CNN-based Model for Face Expression Recognition

Kejun Guo [1,*,†], Shizhe Song [2,†], Qijia Yang [3,†]

[1] Shandong University of Science and Technology, Taian, China

[2] Central South University, Changsha, China

[3] Baofeng County First Senior High School, Pingding shan, China

* Corresponding Author Email: 202023040204@sdust.edu.cn

[†] These authors contributed equally.

Abstract. Face recognition is a biometric technique that uses data on facial features to identify individuals. It is also a key area of study for computer vision researchers. CNN is a subclass of feedforward neural networks with convolutional processing and depth structure and one of the illustrative deep learning techniques. Since the deep learning theory was put forth and computational power increased, CNN has rapidly advanced and is now utilized in computer vision, natural language processing, and other fields. Our research is focused on face recognition, and because the mini-Xception model has a condensed volume and few parameters, it is used in this study. The dataset we used is fer2013, which is a classical dataset among CNN algorithms and is used in many studies. We also used data augmentation methods, and Keras' ImageDataGenerator image generator was the optimal data augmentation method we came up with after reading the paper. Finally, we came up with a final model with 61% accuracy, which we are satisfied with and within the error results of the papers we reviewed.

Keywords: Emotion Recognition, CNN, Xception, ImageDataGeneration.

1. Introduction

The face is a powerful indication of the mentality since actions speak louder than words. One of the most active fields of computer vision research right now is human behavior analysis. The most important factors in deciphering human behavior are body language, hand gestures, and facial expressions, which are equally important in the semantic analysis of multimedia content [1]. Humans are able to guess another person's emotions by seeing the spatiotemporal fluctuations in their face muscles [2]. The 43 facial muscles can create 10,000 different facial expressions, based on Paul Ekman's research. Facial expression recognition (FER) has several applications in computer vision and human-computer interaction. It can be used specifically in monitoring driver weariness, home medical robots, evaluating students' emotions in class, and other applications. Along with the fields mentioned above, there are increasingly requirements for facial expression recognition technique.

Facial expression recognition generally aims to categorize facial emotions into seven groups, including sad, angry, joyful, surprised, fearful, disgusting, and neutral. Facial expression recognition in photographs is a difficult task that has drawn increasing attention [3]. In some cases, this work is made more difficult by the backgrounds and the low-resolution faces.

Numerous studies in this area have been conducted over the last ten years. Generally speaking, there are two types of methodologies that have been tried: standard methods and deep learning methods. High recognition rates can be attained using conventional feature extraction and classification methods, but doing so is computationally difficult and difficult. Convolutional Neural Networks (CNN), on the other hand, have gained significant experimental and practical advancement as deep learning has grown in popularity. In the recent years, deep learning techniques, particularly CNN, have taken over the field of facial expression detection. As a result, traditional techniques won't be covered in great length in the following; instead, we'll concentrate on CNN's use in this area.

Convolutional neural networks have evolved into the de facto paradigm for computer vision in recent years. The first iterations of convolutional neural network design were LeNet-style models, which were simple stacks of convolutions for feature extraction and max-pooling operations for

spatial subsampling in 1995 [4]. Convolution operations were regularly performed in between max-pooling processes in the AlexNet design [3] from 2012, allowing the network to acquire richer information at all spatial scales. This type of network significantly deepened during the ensuing years, largely as a result of the yearly ILSVRC competition.

A brand-new network architecture called Inception was released in 2014 as GoogleLeNet by Szegedy et al (Inception V1), which has several convolutional layers and is novel since it uses Inception blocks [5]. Each inception block contains a number of convolution layers that are independently connected to one another. At the final step, the concatenated results of the convolution layers are passed on to the following convolution layer outside the inception block. In the 2014 ImageNet Large-Scale Visual Recognition Challenge, their revolutionary deep convolutional neural network architecture sets a new standard for classification and detection and performs admirably (ILSVRC14). Inception-ResNet first appeared in 2015, and it was later improved as Inception V3 [6, 7]. Network-In-Network architecture [8], an older game, served as an inspiration for Inception. Inception has continuously ranked among the best model families on the ImageNet dataset [9], and other internal Google datasets, especially JFT [10], since its initial release. The Xception convolutional neural network design, which is just made up of depth-wise separable convolution layers, was then published by F. Cholet in 2017. In two significant picture classification tasks, Xception performs better than Inception V3 while using a same number of parameters [11]. In 2020, A. Fatima put into practice the Mini-Xception model, an improved Xception architecture that makes use of residual networks for emotion expression and recognition. The Mini-Emotion Xception has a 95.60 percent expression and identification accuracy score. The accuracy and recall rates are 93% and 90%, respectively [12].

The correctness of this system can still be promoted, although there are still flaws in it. Poor facial expression recognition outcomes are caused by the gradient's propensity to vanish throughout the training process and the difficulty of initializing the network parameters. By modifying the learning rate, batch size, optimizer settings, and other hyperparameters, as well as by optimizing the size of the number of hidden layers, we hope to improve the model. Accuracy can be raised by increasing the size of the filter and the quantity of convolutional layers. Additionally, we use Keras to expand the dataset and perform operations like rotating the images. It will reduce overfitting and improve the model's capacity for generalization. Finally, to visualize our findings, we create a confusion matrix with Matplotlib and a user interface with PyQt5.

2. Data Sets and Models

To begin with, we utilized the fer2013 dataset [13]. This dataset contains 28, 709 preparing tests, 3, 859 approval datasets, and 3,859 test tests, a add up to of 5887 pictures containing seven categories: irate, nauseated, dreadful, cheerful, pitiful, astounded, and ordinary, with a picture determination of 4848. Figure 1 presents a part of sample data. Most of the pictures in this dataset have revolutions in both planes and non-planes, and numerous of them have occlusions of hands, hair, scarves, etc. occlusions. This database is from the 2013 Kaggle competition, and there's a certain sum of blunder as this database is for the most part downloaded from web crawlers. The human precision of this database is 65–55%. Since the FER2013 dataset is more total in terms of information conjointly more in line with real-life scenarios, FER2013 is primarily chosen here to prepare and test the demonstrate. In arrange to avoid the arrange from overfitting as well rapidly, a few picture changes, such as flipping, turning, cutting, etc., can be done misleadingly. The over operations are called information improvement. We moreover chosen to embrace information expansion after seeing the strategy and particular execution code in a few papers, which is able not as it were provide us a better redress rate but too anticipate overfitting.

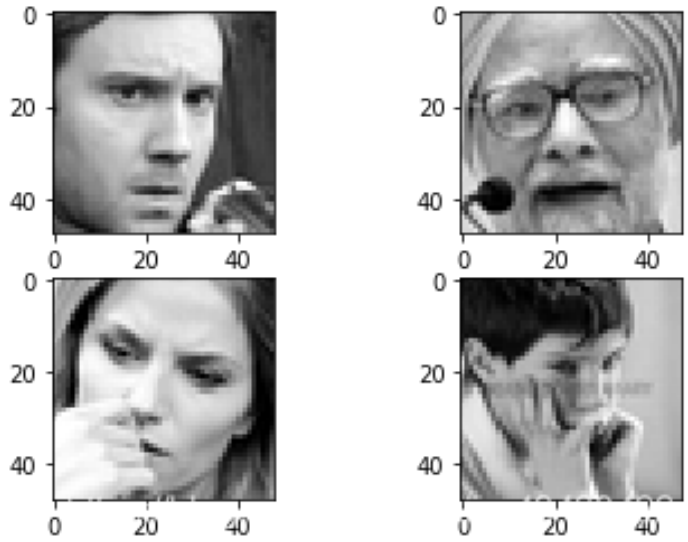

Figure 1. Sample images on the collected dataset.

Following, we chosen to utilize Xception, as it were since our PC is little sufficient to utilize mini-Xception. XCEPTION is another advancement to Initiation v3 proposed by Google after Initiation, primarily by utilizing depthwise distinguishable. The organize structure of XCEPTION somewhat beats Beginning v3 on the ImageNet dataset (the plan arrangement target of Beginning v3) and essentially outflanks Beginning v3 on picture classification datasets containing 350 million pictures or indeed bigger. Beginning v3 keeps up the same number of parameters, with execution picks up coming from more proficient utilize of show parameters.

We utilize the xception show with a depth-separable convolution, which may be a profound convolution on the separable convolution. For case, within the initiation may be a distinct convolution, and the beginning that presents the depth-separable convolution is xception. For depth-separable convolution, the operation is to to begin with perform the standard convolution on the highlight outline of each channel The operation of deep separable convolution is to to begin with perform the standard convolution on the include outline of each channel, and after that utilize the 1×1 convolution to intertwine the data of these channels, which can diminish the number to operations.

Essentially a hybrid mapping of channel correlation and spatial correlation, the basic convolution process. Think of the convolution kernel as a three-dimensional filter that has both a spatial and a channel dimension (width and height of the Feature Map). The Inception module is based on the idea that the spatial and temporal correlations of convolutional channels can be decoupled and separated to get superior results. The components of the Inception architecture are shown in Figures 2, 3, 4, and 5.

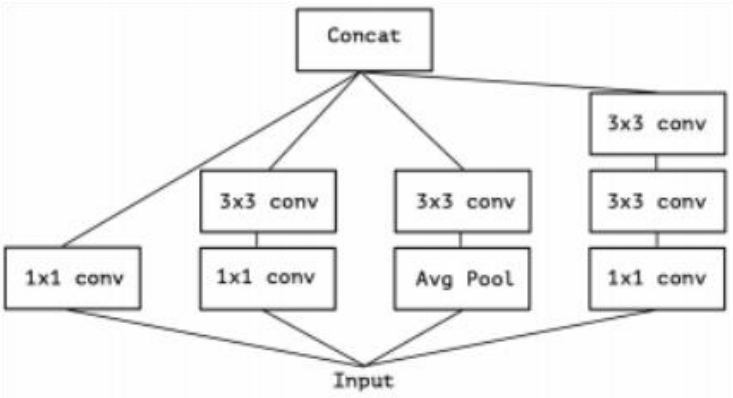

Figure 2. One module Canonical Inception (Inception V3)

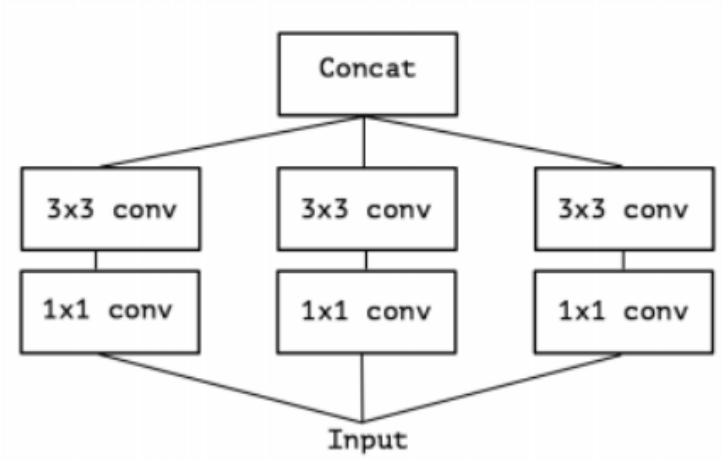

Figure 3. A Simplified creation module.

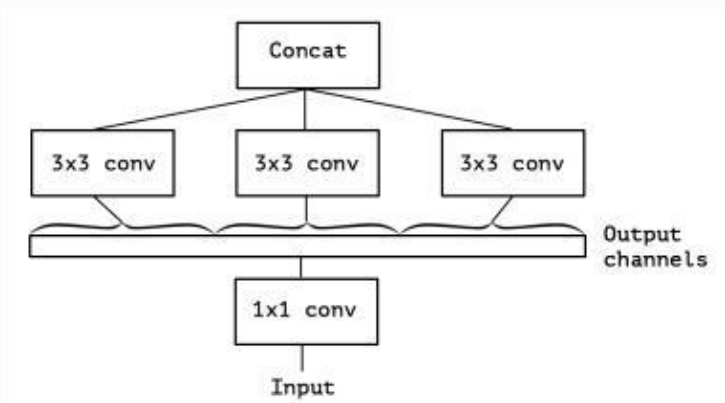

Figure 4. Strictly equivalent re-wording of the Streamlined Inception module.

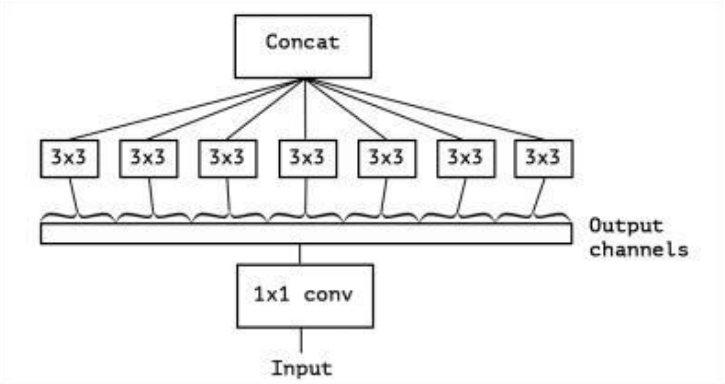

Figure 5. An "extreme" version of our Inception module, with a spatial convolution through the 1 1 convolution output channel.

The structure of Inception, called Xception (Extreme Inception), which introduces a deep separable convolution, is shown below, in which SeparableConv is the deep separable convolution, replacing the convolution layer in Inception with a deep separable convolution. In addition, it can be seen that each module is connected using residuals (except for a few at the beginning and end).

Then we further process the data, that is, data augmentation. The Kaggle Fer2013 data set is only 30,000 records, and there are a lot of blocking, angle, and other external factors. We employ technology to address this issue because collecting data requires a lot of effort and resources To avoid repeated development, first see if there is a well-written library. We searched for relevant papers and found through them that a means of data enhancement could be ImageDataGenerator's image generator.

We looked at the Keras official documentation and came to the conclusion that ImageDataGenerator is an image generator that can also improve data in a batch, increase the size of the dataset (for things like rotation, deformation, normalization, etc.), and improve the model's capacity for generalization.

The parameters we used to train the model are as follows, firstly batch_size is 48, secondly num_epochs is 1000. input_shape is (24,24,1), then validation_split is 0.2. finally verbose is 1, num_classes is 7, patience is 50.

3. Results and Discussion

The results obtained by this study were very successful, first of all we obtained a visualization of the results, which is partially presented by the confusion matrix, which means that our model was trained successfully and with a good accuracy rate of 61%, which is satisfactory for us, although we still have some improvements compared to others' models, but it is certainly a good start. It took a total of 4 days to train the model, and we ran it on our own PC, but then we realized that we could rent a server on our website and run the program much faster, which would not only save a lot of time, but also allow us to use that time to improve our model, which is certainly a huge mistake in this study, and we will avoid it in future studies. Figure 6 shows the normalized confusion matrix based on the result.

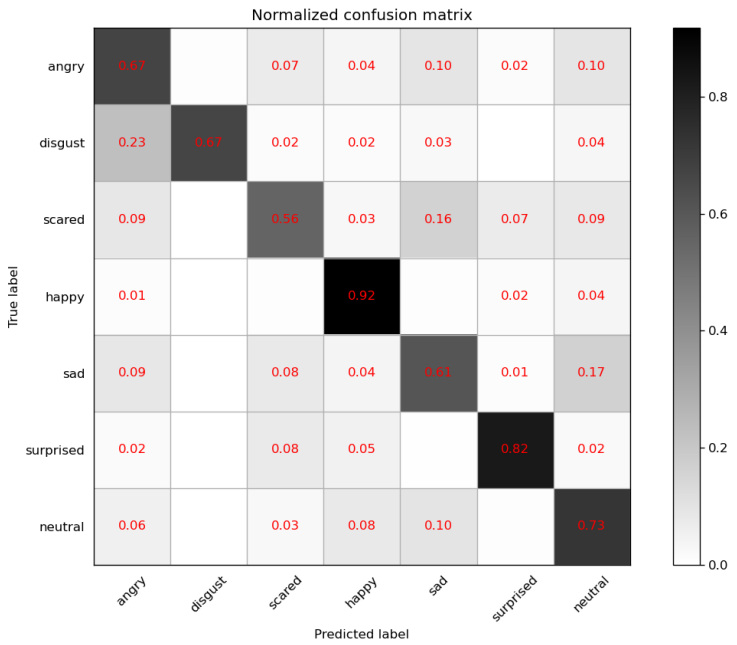

Figure 6. The normalized confusion matrix based on the result.

This is the visualization result of the model. We found that the cross part of several expressions were not shown, and we initially judged that the model did not recognize this as a cross expression, but we did not have the time to check carefully because of the time, we will try our best to avoid this situation in the future and make the model recognize every expression.

4. Conclusions

Based on convolutional neural network theory, this paper adopts Mini xception model and data set fer2013 and adopts data enhancement method. Keras' imagedatagenerator's image generator is the best data enhancement method we have obtained after consulting the data. Finally, we get the final model with an accuracy of 61%. Using better servers to process larger data sets can shorten the training cycle and achieve higher accuracy. In the future research, we will continue to study

convolutional neural network theory in depth and realize visual results in combination with programming language.

References

[1] Ijjina E P. Facial Expression Recognition Using Kinect Depth Sensor and Convolutional Neural Networks [C], 2014 13th International Conference on Machine Learning and Applications, 2014, pp. 392-396

[2] Wang Z. Capturing complex spatiotemporal relations among facial muscles for facial expression recognition [C], in 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2013, pp.3422-3429.

[3] Krizhevsky A. Imagenet classification with deep convolutional neural networks [J], in:Advances in neural information processing systems.2012, pp.1097-1105.

[4] LeCun Y et al. Learning algorithms for classification: A comparison on handwritten digit recognition [J]. Neural networks: the statistical mechanics perspective, 261:276, 1995.

[5] Szegedy C. Going deeper with convolutions [C], in:Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2015,pp.19.

[6] Szegedy C. Rethinking the inception architecture for computer vision [R]. arXiv preprint arXiv:1512.00567, 2015.

[7] Szegedy C. Inception-v4, inception-resnet and the impact of residual connections on learning [R]. arXiv preprint arXiv:1602.07261, 2016

[8] Lin M. Network in network [R]. arXiv preprint arXiv:1312.4400, 2013.

[9] Russakovsk O et al. Imagenet large scale visual recognition challenge [J]. International journal of computer vision 115.3 (2015): 211-252.

[10] Hinton G et al. Distilling the knowledge in a neural network [R]. arXiv preprint arXiv:1503.02531 2.7 (2015).

[11] Chollet F. Xception: Deep Learning with Depthwise Separable Convolutions [C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1800-1807

[12] Fatima S A. Real time emotion detection of humans using mini-Xception algorithm [C]. IOP Conference Series: Materials Science and Engineering. Vol. 1042. No. 1. IOP Publishing, 2021.

[13] Jain D K Extended deep neural network for facial emotion recognition [J]. Pattern Recognition Letters, 2019, 120: 69-74.