Facial Expression Recognition Using Convolutional Neural Network

Yijun Gan
Ball State University
2217 w bethel Ave apt#131
Muncie, Indiana, United States
+1 7657604761
ganyijun0331@gmail.com

ABSTRACT

Facial expressions are part of human language and are often used to convey emotions. With the development of human-computer interaction technology, people pay more and more attention to facial expression recognition (FER) technology. Besides, in the domain of FER, human beings have made some progress. In this paper, we reviewed the development of FER: VGGNet, ResNet, GoogleNet, and AlexNet. Besides, we studied some ideas of CNN (Convolutional Neural Network), and we used FER2013, which is one of the most significant databases of human faces, as the dataset to be considered. Furthermore, we made some improvements based on the original methods of FER. By training the FER2013 dataset with different revised ways, the best result of accuracy we got is 0.6424. At last, we generated and summarized the progress and deficiencies in this study.

Keywords

Facial Expression Recognition; CNN; FER2013.

1. INTRODUCTION

Faces are much more than keys to individual identity. Furthermore, facial expression, in other words, is the direct way for people to express the intuitive sense. The fundamental purpose of Facial Expression Recognition (FER) is to classify a facial emotion into several types of emotion: joy, fear, sadness, disgust, anger, surprise and so on. Since it is hard for human beings to detect the nuance of facial expression from others, the machine becomes a significant medium to capture the individual facial emotion from face-to-face interaction. Due to the critical role of facial expressions in human interaction, the ability to perform FER automatically via computer vision enables a range of novel applications in fields such as human-computer interaction and data analytics [1].

Due to the high accuracy of face detection, this technology has been used in many industries such as medicine, e-learning, monitoring, entertainment, law, and marketing. For example, in building security, a face recognizer could be used at the front entrance for automatic access control. They could be used to enhance the security of user authentication in ATMs by recognizing faces as well as requiring passwords [3]. The initial idea of face detection is to find the faces in a picture, and localize them with bounding boxes, as illustrated in Figure 1.

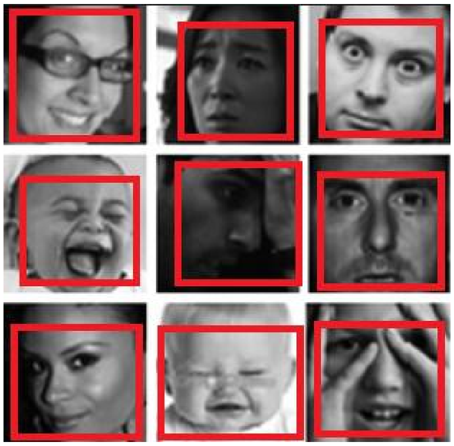

Figure 1. Human face detection. A red bounding box marks all the people in this image. This picture is from the FER2013 dataset.

Face recognition, which is an identification technology based on human facial information, is a further step of face detection. Face recognition consists of several steps: face image detection, face image preprocessing, facial feature extraction and face image matching and identification. The system always needs to input a series of face images which is unknown identity, then output a set of similarity scores that indicate the identity of the human.

For instance, to recognize the one who is on the blacklist of crimination, machines have significant effects on face identification since human cannot remember all the faces of criminals. Face detection, which is the primary step of facial expression recognition, can be solved by many algorithms. The first type aims to convert two-dimensional human faces to one-dimensional features using statistical methods. Eigenfaces and adaboosting fall into this category. Another one is to extract the facial features and sent to a classifier for facial expression recognition. The typical example is color-based skin detection.

This paper mainly focuses on facial expression recognition. There are two main types of feature extraction: feature

ICVISP 2018, August 27–29, 2018, Las Vegas, NV, USA
© 2018 Association for Computing Machinery.
ACM 978-1-4503-6529-1/18/08…$15.00
https://doi.org/10.1145/3271553.3271584

extraction of geometric feature and a method based on overall statistical characteristics, as shown in Figure 2.

Facial recognition is no longer a technical issue today, but computers still cannot quickly figure out human expression. If computers can understand the human emotion, they will offer a better service to humans. It has some main steps: image acquisition, image preprocessing, feature extraction and classification. In this paper, we suppose that the first two steps have completed, to focus on using CNN (Convolutional Neural Network) in facial expression recognition.

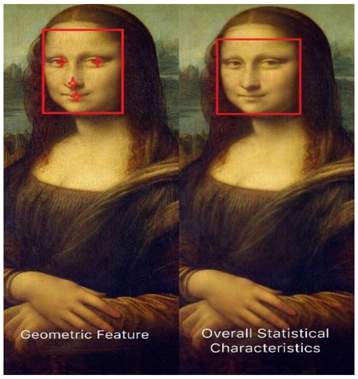

Figure 2. Mona Lisa. Left side: a bounding box marks the woman's face, and the five sense organs are marked as well. The method of geometrical features can recognize it. Right side: a bounding box marks the woman's face, and it can be identified by the process based on overall statistical characteristics. We can attempt to use facial expression recognition to deal with the problem which puzzles human beings for thousands of years: Is the woman in this painting smiling?

Moreover, CNN and DNN (Deep Convolutional Neural Network), which are the feature-based approach for face recognition, are popular algorithms in recent year. There is a convolutional layer, pooling layer, fully connected layer and output layer in CNN model. There are some other nonlinear activation functions used in CNN such as Sigmoid, Tanh and so on. In DNN model, it has an input layer, an output layer, and many hidden layers.

With the rise of face recognition in the past few years, there are increasing datasets established for individuals to use, just like Kaggle, ORL, and FERET. Kaggle provides a platform for developers and scientists to host some machine learning contests or to share code. In this paper, we will focus on Kaggle dataset.

In this work, we mainly focused on applying CNN to solve the FER problem. We used different architectures such as VGG16, ResNet, and GoogLeNet to facial expression recognition. The organization of this paper is as follows. First, the overview of related works is described in Section 2. Then, the detailed methods are introduced in Section 3. The exact experimental results are reported in Section 4. Finally, we made conclusions in Section 5.

2. RELATED WORK

In the process of facial expression feature extraction, it is required to accurately extract the features of facial expression in the image of human expression, since the accuracy of facial expression will impact the results of the classification of facial expression, which is the subsequent step of the extraction of facial expression.

Though there are plenty of researches in the history and development of face recognition, we find that there are several types of approaches that can be used in face recognition such as face recognition based on geometrical feature points extraction and eigenface.

The method of geometrical features based is generally used to extract the location of facial organs as the characteristics of classification. Roberto Brunelli and Tomaso Poggio mentioned that the idea is to obtain relative position and other parameters of distinctive features such as eyes, mouth, nose, and chin [8]. This function is old and dull, but there are still two main weaknesses in geometrical features based function: the first one is that in energy function, the weighting coefficients, which is difficult to summarize, can only be determined by experience. Another disadvantage is the process of optimization of energy function is time-consuming. On the side, the detection technology of feature point cannot be precisive, and the computation is expensive as well.

Eigenface, which is introduced initially by Matthew Turk and Alex Pentland in 1991, has become one of the most popular algorithms during these years. It is known as simple and effective. A simple approach to extracting the information contained in an image of a face is to somehow capture the variation in a collection of face images, independent of any judgment of features, and use this information to encode and compare individual face images [5].

In other words, the main idea of eigenface is to transform a face image from pixel space to another space, then do a calculation of similarity in another space. The eigenface uses PCA function to get the composition of human faces. Otherwise, in the training set, we desire to obtain the eigenvalue decomposition of the covariance matrix of human faces' images and the corresponding eigenvectors. All the eigenvectors which we achieve through computation are called feature faces.

CNN is a profound learning method which is developed for image classification and recognition. Also, it has been widely used in the field of FER nowadays. CNN, as a deep learning architecture, can reduce the complexity of the model and accurately extract the image features [7].

3. METHODS

The original structure contains six steps which are input image, training data, template library, feature extraction, comparison and output result, as shown in Figure 3. However, a simplified structure that uses in this paper only has four steps after we combine the step of template library, feature extraction and comparison to facial expression recognition, as shown in Figure 4. It will greatly increase the efficiency and reduce the running time.

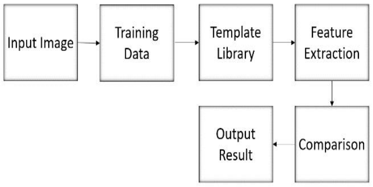

Figure 3. The original schema of facial expression recognition.

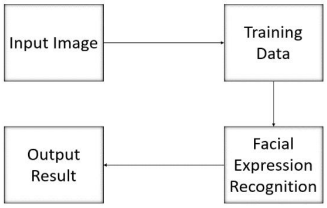

Figure 4. The approach of facial expression recognition that we used in this paper. We simplified the original method. This flow diagram shows only four steps: input image – training data - facial expression recognition - output result.

In the evolution of CNN network structure, there have been many excellent CNN networks such as AlexNet, Vgg, GoogLenet, Resnet, and Densenet.

3.1 AlexNet

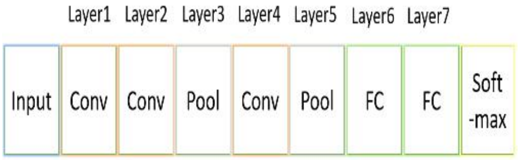

Figure 5. The structure of Alexnet.

As shown in Figure 5, Alexnet was trained on ImageNet, which is a dataset that contains more than 1.2 million images for classification [9]. AlexNet has five convolutional layers and three fully connected layers, which proves the validity of CNN in the complex model. Besides, it will output the 1000*1 vector, and transfer the SoftMax classifier of 1000 classes, and then the results of classification are obtained. In contrast to the earliest traditional method LeNet, AlexNet has some innovations. For instance, AlexNet uses ReLu as its activation function, and now ReLu is widely used in various CNN structures. AlexNet not only establishes the dominance of deep learning in computer science but also has promoted the development of deep learning in the field of facial expression recognition. In our paper, we modified the SoftMax layer to output seven classes.

3.2 VGG-net

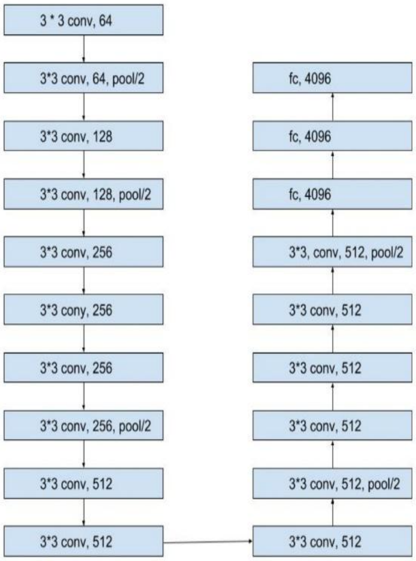

Figure 6. The structure of VGG-net.

As shown in Figure 6, a series of VGG-net have the same structure at the last three levels of the full connection layer, and the overall structure contains five sets of convolutional layers, which are followed by the Max pool layer. VGG-net has improved based on the previous architectures. VGG-net uses 3*3 filters throughout the convolutional layers of the network. Each convolutional layer was followed by a ReLU activation function [2]. In the process of training, VGG-net first trains the simple system of level A; then it reuses the weight of level A's network to initialize several complex models later, to accelerate convergence. Thirdly, to introduce the convolutional kernel of 1*1 to reduce the computation in the convolutional structure of VGG-net.

3.3 GoogleNet

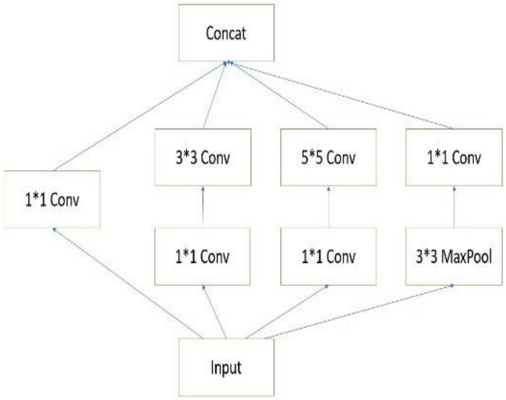

Figure 7. The Inception Structure of GoogleNet. GoogleNet is compositing by many inception structures.

As shown in Figure 7, one significant characteristic of GoogLeNet is that it is designed very deep. The optimization methods which GoogLeNet used are worthy of further study. For example, GoogLeNet has adopted a modular approach to standardizing the results, which makes it easier to modify. Besides, the average pooling was used to replace the whole connectional layer at the end, which makes the rate of success increased by 0.6.

3.4 Resnet

ResNet was proposed in 2015 and won the first place in the competition of ImageNet. ResNet with hundreds or even thousands of layers have become the most successful image recognition model in the computer vision community [4].

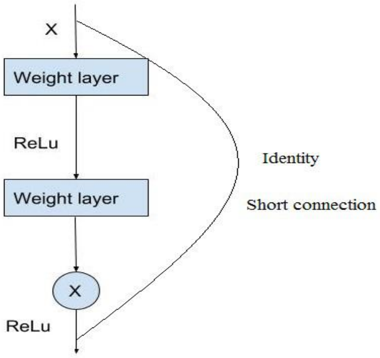

Figure 8. The thumbnail of ResNet.

The structure of ResNet is shown in Figure 8. The main idea of this function is that the input data will pass a layer with a small output of 1*1 initially, then it will move to a layer of 3*3, and then using a layer of 1*1 to handle a more significant number of features. Moreover, compared with the traditional CNN such as VGG, ResNet reduces the required parameters. Besides, ResNet can be more in-depth, without the problem of gradient dispersion. ResNet was regarded as a vast improvement at that time. This algorithm will be implemented and proven in the subsequent articles.

4. EXPERIMENTAL RESULT

4.1 Dataset

The dataset we use is called emotion recognition dataset (FER2013) which is published in Kaggle. It is a dataset about human expressions, and it contains 35887 different face images with expressions belongs to seven categories. FER2013 was published in International Conference on Machine Learning. The usage of this dataset is divided into the private test, public test, and training, and the number of training dataset is 28709. We used different methods such as AlexNet, GoogLeNet, VGGNet, and ResNet to train the model and the results are shown in the next section.

4.2 Results

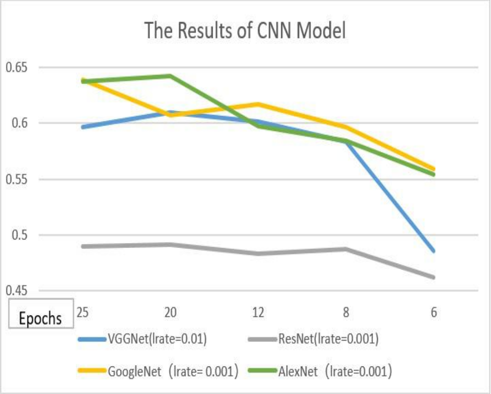

Figure 9. The results of CNN model. We made a curve chart based on the improved CNN models (VGGNet, ResNet, GoogleNet, and AlexNet) so that it can be perceived as merely the changes of precisions.

As shown in Figure 9, we choose five epochs: 25, 20, 12, 8 and 6. The best result of accuracy on FER2013 in machine learning competitions is 0.7116 which is computed by Yichuan Tang [6]. Overall, except the accuracy of ResNet, the average of other methods is above 0.55 to 0.6, which indicates that the results of our methods are good. Besides, in the period of 25 to 20, the accuracy of VGGNet and AlexNet is on the rise, and the accuracy is on the peak when the epochs equal 20. Overfitting can be one of the reasons. In other words, the final accuracy cannot be determined merely by the number of epochs. Also, the overall results of ResNet are not very well. Although ResNet is mostly a collection of shallow networks, ResNet can achieve state of the art performance on many other tasks. Besides, the possible reason is that we do not have enough data to train a reasonable ResNet model. In the later work, we can make some data by flipping and cut the same image for FER to increase the size of the dataset.

Table 1. The results of accuracy using improved models.

Epochs	VGG lrate= 0.01	ResNet lrate= 0.01	Google Net lrate= 0.001	AlexNet lrate=0.001
25	0.5964	0.49	0.6391	0.6374
20	0.6098	0.4916	0.6073	0.6424
12	0.6017	0.4833	0.6171	0.5978
8	0.5836	0.4869	0.5966	0.5844
6	0.4858	0.4621	0.5594	0.5546

With different methods, the highest-level data accuracy of AlexNet, GoogleNet, and VGGNet is above 0.6. Besides, the best accuracy is around 65% of all the previous results of FER2013. There are many reasons why the accuracy is hard to increase. For instance, missing labels in training sets, containing noises in datasets and overfitting can reduce the accuracy. Namely, the same object can look vastly different. As human beings, we can identify whether the items are the same although the status is changed. However, sometimes machines cannot. All the factors can impact the result. In our paper, we get 0.6424 by using AlexNet (epochs =20 and lrate =0.001) as our best result which is close to the best results of FER2013. In general, it is an optimistic data.

5. CONCLUSION

In this paper, we applied CNN with four different architectures for facial expression recognition. Initially, we studied the basic structures of CNN models, and we improved the accuracy of those CNN models. Also, we calculated and compared the accuracy of each model. The four CNN models are GoogleNet, ResNet, VGGNet and AlexNet and they are examined on the same database which is FER2013. FER2013 is one of the famous datasets. Since the dataset is large, it contains some noises. Generally, the limited results we got are also suitable for other general situations. Finally, we found that AlexNet achieved the best overall accuracy which is 0.6424. The results of our approach show that CNN can produce some useful results on FER.

Since FER is a good method that help human beings in daily life in many field, for the future, we will still work on to improve the accuracy of each CNN models. Besides, we will investigate more factors that could impact the accuracy, such as some bias that could affect the FER2013, and we expect to find the efficient ways to solve the problems.

REFERENCES

[1] A. T. Lopes, E. D. Aguiar, and T. Oliveirasantos. A facial expression recognition system using CNN. In Graphics, Patterns and Images, pages 273–280, 2015.

[2] B. E. Bejnordi, J. Lin, B. Glass, M. Mullooly, G. L. Gierach, M. E. Sherman, N. Karssemeijer, J. V. D. Laak, and A. H. Beck. Deep learning-based assessment of tumor-associated stroma for diagnosing breast cancer in histopathology images. In IEEE International Symposium on Biomedical Imaging, pages 929–932, 2017.

[3] C. F. Bobis, R. C. Gonza´lez, J. Cancelas, I. A´ lvarez, and J. Enguita. Face recognition using binary thresholding for features extraction. In International Conference on Image Analysis and Processing, page 1077, 1999.

[4] H. Li, H. Li, H. Li, H. Li, and H. Li. Does resnet learn good general-purpose features? In International Conference on

Artificial Intelligence, Automation and Control Technologies, page 19, 2017.

[5] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, and D. H. Lee. Challenges in representation learning: A report on three machine learning contests. Neural Netw, 64:59–63, 2015.

[6] M. A. Imran, M. S. U. Miah, and H. Rahman. Face recognition using eigenfaces. Proc Cvpr, 118(5):586–591, 2002.

[7] Shen, Dinggang, Guorong Wu, and Heung-Il Suk. "Deep Learning in Medical Image Analysis." Annual review of biomedical engineering 19 (2017): 221–248. PMC. Web. 25 June 2018.

[8] Y. Tu, S. Li, and M. Wang. Intelligent facial expression recognition system r&c-fer. In Intelligent Control and Automation, 2008. Wcica 2008. World Congress on, pages 2501–2506, 2008.

[9] Y. Zhang, F. Chang, L. I. Nanjun, H. Liu, and Z. Gai. Modified alexnet for dense crowd counting. (cii), 2017.