Proceedings of the 5th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/21/20231110

Facial emotion detection based on improved VGG-16

Ran Zhang

College of Engineering, Southern University of Science and Technology, Shenzhen, 510000, China

12011511@mail.sustech.edu.cn

Abstract. In real life, facial emotion recognition is very important because it can convey information, build relationships, and facilitate communication. Therefore, emotion recognition technology is used in medicine, education, entertainment, security, and other fields. In the emotion detection field, the Facial Emotion Recognition 2013 Dataset (FER-2013) is a dataset that has been used in many places, that contains images of seven emotional expressions. In the area of detecting emotions from facial expressions, the deep learning structures, especially the convolutional neural networks (CNNs), have demonstrated significant potential since they have the ability to extract features and their computational efficiency. In this paper, the author constructs a model named Improved VGG-16 based on Visual Geometry Group Network of 16 weight layers (VGG-16). To be specific, first, the author adds two dense layers to improve the complexity and expressiveness; second, two dropout layers are used in order to reduce overfitting. An accuracy of 68.0% is achieved by this model on the test dataset of FER-2013. The result is better than some previous methods and shows that the improved VGG-16 model can recognize facial expressions effectively. In conclusion, this work aims to increase the accuracy and reliability of facial emotion recognition, providing support for research and application in related fields.

Keywords: convolutional neural network, VGG-16, FER-2013, facial emotion recognition.

1. Introduction

Facial emotion is one nonverbal method of communication in human society. Facial emotions can be classified into many types, such as disgust, fear, anger, happiness, sadness, surprise, and neutrality. Mehrabian says that 7% of information is conveyed by spoken language, 38% by intonation, and 55% by facial expression [1]. Here, several actions or states of facial muscles produce facial expressions. Since facial expression accounts for a large proportion of information in communication, it plays an significant role in the daily lives. Facial emotion detection is useful in human facilities and clinical practice. Facial expression analysis is important for applications based on emotion recognition, such as psychology, social robotics, animation, alarm systems, and patient pain monitoring [2]. Specifically, in terms of psychology, the accurate interpretation of facial emotion contributes to reasonable communication methods and efficient communication. And doctors can check the facial expressions of patients to see whether they are in pain. In addition, facial expressions can reveal a person's intention, emotional state, cognitive activity, psychopathology, and personality. In offline communication, facial expressions convey a lot of meaningful messages. These messages can help both people understand what they want to express. Facial emotion detection includes four steps. The first is the phase of face

DOI: 10.54254/2755-2721/21/20231110

detection, which detects faces from still images or videos. The second is the normalization phase. It removes the noise and normalizes the face based on brightness. During the third stage, relevant features are identified and unnecessary ones are removed. Then, in the final step, the basic facial expressions are categorized into seven distinct emotions [3].

Computer vision is a technology which uses computers along with algorithms to simulate and automate the process of human vision. It involves the processing, analysis, and understanding of images and videos, as well as the ability to extract useful information from them. Computer vision provides an efficient method to process the image of facial emotion. Images captured by a camera are processed and transformed into numerical data. Traditional algorithms or algorithms based on deep learning are applied to analyze facial emotions. In 1978, psychologists Paul Ekman and Wallace Friesen were interested in studying facial emotions and participated in the development of the Facial Action Coding System (FACS) [4]. Also, there are some experiments involving using Gabor wavelets and sparse representation to extract features from facial images. However, their performance is not very good. Subsequently, convolutional neural networks (CNNs) were proposed for data segmentation, classification, and detection. CNNs have shown great potential in image classification because of their computational efficiency and feature extraction capability [5]. Neural networks are a type of machine learning algorithm whose idea comes from the biological nervous systems, particularly the human brain [6]. At present, CNNs have become a popular and effective approach for analyzing speech and recognizing images. It is the first truly successful learning model for training multi-layer neural networks, and its advantages are more obvious when the network input is multi-dimensional. VGG is the abbreviation of Visual Geometric Group. Visual Geometry Group Network of 16 weight layers (VGG-16) was first shown in ILSVRC 2014 challenge and achieved excellent performance on the task of image classification.

At present, facial emotion detection based on deep learning is mostly processed based on Facial Emotion Recognition 2013 Dataset (FER-2013). FER-2013 dataset is collected by Kaggle and was shown by Aaron Courville and Pierre-Luc Carrier in 2013 at the International Machine Learning Conference (ICML) [7]. Each face was classified based on seven emotional categories, and the corresponding image is grayscale in this dataset. FER-2013 contains images of seven different types in total, labeled according to seven different classifications. On FER-2013, the average human performance is about 65.5% [7]. To improve the performance of emotion recognition, this work introduces CNNs to build recognition models. Transfer learning is applied to the VGG-16 model, and a custom model named improved VGG-16 is constructed. Subsequently, the model is fine-tuned by adjusting its parameters and optimizing its hyperparameters to achieve optimal performance. This study achieves 68% accuracy on the FER-2013 dataset, which significantly outperforms previous methods. These experimental results indicate that this improved VGG-16 can effectively complete the facial emotion detection tasks.

2. Methodology

2.1. Dataset description and preprocessing

The FER-2013 dataset has 35,887 images of facial emotions, including seven different emotions: disgust, fear, anger, happiness, sadness, surprise, and neutrality [7]. These images were collected from various places on the Internet, and each image was labeled with a sentiment. The release of this dataset is intended to provide a standard benchmark for the development and evaluation of emotion detection algorithms. FER-2013 dataset has become an important dataset in facial emotion detection field, and has been used in various research and applications, such as facial emotion detection, emotion analysis, human-computer interaction, etc. The image resolution and quality can vary widely in the FER-2013 dataset, which can make it difficult to accurately extract landmarks and other features from the images. Despite these challenges, FER-2013 has become a widely used dataset for evaluating facial emotion detection algorithms.

DOI: 10.54254/2755-2721/21/20231110

The training set has 436 disgust images, 4097 fear images, 3995 anger images, 7215 happiness images, 4830 sadness images, 3171 surprise images, and 4965 neutrality images. The author here does data augmentation in the training set. The data augmentation can automatically carry out random transformations of the image, such as rotation, scaling, translation, flip, etc., to expand the data set and enhance the model's expression ability.

2.2. Proposed approach

The general process is as follows: First, images with a three-channel RGB of 48x48 size enter the VGG-16 model. Then the images are processed by a series of convolution layers and pooling layers to extract image features. After the feature is processed by the full connection layers, the image classification results are obtained.

2.2.1. Introduction of VGG-16. VGG-16 is a 16-layer neural network architecture that consists of thirteen convolutional layers and three dense layers. The major feature of VGG-16 is the use of a very small convolution kernel whose size is (3x3) and the use of pooling layers after every convolutional layer, resulting in a deeper hierarchy of the network. Figure 1 illustrates the architecture of VGG-16. Note that the input shape whose width, height, and channels are (224, 224, 3) is not fixed, it can be changed into any shape. As a result, the output shape in the last max pooling layer is also not fixed. VGG-16 was trained with the ImageNet dataset. The ImageNet dataset is a massive collection of over 1 million images, each labeled with one of 1,000 categories. It achieved very good performance on the ImageNet data set, and its Top-5 error rate was only 7.3%, making it one of the most excellent image classification models at that time [8].

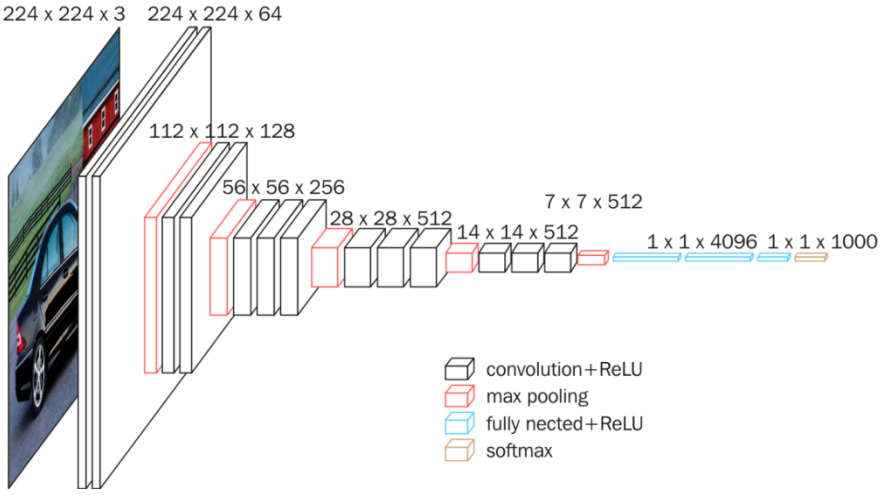

Figure 1. The overall structure of VGG-16.
From https://neurohive.io/en/popular-networks/vgg16/.

Here, to classify FER-2013 dataset, some changes in output layer are needed since there are only 7 classes, which are different from the 1000 classes in the ImageNet dataset. Table 1 shows the structure of Improved VGG-16. Specifically, the input shape is changed from (224, 224, 3) to (48, 48, 3). And all layers after the last max pooling layer are removed. Instead, a Flatten layer and a Dense layer are added, as Table 1 demonstrates. Here the Dense layer acts as the output layer, consisting of 7 units corresponding to the 7 classes in FER-2013 and softmax activation function. Softmax is a common activation function that is widely used in multi-class problems. and it is usually used in the output layer in neural networks.

DOI: 10.54254/2755-2721/21/20231110

Table 1. VGG-16 model.

Layer Type	Input Shape	Output Shape	Units	Activation
VGG-16	48×48×3	1×1×512	-	-
Flatten	1×1×512	512	-	-
Dense	128	7	7	softmax

2.2.2. Improvement. Based on the idea that adding more layers to a neural network can improve the model's expressiveness and robustness, the author constructs the Improved VGG-16. The Improved VGG-16 is demonstrated in Table 2. Except for the changes in the VGG-16 model, there are two additional fully connected layers and two dropout layers. The first fully connected layer consists of 256 neurons using the relu activation function and the L2 kernel regularizer. The second fully connected layer, which consists of 128 neurons, also uses the relu activation function and the L2 kernel regularizer. The first dropout layer is the same as the second dropout layer, both have a 0.3 rate. The function of relu activation is to map the input values nonlinearly to improve the neural network's expression ability. The L2 kernel regularizer and dropout layer are used to prevent overfitting.

Table 2. Improved VGG-16 model.

Layer Type	Input Shape	Output Shape	Units	Activation	Kernel Regularizer	Dropout Rate
VGG-16	48×48×3	1×1×512	-	-	-	-
Flatten	1×1×512	512	-	-	-	-
Dense	512	256	256	relu	L2	-
Dropout	256	256	-	-	-	0.3
Dense	256	128	128	relu	L2	-
Dropout	128	128	-	-	-	0.3
Dense	128	7	7	softmax	-	-

2.2.3. Loss function. The author uses categorical cross entropy here as the model's loss function [9]. It calculates the difference between the true label and the predicted label of a model and is used to update the parameters of the model through the backpropagation algorithm. In multi-classification tasks, the cross-entropy loss is often used in conjunction with the activation function of softmax, which transforms the outputs of the model into the probabilities among the possible class labels. The formula is as follows:

$$H(p,q) = -\sum_{i=1}^{n} p_i \log q_i \tag{1}$$

where means the number of classes, means the probability of the true label falls into class i and means the probability of the predicted label falls into class i.

2.3. Implementation details

The data augmentation involves re-zooming the image to ±10% of its original scale, moving the image horizontally and vertically to ±10% of its size, and rotating it to ±10 degrees. Each pixel is divided by 255 to do normalization. The author runs both models for 100 epochs with an early stopping of 70 patience. The optimizer used here is Adam. The author adopts a fixed learning rate of 0.0001 on Adam. And other parameters in Adam are the default.

3. Experiment results

This section analyzes the experiment results of VGG-16 model and improved VGG-16 model. Also, the author compares some other previous methods with them.

Figure 2 shows the accuracy, loss, AUC, precision, and F1-score in training process of VGG-16 and Improved VGG-16. To sum up, they have similar performances since they both use VGG-16 as their

DOI: 10.54254/2755-2721/21/20231110

basic structure. But still, there are mall differences. For instance, here it can be seen in the second column that the loss curve of Improved VGG-16 model is not as steep as VGG-16 model when the curve goes up. This is due to the dropout layer and L2 kernel regularizer.

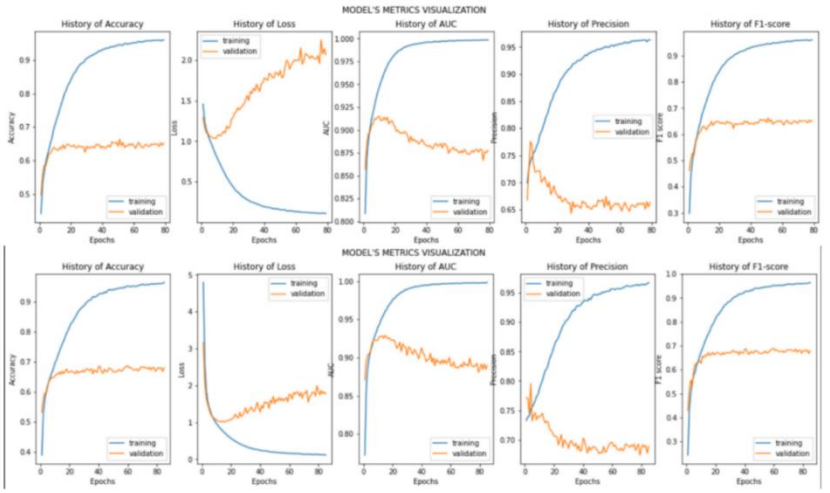

Figure 2. Results in VGG-16 and Improved VGG-16. The first row is VGG-16's result. The second row is Improved VGG-16's result. From left to right are accuracy scores, losses, AUC scores, precision scores, and F1 scores, respectively (Picture credit: Original).

Figure 3 depicts the confusion and normalized confusion matrices corresponding to the VGG-16 and Improved VGG-16. It can be clearly illustrated in the second column that the "happy" and "surprised" labels have greater accuracy than most other labels in both VGG-16 and Improved VGG-16. Meanwhile "disgust", "fear", and "sad" have a relatively bad performance. The reason for the low accuracy of "disgust" can be the small size of the samples in the training images. And the reason for bad performance of "sad" may be that some sad expressions are not very obvious, so they are classified as neutral.

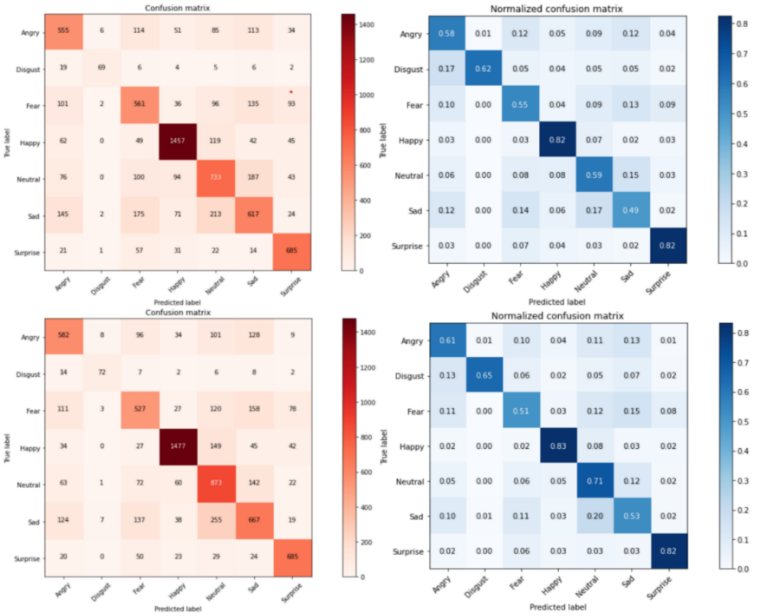

Figure 3. The matrices of VGG-16 and Improved VGG-16. The first row is VGG-16's matrices. The second row is Improved VGG-16's matrices. The first column is confusion matrices. The second column is normalized confusion matrices (Picture credit: Original).

DOI: 10.54254/2755-2721/21/20231110

Finally, the author does performance tests on the test images. The results are demonstrated in Table 3. Here reaches 68% accuracy in Improved VGG-16, better than 65.2% accuracy for VGG-16. And at the same time, Improved VGG-16 has a greater performance in precision, AUC, and F1 score. Specifically, Improved VGG-16 is 2.6%, 1.2%, 3.3% higher than VGG-16 in Precision, AUC and F1 score.

Table 3. Accuracies on FER-2013 test data corresponding to different architectures.

Model	Accuracy	Precision	AUC	F1-score
VGG-16	65.2%	66.3%	87.8%	64.9%
Improved VGG-16	68.0%	68.9%	89.0%	68.2%

Table 4 illustrates some previous networks' performance on the FER2013 classification. Most networks have a better performance than the average human (65.5%). In this work, the author achieves an accuracy of 68.0%.

Table 4. The comparison results of different models in the FER-2013 dataset.

Network	Test Accuracy
CNN [10]	62.4%
GoogleNet [11]	65.2%
VGG-16 (Author's work)	65.2%
Conv + Inception layer [12]	66.4%
Bags of Words [13]	67.4%
Improved VGG-16 (Author's work)	68.0%

4. Conclusion

This paper proposes a model, named Improved VGG-16, to do classification tasks in the face emotion detection field. The process includes training Improved VGG-16 and VGG-16 and comparing their results. This experiment is built on the FER-2013 dataset which is an important dataset in the facial emotion detection field. In the testing phase, the Improved VGG-16 reaches an accuracy of 68.0% which is higher than the 65.2% accuracy of VGG-16. This experimental result has surpassed some traditional methods on the FER-2013 dataset, such as CNN and GoogleNet, which indicates the potential of using VGG-16 in the field of facial emotion detection. However, this performance on the FER-2013 dataset is not very excellent overall since some of the prevailing methods have achieved over 70% accuracy. More complex network structures and some better hyperparameters may be needed to improve the accuracy. In the future, the author plans to use some other processors to deal with images and try to employ neural networks with deeper structures at the same time to achieve better experimental results in facial emotion detection.

References

[1] Mehrabian A 2017 Communication without words Communication theory Routledge pp 193-200
[2] Dubey M Singh L 2016 Automatic emotion recognition using facial expression: a review International Research Journal of Engineering and Technology 3(2): pp 488-492
[3] Rani J Garg K 2014 Emotion detection using facial expressions-A review International Journal of Advanced Research in Computer Science and Software Engineering 4(4)
[4] Alkawaz M Mohamad D Basori A Saba T 2015 Blend shape interpolation and FACS for realistic avatar 3D Research 6: pp 1-10

DOI: 10.54254/2755-2721/21/20231110

[5] Krizhevsky A Sutskever I Hinton G 2017 Imagenet classification with deep convolutional neural networks Communications of the ACM 60(6): pp 84-90
[6] Bishop C 1994 Neural networks and their applications Review of scientific instruments 65(6): pp 1803-1832
[7] Goodfellow I Erhan D Carrier P Courville A Mirza M Hamner B Bengio Y 2013 Challenges in representation learning: A report on three machine learning contests Neural Information Processing: 20th International Conference (ICONIP) Springer berlin heidelberg 20: pp 117-124.
[8] Simonyan K Zisserman A 2014 Very deep convolutional networks for large-scale image recognition arXiv preprint arXiv: 1409.1556
[9] Ho Y Wookey S 2019 The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling IEEE access 8: pp 4806-4813
[10] Liu K Zhang M Pan Z 2016 Facial expression recognition with CNN ensemble 2016 international conference on cyberworlds (CW) IEEE pp 163-166
[11] Giannopoulos P Perikos I Hatzilygeroudis I 2018 Deep learning approaches for facial emotion recognition: A case study on FER-2013 Advances in Hybridization of Intelligent Methods: Models, Systems and Applications pp 1-16
[12] Mollahosseini A Chan D Mahoor M 2016 Going deeper in facial expression recognition using deep neural networks 2016 IEEE Winter conference on applications of computer vision (WACV) IEEE pp 1-10
[13] Ionescu R Popescu M Grozea C 2013 Local learning to improve bag of visual words model for facial expression recognition Workshop on challenges in representation learning (ICML)