Proceedings of the 5th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/22/20231175

Parameter optimization in convolutional neural network for improving performance of facial expression recognition

Jiyin Zhang

Ocean College, Zhejiang University, Zhoushan, 316021, China

3190101225@zju.edu.cn

Abstract. The recognition of similar facial expressions presents a notable challenge, necessitating a focus on the parameters within the fundamental Convolutional Neural Network (CNN) architecture, which serves as a cornerstone in the field of image classification. This research endeavor aims to enhance the model's capacity for facial expression recognition by employing a controlled variable method to examine two specific parameters in a self-designed small CNN: the number of filters and convolutional layers. More specifically, while the filters were fixed at 3, the layers varied from 3 to 6 to 9. Similarly, as the number of the filters rose to 6, the number of the layers also incremented from 3 to 6 to 9. Furthermore, while the number of the filters reached 12, the number of the layers went from 3 to 6 to 9 too. Finally, with the filters increasing to 24, the layers rose from 3 to 6 to 9 as well. Experimental results indicate that both increasing the number of filters and convolutional layers can increment the performance of model in facial expression recognition. Furthermore, increasing the number of filters can exert a more prominent influence on improving the accuracy of facial expression recognition.

Keywords: CNN, facial expression recognition, machine learning.

1. Introduction

Facial expression serves as a prominent means of demonstrating a person's internal emotion. The ability to discern and comprehend these emotions i.e. Facial Expression Recognition (FER) has gained significant traction across various domains [1-3], including health care, transportation and so force. For instance, in hospitals, doctors can use it to diagnose patients at an early stage by detecting any abnormal facial expression that might indicate some lesions. And by detecting drivers' facial expression, people nearby can remind them to calm down, thus reducing traffic congestion. FER can be used to analyze customer preferences, satisfaction and purchase intentions in order to provide better service and recommendations. Furthermore, FER can be applied to monitor students' learning status, engagement and emotional feedback to improve teaching effectiveness and personalize learning.

In order to better implement facial expression recognition, artificial intelligence has emerged as a powerful tool. The field of FER has witnessed significant progress recently, with Convolutional Neural Networks (CNNs) playing a pivotal role in successfully addressing diverse FER tasks since CNN is a deep learning network that requires fewer pre-processing steps than other conventional image classification methods [4]. For example, the facial expression recognition competition was initiated by International Machine Learning Conference in 2013 [5], where fabulous results are gained through a simple CNN with Support Vector Machine (SVM) layer for facial expression classification. In human-

DOI: 10.54254/2755-2721/22/20231175

computer interaction, CNN-based facial expression recognition can be used to achieve more natural, intelligent and humane human-computer interaction, such as voice assistants, virtual characters, robots, etc [6]. In emotional computing, CNN-based facial expression recognition can be used to analyze human emotional states, personality traits, mental health, etc., in order to provide more personalized and attentive services and advice. In terms of security monitoring, CNN-based facial expression recognition can be used to identify suspicious people, abnormal behavior, dangerous situations, etc., in order to improve security prevention and emergency handling capabilities. However, FER remains a challenging task for computers, particularly when differentiating between similar expressions such as "sad" and "fear." In addition, facial features derived from one person in two different expressions may be very close in the feature space, while there is very good chance that facial features obtained from two individuals that share the same expression could be distant from one another. All in all, CNN is capable of extracting features at various levels of generality and it has a not-bad performance on FER. Usually, more depth in the network leads to more semantic information acquisition. However, the problem of gradient vanish and degradation will arise from merely increasing the depth of the model. Although gradient vanish problem can be solved by the conventional approaches such as normalized initialization and batch normalization, they are not able to address the network degradation problem. Besides, training large deep networks directly on a relatively small facial expression dataset is prone to overfitting. Therefore, there is an urge to study the number of layers of CNN. Besides, since any input filter undergoes convolution [7], another important parameter in CNN is the number of filters. Therefore, some emphasis should be laid on these two parameters in CNN.

In this regard, this study started from scratch to perform facial expression recognition on a self-built small network. Specifically, two important parameters were focused on in CNN, the number of convolutional layers and filters respectively. This study plans to use FER-2013 dataset and implement control variable method. The number of convolutional layers will be increased from 3 to 6 to 9, and the number of filters will be increased from 3 to 6 to 12 to 24 to observe the change in the performance of the model.

2. Method

2.1. Dataset description and preprocessing

In this project, FER2013 dataset related to the facial expressions was used provided by a dataset on Kaggle [8]. It is composed of grayscale images of faces with 48×48 pixels and labels for 7 emotions, namely angry, disgust, fear, happy, sad, surprise, and neutral. Assigning each facial expression to one of seven categories, where 0=Angry, 1=Disgust, 2=Fear, 3=Happy, 4=Sad, 5=Surprise, 6=Neutral, is the objective of the FER challenge. The faces have been automatically adjusted so that the face is approximately centred and fills approximately the same amount of space in each image. The total number of samples is 32,298, with 28,709 examples in the training set and 3,589 examples in the public test set. Figure 1 shows some example images.

In terms of image pre-processing, it is consisted of two parts. Initially, the dataset was normalized by dividing 255. Subsequently, to mitigate the risk of overfitting, the images were modified using data augmentation techniques. They were rotated randomly by 20 degrees, shifted horizontally by 10% of the width and vertically by 10% of the height, zoomed randomly by 10%, and flipped horizontally to increase the diversity and robustness of the data. Furthermore, 20% of the data was used for validation. The data of training generator was shuffled while the data of validation generator and test generator was not.

DOI: 10.54254/2755-2721/22/20231175

Figure 1. Visualizations of sample images of FER2013 dataset [8].

2.2. CNN-based model for facial expression recognition

In this study, a CNN architecture was implemented due to its superior performance in various tasks [9, 10]. The CNN model constructed in this study composes several convolutional layers with a set of 3×3 filters, "relu" activation function, "he_uniform" kernel initializer and "same" padding to identify patterns and high-level features. Then, the output of the convolutional layers is passed through a 2×2 max-pooling layer, which keeps the most relevant features while shrinking the size of the feature maps. Then, a flatten layer is connected with the pooling layer. Finally, the CNN model is connected with a fully connected layer with 128 nodes and an output layer with seven nodes and a "softmax" activation function, which maps the output of the fully connected layers to the seven emotion categories.

The key step in this study is changing the number of the filters in convolutional layers and the number of the convolutional layers. The most important thing is to adopt a control variable approach when conducting experiments, whereby specific parameters are held constant to isolate the impact of variables of interest. In other words, in order to show the effect of these two parameters, the number of the convolutional layers should stay the same when the number of the filters is varying and vice versa This study carried out a total of 12 experiments. In experiments No.1~No.4, the number of the convolutional layers was fixed at 3, while the number of the filters varied from 3 to 6 to 12 to 24. In experiments No.5~No.8, the number of the convolutional layers was fixed at 6, while the number of the filters varied from 3 to 6 to 12 to 24. In experiments No.9~No.12, the number of the convolutional layers was fixed at 9, while the number of the filters varied from 3 to 6 to 12 to 24. The loss and accuracy of training, validation and test dataset were recorded in each experiment. As for other parameters such as the number of epochs, learning rates and batch size, they remained the same in all of the experiments.

2.3. Implementation details

The model used categorical crossentropy loss function, Stochastic Gradient Descent (SGD) optimizer with 0.001 learning rate and 0.9 momentum, 32 batch size, and 20 epochs to train. Evaluation metric adopted is accuracy. Tensorflow, Python and Keras are the used as the main software tools. All experiments were implemented on Google Colab.

3. Results and discussion

3.1. Classification performance of different convolutional layers

A series of experiments were conducted using a self-designed CNN model, where the configuration of convolutional layers was adjusted to assess the performance on the dataset. The performance of the model based on various conditions based on Table 1 and Figure 2. For the case of 3 filters, the test accuracy changed from 0.43994 (3 layers) to 0.43715 (6 layers) to 0.44204 (9 layers). When employing 6 filters, the test accuracy varied from 0.46508 (3 layers) to 0.47346 (6 layers) to 0.46997 (9 layers). Similarly, with 12 filters, the test accuracy ranged from 0.47416 (3 layers) to 0.4986 (6 layers) to 0.50628 (9 layers). Lastly, using 24 filters, the test accuracy showed a progression from 0.49791 (3

DOI: 10.54254/2755-2721/22/20231175

layers) to 0.5014 (6 layers) and reached 0.52025 (9 layers). It can be observed that the overall trend is increasing when the layers are incrementing.

Table 1. The performance of the model based on various conditions.

Layer	Filter	Loss of train	Acc of train	Loss of val	Acc of val	Acc of test
3	3	1.4318	0.4474	1.4690	0.4358	0.4399
3	6	1.3905	0.4687	1.4302	0.4541	0.4651
3	12	1.3218	0.4978	1.3629	0.4787	0.4742
3	24	1.3260	0.5032	1.3767	0.4795	0.4979
6	3	1.4615	0.4375	1.4768	0.4290	0.4372
6	6	1.3794	0.4718	1.4197	0.4438	0.4735
6	12	1.3060	0.4987	1.3478	0.4856	0.4986
6	24	1.2672	0.5139	1.2960	0.4971	0.5014
9	3	1.4365	0.4456	1.4741	0.4325	0.4420
9	6	1.3838	0.4688	1.4109	0.4515	0.4700
9	12	1.3081	0.5010	1.3532	0.4886	0.5063
9	24	1.2688	0.5195	1.3147	0.4978	0.5203

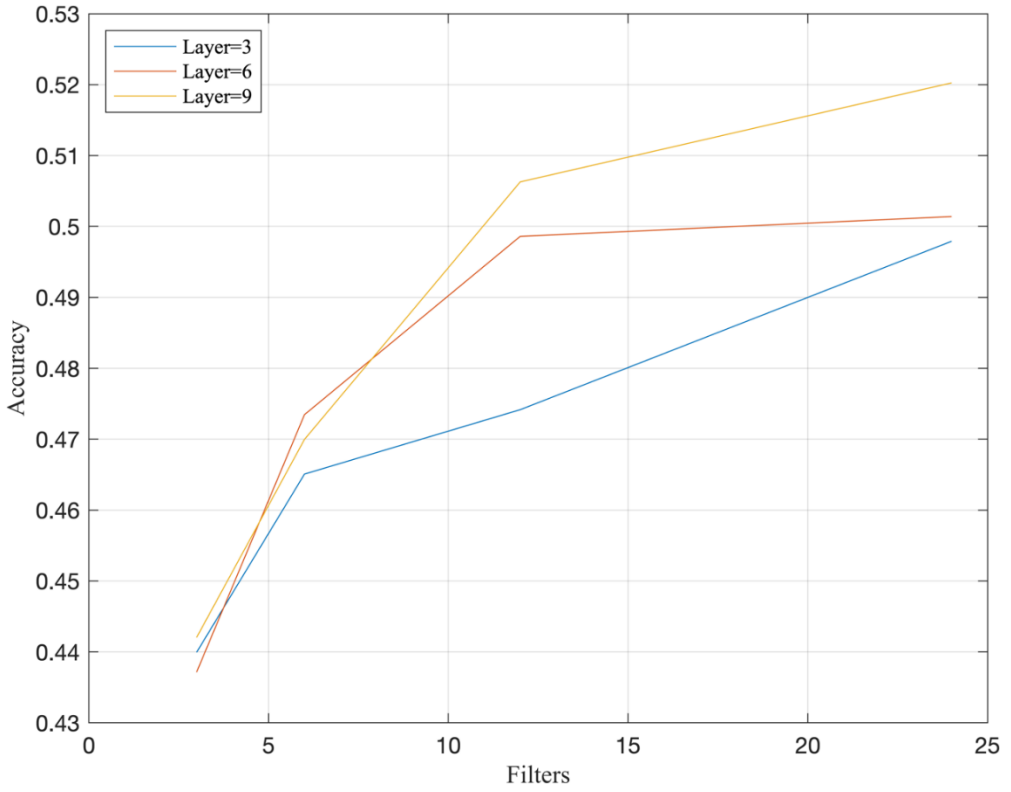

Figure 2. The performance of the model based on various filters and layers (Photo/Picture credit: Original).

DOI: 10.54254/2755-2721/22/20231175

3.2. Classification performance of different filters

With 3 filters, the average accuracy of the test dataset was 0.43971. Increasing the number of filters to 6 resulted in an average test dataset accuracy of 0.46950333. Further increasing the number of filters to 12 yielded an average test dataset accuracy of 0.49301333. Finally, with 24 filters, the average test dataset accuracy reached 0.50652. Apparently, a direct relationship links the number of filters with the classification performance of facial expressions, as an increase in the number of filters led to improved accuracy. Clearly, the classification accuracy improves more significantly with more filters than with more layers.

The empirical findings demonstrate that the performance of the model improves as the number of filters and layers increases, albeit with a more pronounced impact observed when increasing the number of filters compared to increasing the number of layers. This discrepancy can be attributed to the fact that augmenting the filters allows the model to capture a wider range of features, while expanding the layers enhances the intricacy of these features. In the context of facial expressions, the results suggest that the presence of diverse features holds greater significance than the complexity of features. Moreover, diverse features potentially encompass subtle information that contributes to the accurate recognition of facial expressions.

4. Conclusion

In this article, self-designed CNN architecture was used to perform facial expression recognition. Control variable method was implemented to test the effect of two parameters, namely the number of filters and convolutional layers related to the performance of the model in facial expression recognition. Extensive sets of comparable experiments were conducted to evaluate the effect of increasing the number of filters and layers. Experimental results showed that both increasing the number of filters and convolutional layers can improve the accuracy of facial expression recognition and incrementing the number of filters can have a more remarkable effect. In the future, more tests can be done to find the limit of the number of filters and layers where the model's accuracy stops improving. Additionally, other hyperparameters also can be experimented to observe whether increasing the number of those hyperparameters can make a difference to the performance of the model.

References

[1] Tian Y Kanade T Cohn J F 2011 Facial expression recognition Handbook of face recognition pp 487-519.
[2] Bettadapura V 2012 Face expression recognition and analysis: the state of the art arXiv preprint arXiv:1203.6722.
[3] Li S Deng W 2020 Deep facial expression recognition: A survey IEEE transactions on affective computing 13(3) pp 1195-1215.
[4] Fei Z et al 2020 Deep convolution network based emotion analysis towards mental health care Neurocomputing 388 pp 212-227.
[5] Goodfellow I J et al 2015 Challenges in representation learning: A report on three machine learning contests Neural Netw vol 64 pp 59–63.
[6] González-Lozoya S M et al 2020 Recognition of facial expressions based on CNN features Multimedia tools and applications 79 pp 13987-14007.
[7] Meryl C J 2020 Deep Learning based Facial Expression Recognition for Psychological Health Analysis 2020 International Conference on Communication and Signal Processing (ICCSP) pp 1155-1158.
[8] Wolfram Research 2018 FER-2013 Retrieved from: https://www.kaggle.com/datasets/msambare/fer2013.
[9] Qiu Y Yang Y Lin Z et al 2020 Improved denoising autoencoder for maritime image denoising and semantic segmentation of USV China Communications 17(3) pp 46-57.
[10] Kayalibay B Jensen G van der Smagt P 2017 CNN-based segmentation of medical imaging data arXiv preprint arXiv:1701.03056.