Proceedings of the 3rd International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/4/2023394

Facial expression recognition for a seven-class small and medium-sized dataset based on transfer learning CNNs

Zixuan Chu

School of Computer Science, University of Nottingham, University Park, Nottingham NG7 2RD, UK

psxzc5@nottingham.ac.uk

Abstract. As one of the most common biological information for human to express their emotions, facial expression plays an important role in biological research, psychological analysis and even human-computer interaction in the computer field. It is very important to use the efficient computing and processing power of computers to realize automatic facial expression recognition. However, in the research process of this field, most people's technology and model achievements are based on large datasets, which lack the universality of small and medium-sized datasets. Therefore, this project provides an optimization model on a specific seven-class small and medium face image dataset and provides a possible technical optimization reference direction for facial expression recognition models on similar small and medium-sized datasets. During the experiment, the training performance of VGG16 and MobileNet is compared. A comparative experiment is set up to observe the effect of transfer learning mechanism on training results. The results show that transfer learning has a significant effect on the performance of the model, and the accuracy of the optimal test set is more than 90%. Regardless of whether transfer learning mechanism is used, the training performance of VGG16 model structure is better than that of MobileNet structure in the same dataset.

Keywords: Facial expressions recognition, Transfer learning, Convolutional Neural Network.

1. Introduction

Given the recent acceleration in the development of artificial intelligence, how to liberate human and material resources by using efficient computing and helped by deep learning models of computers has become the main research goal of one. In today's information age, facial expression, as a kind of human biological information, is undoubtedly the most powerful daily expression of human emotions. As early as 2000 Maja Pantic and Leon J.M. Rothkrantz proposed the concept of an automated facial expression analysis system [1]. They believe that an ideal facial expression system is a combination of facial image detection, facial feature extraction and finally emotion classification. Compared with fingerprint recognition, iris recognition and other technologies, facial expression recognition can not only make judgment and recognition of human information, but also has a wide range of application value in psychology, biology, even human-computer interaction, virtual reality and other aspects.

Ekman and Friesen defined six basic emotions in the 20th century [2]. These typical facial expressions were anger, happiness, disgust, fear, sadness and surprise. According to their cross-cultural research, human perception of basic emotions is not influenced by culture [3]. In addition,

DOI: 10.54254/2755-2721/4/2023394

with the development and application of deep learning, Shan Li and Weihong Deng proposed a neural network training strategy based on static and dynamic image sequences [4]. A training set evaluation method and a complete set of training sets and algorithms are presented to solve the overfitting problem caused by irrelevant information related to expression. Another important point is that for different models, model building and parameter tuning are essential. In 2020, Xuewen Rong et al. analyzed and compared the common activation functions of five different convolutional neural networks in facial expression recognition task, and concluded that the improved activation functions have better classification performance than the advanced activation functions [5]. And in terms of model selection and optimization, Yong Li et al. proposed an attention mechanism of convolution neural network in obscured facial expression recognition problem has the very good performance in 2019 [6]. In the same year, Ismail Oztel et al. also compared the performance of using transfer learning models with basic deep learning models and pointed out the positive role of transfer learning concept in the process of improving machine learning models [7].

Although the above related works have provided some inspiration and technical support for this study, their experiments and results were based on some large datasets and some public datasets. These larger datasets had better performance in the training process may be due to data size and epoch times. They did not consider the learning performance and generalization ability of the model in small and medium-sized datasets. Therefore, in order to cope with this limitation, this project will explore the optimal model of facial expression recognition based on a seven-category small and medium-sized dataset. Different from Ekman and Friesen's six emotion classifications [2], the dataset used in this project added a seventh classification label for neutral emotion expression (neutral) to enhance generalization. By comparing the performance of the basic convolutional neural network model trained on this specific dataset and the model modified by adding transfer learning optimization, the optimization model will be obtained, and the reference standard will be provided for other similar tasks. The main machine learning model structures of this study are MobileNet and VGG16. By controlling whether transfer learning weights in ImageNet are used to design comparative experiments, it showed that the performance of the two models with transfer learning in the same small and medium-sized dataset is better than that without transfer learning in the same model.

2. Methodology

2.1 Dataset description and preprocessing

The dataset used in this project is an open source data set downloaded from Kaggle platform [8]. The entire dataset is made up of gray images of facial expressions. An example of six single images randomly selected from the dataset is shown in Figure 1 below.

Figure 1. An example of six single images extracted from the dataset randomly.

DOI: 10.54254/2755-2721/4/2023394

This dataset has been divided into seven categories and labelled the following seven facial emotions: anger, disgust, fear, happiness, sadness, surprise and neutral. Faces have been automatically registered, so faces are more or less centred and take up about the same amount of space in each image. Each individual image is 48 by 48 pixels in size. The whole dataset is automatically divided into training set and test set. The training set contains 28, 798 images and the test set contains 7, 178 images.
In the process of data preprocessing, the image was normalized first. After dividing the size of each pixel in the training set by 255, the newly obtained floating point value was passed into the model for training, which can improve the training efficiency. After that, the images were randomly selected and rotated by 10 degrees, the horizontal and vertical directions were shifted by 0.1 times, and were scaled by a factor of 0.1 after doing a horizontal mirror flip. Different from the training set, only normalization was carried out for the test set. And finally, the size of all pictures is changed to 48 by 48 pixels in preparation for the subsequent input of the model.

2.2 Proposed model

Convolutional neural network (CNN) is a classical feedforward network structure with deep network structure and convolutional computation, which is a very representative and widely used algorithms in deep learning field [9]. It has played a very significant part in the field of artificial intelligence (AI) in recent years. The main structure of CNN consists of the layer of input, the hidden layer and the layer of output. The main difference between the various models is the different construction structures for the hidden layers, which mainly includes the convolutional layer, the layer of pooling and the fully connected layer. Firstly, the convolutional layer will extract features from the input data. There are multiple convolution kernels inside the convolutional layer. Each element in the convolution kernel has its corresponding weight and bias. The second is the layer of pooling. In pooling layers, the parameter matrix size will be effectively reduce, thereby enabling the number of parameters to be reduced in the final layer. The pooling layer can speed up the process of calculation and prevent the over-fitting issue to some extent. At the end of the CNN is the fully connected layer. Only the signals are passed to the other fully connected layer here. The processed feature map will also lose its spatial structure here and be transformed into vector form, and then passed to the output layer through the excitation function.

Transfer learning mechanism is an application of using the knowledge which learned in some tasks to another similar task. Its core is to find the similarity between the existing knowledge and the new knowledge and achieve the purpose of transfer learning through the transfer of this similarity. Transfer learning mainly solves the problem that it is too costly to directly learn the target domain from scratch in some machine learning scenarios. In machine learning scenarios, transfer learning preserves the first few layers of trained neural networks and uses information such as weights and deviations of trained models to solve new learning problems. Jia Deng et al., provided a large visual database ImageNet for the research of visual object recognition software in 2009 [10]. More than 14 million images are annotated in the database. As a well-known dataset, this database is also used for training and learning of machine learning models such as image classification and object recognition.

In this project, the main transfer learning models used are MobileNet and VGG16. MobileNet is a lightweight convolutional neural network that convolves with far fewer parameters than standard convolutions. VGG16 is a neural network architecture based on CNN. This algorithm was devised by Oxford University's Geometric Vision teamand won the prize of ILSVRC object recognition algorithm in 2014 [11]. The main core idea is to use the weights obtained from the training of these two models on ImageNet database to transfer to the dataset of the project to complete the final recognition task. The main method of the project is to use two groups of comparative experiments and use the parameter weight in the model to control whether the weight of transfer learning based on ImageNet dataset is used, so as to compare whether the use of transfer learning in the same model will cause significant influence on the results.

DOI: 10.54254/2755-2721/4/2023394

2.3 Implementation details

The main research environment of this project is based on Tensorflow platform. The batch size of the machine learning model used is 64 and the epoch is 10. The optimizer chose Adam optimizer, which the learning rate was 0.01%, and using categorical cross entropy function as the loss function. As for the experimental results, there are four main evaluation indicators, namely accuracy, precision, recall and Area Under Curve (AUC). These four indicators were used to evaluate how well different models performed.

3. Results and discussions

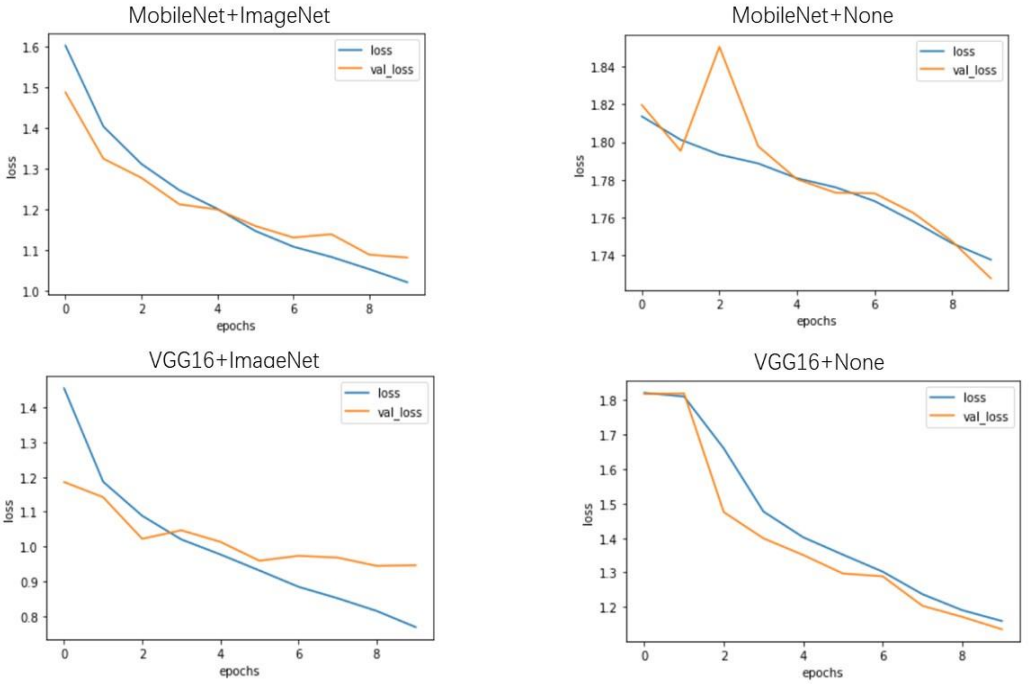

Figure 2. A combination chart of loss curves trained in four different models.

Figure 2 illustrates four different loss broken lines of training set and testing set trained in four structure models.

Table 1. A table of performance trained in four different models.

Model	Performance			
	Training loss	Training accuracy	Testing loss	Testing accuracy
MobileNet+ImageNet	1.0200	0.9021	1.0808	0.8987
MobileNet+None	1.7376	0.8572	1.7277	0.8578
VGG16+ImageNet	0.8144	0.9203	0.9441	0.9094
VGG16+None	1.1593	0.8921	1.1354	0.8939

Table1 shows the different performance of four machine learning models under the same specific dataset in this project. It is obvious that the VGG16 and MobileNet models using transfer learning perform very well on specific small and medium-sized datasets, with the accuracy of the test set exceeding 85%. In addition, by comparing the training performance under the same network structure,

DOI: 10.54254/2755-2721/4/2023394

it can be found that whether transfer learning is used, namely using the trained weights in the ImageNet dataset, has a significant influence on the accuracy of the model. In MobileNet model, using transfer learning will improve the accuracy by about 5%, while in VGG16 model, it will improve by about 2%. Another point is that by comparing the performance of different network structures, it can be found that VGG16 model performs better than MobileNet model regardless of whether transfer learning mechanism is used.

The main reason for this result is probably because the Imagenet database contains most natural images. Images of humans were low and faces with different emotions were even lower. However, because there are lots of basic and universal texture features, it can be easily transferred to facial expression recognition. In terms of models, MobileNet is a lightweight model. It contains less structured information such as parameters compared to VGG16. Therefore, for the more complex data such as face images, the model with more parameters will perform better.

4. Conclusion

This project mainly proposes a machine learning model for facial expression recognition based on small and medium-sized datasets. It makes up for the lack of research on the optimization of recognition models for small and medium-sized datasets in the field of expression recognition and gives possible technical directions to improve the models' performance. This project set up a comparative experiment to compare the training performance of two mainstream network structures the MobileNet and VGG16 in the same specific small and medium-sized dataset. By controlling whether transfer learning weights trained in ImageNet database are used, the influence of transfer learning on recognition models of small and medium-sized datasets is explored. Experimental results show that transfer learning mechanism can improve the performance of the model obviously. At the same time, whether transfer learning is used, the training performance of VGG16 model is better than that of MobileNet structure on the same small and medium-sized dataset. In the future, further study plans to carry out more accurate division, collection and processing of datasets, so as to solve the problem of facial expression recognition in different situations such as natural state and image with noise state.

References

[1] Pantic M and Rothkrantz L 2000 Automatic analysis of facial expressions: the state of the art in IEEE Transactions on Pattern Analysis and Machine Intelligence vol. 22 no 12 pp 1424-1445

[2] Ekman P and Friesen W V 1971 Constants across cultures in the face and emotion Journal of personality and social psychology vol 17 no 2 pp 124–129

[3] Ekman P 1994 Strong evidence for universals in facial expressions: a reply to russell's mistaken critique Psychological bulletin vol 115 no 2 pp 268–287

[4] Li S and Deng W 2020 Deep Facial Expression Recognition: A Survey in IEEE Transactions on Affective Computing

[5] Wang Y et al. 2020 The Influence of the Activation Function in a Convolution Neural Network Model of Facial Expression Recognition Applied Sciences 10 5

[6] Li Y et al. 2019 Occlusion Aware Facial Expression Recognition Using CNN With Attention Mechanism in IEEE Transactions on Image Processing vol 28 no 5 pp 2439-2450

[7] Oztel I et al. 2019 Performance Comparison of Transfer Learning and Training from Scratch Approaches for Deep Facial Expression Recognition 2019 4th International Conference on Computer Science and Engineering (UBMK) pp 1-6

[8] Manas S 2013 FER-2013 Learn facial expressions from an image https://www.kaggle.com/datasets/msambare/fer2013

[9] Goodfellow I et al. 2016 Deep learning. Cambridge, Mass London MIT Press

[10] Deng J et al. 2009 ImageNet: A large-scale hierarchical image database in 2009 IEEE Conference on Computer Vision and Pattern Recognition IEEE pp 248–255

DOI: 10.54254/2755-2721/4/2023394

[11] Karen S et al. 2015 Very Deep Convolutional Networks for Large-Scale Image Recognition arXiv.org [Preprint]