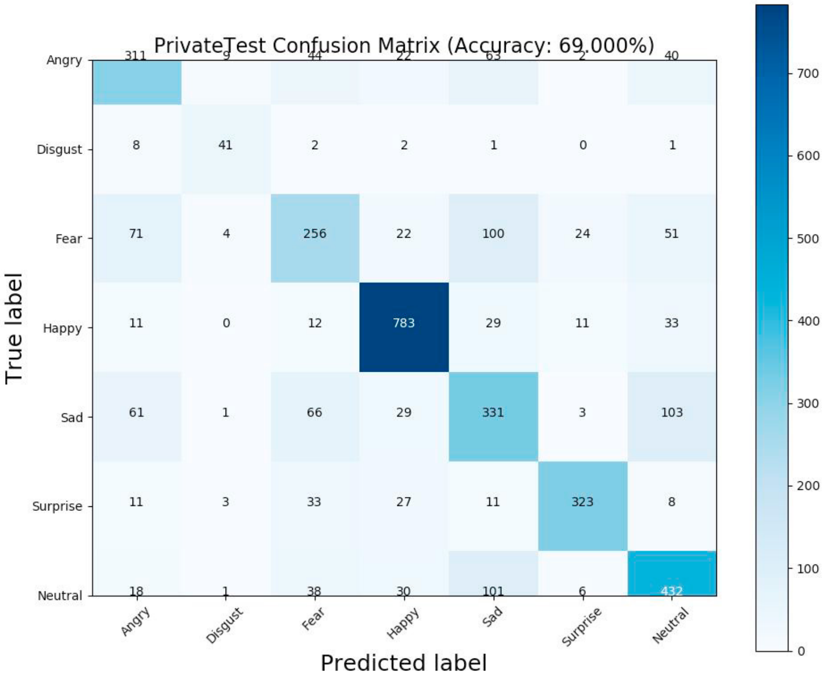


Figure 15. The private test confusion matrix of VGG19 network without CBAM embedded on FER-2013 dataset

And we put the all the comparative results into a table shown in Table 2.

Table 2. The comparative results on CK+ dataset

	VGG_CBAM	VGG
Validation (Public) Accuracy	69%	68%
Test (Private) Accuracy	71%	69%
Validation (Public) Loss	0.00908	0.00911
Test (Private) Loss	0.00822	0.00850

4.3 Practical Application

At last we apply our models to recognize facial expressions besides our datasets. We select some classic facial expression images according to our expression categories from the Internet, and then plot the classification results. Because there are several kinds of facial expressions that share many common features like surprise and fear, angry and disgust, the results may have a little bias. However, most of them are accurate, conforming to human cognition. Fig. 16 includes two of the results.

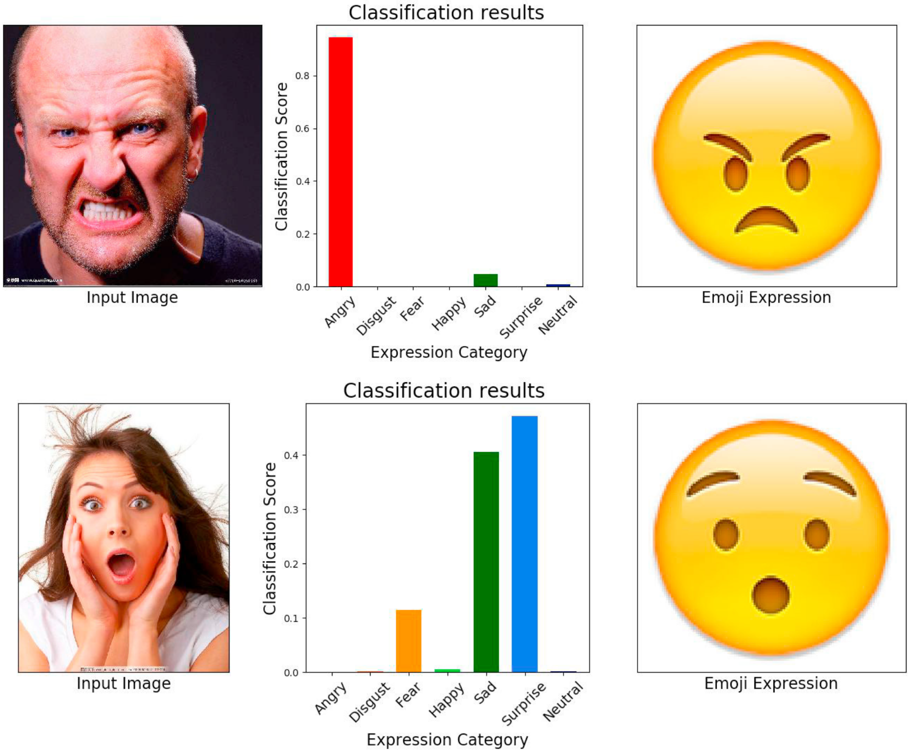


Figure 16. Classification results

5. Conclusions

In this paper, we have proposed a new approach for facial expression recognition, which is based on a deep convolutional neural network VGG19 with CBAM embedded. Our model further boosts the performance of the original model, achieving higher accuracy and stability on both CK+ and FER-2013 datasets, indicating the possible application and potential of our facial expression recognition approach.

Acknowledgements

The completion of the thesis is attributed to many people's support and encouragement. Without of the help of many people,it could never be completed, thus here we would like to express my sincere gratitude towards them.
First and foremost, we owe our heartfelt thanks to my distinguished and cordial supervisor,Professor Xueen Li, who influenced us with his insightful ideas and meaningful inspirations, guided us with practical academic advice and feasible instructions, and enlightened us while we was confused during the writing procedure. His thought-provoking comments and patiently encouragements are indispensable for my accomplishment of this thesis.
Ultimately, thanks go to our parents who are our mentor and guardian from the very beginning in primary school. Without their refined education and care,we could never grow up in such a joyous and cozy environment nor have the courage to confront any obstacles on our way to success.

References

[1] Xiaofei He, Shuicheng Yan, Yuxiao Hu, P. Niyogi and Hong-Jiang Zhang, "Face recognition using Laplacianfaces," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 3, pp. 328-340, March 2005.

[2] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry and Y. Ma, "Robust Face Recognition via Sparse Representation," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 2, pp. 210-227, Feb. 2009.

[3] G. Zhao and M. Pietikainen, "Dynamic Texture Recognition Using Local Binary Patterns with an Application to Facial Expressions," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, no. 6, pp. 915-928, June 2007.

[4] P. N. Belhumeur, J. P. Hespanha and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: recognition using class specific linear projection," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, no. 7, pp. 711-720, July 1997.

[5] Chih-Wei Hsu and Chih-Jen Lin, "A comparison of methods for multiclass support vector machines," in IEEE Transactions on Neural Networks, vol. 13, no. 2, pp. 415-425, March 2002.

[6] A. C. Braun, U. Weidner and S. Hinz, "Support vector machines, import vector machines and relevance vector machines for hyperspectral classification — A comparison," 2011 3rd Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS), Lisbon, 2011, pp. 1-4.

[7] V. Bruce and A. Young. Understanding face recognition. British journal of psychology, 77(3):305–327, 1986.

[8] A. J. Calder and A. W. Young. Understanding the recognition of facial identity and facial expression. Nature Reviews Neuroscience, 6(8):641–651, 2005.

[9] H. Ding, S. K. Zhou, and R. Chellappa. Facenet2expnet: Regularizing a deep face recognition net for expression recognition. In Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on, pages 118–126. IEEE, 2017.

[10] J. Gauthier. Conditional generative adversarial nets for convolutional face generation. Class Project for Stanford CS231N: Convolutional Neural Networks for Visual Recognition, Winter semester, 2014(5):2, 2014.

[11] S. Lawrence, C. L. Giles, Ah Chung Tsoi and A. D. Back, "Face recognition: a convolutional neural-network approach," in IEEE Transactions on Neural Networks, vol. 8, no. 1, pp. 98-113, Jan. 1997.

[12] S. K. Pal and S. Mitra, "Multilayer perceptron, fuzzy sets, and classification," in IEEE Transactions on Neural Networks, vol. 3, no. 5, pp. 683-697, Sept. 1992.

[13] A. Mollahosseini, D. Chan and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Placid, NY, 2016, pp. 1-10.

[14] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchicalimagedatabase. In:Proc.ofComputerVisionandPatternRecognition (CVPR). (2009)

[15] Krizhevsky, A., Hinton, G.: Learning multiple layers of features from tiny images

[16] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proc. of Computer Vision and Pattern Recognition (CVPR). (2016)

[17] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137-1149, 1 June 2017.

[18] H. Yang, U. Ciftci and L. Yin, "Facial Expression Recognition by De-expression Residue Learning," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, 2018, pp. 2168-2177.

[19] Sanghyun Woo, Jongchan Park, Joon-Young Lee, In So Kweon: CBAM: Convolutional Block Attention Module, Cornell University, Wed, 18 Jul 2018.

[20] M. Pantic and L. J. M. Rothkrantz, "Automatic analysis of facial expressions: the state of the art," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 12, pp. 1424-1445, Dec. 2000.

[21] J. J. Bazzo and M. V. Lamar, "Recognizing facial actions using Gabor wavelets with neutral face average difference," Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004. Proceedings., Seoul, South Korea, 2004, pp. 505-510.

[22] J. Gauthier. Conditional generative adversarial nets for convolutional face generation. Class Project for Stanford CS231N: Convolutional Neural Networks for Visual Recognition, Winter semester, 2014(5):2, 2014.

[23] 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, pages 2983–2991, 2015.

[24] Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. arXiv preprint arXiv:1709.01507 (2017)

[25] Karen Simonyan, Andrew Zisserman: Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv preprint arXiv:1409.1556(2014)

[26] 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, pages 94–101. IEEE, 2010.

[27] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, et al. Challenges in representation learning: A report on three machine learning contests. In Neural Information Processing, pages 117 – 124. Springer, 2013.