Facial expression recognition based on improved VGG convolutional neural network

Facial expression recognition based on improved VGG convolutional neural network

Cui Dong [1, a], Rongfu Wang [1], Yuanqin Hang [2]

[1] School of Guangdong University of Science and Technology, Dongguan City, Guangdong Province, China
[2] College of Electronic Engineering Guangxi Normal University Guiling, China

[a] dongcui@gdust.edu.cn

Abstract. With the development of artificial intelligence, facial expression recognition based on deep learning has become a current research hotspot. The article analyzes and improves the VGG16 network. First, the three fully connected layers of the original network are changed to two convolutional layers and one fully connected layer, which reduces the complexity of the network; Then change the maximum pooling in the network to local-based adaptive pooling to help the network select feature information that is more conducive to facial expression recognition, so that the network can be used on the facial expression datasets RAF-DB and SFEW. The recognition rate increased by 4.7% and 7% respectively.

Keywords: Artificial intelligence; Convolutional Neural Network; Facial expression recognition formatting; Pooling.

1. Introduction

Because different facial expressions can reflect people's emotional changes and psychological changes in different situations, facial expression recognition has very important research significance and practical application value for the study of human behavior and psychological activities. In recent years, with the rapid development of computer vision, the rise of deep learning, and the improvement of related theoretical systems such as machine learning, domestic and foreign researchers have developed rapidly on facial expression recognition, and different applications in facial expression recognition have emerged. Method, and achieved remarkable results. With the rise of deep learning, neural networks are favored by research scholars due to their high recognition rate. Therefore, facial expression recognition algorithms based on deep learning have become a research hotspot. For example, Liu et al. [1] proposed the deep network AUDN (AU-inspired Deep Networks), which divides facial expressions into different facial expression units according to facial actions, and uses deep neural networks for deeper feature extraction. So that the network model can achieve better facial expression recognition. Lopes et al. [2] used preprocessing to extract specific facial expression features. Researchers have developed different algorithms to extract facial expression feature information, and then use different methods to recognize and classify facial expressions, so as to better use emotional information to apply facial expression recognition technology to actual production, work and life.

How to extract effective features from facial expression images has become a hot and difficult point of current research. In response to this problem, this paper uses the VGG16 network as the framework for facial expression feature extraction to improve the network and further enhance the network's ability to extract facial expression features, thereby improving the recognition rate of the network on the facial expression dataset. First, change the first two fully connected layers of the three fully connected layers of the VGG16 network to a convolutional layer, thereby reducing the number of parameters of the network model and reducing the complexity of the network model; Secondly, the down-sampling method in the VGG16 network is changed from maximum pooling to adaptive pooling. Because when using the maximum pooling, there will be some feature information extracted by the network model is lost, so by using the adaptive pooling method, the feature information that is important for facial expression recognition is automatically selected, thereby improving the network model's impact on people. Recognition ability of facial expression dataset. Finally, this paper uses the adaptive channel attention model at the end of the network model, which helps the network model to further assign weights to the features that are conducive to facial expression recognition, thereby improving the facial expression recognition ability of the network model.

2. Improved network model structure

2.1. VGG16 network model

VGG-Nets (Visual Geometry Group Networks) [3] is the network model that Oxford University won the first place in the positioning task and the second in the classification task in the ImageNet competition task in 2014. The input size of the original image of VGG-Nets is fixed at 224×224, using a color three-channel image. After the image passes through a series of 3×3 convolutional layers, the feature information of the image is extracted to generate a series of feature maps and network models. The step size and padding are set to 1 pixel. The convolutional layer is followed by a batch normalization layer and an activation function named Relu, and Max pooling is used to reduce the dimensionality of the extracted feature maps. After the convolutional layer, three fully connected layers are connected, and the first two fully connected layers each have 4096 channels, The third fully connected layer is used to classify categories, and the parameter configuration of the fully connected layer is the same. Finally, the network is connected to the Softmax function, and then the final output is connected to the classifier for identification and classification. There are many series of VGG network models, such as VGG11 network model and VGG16 network model. The number of 3×3 convolutions in different VGG network models is different. This paper uses the VGG16 network model as the basic network for facial expression feature extraction. On this basis, the network model is improved and optimized, and then the task of facial expression recognition is studied. The VGG16 network model has 16 layers, including 13 convolutional layers and 3 fully connected layers. The structure of the VGG16 network model is as shown in Figure 1. Since the parameters of the fully connected layer are very many, this article will change the last three fully connected layers of the VGG16 network model to two convolutional layers and one fully connected layer, and the last fully connected layer is used For the classification of facial expression images, this reduces the number of parameters of the network model and saves computer memory while maintaining network performance.

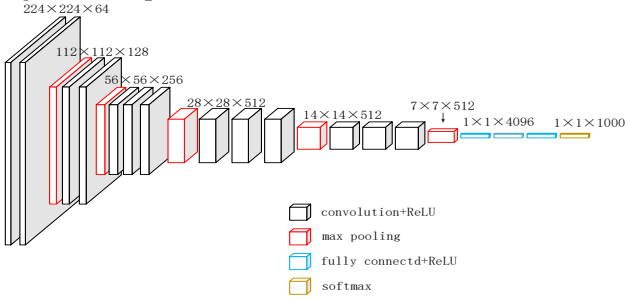

Fig. 1 VGG16 network structure

2.2. Max pooling

For the identification task of image classification [4], the modern structure of convolutional neural networks (CNNs) mostly uses spatial down-sampling (pool) layers to reduce the spatial size of feature maps in the hidden layer. Such a pooling layer is used for larger receiving fields and less memory consumption, especially in extremely deep networks, the widely used maximum pooling, average pooling, etc. This article divides the convolutional layer of the VGG16 network model into five blocks. Each block is composed of a convolutional layer and a relu layer. After each block, there is a maximum pooling, as shown in Figure 1 above. The convolutional layer is used to extract the features of the input facial expression image, generate a series of facial expression feature maps, and then use maximum pooling to reduce the dimensionality of the facial expression feature maps. The pooled matrix is 2×2, the default is the step size of 2×2. After the pooling, the length and width of the facial expression feature map are reduced to half of the original.

The assumption is that the most distinguishing feature should be the maximum activation. This assumption has two main disadvantages. 1) The prior knowledge that the maximum activation represents the most discriminative details may not always be correct; 2) The maximum operator on the sliding window hinders gradient-based optimization, because in backpropagation, the gradient is only Assign to a local maximum, as discussed in [6].

2.3. Local based adaptive pooling

Because the maximum pooling is used when performing Jiang Wei operation on the feature map in the VGG16 network, and the maximum pooling will cause part of the information of the facial expression feature map to be lost. Therefore, this paper uses a new pooling method to replace the maximum pooling method in the original VGG16 network, called local importance-based pooling (LIP) [7], which can automatically learn importance weights. LIP automatically learns the importance metric based on the subnet of the input features, which can adaptively determine which features are more important, and maintain it through downsampling, so that the network can retain the features of small targets In a sense, it can simulate the behavior of average pooling, maximum pooling, and pooling to keep details [5]. The principle is to use the importance function F, which has the characteristic of discriminating and selecting characteristics, in the down-sampling process to deal with the smaller displacement and distortion as much as possible, so that the fixed interval sampling scheme should be avoided. The principle is shown in the following formula.

$$\mathrm{F}(I)=\exp(\vartheta(I)) \tag{1}$$

$$O_{x',y'}=\frac{\sum_{(\Delta x,\Delta y)\in\Omega} I_{x+\Delta x,y+\Delta y}\exp(\vartheta(I))_{x+\Delta x,y+\Delta y}}{\sum_{(\Delta x,\Delta y)\in\Omega}\exp(\vartheta(I))_{x+\Delta x,y+\Delta y}} \tag{2}$$

In the above formula, ϑ and $\vartheta(I)$ are named as logit module, In the process of downsampling, this module automatically emphasizes the discriminative features by learning the larger $\vartheta(I)$ value of the corresponding position, so that the network learning is better and more important. Compatible facial expression image features. In order to make the importance weight non-negative and easy to optimize, $\exp(\bullet)$ operations are added on the basis of ϑ. The input feature map I, the core index set Ω is composed of the relative sampling position $(\Delta x,\Delta y)$ in the sliding window, and the top left position in the input feature map corresponding to the output position (x,y), and $O_{x',y'}$ is the final output.

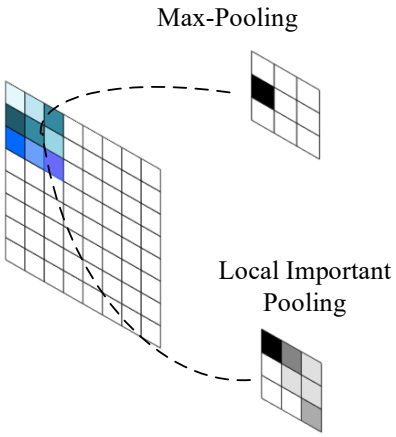

Fig. 2 Pooling comparison chart

As can be seen in Figure 2, the activation in the input feature map is blue, and the darker color means larger. The maximum pooling only retains the maximum activation. The pooling method used in this article will automatically learn the feature map. With different importance, the feature information of facial expression pictures is more comprehensively retained, which is more conducive to the recognition of facial expressions on the network.

This article replaces the 2×2 maximum pooling layer in the VGG16 network with a LIP layer with the same window size configuration. The structure diagram is compared as shown in Figure 3. At the same time, this article changes the three fully connected layers of the VGG16 network as shown in Figure 1 to two convolutional layers and one fully connected layer. The convolutional layer is used to further extract facial expression features. The fully connected layer Used for the final classification of the network, which reduces the complexity of the original network.

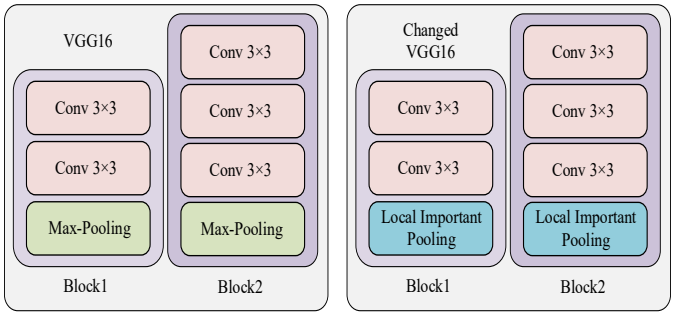

Fig. 3 Comparison of network pooling after improvement

3. Experiment Prepare

3.1. Dataset and data processing

RAF-DB [8] dataset: The real-world facial emotion database RAF-DB is a large-scale facial expression database containing about 30,000 annotated facial images downloaded from the Internet. The RAF-DB dataset has a large number of images and rich annotations, including seven basic expressions and twelve composite expressions independently annotated by 40 trained coders. The seven basic expressions include six basic expressions (happiness, surprise, sadness, anger, disgust, fear) and images of neutral expressions. We mainly conduct experiments on seven basic expressions to verify the performance of the network model. The dataset has been divided into a training set and a test set. There are 12,271 facial expression images for training and 3068 for testing.

SFEW dataset: The wild static facial expression (SFEW) dataset is constructed by selecting frames from AFEW [9], covering unconstrained facial expressions, various head poses, large age range, occlusion, different focus, and face discrimination Rate and real lighting. This article uses the latest version of the decomposed SFEW in literature [10], which divides the images of the SFEW dataset into three groups: training (958 images), verification (436 images), and testing (372 images). Each image is marked by two independent labelers, and is marked as one of seven emoticons. The seven emoticons are: surprise, fear, disgust, happiness, sadness, anger, and neutrality. This article mainly tests on the validation set.

3.2. Data Processing

In order to prevent the over-fitting phenomenon of the network model, during the training process, the data is randomly horizontally flipped, cut, or angularly rotated. This method is called data enhancement, so we use horizontal, cropping and angular rotation at the same time The method of data enhancement allows the network model to learn expression features from multiple angles to enhance the robustness of the network model. Data reduction can expand the number of images in the dataset, so that the convolutional neural network can learn more comprehensive feature information, so the dataset is usually cut randomly during experiments. Since the facial expression dataset RAF-DB, the size of the original aligned facial expression image is 100×100, and the size of the image is moderate. Therefore, the image will not be enlarged in the experiments in this chapter to preserve the original information of the image to the greatest extent. When training the network, the method of random cropping is adopted. The original 100×100 image is randomly cropped into a sub-image of 90×90 size, and then the image is randomly mirrored, and the processed image is sent to the network model for training. In the test phase, the image is cut by ten times in the upper left corner, the lower left corner, the upper right corner, the lower right corner, and the center. At the same time, the mirror operation is used to enlarge the number of pictures by 10 times, and the processed pictures are tested again. Then take the average value of the obtained probability, and get a maximum output value of the corresponding expression, which reduces the classification error rate. For the SFEW dataset, first we set the image size to half of the original input of the original VGG16 network model, that is, the size of 112×112, and then randomly mirror the image and send it to the network model for training.

3.3. lab Environment

Operating system: Ubuntu16.04; GPU: GTX 1080TIi; memory size is 128G; CPU: E5-2637 v4;
Acceleration library: CUDA9.0; programming language: Python3.6;
Deep learning framework: PyTorch0.4A subsubsection.

3.4. Comparison of experimental results of RAF-DB dataset.

In order to verify the improvement of the expression recognition effect of the improved VGG16 network model designed in this chapter, firstly, a comparative experiment was carried out on the facial expression dataset RAF-DB with the original VGG16 network model. The experimental results are shown in Table 1, and then other comparisons are also made. The expression recognition rate of the advanced method on this dataset proves the effectiveness of the improved network model. The comparison results are shown in Table 2.

Table 1. Comparison of facial expression recognition rate on RAF-DB dataset

Networks	Local adaptive pooling	Ten times reduction	Pre-training	Recognition rate %
VGG16	-	-	-	81.68
This article	✓	-	-	83.15
	✓	✓	-	84.97
	✓	✓	✓	86.31

Table 2. Comparison of facial expression recognition rate between SFEW and existing methods

Networks	Recognition rate %
MRE-CNN[11]	82.63
PAT-VGG-F-(gender,race) [12]	83.83
DLP-CNN[13]	84.13
This article	86.31

It can be concluded from Table 1 that under the same experimental conditions, the facial expression recognition rate of the original VGG16 network model on the facial expression dataset RAF-DB is 81.68%, The improved network in this paper has a facial expression recognition rate of 87.84% in this dataset. 83.83%. The improved network is nearly 1.5% higher than the original network on RAF-DB. In the training process, load the network parameters of the VGG16 network model pre-trained on ImageNet for fine-tuning, and at the same time use the ten-fold reduction method to enhance the data during the test, and finally the improved network model is in the facial expression dataset RAF-DB The recognition rate reached 86.31%. The improved network Compared with the original network's facial expression recognition accuracy rate, the model has increased by 4.7%. In the comparison with the existing methods in Table 2, it can be seen that the improved network's facial expression recognition is efficient.

3.5. Comparison of experimental results of SFEW dataset.

Since the number of facial expression pictures in the SFEW dataset is small and the recognition of the pictures is difficult, the pre-training dataset will be used to expand the dataset when performing experiments on the dataset. This article uses the model trained on RAF-DB, and then performs experiments on the SFEW dataset after loading the training model parameters. Finally, the experimental results are shown in Table 3 and Table 4.

Table 3. Comparison of facial expression recognition rate on SFEW dataset

Networks	dataset	Pre-training dataset	Recognition rate %
VGG16	SFEW	ImageNet	50.00
This article		RAF-DB	57.34

Table 4. Comparison of facial expression recognition rate between SFEW and existing methods

Networks	Pre-training dataset	Recognition rate %
Identity-aware CNN [14]	FER2013	50.98
Island Loss [15]	FER2013	52.52
RAN (VGG16+ResNet18) [16]	MS Celeb 1M	56.40
This article	RAF-DB	57.34

It can be obtained from Table 3 that after the original VGG16 network model is loaded on the ImageNet pre-trained model parameters and fine-tuned, the recognition rate on the facial expression dataset SFEW is 50%. For the improved network model designed in this chapter, we load the model parameters of the improved model trained on the facial expression dataset RAF-DB, and fine-tune the parameters of the network model during the training of the facial expression dataset SFEW, and finally The recognition rate on this dataset reached 57.34%%, which is 7.34% higher than the recognition rate of the original VGG16 network model for facial expressions. I It can be obtained from Table 4, that the improved network model and existing methods designed in this paper still have a relatively high improvement in the facial expression recognition rate of the facial expression dataset SFEW. For example, the RAN method in literature [60] uses the combination of VGG16 and residual 18 network models to recognize expressions. This article only uses a single VGG16 network for improvement. The

improved network model is still better than the expression recognition rate of the above method. High, which proves the effectiveness of the improved network model in this article.

Confusion matrix diagrams of the VGG16 network and the improved network of this article on the RAF-DB and SFEW facial expression datasets. As shown in Figure 4 below.

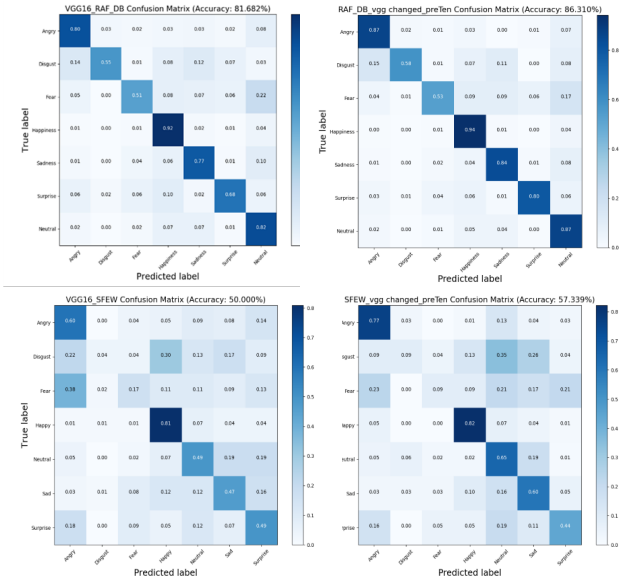

Fig. 4 Confusion matrix diagram of VGG network and improved network

4. Organization of the Text

In this paper, the last two fully connected layers in the VGG16 network are changed to convolutional layers to reduce the parameters of the network, change the pooling method in the original network, and load the parameters of the original VGG16 network pre-trained on ImageNet during training. In the test, a tenfold cropping method was used, and a relatively good facial expression recognition effect was achieved on the facial expression datasets RAF-DB and SFEW, and the facial expression recognition rate of the original network was increased by 4.7% and 7%, respectively.

Acknowledgments

Fund Project: Key Project of Natural Science of Guangdong University of Science and Technology (GKY-2021KYZDK-7).

References

[1] Liu M, Li S, Shan S, et al. Au-inspired deep networks for facial expression feature learning [J]. Neurocomputing, 2015, 159 (Jul.2): 126-136.

[2] Lopes A T, De Aguiar E, De Souza A F, et al. Facial expression recognition with convolutional neural networks: coping with few data and the training sample order [J]. Pattern Recognition, 2017, 61: 610-628.

[3] Qian Yongsheng, Shao Jie, Ji Xinxin, et al. Multi-view facial expression recognition based on improved convolutional neural network [J]. Computer Engineering and Applications, 2018, 54(24):12-19.

[4] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition [J]. Computer Science, 2014, PP: 1409-1556.

[5] Miami, Florida, ImageNet: a Large-Scale Hierarchical Image Database [C]// 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009,, USA. IEEE, 2009..

[6] Saeedan F, Weber N, Goesele M, et al. Detail-Preserving Pooling in Deep Networks [C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2018.
[7] Gao Z, Wang L, Wu G . LIP: Local Importance-based Pooling [C]// International Conference on Computer Vision. 0.
[8] Li S, Deng W. Reliable Crowdsourcing and Deep Locality-Preserving Learning for Unconstrained Facial Expression Recognition[J]. IEEE Transactions on Image Processing, 2018, PP: 1-1.
[9] Dhall A, Goecke R, Lucey S, et al. Collecting Large, Richly Annotated Facial-Expression Databases from Movies[J]. IEEE Multimedia, 2012, 19(3):0034.
[10] Mao Bo, Xu Ke, Jin Yuehui. Deep Home: A Smart Home Control Model Based on Deep Learning [J]. Chinese Journal of Computers, 2018, 41(12): 2689-2701.
[11] Fan Y, Lam J, Li V. Multi-Region Ensemble Convolutional Neural Network for Facial Expression Recognition [J]. 2018. PP: 84-89.
[12] Cai J, Meng Z, Khan A S, et al. Probabilistic Attribute Tree in Convolutional Neural Networks for Facial Expression Recognition [J]. 2018.
[13] Li S, Deng W. Reliable Crowdsourcing and Deep Locality-Preserving Learning for Unconstrained Facial Expression Recognition [J]. IEEE Transactions on Image Processing, 2018. PP: 1-1.
[14] Meng Z, Ping L, Jie C, et al. Identity-Aware Convolutional Neural Network for Facial Expression Recognition [C]. IEEE International Conference on Automatic Face & Gesture Recognition. IEEE Computer Society, 2017, PP: 558–565.
[15] Cai J, Meng Z, Khan A S, et al. Island Loss for Learning Discriminative Features in Facial Expression Recognition [J]. IEEE, 2018.10, PP: 302–309.
[16] Wang K, Peng X, Yang J, et al. Region Attention Networks for Pose and Occlusion Robust Facial Expression Recognition [J]. IEEE Transactions on Image Processing, 2020, PP(99): 1-1.