Proceedings of the 3rd International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/5/20230538

Human emotion recognition with convolutional neural network

Yu Zhang

Computer Science, Tianjin Ren'ai College, Tianjin, China.

2019302060185@cuc.edu.cn

Abstract. The interaction between intelligent robots and humans has always been a hot issue, and researchers hope to make human-robot interaction as harmonious as human-human interaction. To achieve this, it is particularly important to enable robots to recognize human facial emotions automatically. However, many intelligent robots can already understand people's emotions through vocal communication. However, some people do not like to express their feelings through words, so it would be more convenient to let machines can automatically analyze people's facial emotions. This paper aims to make the machine recognize people's facial expressions and automatically analyze their emotions to make human-computer interaction more harmonious. The convolutional neural network has shown great influence on image feature extraction in the development of the machine learning field today. Therefore, this paper will adopt the advanced method of CNN to train the model on the FER2013 dataset. The abundant experiments demonstrate that the final trained model has good accuracy in recognizing three emotions: happy, surprise, and neutral.

Keywords: Convolutional neural network, facial expression recognition.

1. Introduction

As artificial intelligence continues to advance, people increasingly want to be able to communicate with machines in the same manner that humans communicate with other people. In the process of people's daily communication, not only words but also the meaning conveyed by expressions cannot be ignored. Because expressions are the expression of human emotions, different expressions can express different emotions. Therefore, in order to facilitate human-computer interaction, it is essential to allow machines to recognize human expressions.

One example of non-verbal communication is facial expressions, which are a primary approach to expressing human emotions, and it is almost impossible for people to avoid expressing facial expressions. There are many kinds of facial expressions; by the 1970s, two American psychologists named Ekman and Friesen identified six universal human expressions: happy, anger, surprise, fear, disgust, and sad through adequate experiments [1]. Subsequently, Ekman further refined facial emotions, and then the Facial Action Coding System (FACS) was proposed, which is founded on Action units (AUs) in conveying facial emotions [2].

In modern society, due to the ongoing advancement of technology, intelligent robots are gradually appearing in people's daily life. The ability of robots to communicate with people like humans and understand human emotions has always been a hot issue when it comes to artificial intelligence. At

DOI: 10.54254/2755-2721/5/20230538

present, many intelligent robots mainly analyze people's words to perform emotional analysis. However, in some cases, for example, some people may feel it awkward to express their emotions to robots, or children may not be able to express their emotions through words. A better way is to get people's emotions directly on the basis of their facial expressions. Therefore, it is very crucial to implement the identification of facial expressions, and this approach can make human-computer interaction more natural and harmonious.

In this research, the extraction of facial expression features is done by using the convolutional neural network (CNN). Compared with traditional methods, such as Eigen Face [3], NN, and Adaboost [4], the advantages of convolutional neural networks include local perception, weight sharing, preservation of the image's spatial features information, no explicit feature extraction, etc. As a result, the recognition of facial expressions also frequently uses convolutional neural networks. There are many classical CNN models, such as LeNet5, AlexNet and VGG, which are widely used to extract image features.

2. Related work

Convolutional neural networks have demonstrated significant promise in image processing since the late 1990s. Before that, all feature extraction methods in image processing were manual, but the advent of CNNs made the manual feature extraction process redundant. For example, the most classic CNN model, LeNet5, effectively solved the vanishing gradient problem in the nonlinear classification and achieved an excellent recognition accuracy rate in the handwritten digit recognition problem, which is regarded as the pioneer of the convolutional neural network by later generations [5].

Before this paper, there had been vast amounts of research on how to utilize the convolutional neural network to classify and recognize images. Some previous research results are listed below. Ravi et al. compared the recognition accuracy of the support vector machine (SVM) classifier, which is applied to classify the local binary pattern (LBP) manually extracted characteristics, and the convolutional neural network's recognition accuracy, which is automatically extracted features [6]. Xie et al. proposed a regularization method that embeds feature sparsity into the loss function to improve the generalization capability of neural networks [7]. With the excellent research experience of our predecessors, we are able to further investigate and apply facial expression recognition in practice.

3. Method

The convolutional neural network, the recurrent neural network (RNN), and the generative adversarial network (GAN) are now really well-known networks when it comes to deep learning. Among them is a deep neural network with a convolutional structure called a convolutional neural network, which is distinguished by having a deep structure and incorporating convolutional operations. It is also the most widely used network now and has shown excellent performance in image classification problems. Convolutional layers, pooling layers, activation functions, and fully-connected layers make up the majority of CNN [8]. A simple convolutional neural network structure for the handwritten digits dataset is illustrated in figure 1.

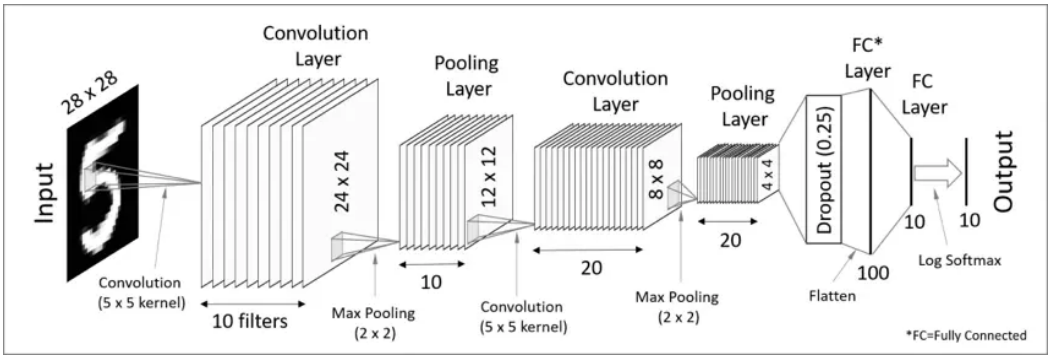

Figure 1. A simple CNN structure.

DOI: 10.54254/2755-2721/5/20230538

The primary function of the convolutional layer is to use convolutional kernels to obtain features from the input data, which are the core of convolutional neural networks. Each convolutional layer contains at least one convolutional kernel, which performs convolutional operations on the region of the image of the same size as the convolutional kernel. The convolutional operations are actually a matter of mathematical feature mapping, and the results of the operations are passed to the next layer as the feature values of the regions in the image.

Reducing the feature map's dimensionality is the primary purpose of the pooling layer, which is frequently positioned after the convolutional layer. In image recognition or image classification tasks, average pooling and maximum pooling are used most frequently [9]. Because the high dimensionality of the feature map after the convolution process leads to the low efficiency of network training, the dimensionality of the feature map needs to be reduced by the pooling layer, speeding up the operation speed and enhancing the network's capacity for generalization.

To improve the convolutional neural network's non-linear representation, activation functions are introduced as a portion of the neural network to improve the model's generalization ability. Without the introduction of activation functions, the convolutional neural network would only perform simple linear weighting of the weights of the neurons. The activation functions used more frequently are Sigmoid, Tanh, ReLU, etc. Currently, the ReLU activation function is the most widely used and is used most frequently [10]. In figure 2, the three functions are shown separately.

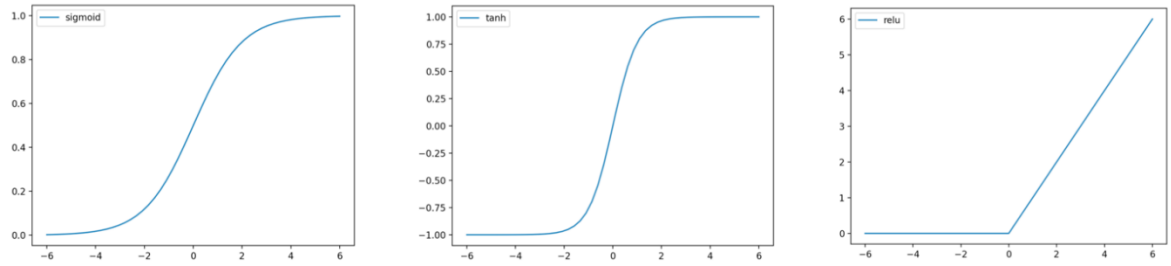

Figure 2. Three types of activation functions.

At the very bottom of the convolutional neural network is the fully-connected layer. Every neuron between adjacent layers in the fully-connected layer is connected to each other, integrating and mapping the features obtained from the previous network layer to the latter network layer and output to the sample label space finally. The fully-connected layer weights and sums the output of the features from the earlier layer and finally completes the target classification by feeding the results into the activation function. A diagram of the relationships between the fully-connected layers is shown in figure 3.

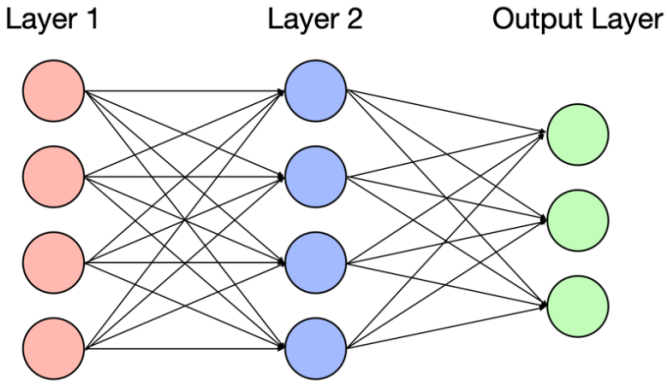

Figure 3. Fully-connected layers.

DOI: 10.54254/2755-2721/5/20230538

4. Experiment

4.1. Dataset and Preprocessing

In this research, the FER2013 dataset was chosen; it consists of 48x48 pixel grayscale images of faces. There are 35,888 images of seven emotions: neutral, anger, surprise, fear, happiness, sadness, and disgust [11]. The division of the training dataset and test dataset follows the official standard.

About data preprocessing: In the data visualization analysis, we found that the data distribution was not uniform, and the disgust emotions were very sparse, as shown in figure 4. We also reviewed the distribution of disgust in the training dataset and test dataset and found that there were no disgust results in the test dataset. So we removed all 436 disgust features and only recognized six emotions: neutral, anger, surprise, fear, happiness, and sadness. Finally, the training dataset has 28,273 images, whereas the test dataset contains 7,178 images.

4.2. VGGNet

A traditional convolutional neural network for large-scale image recognition is called VGGNet [12]. One of the improvements of VGGNet over AlexNet is to replace the larger convolutional kernels of AlexNet with several consecutive 3x3 convolutional kernels. This approach allows VGGNet to have a higher network depth and more nonlinear layers than AlexNet, allowing the model to learn more complex patterns. The model in the paper is trained on VGG19, which network structure with sixteen convolutional layers and three fully-connected layers. Batch normalization (BN) is applied and the activation function utilized is ReLU to speed up the learning speed of the neural network and avoid vanishing gradient or gradient explosion.

4.3. Result

The experiments were conducted with 200 epochs, using the Adam optimizer to reduce the loss value of the cross-entropy loss function. In the end, it achieved a 97.86% accuracy rate on the training dataset and a test dataset accuracy rate of 63.99%. Figure 4 depicts how the loss values and accuracy rates changed throughout the training phase.

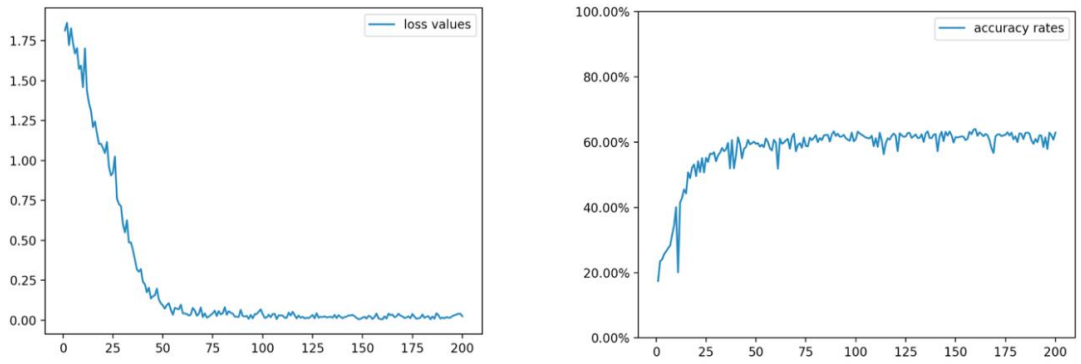

Figure 4. The variation of the training phase.

And then, the recognition accuracy of each category of emotions in the test dataset is shown in figure 5. It can be found that the recognition accuracy is higher for happy, surprise, and neutral but worse for anger, fear, and sadness.

DOI: 10.54254/2755-2721/5/20230538

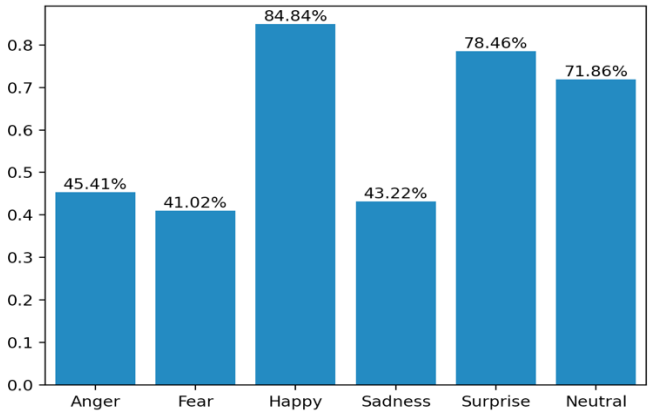

Figure 5. The accuracy rate of six emotions.

5. Conclusion

Through the study of facial expression recognition, this paper successfully demonstrates the capability of CNN for facial expression recognition through experiments. This recognition method is different from the traditional methods; not only can it produce better results, but the feature extraction is also nonmanual. The trained model performs well in some emotions but not well in others. The reason for this result may be due to the fact that different expressions require different feature information. For example, for happy emotion, only the mouth features of the face are needed to be extracted; but for fear emotion, both the mouth and eye features are required. In the case of CNN, some features may be missing in automatic feature extraction, which leads to unsatisfactory recognition results. In a subsequent study, it may be possible to achieve a better result by first recognizing key facial traits like eyes and mouth and then handing these characteristics to the CNN for classification.

In the future, there is still huge research potential for facial expression recognition technology, such as: improving the recognition accuracy, expanding the recognition expression category, and dealing with unfavorable factors such as lighting, pose, and occlusion. As these challenges are overcome, facial expression recognition technology can be gradually popularized.

References

[1] Ekman P and Friesen WV 1978 Facial action coding system *Environmental Psychology & Nonverbal Behavior*
[2] Ekman P 1993 Facial expression and emotion *American psychologist* 48(4) 384-92
[3] De A, Saha A and Pal MC 2015 A human facial expression recognition model based on eigen face *Procedia Computer Science* 45 282-9
[4] Owusu E, Zhan Y and Mao QR 2014 An SVM-AdaBoost facial expression recognition system *Applied Intelligence* 40(3) 536–45
[5] Lecun Y, Bottou L, Bengio Y and Haffner P 1998 Gradient-based learning applied to document recognition *Proceedings of the IEEE* 86(11) 2278–324
[6] Ravi R and Yadhukrishna SV 2020 A face expression recognition using CNN & LBP *Fourth International Conference on Computing Methodologies and Communication (ICCMC)* 684-9
[7] Xie W, Jia X, Shen L and Yang M 2019 Sparse deep feature learning for facial expression recognition *Pattern Recognition* 96 106966
[8] Yamashita R, Nishio M, Do RKG and Togashi K 2018 Convolutional neural networks: an overview and application in radiology *Insights into Imaging* 9(4) 611–29
[9] Zafar A, Aamir M, Mohd Nawi N, Arshad A, Riaz S, Alruban A, Dutta AK and Almotairi S 2022 A comparison of pooling methods for convolutional neural networks *Applied Sciences* 12(17) 8643

DOI: 10.54254/2755-2721/5/20230538

[10] Ramachandran P, Zoph B and Le QV 2017 Searching for activation functions *arXiv preprint arXiv:1710.05941*
[11] Goodfellow IJ, et al. 2013 Challenges in representation learning: a report on three machine learning contests *International conference on neural information processing* 117–24
[12] Simonyan K and Zisserman A 2014 Very deep convolutional networks for large-scale image recognition *arXiv preprint arXiv:1409.1556*