Review of CNN based face emotion recognition models

Youdan Zhang, Xin Chen

Works are evenly distributed

Youdan Zhang: finished writing the project report, testing and evaluation of the model and quantitative analysis

Xin Chen：completed the model architecture, model testing and project report review and report layout

Review of CNN based face emotion recognition models

YOUDAN ZHANG, Syracuse University, USA

XIN CHEN, Syracuse University, USA

Facial expression recognition is now a hot topic, it is one of the most important features of human beings, and expression recognition has a great application space in life. At present, many expression recognition projects are based on CNN (Convolutional Neural Network) algorithm. How to choose the best algorithm and algorithm optimization has become a topic worthy of research. In this review, we will compare various optimization models based on CNN models, including DCNN, CNN-LSTM, and CNN-biLSTM. We will cite two datasets to train and test different models and their accuracies, fer2013 and CK+ [1], to analyze the various models based on their characteristics and experimental results, and to provide some basis and direction for future research.

Additional Key Words and Phrases: emotion recognition; deep learning; Convolutional Neural Network

ACM Reference Format:

Youdan Zhang and Xin Chen. 2021. Review of CNN based face emotion recognition models. 1, 1 (December 2021), 8 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

1.1 Facial emotion recognition

The expression recognition is mainly performed by the following process as shown in Fig 1.

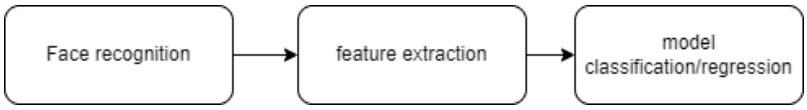


Fig. 1. Expression recognition process

Now face recognition is already a mature technology we will not elaborate too much, the computer needs us to give it features to distinguish different expressions, so we should choose features with high distinction to the computer. The good or bad feature extraction will directly affect the subsequent classification or regression results. The specific applications of facial expression recognition are human-computer interaction, sentiment analysis, driver fatigue detection.[1] The dataset used in this study is mainly static expression recognition, i.e., using static images for recognition. In facial emotion recognition, there are two types of approaches: holistic and local-based approaches. Holistic approaches focus on global modeling of human facial deformations, encoding the whole face as a whole. Local approaches concentrate on the observation of eyes, eyebrows, nose, mouth, etc., and their geometric relationships, which are used to build descriptive and expressive models. [2].

Authors' addresses: Youdan Zhang, yzhan143@syr.edu, Syracuse University, 900 South Crouse Ave, Syracuse, New York, USA, 13210; Xin Chen, xchen210@syr.edu, Syracuse University, 900 South Crouse Ave, Syracuse, New York, USA, 13210.

Manuscript submitted to ACM

1.2 Data pre-processing

Fer2013 face expression dataset consists of 35886 face expression images, including 28708 test images, 3589 public and 3589 private validation images, each image is composed of gray scale images with fixed size of 48×48. There are 7 expressions, corresponding to the numerical labels 0-6, and the specific expressions correspond to the following labels: 0 angry; 1 disgusted; 2 fearful; 3 happy; 4 sad; 5 surprised; 6 neutral. Because the training set has been preprocessed with the training and test sets, we do the following work: convert pixels to list, reshape and normalize data (normalize: test/255) create one-hot encording label (fig 2)

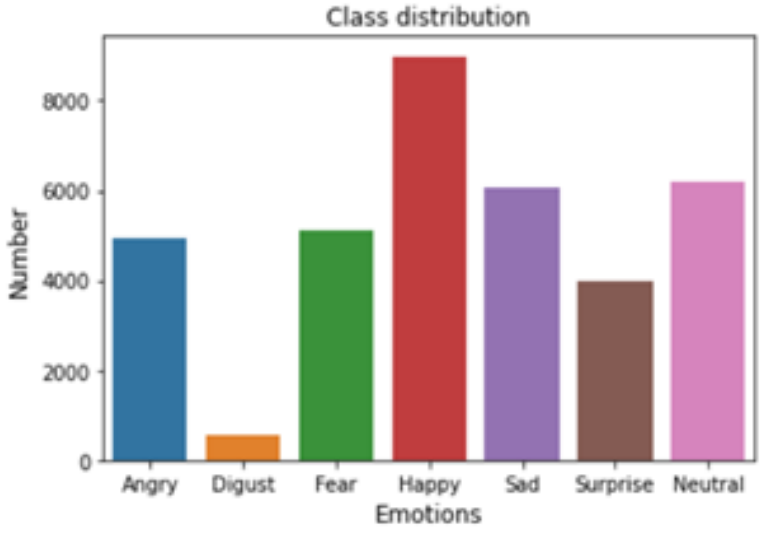


Fig. 2

For the CK+ dataset consists of 981 face expression images, which also correspond to seven expressions sadness corresponds to 87 images; contempt corresponds to 54 images; surprise corresponds to 249 images; disgust corresponds to 177 images; angry corresponds to 135 images fear corresponds to 75 images; happy corresponds to 207 images. The data were first labeled and classified as 0: 'anger', 1: 'contempt', 2: 'disgust', 3: 'fear', 4: 'happy', 5: 'sadness', 6: 'surprise'. Since this dataset is not classified as a training set, we use the train_test_split function to divide it into a train set, a test set, and a valid set. Finally, the data normalization is performed in the same way as fer2013.

2 CNN BASED TRAINING MODELS

2.1 CNN DCNN

A convolutional neural network (CNN) is a feed-forward neural network with artificial neurons that respond to a portion of the surrounding units in the coverage area and excels for large image processing. So we decided to choose CNN model for feature extraction and training. A convolutional neural network consists of these layers: input layer, convolutional layer, ReLU layer, pooling layer and fully connected layer (fully connected layer is the same as in a regular neural network). By stacking these layers together, a complete convolutional neural network can be constructed.

Currently, convolutional neural networks are widely used in FER because of their properties, and we have established a CNN model based on expression recognition with reference to the literature [2]. CNN will also be the base training model for this study. The schematic diagram is shown in Fig 3.

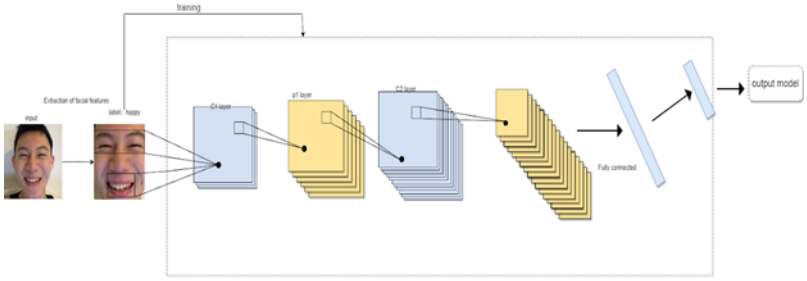


Fig. 3. Schematic diagram

We used Relu function as the activation function, max pooling as the pooling method and 6-layer convolutional layer construction function. We also used two activation functions, Relu and elu, to test the accuracy of the CNN respectively.

Deep Convolutional Neural Network (DCNN) achieves far better results than traditional methods in feature recognition related tasks. So in order to compare the relationship between the number of layers and the accuracy, we also used a deep convolutional neural network and constructed 20 convolutional layers for the calculation, again using Relu as the activation function and maxpooling as the pooling layer The specific results will be shown in Section 3.

2.2 RNN and LSTM

Unlike CNNs, CNNs as feedforward neural networks have no correlation between the output and the model itself. A recurrent neural network (RNN) is a neural network with feedback between the output and the model, while an RNN is a sequence-based neural network that carries over the previous value to the next training, more closely resembling the human thought pattern. A simple RNN schematic is shown in Fig 4.

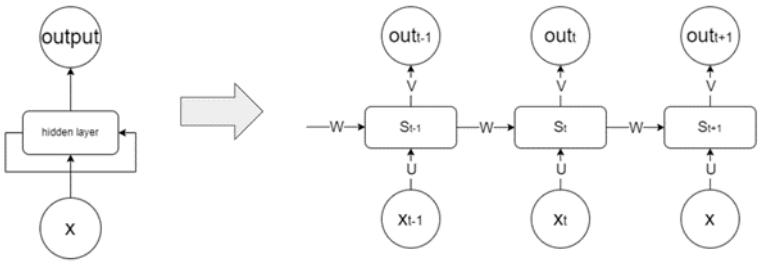


Fig. 4. RNN schematic

For better training results, bidirectional RNN is also an attempt, and bidirectional RNN is very similar to unidirectional RNN, as shown in Fig 5.

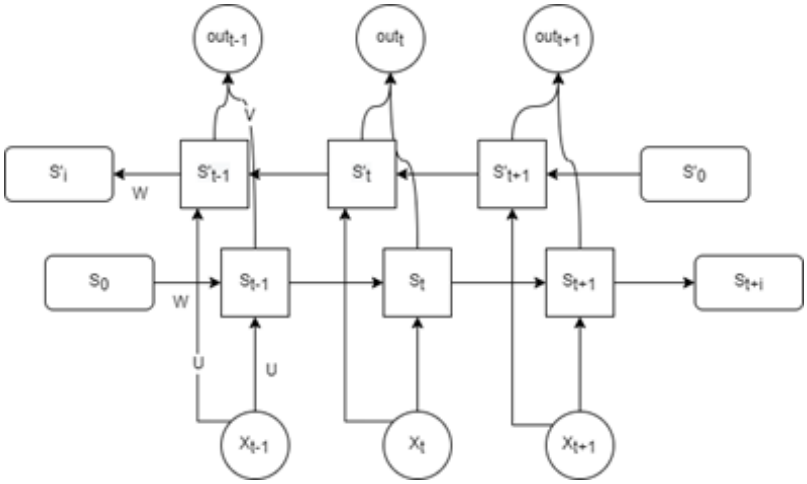


Fig. 5. Bidirectional RNN

However, due to the short memory period of RNN we use long short-term memory (LSTM) and bidirectional LSTM for optimization, a simple LSTM structure as shown in Fig 6.

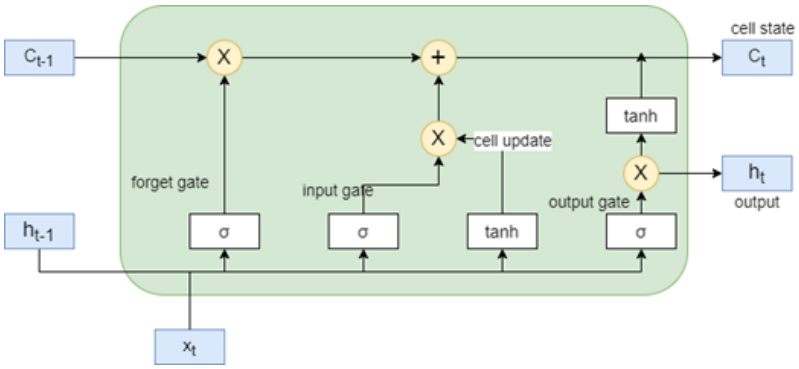


Fig. 6. LSTM structure

The LSTM adds forgetting gates, input gates, and output gates to the RNN. The forgetting gate determines how much of the unit state from the previous moment is kept to the current moment; the input gate determines how much of the input to the network at the current moment is saved to the unit state. The output gate is used to control how much of the cell state is output to the current output value of the LSTM.[3] The bidirectional LSTM is very similar to the bidirectional RNN, only the LSTM is added into the bidirectional RNN, so we will not go into too much detail here.

In using the CNN+LSTM model, we refer to the literature [4] to transfer the results of CNN training of the dataset directly into the LSTM. In the CNN we made new optimizations, first we used dropouts to generalize the data, we used 8 convolutional layers, and we used ELU instead of Relu as the activation function because it avoids the death problem. We tested the LSTM and bidirectional LSTM separately, and the specific results will also be shown in Section 3.

3 EXPERIMENTAL MODEL AND EXPERIMENTAL RESULTS

3.1 CNN

For the CNN we used two activation functions, Relu and elu, there was no obvious difference between these two activation functions. However , elu can avoids the death problem, therefore, we use elu in CNN+LSTM model. The following figure shows the comparison of elu activation function and relu activation function.

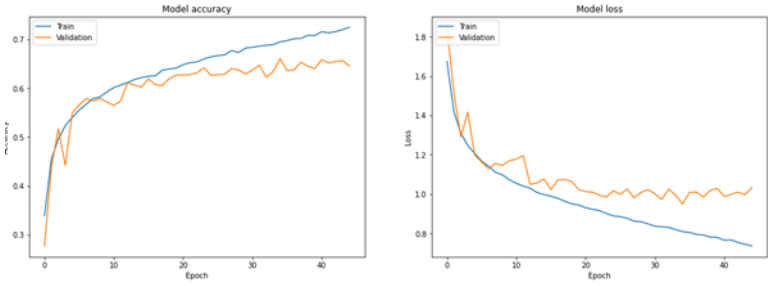


Fig. 7. accuracy under Elu

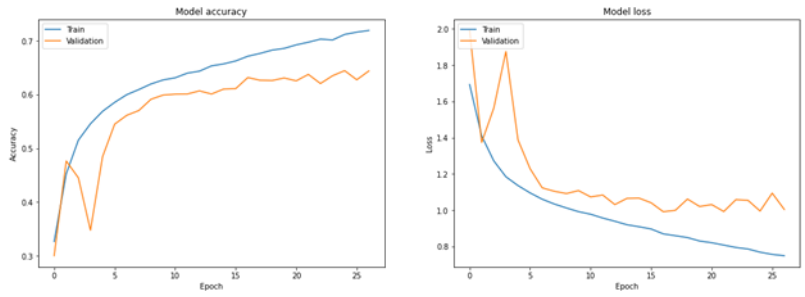


Fig. 8. accuracy under Relu

FER2013 test	Elu	Relu
Training accuracy	0.7247	0.7132
Test accuracy	0.6487	0.6556

Fig. 9. Comparison

3.2 DCNN

For the DCNN we used a 20-layer convolutional neural network with Relu as the activation function, and the final training accuracy for FER2013 was 0.6813 and the test accuracy was 0.6255, which is slightly smaller than the training result of CNN, thus, we believe that the training accuracy and the number of neural network layers do not constitute a proportional relationship.

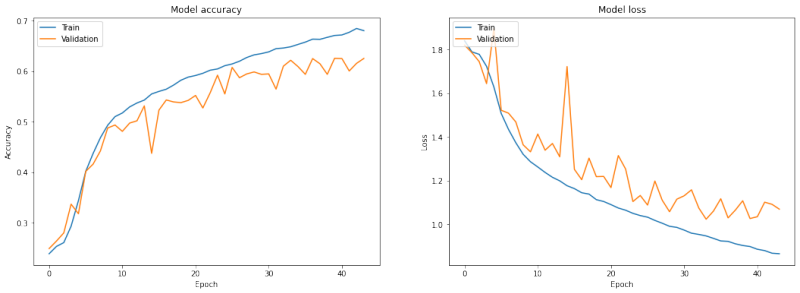


Fig. 10. DCNN accurancy

3.3 CNN+LSTM

Before the test was conducted, we predicted that LSTM should have a great improvement on the accuracy, however, the test results were 0.7032 training accuracy and 0.7313 test accuracy with FER2013, 0.8057 training accuracy and 0.76 test accuracy with CK+ test, for the FER2013 test accuracy was higher than the CNN model, with about 10

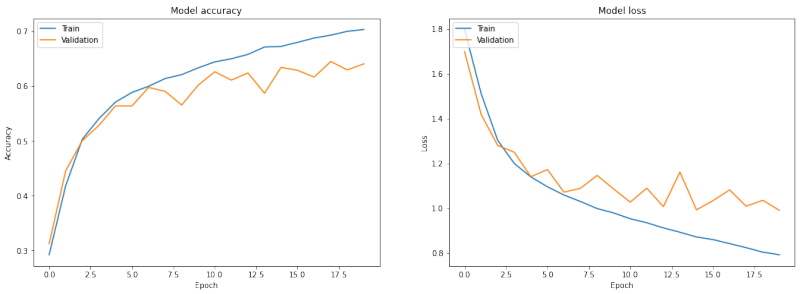


Fig. 11. CNN+LSTM fer2013

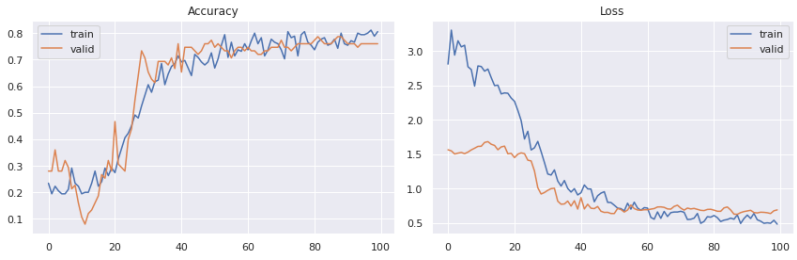


Fig. 12. CNN+LSTM CK+

3.4 CNN + bidirectional LSTM

In CNN+biLSTM, the training accuracy for FER2013 is 0.7602 and the test accuracy is 0.6694, which are both slightly higher than the test results of CNN, and the training accuracy of CNN+biLSTM for CK+ is 0.84 and the test accuracy is

0.7800, which are also higher than the training results of CNN for CK+. This model is the best result in general for all of the models we trained.

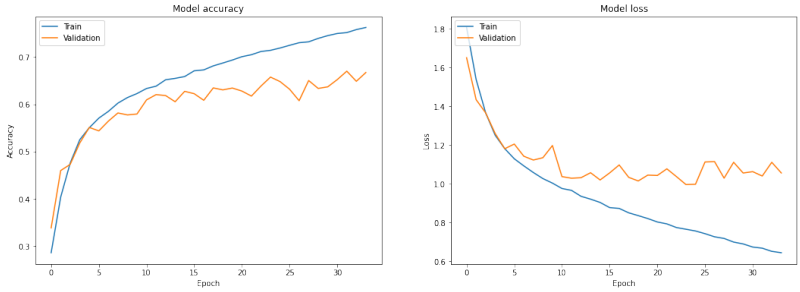


Fig. 13. CNN+biLSTM fer2013

Fig. 14. CNN+biLSTM CK+

4 CONCLUSION

In this study, we constructed and tested four models, which were quantitatively analyzed and studied according to their accuracy. Based on the quantitative results, we have the following conclusions Firstly, the expression recognition accuracy has no proportional relationship with the number of layers of convolutional neural network. Secondly, for different activation functions, there is no effect on the accuracy rate unless to avoid death. Finally, the combination of CNN and LSTM can improve the accuracy to some extent, but the improvement is not huge. We believe that the characteristics of LSTM itself are more suitable for sentiment analysis of text, and the effect is not significant for facial feature analysis. The analysis of facial features is not exactly the same as the "context" of text, because small muscle changes lead to different facial expressions.

For future work, we believe that we could apply different pooling layers to test the accuracy rate, in this study we used maximum pooling as the pooling method. We speculate that average pooling may be able to balance the bias of the training set and thus bring some changes in accuracy. Also, future researchers can try to combine more models. In addition, in the confusion matrix, we found that the difficulty level of recognition is not the same for different facial expressions, which we believe would be a valuable research topic in the future.

The Quantitative analysis and Data Visualization is in the appendix , we used the confusion matrix and line chart as a visualization result. The confusion matrix is one of the most widely used evaluators in multi-class classification.

REFERENCES

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

[2] Giannopoulos, P., Perikos, I., & Hatzilygeroudis, I. (2018). Deep learning approaches for facial emotion recognition: A case study on FER-2013. In Advances in hybridization of intelligent methods (pp. 1-16). Springer, Cham.

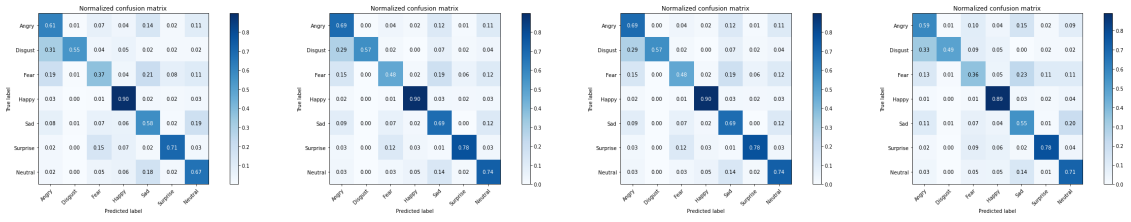
[3] Liu, K., Zhang, M., & Pan, Z. (2016, September). Facial expression recognition with CNN ensemble. In 2016 international conference on cyberworlds (CW) (pp. 163-166). IEEE.

[4] Gers, F. A., Schmidhuber, J., & Cummins, F. (2000). Learning to forget: Continual prediction with LSTM. Neural computation, 12(10), 2451-2471.

[5] Hung, B. T., & Tien, L. M. (2021). Facial expression recognition with CNN-LSTM. In Research in Intelligent and Computing in Engineering (pp. 549-560). Springer, Singapore.

[6] Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010). The Extended Cohn-Kanade Dataset (CK+): A complete expression dataset for action unit and emotion-specified expression. Proceedings of the Third International Workshop on CVPR for Human Communicative Behavior Analysis (CVPR4HB 2010), San Francisco, USA, 94-101

A QUANTITATIVE ANALYSIS AND DATA VISUALIZATION



(a) CNN-FER2013 (b) CNN+LSTM-FER2013 (c) CNN+LSTM-FER2013 (d) CNN+biLSTM-FER2013

Fig. 15. CNN based models with fer2013

(a) fer2013 (b) Ck+48

Fig. 16. Accuracy Comparison Histogram