Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 5
AIDML 2024

Research on Facial Expression Recognition Based on Improved VGG19

Wei Liang *

School of Computer Science, South China Normal University, Guangzhou, Guangdong, 510631, China

* Corresponding Author Email: wayneleung22@gmail.com

Abstract. Facial expression recognition is of great significance in the fields of computer vision and human-computer interaction. Given the problems of over-fitting, high computational complexity, and vanishing gradient in VGG19, this paper proposes an improved VGG19 network model. By introducing depth wise separable convolution, the number of parameters of the model is significantly reduced and the computational efficiency is improved. This paper applies Batch Normalization and Dropout technology to effectively alleviate the over-fitting problem and improve the model's ability for generalization. In addition, this paper introduces residual connections into the model, which solves the vanishing gradient problem and improves the training speed and model stability. This paper uses the FER2013 dataset to conduct comparative experiments on the improved VGG19 network model, the original model, and other network models. The results show that the parameter amount of the improved VGG19 decreased by approximately 45.26%, and the accuracy increased by 4.52%. In addition, the improved VGG19 model is also better than other network models implemented in this paper and has better facial expression recognition results.

Keywords: Facial Expression Recognition; Convolutional Neural Network; Depth wise Separable Convolution; Residual Connection; Deep Learning.

1. Introduction

Human emotional expression is complex and subtle, of which facial expressions are one of the most direct and universal ways. In interpersonal communication, expressions convey emotional states and contain important information for social interaction [1]. For example, a smile can convey a friendly and open attitude, while frowning may indicate dissatisfaction or confusion. These subtle facial changes are crucial to understanding the emotions and intentions of others. Therefore, the recognition of facial expressions is of great research significance and value for studying human behavior and psychological activities.

With the rapid development of artificial intelligence technology, facial expression recognition technology has become one of the hot spots of research [2]. This technology plays a wide range of applications in mental health, classroom status monitoring, security monitoring, smart home, and other applications [3, 4]. For example, in the realm of mental health, expression recognition can help doctors monitor patients' emotional changes and provide important information for treatment. In the field of security monitoring, by identifying the expressions of suspicious persons, potential threats can be discovered in advance and public safety can be improved.

Facial expression recognition is a technology that uses computers to extract image features and classify them. It is mainly divided into facial expression recognition algorithms based on traditional machine learning algorithms and deep learning-based facial expression recognition algorithms [5]. Early methods relied on traditional machine learning methods, such as support vector machines, decision trees, etc. Although certain results have been achieved, it is limited by the limitations of manual feature extraction and is difficult to cope with complex and changeable practical application scenarios. In recent years, the rise of deep learning, especially convolutional neural networks, has provided new solutions for facial expression recognition. The VGG model has been widely studied and applied due to its excellent performance in image classification tasks [6].

Among them, VGG19, as a deep variant of VGG, further increases the depth of the network to improve feature extraction capabilities. Although VGG19 performs well in feature extraction, it suffers from issues such as sensitivity to overfitting, high computational complexity, and long training time in expression recognition tasks.

In view of the limitations of VGG19 in facial expression recognition, this paper proposes an improved VGG19 network model. First, depthwise separable convolution technology is introduced, which reduces the number of parameters and the amount of calculation while maintaining the expressive ability of the model. Secondly, by introducing Dropout and Batch Normalization technology, the over-fitting problem is effectively alleviated and the generalization ability of the model is improved. Finally, the residual connection is introduced to solve the problem of vanishing gradients and further improve the learning ability of the model.

2. Improved VGG19 Network Model

VGGNET is a deep convolutional neural network model proposed by the Oxford University Computer Vision Group VGG. This model achieved excellent results in the 2014 ImageNet Challenge, ranking first in the localization task and second in the classification task. It is popular for its excellent performance and relatively simple structure [7]. VGGNET consists of multiple convolution blocks, pooling layers, fully connected layers and softmax layer. It is characterized by using a small size 3x3 convolution kernel in the convolution layer and increasing the depth of the model by stacking more convolution layers, thereby improving feature extraction capabilities. The structure of this model is relatively deep, usually 16-19 layers. This paper uses the VGG19. The VGG19 structure is shown in Figure 1. Although VGG19 has many advantages, it also has some shortcomings. First, VGG19 has a large number of parameters, resulting in high computational cost and long training time. Secondly, because the model is deep, overfitting is prone to occur during the training process. In addition, as a deep convolutional neural network, VGG19 has a vanishing gradient problem that may affect the training effect of the model. In response to the above shortcomings of this network model, this paper improves the VGG19 network model in the following three aspects: introducing depthwise separable convolution, applying Batch Normalization and Dropout, and introducing residual connections.

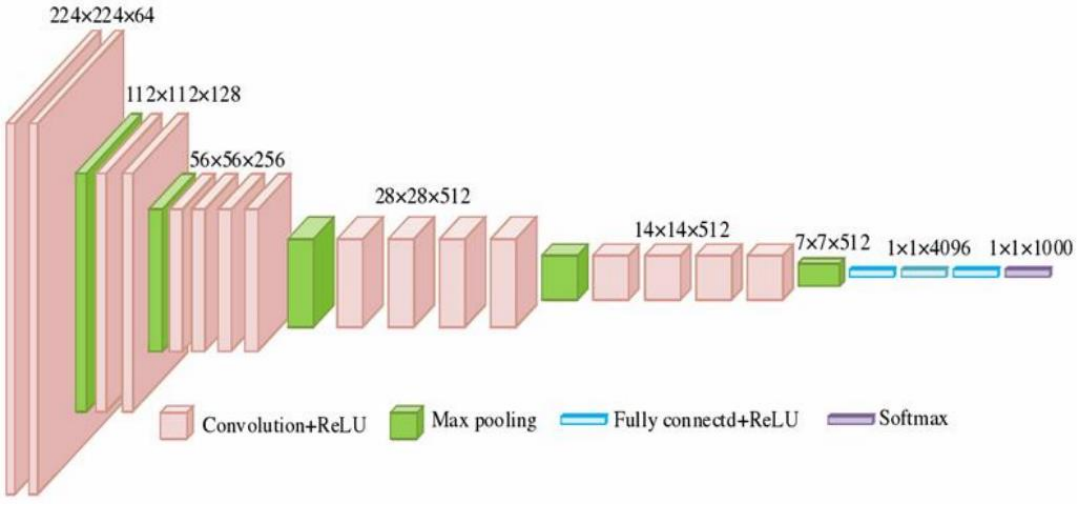

Figure 1. VGG19 network structure diagram [8].

2.1. Depthwise Separable Convolution

One challenge of the VGG19 network model is its huge number of parameters, which not only increases the computational cost but also exacerbates the risk of overfitting [9]. To address this issue, this paper introduces depthwise separable convolution technology into the improved VGG19 network model. Depthwise separable convolution is an optimization technique used for convolutional neural networks in deep learning, which consists of two parts: depth convolution and point-wise convolution

[10]. In the depth convolution stage, each input channel is independently convolved with a trainable convolution kernel to obtain an intermediate feature map; in the point-wise convolution stage, the intermediate feature map is combined through a 1x1 convolution kernel to get the output feature map.

In the improved VGG19 network model, the original standard convolutional layers are replaced with depthwise separable convolutional layers. In each convolutional block, depthwise separable convolutional layers are used instead of 3x3 convolutional layers. This adjustment maintains the spatial structure characteristics of the convolution operation, while also reducing the amount of parameters and computational complexity of each layer, shortening the training time of the model, and reducing memory usage. In addition, reducing the number of parameters can also help improve the generalization ability of the model and reduce the risk of overfitting.

2.2. Batch Normalization and Dropout

Due to its depth and complexity, the VGG19 network model is prone to overfitting during the training process. To alleviate this problem, this paper introduces Dropout and Batch Normalization technology into the model. Dropout is a commonly used regularization technique to prevent overfitting of neural networks. It enhances the robustness of the network by randomly discarding the output of a part of neurons during the training process. Batch Normalization is a standardization technique used to speed up training, reduce overfitting, and improve model stability. Its main idea is to speed up the training process by performing normalization processing on each layer so that the input of each layer maintains a stable distribution.

In the improved VGG19 network model, a Batch Normalization layer is added after each depthwise separable convolution layer to stabilize the input distribution of the network. In the original VGG19 network model, there are only ReLU activation functions and pooling layers between each convolutional layer. In the improved model, a Batch Normalization layer is inserted inside and behind each depthwise separable convolution block to solve the internal covariate offset problem by normalizing each input layer to improve the training convergence speed of the VGG19 network model. [7, 11]. In addition, this paper introduces a Dropout layer after each fully connected layer with a dropout rate of 0.1, that is, only one-tenth of the neuron output is retained in each training to simulate different network structures and increase the model's generalization ability.

2.3. Residual Connection

As the number of layers of a deep neural network increases, the performance of the network will improve to a certain extent. But it also faces some challenges, one of which is the problem of gradient disappearance and gradient explosion, which limits the further improvement of network depth. To address this issue, this paper introduces residual connection technology into the improved VGG19 network model.

Residual connections are a technique that alleviates the vanishing gradient problem in deep neural network training by adding skip connections in the network [12]. It allows the direct transfer of information from the input layer to the output layer, allowing the network to learn the residual mapping between input and output, rather than directly learning the mapping itself. This not only facilitates the backpropagation of gradients but also enables the network to be trained deeper, thereby improving training efficiency and network performance.

In the improved VGG19 network model, this paper introduces residual connections between each convolution block, as shown in Figure 2. The specific implementation method is to add a skip connection between the input and output of each depthwise separable convolution block, and the output of the skip connection is added to the output of the convolution block to form a residual connection. In this way, even if the network is deepened, the effective transmission of information can be ensured and the training efficiency and performance of the network can be improved.

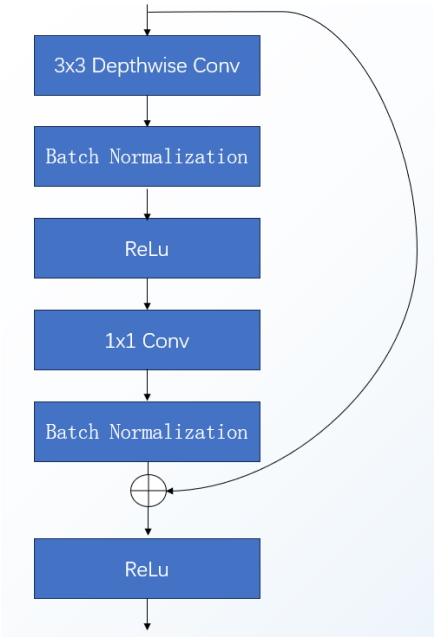

Figure 2. Depthwise separable convolution block with residual connections (DSConvRBlock) (Picture credit: Original).

Figure 3 illustrates the specific structure of the improved VGG19 network model. The network contains 5-segment convolutions with a total of 16 DSConvRBlocks with residual connections. The first two convolution stages each contain two DSConvRBlocks with residual connections, and the last three convolution stages each contain four DSConvRBlocks with residual connections. Each convolution is followed by a max pooling layer to reduce the image size. To avoid overfitting during the training process, a Dropout layer is added after the fully connected layer.

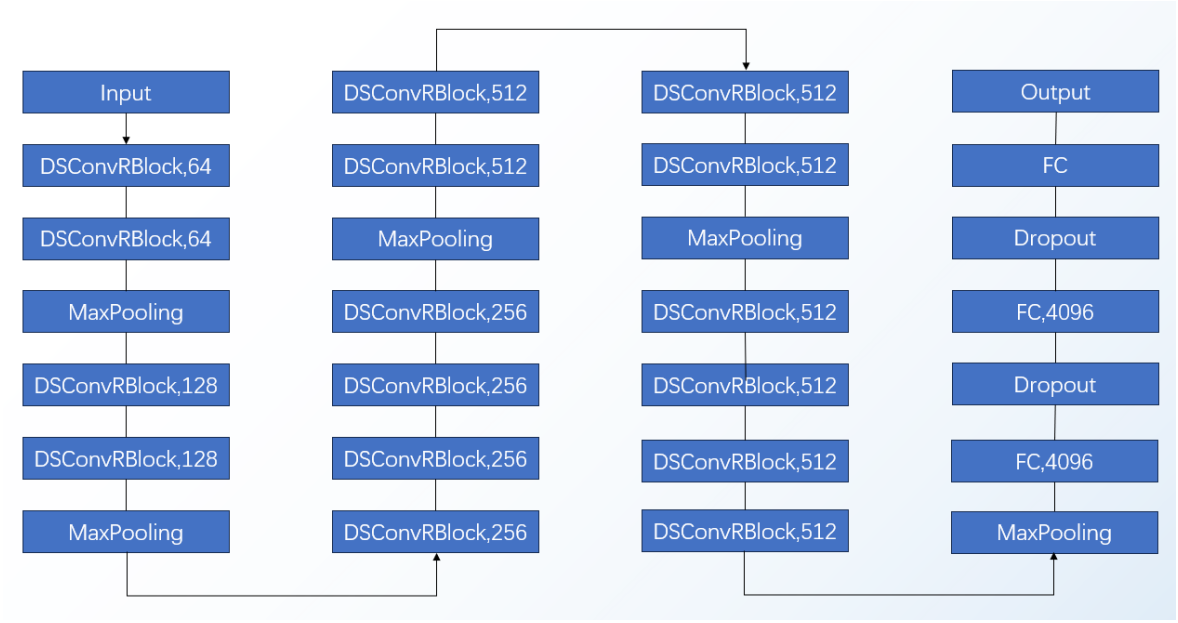

Figure 3. Improved VGG19 network model structure (Picture credit: Original).

3. Experiment

3.1. Dataset

This paper uses the public facial expression recognition dataset FER2013 for experiments. The large scale of this dataset helps train and verify the improved VGG19 network model. At the same time, this dataset is often used in facial expression recognition research and is a basic dataset for easy

comparison with other research. The dataset contains 35,887 grayscale images with a size of 48x48 pixels, which are labeled with seven different emotions: Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral. The dataset is divided into three parts: training (28709 images), validation (3589 images), and testing (3589 images).

3.2. Data Enhancement

In the realm of deep learning, data augmentation is a common technique used to improve the generalization ability and robustness of the model, especially in facial expression recognition tasks. This paper uses the ImageDataGenerator class for data enhancement during the training process, and introduces the following enhancement technologies:

1. Zoom Range: zoom_range=0.2, the image will be randomly scaled during the training process, and the scaling ratio is between 80% and 120% to simulate facial expressions at different distances.

2. Width Shift Range: width_shift_range=0.2, the horizontal position of the image will change randomly, with a range of 20% of the image width, which helps the model learn expression features at different positions.

3. Height Shift Range: height_shift_range=0.2, the image to be displaced in the vertical direction, enhancing the model's ability to recognize expression features, even when the expression is partially blocked.

4. Rotation Range: rotation_range=10, the image will be randomly rotated up to 10 degrees to simulate expression recognition at different angles.

5. Horizontal Flip: horizontal_flip=True enabled, the image will have a 50% probability of being flipped horizontally, which increases the symmetry diversity of expressions in the dataset.

3.3. Learning Strategies

This paper adopts the "Reduce Learning Rate on Plateau" strategy, that is, when the performance indicators on the verification set no longer improve, the learning rate is reduced accordingly [11]. This strategy is based on an observation: once the learning progress of the model has stalled, further learning of the model can often be promoted by reducing the learning rate by 2 to 10 times. This paper sets a "patience period". If the performance indicators do not improve during this period, the learning rate adjustment will be triggered.

4. Results and Discussions

4.1. Comparative Experimental Analysis Between Improved VGG19 and VGG19

To comprehensively evaluate the effect of the model, this paper uses the number of model parameters (M), the epoch required before early termination of training, and the accuracy of the test set (%). The results are shown in Table 1. Experimental results show that the improved VGG19 model in this paper is superior to VGG19 in terms of model parameter quantity, epoch required before early termination of training, and accuracy. From an accuracy perspective, the accuracy of the improved VGG19 has increased by 4.52%. From the perspective of model parameter volume, the model parameter volume dropped from 148.61M to 80.90M, a decrease of approximately 45.26%. From the perspective of the epochs required before early termination of training, the improved VGG19 trains faster, converges faster and reduces training time and resource consumption. This is mainly due to the introduction of depthwise separable convolution, which lowers the model's parameter count while increasing the model's efficiency during training and inference. At the same time, Batch Normalization and Dropout are used to alleviate the phenomenon of over-fitting to a certain extent and improve the model's ability for generalization. In addition, the introduction of residual connections improves the convergence and generalization capabilities of the model, thereby further improving the accuracy and robustness of facial expression recognition.

Table 1. Comparative experimental results between Improved VGG19 and VGG19

	Number of model parameters (M)	Epoch	Accuracy(%)
VGG19	148.61	96	63.47
Improved VGG19	80.90	74	67.99

4.2. Comparative experimental analysis between the improved VGG19 and other network models

This paper also uses other network models to test on the dataset and compares the test results with the improved VGG19 model. The results are shown in Table 2. According to the comparison results, it can be seen that the improved VGG19 model in this paper has advantages compared with ResNet34 and DenseNet121. The results of DenseNet121>Resnet34>VGG19 are predictable. The reason why ResNet34 is better than VGG19 is that it introduces residual connection and residual learning. DenseNet121 is better than Resnet34 because each layer in DenseNet121 is directly connected to all previous layers, forming a dense connection structure, which makes information more easily spread and reused in the network, helping to alleviate the gradient disappearance problem. The improved VGG19 model learns the excellence of DenseNet121 and Resnet34, introduces depthwise separable convolution, and introduces residual connections, making the model even better.

Table 2. Comparison of results between improved VGG19 and other network models

Network Model	Accuracy(%)
VGG19	63.47
ResNet34	64.24
DenseNet121	65.48
Improved VGG19	67.99

4.3. Improved VGG19 Experimental Analysis

Figure 4 shows the confusion matrix diagram of the improved VGG19 experiment on the dataset FER2013. According to Figure 4, Happy has the highest recognition success rate, and Disgust has the lowest recognition success rate. In the dataset, the Happy class has the largest amount of data, while the Disgust class has the least amount of data. This reflects that to a certain extent, the larger the amount of data, the higher the recognition rate. In addition, it can be seen from Figure 4 that the probability of disgusting images being recognized as Angry exceeds 50%. The reason may be that the amount of disgust image data is too small, resulting in the model not fully learning enough features. The two expressions are also possible to be visually similar, making it difficult for the model to distinguish them.

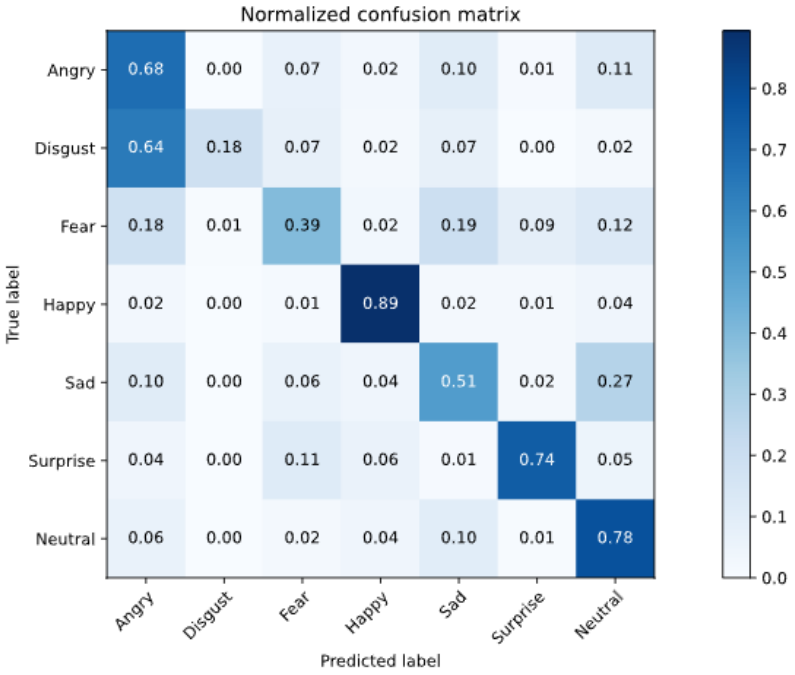

Figure 4. Confusion matrix diagram of the improved VGG19 experiment on the dataset FER2013 (Photo/Picture credit: Original).

5. Conclusion

This paper proposes an improved VGG19 network model, which makes three main improvements to the classic VGG19 model. Improvements include introducing depthwise separable convolutions, applying Batch Normalization and Dropout, and introducing residual connections. The results of the experiment show that the improved VGG19 model can effectively solve the problems of overfitting, high computational complexity, and gradient disappearance in the VGG19 model. There are also some limitations in the research of this paper. For example, although this paper has achieved good results on the FER2013 dataset, the performance of these improvements on other datasets needs further verification. In addition, the performance of the improved VGG19 network model in this paper needs to be improved in some complex expression recognition tasks. Based on the findings of this paper, future research can consider the following two aspects. First, the FER2013 dataset can be processed. Because the amount of picture data of the 7 emotions in the dataset is not close, and some are very different. Try other technical means to expand data samples with small amounts of data. Secondly, explore more optimization strategies, such as automated hyperparameter tuning and further structural improvements, to verify the versatility and effectiveness of these techniques in a wider range of image recognition tasks.

References

[1] Binyu Z. Analysis of facial expression recognition technology based on deep learning. Electronic Technology, 2024, 53 (03): 68 - 71.

[2] Jiarong Q, Changwei Z. Research and development of behavior recognition technology. Intelligent Computers and Applications, 2017, 7 (04): 24 - 26+30.

[3] Ning Z. Student classroom emotion analysis based on face recognition. Xi'an University of Technology, 2022.

[4] Bo M, Ke M, Yuehui J, et al. DeepHome: A smart home management and control model based on deep learning. Journal of Computer Science, 2018, 41 (12): 2689 - 2701.

[5] Mengyang X. Improved expression recognition method based on attention mechanism. Modern Information Technology, 2024, 8 (08): 102 - 105+110.

[6] Zhuangzhuang G, Xuebin X, Longbin L, et al. Research on facial expression recognition method based on DenseNet. Computers and Digital Engineering, 2023, 51 (10): 2425 - 2430.

[7] Ye Z, Cihui Y, Jiemei Z, et al. Research on facial expression recognition algorithm based on SR-VGG19. Computers and Digital Engineering, 2021, 49 (09): 1889 - 1894+1898.

[8] Nan Z. Research on facial expression recognition algorithm based on VGG19. Chang'an University, 2023.

[9] Kowsar I, Zaman M S, Sakib M F R. Facial recognition expression: convolutional attentional masking network and ensemble approach. Brac University, 2021.

[10] Yan W, Zhenyu W. Hyperspectral image classification based on improved SE-Net and depthwise separable residuals. Journal of Lanzhou University of Technology, 2024, 50 (02): 87 - 95.

[11] Khaireddin Y, Chen Z. Facial emotion recognition: State of the art performance on FER2013. arXiv preprint arXiv:2105.03588, 2021.

[12] Gang L, Taibing C, Zhibo Y, et al. MBRNet: Multi-branch handwritten character recognition network integrating residual connections. Computer Engineering and Applications, 2024, 1 - 12.