CrossNet: Lightweight Deep Learning Subnetwork with Embedded Inception Modules for Facial Emotion Recognition

Savina Jassica Colaco[1], Dong Seog Han[*]

[1]Graduate School of Electronic and Electrical Engineering

[*]School of Electronics Engineering

Kyungpook National University, Daegu, Republic of Korea

[1]savinacolaco@knu.ac.kr, [*]dshan@knu.ac.kr

Abstract

Complex patterns on human faces make understanding emotions more challenging for machines than for humans. The automatic recognition of basic emotions is critical for a wide range of applications, including medical imaging, virtual reality, and security. Since recent deep learning approaches are achieving outstanding performance in facial emotion identification, the best performance has yet to be achieved. This paper proposes CrossNet, a lightweight deep learning-based subnetwork for recognising facial emotions using depthwise separable layer with embedded inception modules for better detection of important emotion features. The model is trained with a benchmark dataset called FER2013 to achieve 63.77% classification accuracy.

Ⅰ. Introduction

Human communication is strongly intertwined to emotions. They can be expressed in a variety of ways, which may or may not be easily detected by the human eye. Automatic or real-time emotion detection is now employed in a variety of applications, including human identification, healthcare, virtual reality, cognitive research, and so on [1]. Several systems for recognizing these emotions have been developed. However, as compared to conventional methods, deep learning systems performed better for automatic identification [2]. In this paper, we propose a deep learning model to classify different facial emotions with cross-connected depthwise layers and Inception modules. The Inception module is added to focus on the vital features such as mouth, eyes, *etc*, for better identification of the emotions.

Ⅱ. Experiment & Discussion

1. Model

The proposed model, CrossNet, which is subnetwork of cross-connected depthwise separable layers with inception modules. The depthwise separable layers are followed by a max pooling layer with 2×2 filter to reduce the spatial dimension of the feature map. The model adopts tunable Inception module [3], which consists of different filters such as 1×1, 3×3 and dilation filters to extract features. The different feature extraction from filters helps to focus on the different parts of face images to detect facial emotion patterns. Hence emotions can be detected with important facial features such as mouth, eyes, *etc*. The Inception module is added after every max pooling. The inception module also consists of skip

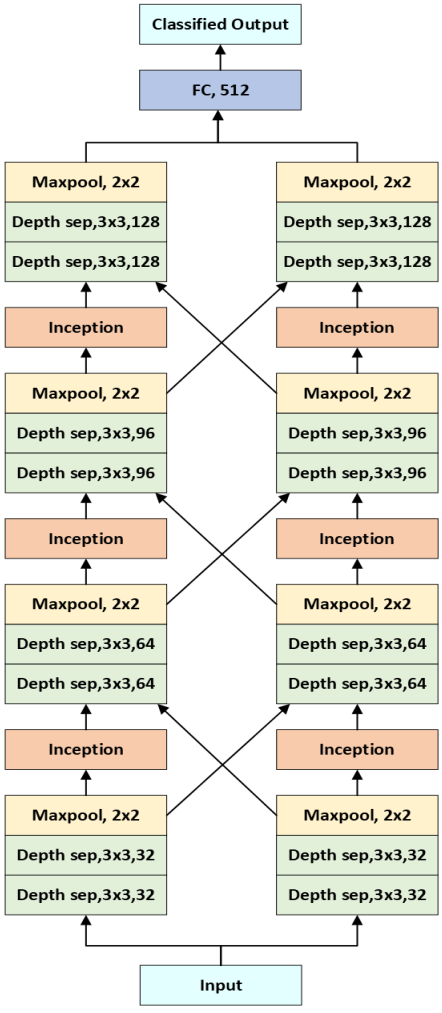

Fig. 1 CrossNet architecture

connections for identity mapping. The output from the max pooling is cross-connected with depthwise separable layers of the subnetwork and concatenated with the Inception module output which is fed as input to the next layer. The fully connected layer will process the class scores, resulting in volume in size, where each of the seven numbers corresponds to a class score. The number of filters results in the same number of feature maps which becomes the parameter for the model to be learnt. The neurons in the fully connected layer are connected to every other neuron in the previous layer. Fig. 1 provides the overall architecture of CrossNet.

2. Experiment and Results

The CrossNet model has been trained with the FER2013 [4] dataset, which consists of 28,709 training images and 7,178 test images with 7 emotion classes. The emotion classes are angry, disgust, fear, happy, sad, surprise, and neutral. The images are resized to 48×48 size. The model is trained with a batch size of 64 and epochs of 100. The Keras framework is used for the model implementation and training. The model is optimized with the Adam optimization technique [5] with a learning rate of 10^{-3}.

The CrossNet model uses the ability of the Inception module to focus on different parts of facial images to find patterns to associate them with label emotion. The inception allows working with different filters to capture the level of abstraction. Hence not being restricted to a single filter size, in a single image block, which is then concatenated and passed onto the next layer. The accuracy plot of the proposed model is shown in Fig. 2. The proposed model was able to achieve 63.77% of validation accuracy with 2.1 million parameters which needs further improvements to close the generation gap between the training and testing accuracy. The addition of dropout avoids the model to overfit. The variation of the emotion could depend on the accuracy of the model. The model stopped increasing or decreasing after a certain epoch which could be learning particular images in the training dataset rather than generalizing the model for better results. Fig. 3 shows the confusion matrix of the model.

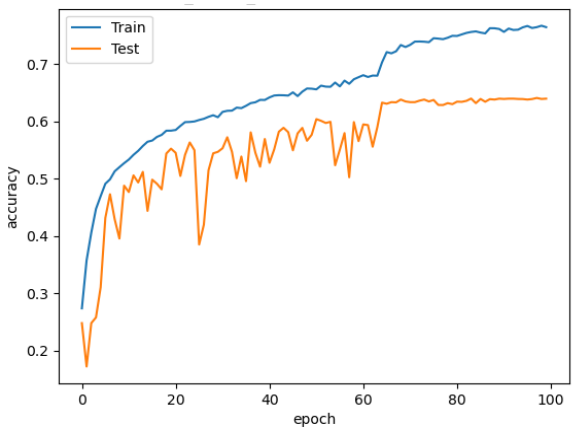

Fig. 2 Accuracy plot of CrossNet

Fig. 3 Confusion matrix of CrossNet

Ⅲ. Conclusion

This paper proposed CrossNet, consisting of depthwise separable layer and embedded Inception module. It allows extracting different levels of features on the face images. The model still needs to improve to get better generalization. The proposed model can be fine-tuned for further improvement with the help of inception modules. Since all features are not important for emotion recognition, concentrating on specific regions to detect basic, as well as complex emotions, can be achieved in the future.

ACKNOWLEDGMENT

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education (2021R1A6A1A03043144).

REFERENCES

[1]B. Fasel and J. Luettin, "Automatic facial expression analysis: a survey," Pattern recognition, vol. 36, pp. 259-275, 2003.

[2]D. K. Jain, P. Shamsolmoali, and P. Sehdev, "Extended deep neural network for facial emotion recognition," Pattern Recognition Letters, vol. 120, pp. 69-74, 2019.

[3]S. Colaco, Y. J. Yoon, and D. S. Han, "UIRNet: Facial Landmarks Detection Model with Symmetric Encoder-Decoder," in 2022 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), 2022, pp. 407-410.

[4]I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, et al., "Challenges in representation learning: A report on three machine learning contests," in International conference on neural information processing, 2013, pp. 117-124.

[5]D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.