Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 2, No. 1, 2022

A facial expression recognition method based on Convolutional Neural Network

Hongbin Huang*

School of Computer Engineering, Jimei University, Xiamen, Fujian,361021, China
* **Corresponding author**: Email: 15980839993@163.com

Abstract: This project is the implementation method of a simple static facial expression recognition (FER) project. Under the environment of Python3, the deep learning model is used to compare with the traditional model. Finally, CNN (Convolutional Neural Networks) is actually used to construct the entire system, and model evaluation is carried out on the three FER datasets FER2013, JAFFE and CK+. The project's main functions include "get a picture of a person's face" and "recognize expressions". The principles of the above functions include the following: the establishment of neural network structure, the acquisition of data sets, the model training based on data sets, the use of OpenCV to obtain the face, and the use of the model to recognize the expression. The experimental results reproduce the simple deep learning model to realize FER and verify the different effects of different data sets on the results. Face recognition has been widely used in all aspects of life, but different purposes need different models and data collection methods, and the laboratory collection cost is high, and the amount of data is limited, so this project also discusses the data set selection methods under different purposes.

Keywords: CNN; FER; Jaffe; CK+; FER2013.

1. Introduction

1.1. Facial expression recognition

Facial Expression Recognition (FER) is one of the most effective, natural, and universal human cues for conveying emotional reactions and intentions. It is the most natural way to express the inner world, and it plays a crucial role in social interactions [1]. FER generally includes static image FER and dynamic sequence FER. The former one generally refers to image expression recognition, and the latter refers to analysis and modelling based on video sequence.

1.2. Literature review

Ekman and Friesen named six fundamental emotions in the twentieth century based on cross-cultural research indicating that people experience certain basic emotions in the same manner regardless of culture. [2] Anger, disgust, fear, happiness, sadness, and surprise are the classic facial expressions. Subsequently, contempt was listed as one of the fundamental emotions.

The use of CNN set can outperform a single CNN classifier, which is often limited by certain conditions in the application. While the ensemble CNN fuses the discriminative information of each single classifier, it can realize the complementarities between the advantages and disadvantages of each classifier. Therefore, it is very important to find a method to improve the classification performance. [3] In addition, the facial micro-expression can sometimes express opposite emotions according to psychological changes. It is necessary to accurately judge the emotions to be expressed by the other person based on the context relationship such as body movements, language and events. [4]

This paper presents a FER method based on CNN, which can meet the requirements of FER. This research can make FER applied to many fields.

2. Materials and methods

2.1. Program environment

This FER is based on Python3 and Keras2 (TensorFlow backend) with the following installations (Conda virtual environment is recommended). Environments: anaconda, python3.6, tensorflow3.x, PyCharm; Equipment: Personal computer PC, Windows 10; The construction of the model mainly refers to the following network structure designed by Going Deeper. After the input layer, the (1,1) convolutional layer is added to increase the nonlinear representation, and the model has a shallow level with fewer parameters.

2.2. Data processing and analysis

In this paper, the datasets used to train the model are 'CK+', 'JAFFE' and 'FER2013'. We chose these three datasets over a single one mainly because a single dataset is either too small or inaccurate, and the trained model is not good. Therefore, we decided to take turns using three datasets for training and comparison through several trials, and tried to analyse the impact of different datasets on model training.

In addition to the dataset selection, we also need to preprocess the target images to be recognized so that the trained model can be used. The specific preprocessing includes the following steps: the location of the face, the geometric normalization of the face region and the gray level normalization [5]. Geometric normalization is to determine the main rectangular feature region according to the sum of feature points and geometric model of facial expression image, and then cut the original size of 256×256 image into 150×100 image, so that the size of the face in the image is consistent.

Table 1. An overview of the facial expression datasets

Database	Samples	subject	Collection condition	Expression dirstribution	Access
CK+	593 image sequences	123	Lab	Seven basic expressions plus contempt	http://www.consortium.ri.cmu.edu/ckagree/
JAFFE	213 images	10	Lab	Seven basic expressions	http:/www.kasrl.org/jaffe.html
FER-2013	35,887 images	N/A	Web	Seven basic expressions	https://www.kaggle.com/c/chal enges-in-representation- learning-facial-expression-

2.3. Theoretical Analysis of the CNN and VGGNet

2.3.1. Definition of CNN

Convolutional Neural Network (CNN) structure is mainly composed of convolutional layer, pooling layer and fully connected layer(fc)[6]. The function of the convolution layer is to realize the feature extraction composed of some convolution kernels, do convolution operation on the input image, add the offset, and output the result to the activation function to obtain the output, which reduces the number of network parameters and reduces the complexity of parameter selection. The image can be directly used as the network input, avoiding the complex process of feature extraction and data reconstruction in traditional methods. Pooling layers can be used to preserve invariance. The role of the convolutional layer is to detect the local connections of the features of the previous layer to achieve feature extraction, while the role of the pooling layer is to fuse the similar features. Pooling layers are often used together with convolutional layers to reduce the size by down sampling, resulting in invariance of the features. [7]

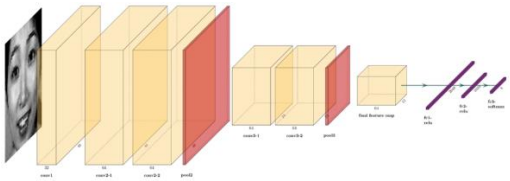

Figure 1. Structure of CNN used in this paper

2.3.2. VGGNet

The VGGNet used in this paper was designed in 2014. According to the change of the feature map size, the VGG16 model can be divided into six parts. The side length of the feature map is reduced to 1/2 for each pooling operation, and the other operations do not affect the feature map size. There is no essential difference between VGGNet16 and VGGNet19, only the network depth is different, with the former having 16 layers (13 convolutions and 3 fully connected layers) and the latter having 19 layers (16 convolutions and 3 fully connected layers). [8]

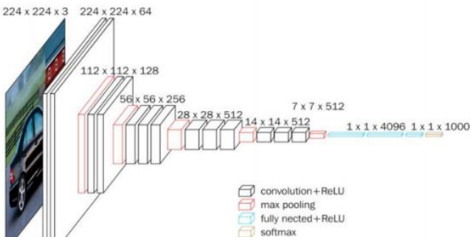

Figure 2. Data configure in CNN

2.4. Sampling and Implementation

We import three datasets (Jaffe, Ck+, and FER2013) and modify the model until all are trained. Normalized datasets and three CNN models are trained. Table 1 compares three CNN subnetwork setups. Three networks ensure network diversity. Various convolutional layers can learn different characteristics. CNN1 to CNN3 are three models. CNN1 controls the range of the receptive field. If the field is too vast, noise will hamper performance. It has three convolutional layers and three max-pooling layers, 32, 64, 128 convolutional filters, and 33 filter windows. CNN2 has three convolutional layers, three max-pooling layers, and 32,32,64 convolutional filters. CNN1's nonlinear representation is improved by adding an 11-convolution layer after the input layer. All three AD hoc networks end up as two dense fully linked layers in CNN3. All pooling layers are 22. The max-pooling layer summarizes the filter region as a nonlinear down sampling, providing translation invariance and lowering computations.

Table 1. Parameters of the models

layer	kernel	k-size	stride	pad	drop	output
Input	0	0	None	None	0	48*48*1
conv1-1	32	1*1	1	0	0	48*48*32
conv2-1	64	3*3	1	1	0	48*48*64
conv2-2	64	5*5	1	2	0	48*48*64
pool2	0	2*2	2	0	0	24*24*64
conv3-1	64	3*3	1	1	0	24*24*64
conv3-2	64	5*5	1	2	0	24*24*64
pool3	0	2*2	2	0	0	12*12*64
fc1	None	None	None	0	50%	1*1*2048
fc2	None	None	None	0	50%	1*1*1024
output	None	None	None	0	0	1*1*8

3. Result and Discussion

The training experimental results are shown in Table 2:

Table 2. CNN training result

loop	train-acc	test-acc
200	32%	28.80%
400	60%	34.90%
600	81%	48.80%
1000	100%	63.10%

3.1. Result analysis

In this paper, we train on FER2013, JAFFE, CK+. Since the JAFFE dataset provides a half-body map, a face detection step has to be performed. We finally achieved about 67% accuracy for both Pub Test and Pri Test on FER2013, and about 99% accuracy for both JAFFE and CK+ with five-fold cross validation.

Training on the FER2013 dataset, when the Batch size is 200, the accuracy is 0.67. We can observe a sharp drop in validation and training accuracy as the Batch size approaches 185, and then pick up to create a gap. The reason for this phenomenon is speculated to be the overfitting phenomenon caused by limited datasets. Compared with the accuracy obtained by other scholars who also used the FER2013 dataset, our accuracy in this paper is not the highest, but it is as expected. Although FER2013 has the largest amount of data in the three datasets, the crawler collection of this dataset has problems such as label errors, watermarks, and animated pictures, which lead to training errors and finally lead to obvious training inaccuracies.

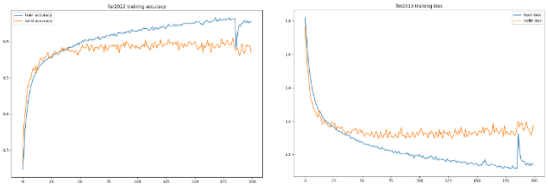

Figure 3. FER2013 training accuracy and loss

However, we used JAFFE and CK+ as datasets for training, and obtained much higher accuracy than the model trained by FER2013. The following figure shows only the training results of JAFFE. The main reason is that the amount of data in this dataset is very small, and the accuracy of the data collected in the laboratory is high enough, and there are almost no mislabeled cases, and only a few unavoidable cases of FER ambiguity, so the accuracy of both training and validation can reach 99%.

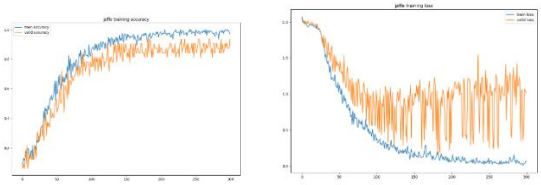

Figure 4. JAFFE training accuracy and loss

We employed a visual training process to evaluate the training results and accuracy of three datasets. The following figure shows the common plot of training on the three datasets with the specified rounds of batch size training.

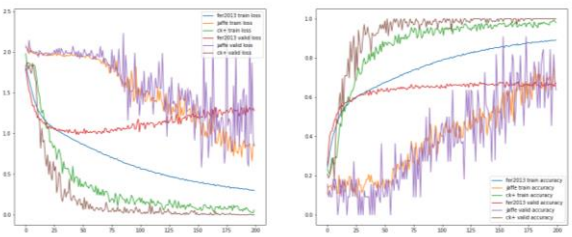

Figure 5. common plot of training

Compared to other methods, this paper's strategy achieves good results on the CK + and JAFFE expression datasets. In the CK+ dataset, the suggested approach's recognition rate is higher than other classic machine learning algorithms, including the fundamental CNN method. In the JAFFE dataset, the proposed technique has a lower recognition rate than traditional methods, but because it uses CNN, the complexity of manually built features is avoided, and a good recognition effect may be reached by training a few layers of the network, saving a lot of training time. JAFFE has fewer original data, which affects its recognition rate.

3.2. Possible solution

In recent years, the most popular strategy to alleviate the problem of expression database size is to transfer the object recognition model or face recognition model to the expression recognition task, namely transfer learning method. In addition to transfer learning strategies, the use of semi-supervised methods is also a possible development trend in the future. The main reasons are as follows, firstly a large-scale face recognition database contains a large number of expressions faces. Also, databases like AffectNet and EmotioNet still have a large proportion of expression faces that are not annotated.

4. Conclusion

This paper studies the use of CNN to realize FER under different data sets, and obtains high accuracy. Under the VGGNET structure of CNN, it is compared with the traditional method, and the implementation process and results are discussed.

4.1. Program result

Our system can also run cross-platform Internet applications. Parallel development of prototypes and code saves time. HDF5 files can iterate and replace Kera's-trained models, facilitating engineering implementation.

By using a simple parallel network model for FER, training and testing may be done in 45 minutes. In future research, can try to optimize the key area extraction accuracy and robustness of the model, which can better meet the needs of real-life scenes, centralized data quantity and accuracy are important, but the huge amount of data access is not the best solution, at the same time need to optimize the model framework, to achieve the most efficient balance.

4.2. Fer datasets

As FER research focuses on difficult in-the-wild environmental conditions, several researchers are employing deep learning to overcome illumination variance, occlusions, non-frontal head orientations, identification bias, and low-intensity emotions. FER is a data-driven task, hence deep FER systems face a lack of both quantity and quality training data.

Different ages, cultures, and genders display and interpret facial expression differently. An ideal facial expression dataset should include abundant sample images with precise face attribute labels, not just expression but also age, gender, and ethnicity, to facilitate research on cross-age range, cross-gender, and cross-cultural FER using deep learning techniques, such as multitask deep networks.

4.3. Dataset Bias & Imbalanced Distribution

Different collecting circumstances and subjective labelling generate bias in facial expression databases. Recent research successfully tests algorithms on a specific dataset. Within-database algorithms lack generalizability on unseen test data,

and cross-dataset performance is worsened by inconsistencies. Due to unequal expression annotations, combining several datasets cannot increase FER performance. FER system evaluation criteria include cross-database performance. Domain adaptability and knowledge distillation minimize prejudice. Sample acquisition causes unbalanced facial class distribution. Identifying disgust, anxiety, and other rare expressions is difficult. Resampling and balancing class distribution during pre-processing is one way. Cost-sensitive loss layer for network training.

References

[1] Tian, Y., Kanade, T., and Cohn, J. F. (2001) “Recognizing action units for facial expression analysis,” IEEE Transactions on pattern analysis and machine intelligence, vol. 23, no. 2, pp. 97–115.

[2] Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. Journal of Personality and Social Psychology, 17(2), 124–129. https://doi.org/10.1037/h0030377.

[3] M. Shi, L. Xu and X. Chen, "A Novel Facial Expression Intelligent Recognition Method Using Improved Convolutional Neural Network," in IEEE Access, vol. 8, pp. 57606-57614, 2020, doi: 10.1109/ACCESS.2020.2982286.

[4] Lisa F. B, Kristen A. Lt, Maria G. Language as context for the perception of emotion, Trends in Cognitive Sciences, Volume 11, Issue 8, 2007, Pages 327-332, ISSN 1364-6613, https://doi.org/10.1016/j.tics.2007.06.003. (https://www.sciencedirect.com/science/article/pii/S1364661307001532).

[5] Weimin Huang and R. Mariani, "Face detection and precise eyes location," Proceedings 15th International Conference on Pattern Recognition. ICPR-2000, 2000, pp. 722-727 vol.4, doi: 10.1109/ICPR.2000.903019.

[6] Alzubaidi, L., Zhang, J., Humaidi, A.J. et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. J Big Data 8, 53 (2021). https://doi.org/10.1186/s40537-021-00444-8.

[7] Taylor, G.W., Fergus, R., LeCun, Y., Bregler, C. (2010). Convolutional Learning of Spatio-temporal Features. In: Daniilidis, K., Maragos, P., Paragios, N. (eds) Computer Vision – ECCV 2010. ECCV 2010. Lecture Notes in Computer Science, vol 6316. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-15567-3_11.

[8] H. Jun, L. Shuai, S. Jinming, L. Yue, W. Jingwei and J. Peng, "Facial Expression Recognition Based on VGGNet Convolutional Neural Network," 2018 Chinese Automation Congress (CAC), 2018, pp. 4146-4151, doi: 10.1109/CAC.2018.8623238.