Proceedings of the 5th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/14/20230797

Recognizing facial emotion leveraging convolutional neural network

Zhike Yang[1, 4, †], Zhengqian Zhang[2, †] and Bowen Zheng[3, †]

[1]School of Technology, Xianjiaotong-Liverpool University, Suzhou, Jiangsu, 215000, China
[2]School of Technology, Taiyuan University of Technology, Jinzhong, Shanxi, 030600, China.
[3]School of Computer Science and Engineering, Tianjin University of Technology, Tianjin, 300384, China

[4]Zhike.Yang20@student.xjtlu.edu.cn
[†]These authors contributed equally

Abstract. Facial expression recognition technology is a technology based on the combination of artificial intelligence technology and biology, and it is interdisciplinary research. The appearance of this technology shows the diversity of the development direction and application fields of computer technology, but it also means that the development of facial expression recognition technology needs not only the support of computer technology but also the exploration and progress of biology. At present, to recognize facial expressions a system usually contains three stages, namely, face detection stage, feature extraction stage, and face expression recognition stage. The facial expression recognition algorithm is transformed into a classification problem, and a convolutional neural network (CNN) is adopted, where features of the facial expressions could be automatically extracted. It takes as input a preprocessed image and can directly produce encoded features and predictions, leveraging the end-to-end strategy. It could discard the complicated intermediate modeling process of traditional machine learning. In this work a CNN is implemented for the recognition of facial expression on Fer2013 dataset. Moreover, the effectiveness of different numbers of CNN layers are testified on this problem. It could be concluded that with deeper architectures, the CNN tends to perform better.

Keywords: facial expression recognition, convolutional neural network, deep learning.

1. Introduction

In daily basis on work or business, the main approach for human to express their emotion is as follow: language, voice, physical behaviors, and facial expressions. In these aspects, expressions are regarded as the most informative carrier of people's inward emotions [1,2]. Based on the research, the information of inner activities carried by human facial expressions has the highest proportion among all the above forms which accounting for about 55% [3,4].

Changes in human facial expressions can convey changes in their inner emotions, and expressions could reflect the true human emotions [5]. Based on the theory of a famous psychologist called Paul Aikman, the fundamental expressions of human beings could be separated as disgust, anger, sadness,

DOI: 10.54254/2755-2721/14/20230797

fear, surprise, and happiness after a host of experiments. Examples of the facial expressions are demonstrated in Figure 1. Besides the theoretical outcomes, researchers also proposed an image database collecting various expressions of human faces.

Figure 1. Examples of various facial expressions. (Figure from: https://www.guyuehome.com/22597).

Recognizing these facial expressions are of great importance. It could be applied into multiple fields. In human-computer interactions, by identifying the facial expressions and further analyze the expressions, researchers could find out the feelings of the users to merchandises. In social network analysis, it could seek and extract relations among people, by analysing the interactions based on facial expressions [6]. Besides conventional machine learning solutions, with the deepening of research, neutral expressions have also been added to basic facial expressions by researchers, forming seven basic facial expressions in today's facial expression recognition research.

The main purpose of writing this article is to provide a detailed introduction to the project and basic principles of facial expression recognition based on CNN convolutional neural networks.

2. Method

2.1. Dataset

The training model used data set Fer2013 from Kaggle2013 Facial Expression Recognition Challenge [7,8]. It consists of 28,711 facial expression pictures with the resolution at 48x48. All images are grayscale images. The dataset classifies the faces into 7 different categories and these samples are labelled from 0-6. Specifically, they are: 0-anger, 1-disgust, 2-fear, 3-happy, 4-sad, 5-surprised, 6-neutral.

Image preprocessing includes three steps. Firstly, images in fer2013 are saved as csv file. Each column vector is a picture, the first number is a label, and the remaining number is picture data (picture matrix is stored by column expansion). Use numpy to restore the pixel data in a csv file to an image size of 48x48. Secondly, the data set was divided into 24,000 training sets and 4710 verification sets. Thirdly image pixel information is separated from tag information, and the image name and tag code corresponding to the training set and verification set are written into csv file for the network to read.

2.2. CNN

CNN is leveraged for implementing the recognition. It includes several layers including input, convolutional, activation, pooling, full connection, and output layers. Reasonable to set the layer structure and demand between different levels of dropout rate, NB, such as operation, eventually forming a high efficiency and high accuracy model [9.10].

As illustrated in Figure 2, the CNN network model used in this work, which includes several layers. They follow various calculations, and the corresponding weights could be gradually updated with the training proceed.

DOI: 10.54254/2755-2721/14/20230797

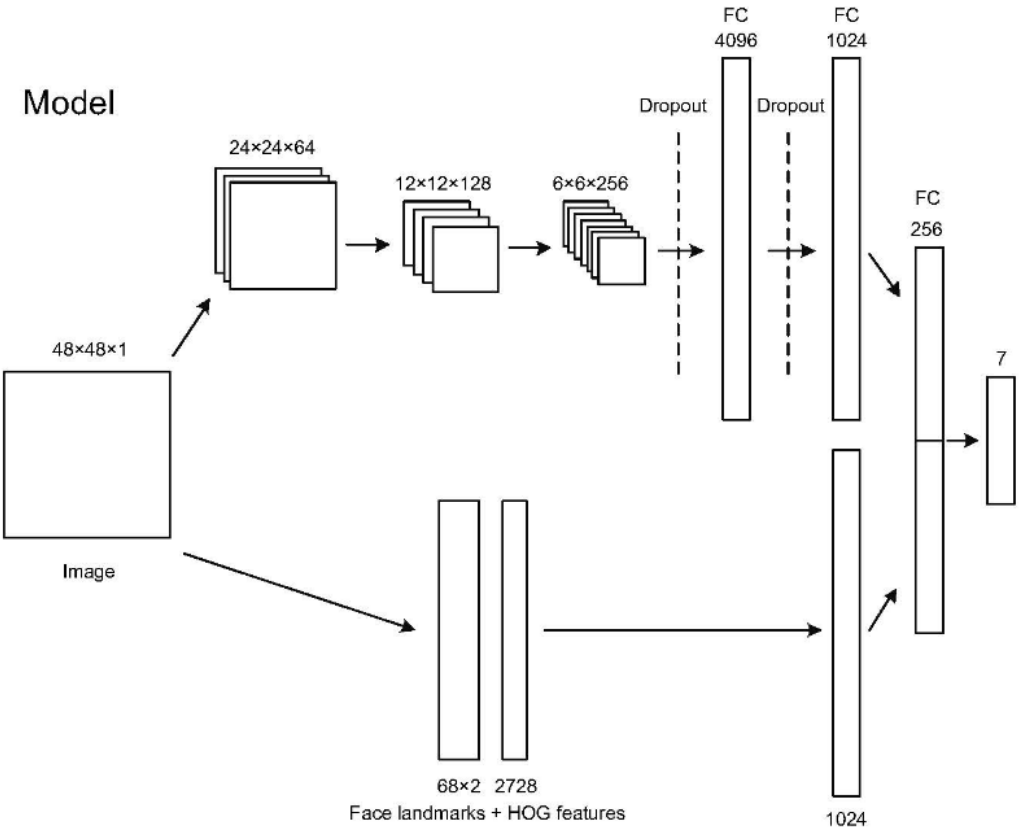

Figure 2. Architecture of the model.

2.2.1. Convolutional layer. It is the major component of the CNN, which is made up of several trainable weights. These weights could be updated by back propagation of the gradients. By optimizing the weights, descriptive features could be learned effectively. As shallow layers of the CNN, low-level features could be learned. At the deeper layers, more complicated and informative high-level features could be learned and aggregated. In this work, these layers are leveraged to learn features from facial images.

2.2.2. Pooling layer. It is the compression of data, it is the input submatrix every n-by-n elements into an element, the pooling layer 2x2 is chosen. The common pooling layer thought thinks that the maximum value or average value represents the local feature and selects the most representative pixel value from the local area to replace the area. It could be used to lower the computational consumption of the model, by decreasing the number of weights, and hence accelerate the calculation and avoid overfitting.

2.2.3. Fully connected layer. The fully connected layer converts the feature map into category output. There is more than one full connection layer. To prevent overfitting, dropout operation will be introduced between each full connection layer to randomly drop some neurons, which can effectively control the sensitivity of the model to noise while preserving the complexity of the architecture.

2.2.4. Advantages and disadvantages. The image is directly input into the model by using CNN for image recognition, and the structure of the image itself can be preserved without the preprocessing and

DOI: 10.54254/2755-2721/14/20230797

feature extraction process of the traditional algorithm, thus reducing the processing complexity of the model. Different from other models, computations layers of CNN are replaced by convolution operation, which takes advantage of deep architectures and the decrease of the number of weights.

However, there are still some limitations and deficiencies in the current neural network. The data of tasks such as face recognition are insufficient and inaccurate, and the biological boundary identification of some facial expressions is not clear. There are basically only positive images in the data, so deep learning is easy to overfit. In addition, the expression of facial features and emotional states varies greatly among different individuals. Different expressions overlap with each other, increasing the difficulty of recognition. The CNN model does not adjust parameters, so it is easy to have expression superposition. All of these will lead to low accuracy of expression recognition results.

2.3. Evaluation index

2.3.1. Indexes. Multiple indexes are leveraged for evaluation, which will be sequentially introduced in this chapter. Accuracy refers to the ratio of the correct predictions over all predictions. A higher value indicates superior model. But it is not always an optimal index, because once the data is seriously unbalanced, accuracy will not work. For example, when dealing with X-rays, the real data is: 99% are disease-free, and only 1% are disease-free. If a classifier only gives it an X-ray, it is judged to be disease-free, then its accuracy rate is also 99%. Obviously, when facing unbalanced data, the defects are obvious.

Precision is translated as precision or accuracy, generally abbreviated as P. It is aimed at the prediction results of the model, indicating how many samples with positive prediction are positive samples.

Recall is translated as recall rate or recall rate, generally abbreviated as R. It indicates the ratio of correctly predicted positive examples.

F1-value is defined based on the Precision and Recall, which is the harmonic average of the two values. It is an index that comprehensively considers Precision value and Recall value. F1 score is mainly used to compare the performance of the two classifiers, with the value ranging from 0 to 1, and the larger the value, the better.

2.3.2. Macro-average method. It directly adds up the evaluation indexes of different categories calculated from various classes. This method assigns equal weights to all categories. However, when dataset is highly biased, the contributions of rare categories could be overlooked.

2.3.3. Weighted-average method. To mitigate the deficiencies of the aforementioned problems, weighted method is proposed. In this method, the weights of various classes are decided by the ratio of the number of samples in each class. During the calculation of the indexes, each category will be multiplied by their corresponding weights.

3. Result

3.1. Training process

To demonstrate the training process and further reveal the intermediate performances, Figure 3 illustrates the accuracy and loss curves. It could be observed that the loss is decreasing with the training proceed.

DOI: 10.54254/2755-2721/14/20230797

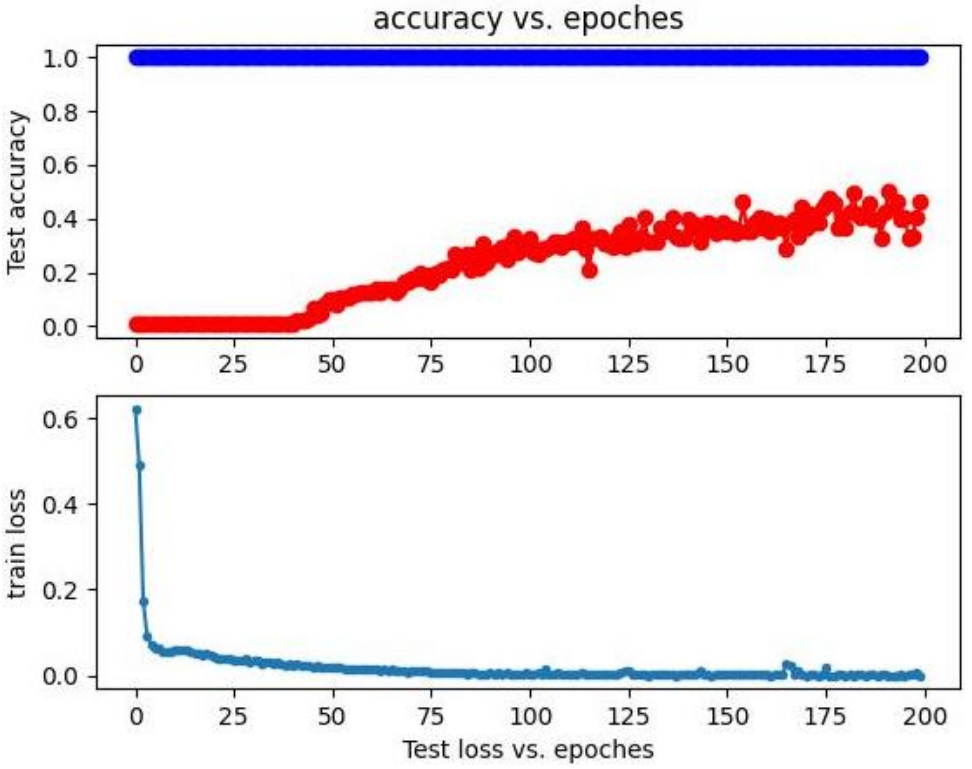

Figure 3. Curves of accuracy and losses.

3.2. Effectiveness of different model structures

In this experiment, the influence of different deep learning layers is demonstrated in the following tables. Table 1 displays the macro average performances and Table 2 demonstrates the weighted average performances. It could be observed that model with 6 layers achieves the optimal performances measured by the F1-score.

Table 1. Macro average performances of different layers.

Macro avg	accuracy	precision	recall	f1-score
3 layers	0.47	0.53	0.56	0.46
4 layers	0.48	0.49	0.56	0.46
5 layers	0.43	0.47	0.51	0.41
6 layers	0.48	0.52	0.56	0.47

Table 2. Weighted average performances of different layers.

Weighted avg	precision	recall	f1-score
3 layers	0.61	0.47	0.51
4 layers	0.56	0.48	0.50
5 layers	0.55	0.43	0.46
6 layers	0.59	0.48	0.51

Besides, to deeply reveal the classification performances and the misclassification status of different classes, the confusion matrix of model with six layers are demonstrated in Figure 3.

DOI: 10.54254/2755-2721/14/20230797

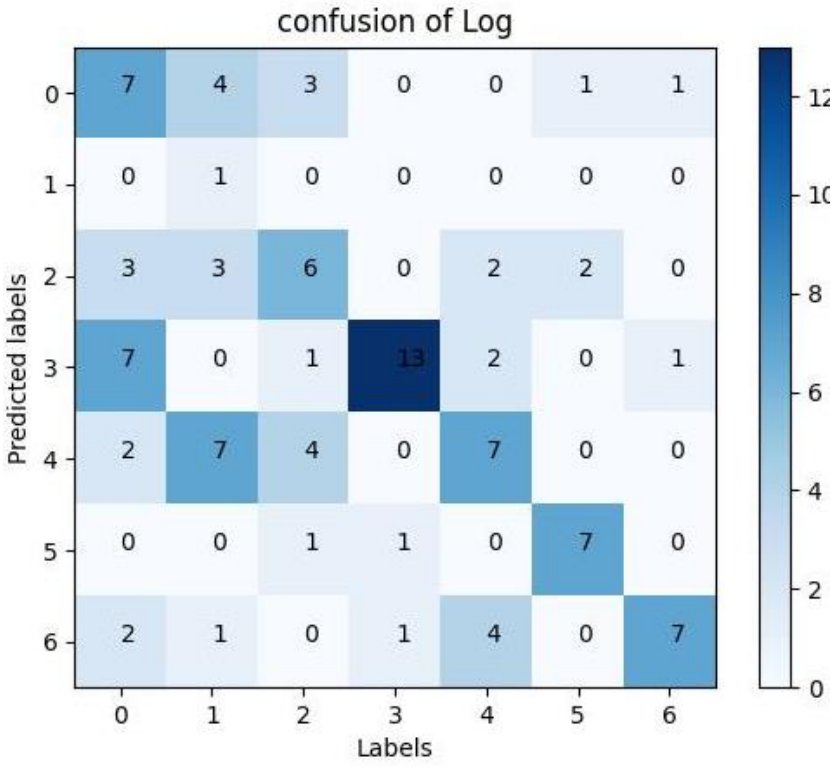

Figure 3. Confusion matrix of model with six layers.

4. Discussion

CNN is widely used for image processing. Features could be extracted through training data. However, the learned features are inexplicit, since the explanation capacities is weak in a CNN. Moreover, from the execution perspective, the model could learn and inference in parallel, which characteristic greatly diminishes the execution time of the model. Besides, the weight sharing mechanism in CNN could strongly decrease the number of weighs used for learning so that saving the storage space when applied on devices, compared with fully connected neural networks.

Convolutional neural networks have been applied to facial expression recognition because of their strong feature learning ability, thus greatly improving the correctness of the recognition capacity. Moreover, its end-to-end nature could simplify the data processing steps, compared with traditional facial expression recognition methods. According to the aforementioned advantages, CNN could perform superior to conventional machine learning algorithms.

But its disadvantages are also particularly obvious, such as: firstly, simple neural networks tend to ignore two-dimensional information of images, leading to inaccurate recognition results. Secondly, the features extracted by shallow convolutional networks have poor robustness.

5. Conclusion

Convolutional neural networks have been applied to facial expression recognition due to their powerful feature learning ability, greatly improving the feature extraction capacity for facial expression recognition. Moreover, convolutional neural networks have greatly simplified the data preprocessing and format compared to traditional facial expression recognition methods. The test experiment uses OpenCV's built-in facial recognition classifier, which leverages the Harr feature to identify the object placements. It is quite an effective detection technology and is often used for face detection, pedestrian detection, and so on. Haar features are widely used in computer vision problems. As a basic feature descriptor, it is also leveraged for the facial expression recognition problem. After loading the XML files of the pkl model and opencv's Haar classifier, predictions can be made by calling the camera and

DOI: 10.54254/2755-2721/14/20230797

extracting files from the path, respectively. CNN is applied to classify the facial emotions. From experiments, it could be concluded that deeper CNN architectures tend to generate better recognition performances. The experiments demonstrate that the model achieves good performances but requires further improvements.

References

[1] Li, S., & Deng, W. (2020). Deep facial expression recognition: A survey. IEEE transactions on affective computing, 13(3), 1195-1215.

[2] Revina, I. M., & Emmanuel, W. S. (2021). A survey on human face expression recognition techniques. Journal of King Saud University-Computer and Information Sciences, 33(6), 619-628.

[3] Fasel, B., & Luettin, J. (2003). Automatic facial expression analysis: a survey. Pattern recognition, 36(1), 259-275.

[4] De la Torre, F., & Cohn, J. F. (2011). Facial expression analysis. Visual analysis of humans: Looking at people, 377-409.

[5] Li, S. Z., Jain, A. K., Tian, Y. L., Kanade, T., & Cohn, J. F. (2005). Facial expression analysis. Handbook of face recognition, 247-275.

[6] Parvanta, C., Hammond, R. W., He, W., Zemen, R., Boddupalli, S., et, al. (2022). Face Value: Remote facial expression analysis adds predictive power to perceived effectiveness for selecting anti-tobacco PSAs. Journal of Health Communication, 27(5), 281-291.

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

[8] Zahara, L., Musa, P., Wibowo, E. P., Karim, I., & Musa, S. B. (2020). The facial emotion recognition (FER-2013) dataset for prediction system of micro-expressions face using the convolutional neural network (CNN) algorithm-based Raspberry Pi. In 2020 Fifth international conference on informatics and computing, 1-9.

[9] Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., et, al. (2018). Recent advances in convolutional neural networks. Pattern recognition, 77, 354-377.

[10] Albawi, S., Mohammed, T. A., & Al-Zawi, S. (2017). Understanding of a convolutional neural network. In 2017 international conference on engineering and technology, 1-6.