ATLANTIS PRESS
Advances in Computer Science Research, volume 74
2nd International Conference on Computer Engineering, Information Science & Application Technology (ICCIA 2017)

Facial Expression Recognition Based on Convolution Neural Network

Yue Duan [a], Linli Zhou [b] and Yue Wu [c]
Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei 230031, China.
[a]13285695962@163.com, [b]linlizhou@iim.ac.cn, [c]wuyue@iim.ac.cn

Keywords: Facial expression recognition, convolutional neural networks, deep learning, graphics processing unit, feature extraction.

Abstract. With the popularity of computer technology in people's daily life, facial expression recognition in the human-computer interaction, home entertainment, public safety and even medical applications in the field more and more widely. In recent decades, the rapid development of deep learning areas has brought new opportunities for breakthroughs in various fields. Unlike traditional methods of extracting features manually, researchers can obtain the characteristics of automatic learning and generalization through the method of machine learning. To avoid the complex explicit feature extraction process in traditional expression recognition, a convolutional neural network (CNN) for the facial expression recognition is proposed. Firstly, the facial expression image is normalized and the implicit features are extracted by using the trainable convolution kernel. Then, the maximum pooling is used to reduce the dimensions of the extracted implicit features. Finally, the Soft max classifier is used to classify the facial expressions of the test samples. Experimental results show the performance and the generalization ability of the CNN for facial expression recognition.

1. Introduction

Face expression can express people's subtle emotional reactions and human mental state, play a very important role in people's exchanges. Facial expression recognition technology has attracted much attention as people pay more and more attention to facial expression information, which has become a hot topic. Facial expression recognition is the process of facial expression image acquisition, expression image preprocessing, facial feature extraction and expression classification by computer. It analyzes the person's expression information through the computer, thus infering the person's mental state, and finally to achieve the intelligent interaction between man-machine. The main application fields of facial expression recognition include human computer interaction, security, robot manufacturing, medical treatment, communication and automobile.

In this paper, we design a CNN architecture for facial expression recognition. Firstly, the facial expression image is normalized and the hidden features are extracted by the trained convolution kernel. Then, the maximum pool method is used to reduce the dimension of the extracted feature, and it has a certain rotation and translation invariance. Finally, the Softmax classifier is used to classify the facial expression of the test samples, and the facial expressions are divided into 6 categories: happy, surprise, anger, sadness, disgust and fear.

2. Deep learning and convolutional neural network

Deep learning was presented by Professor Hinton, a professor of machine learning in 2006, at the University of Toronto, Canada. The typical deep learning model has Deep Belief Networks, DBN[1], Stacked Auto-encoder, SAE[2], Convolutional Neural Networks, CNN[6.7], etc. CNN is a kind of deep neural network which contains the volume layer. Its model is initially inspired by the study of cranial nerves, which mimics the process of visual cells in simple cells and complex cells in the visual cortex. Simple cells respond to the edge information from different directions, and the complex cells accumulate the results of similar simple cells, called the Hubel-Wiesel structure[3]. CNN contains a multi-stage Hubel-Wiesel structure. In CNN, sub block in the image (local experience area) as the lowest level of the input information, and then transmitted to different layers, each layer by a digital

filter to obtain the most significant features of observed data. This method can obtain the remarkable characteristics of the translation, zoom and rotation invariant observation data, because the characteristics of the local area of the image feel neurons or processing unit allowed access to the most basic, such as directional edges or corners.

3. CNN structure design

In this paper, we design a CNN structure for facial expression recognition, as shown in figure 1. Does not include the input layer, the network consists of 7 layers. The input layer is a 96 × 96 face pixel matrix. The convolution layer learns the characterization of the input data. Typical operation of the pool layer includes an average of pooling [4] and a maximum of pooling. The connecting layer will roll layer and Pooling layer stack up, it can form a layer or multi-layer connection layer, so it can realize the thrust ability of higher order to classify the total output connection layer, the facial expression into joy, surprise, anger, sadness, disgust and fear 6 class.

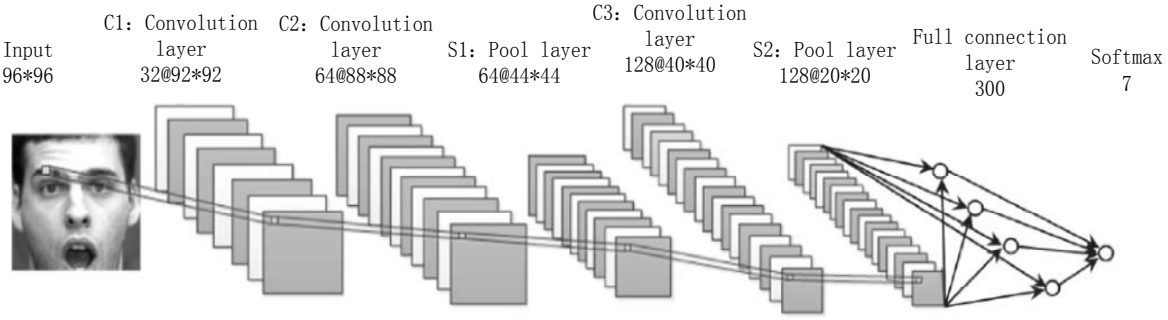

Fig. 1 The CNN structure for facial expression recognition

3.1 Convolution Layer

The statistical characteristics of different sub blocks in natural images usually have consistency, which means that it can be used as the detector from the block to learn a feature image, then traverse the whole image of all sub blocks to obtain other sub blocks activation of the same characteristic value. The convolution in the CNN is the use of the inherent characteristics of the image, with the training of different convolution kernel respectively with a layer of all feature maps of the convolution summation, and biased, then the result is output through the activation function of the formation of the current layer to form a feature map of neurons, different characteristics of the current layer. In this way, the input layer of the upper layer is connected with the feature map of the next layer. The expression of the convolution layer is generally shown in formula (1).

$$y_j^l = \theta(\sum_{i=1}^{N_j^{l-1}} w_{i,j} \otimes x_i^{l-1} + b_j^l), j = 1,2,...,M \quad (1)$$

Where l represents the number of layers; y_j^l represents the j-th feature graph of the current layer; $w_{i,j}$ represents the convolution kernel of the j-th feature map of the current layer and the i-th eigenvector of the previous layer; x_i^{l-1} represents the i-th eigenvector of the previous layer; b_j^l represents the offset of the j-th feature graph; M represents the number of the current feature map.

The convolution layer is a feature extraction layer, and the input of each neuron is connected with the local receptive field (image sub block) of the previous layer, and the local feature is extracted. The convolution layer C1 uses 5 * 5 convolution to check the operation of the input image of 96 * 96 pixels, each neuron specifies a 5 x 5 local receptive fields, so the size (96-5 + 1) x (96-5 + 1) = 92 x 92 feature map is obtained after the convolution operation. Through the convolution operation of 32 different convolution kernels, 32 feature maps are obtained, which are extracted from 32 different local expression features. Each neuron of the same feature map shares the weights (using the same convolution kernel), but they receive input from different local receptive fields. The convolution layer C2 uses 64 * 5 * 5 convolution kernels to convolution the feature graph of the convolution C1 output, 64 feature maps are obtained, each feature map size (92-5 + 1) x (92-5 + 1) = 88 x 88. The

convolution layer C3 uses 128 * 5 * 5 convolution kernels to convolution the feature graph of the pool layer S1 output, 128 feature maps are obtained, each feature map size (44-5 + 1) x (44-5 + 1) = 40x 40.

3.2 Pool Layer (Down Sampling).

The convolution layer is followed by the down sampling operation. The number of input feature layers is consistent with the number of feature layers. But the output of the feature map size will become smaller. The lower sampling formula is generally shown in formula (2).

$$y_j^l = \theta(\beta_j^l down(y_j^{l-1}) + b_j^l) \tag{2}$$

Where down () represents the sampling function. There are many sampling functions, including maximum sampling and mean sampling. Both the maximum sampling or mean sampling, are the first to obtain a pixel in the sampling area (such as the size of the n*n region) , and then find the maximum value, or calculate the sum of the selected sampling area, and then calculate the arithmetic average, and finally the sampling area of the maximum or mean as a sample output. β and B are the parameters of each output feature map.

The pooling layer S1 is obtained by down-sampling a 2 × 2 window for the feature plot of the convolution layer C2, so the resulting feature plot size is 44 × 44. Sampling does not change the number of feature graphs, so the number of feature graphs is still 64. Similarly, the pooling layer S2 uses a 2 × 2 window to perform the down-sampling operation on the feature map of the convolution layer C3, and obtain 128 feature graphs, each with a size of 20 × 20.

3.3 Fully Connected Layer.

The input of the full connection layer must be a one-dimensional array, and the first layer of the pool layer S2 output of each feature map is a two-dimensional array. Therefore, the first two-dimensional array corresponding to each feature map is transformed into a one-dimensional array, and then 128 one-dimensional arrays are concatenated into a feature vector of 51200 dimensions (20 × 20 × 128 = 51 200), which as the input of each neuron in the tethered layer. The output of each neuron is shown in equation (3).

$$h_{w,b}(x) = \theta(w^T x + b) \tag{3}$$

Where $h_{w,b}(x)$ represents the output value of the neuron; x represents the input feature vector of the neuron; *w* represents the weight vector. Rectifier Linear Units (ReLU) function is used in the experiment. ReLU activation function [5] can solve the phenomenon of gradient dispersion. The calculation of the ReLU function is relatively small, and it will make a part of the neural unit output 0, can be achieved with a sparse neural network, equivalent to the pre-training process of unsupervised learning. At the same time, Re LU is also easier to train and optimize. ReLU's mathematical expression is f (x) = max (0, x), ReLU function diagram shown in Figure 2.

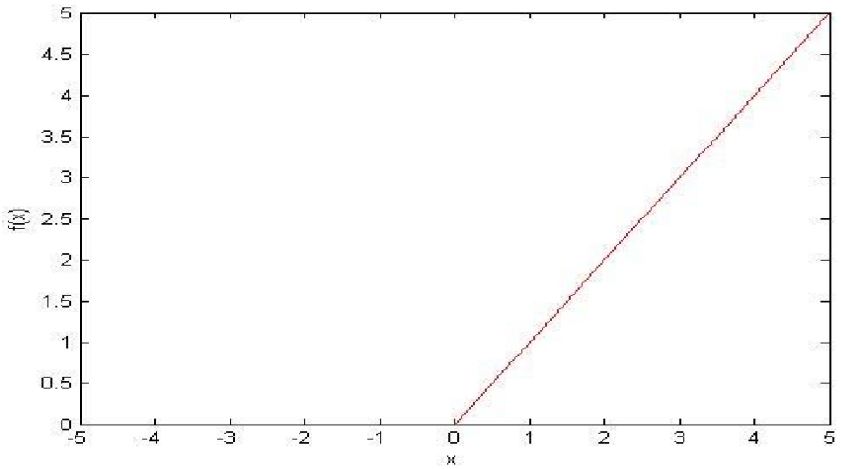

Fig. 2 ReLU function diagram

The full connection layer is fully connected with the pool layer S2, and the number of neurons will affect the training speed and fitting ability of the network. The experimental results show that when the number of neurons is 300, the effect is better.

3.4 Softmax Layer

The last layer of the CNN uses the Softmax classifier. The Softmax classifier is a multi-output competitive classifier. When a given sample is entered, each neuron outputs a value between 0 and 1, which represents the probability that the input sample belongs to that class. Therefore, the category of the neurons with the largest output value is selected as the classification result.

4. Experimental Results and Analysis

The experimental sample selected in this paper consists of two parts: the Cohn-Kanade [8] expression library and the self-timer expression image. Among them, CK expression library contains a total of 1 825 images, divided into anger (386), surprised (360), disgust (265), fear (255), happy (300), sad (259); self-timer expression images contain a total of 1 264 images, divided into anger (211), surprised (215), disgust (204), fear (214), happy (210) and sad (210).

In this paper, three experiments were carried out:

(1) A sample of the CK expression library was used for training and testing. Some of the expressions were training samples, and some of the expressions were test samples and from different people. 90% of each expression was selected as a training sample, 10% as a test sample. Cycle 10 times, the average error of 10 times as a test result. The test results are shown in Table 1.

Table 1. CK expression test method

Expression	Number of training samples	Number of test samples	Average misidentification	Recognition rate/%
anger	385	26	0.6	97.69
surprise	360	40	3.1	92.25
disgust	265	35	0.8	97.71
fear	255	37	3.0	91.89
happy	300	30	1.3	95.67
sad	259	36	5.6	84.44
Average recognition rate: 92.84%				

It can be seen from Table 1 that the recognition rate of CK expression library is more than 90%. It is proved that the convolution neural network expression recognition system without feature extraction has the ability of high accuracy and generalization ability.

(2) In order to verify the robustness of the algorithm, this part of the experiment using self-timer image as a test image. Participate in the expression of the expression image is still from the CK expression library. The test results are shown in Table 2.

Table 2. Self-expression test program

Expression	Number of training samples	Number of test samples	Number of samples	Recognition rate/%
anger	385	211	14	93.36
surprise	360	215	20	90.70
disgust	265	204	12	94.12
fear	255	214	29	86.45
happy	300	210	15	92.86
sad	259	210	54	74.29
Average recognition rate: 88.77%				

It can be seen from Table 2 that the recognition result of the self-timer expression scheme is generally low, and the recognition rate is low due to the lack of sufficient prior knowledge and the differences in the expression of Asians and Europeans.

(3) In order to improve the recognition rate, the self-image is merged with the image of the CK expression database, and then the first experiment is repeated. 90% of the combined expression image

is used as training samples, 10% as test samples. Cycle 10 times, the average error of 10 times as a test result. The test results are shown in Table 3.

Table 3. CK and self-expression test method

Expression	Number of training samples	Number of test samples	Average misidentification	Recognition rate/%
anger	596	26	0.5	98.08
surprise	575	40	2.7	93.25
disgust	469	35	1.8	94.86
fear	469	37	2.6	92.97
happy	510	30	2.0	93.33
sad	469	36	5.1	85.83
Average recognition rate: 92.79%				

As can be seen from Table 3, some facial expression recognition rate has increased, while the other part has declined, the average recognition rate was essentially flat. Experiments show that convolution neural networks can learn the similarity of samples.

5. Conclusion

Convolution neural network has many unique advantages in dealing with two-dimensional images: (1) without complex feature extraction; (2) two-dimensional images can be directly input to the neural network, greatly reducing the difficulty of pre-treatment; (3) The down-sampling operation of the pooled layer enhances the robustness of the convolution neural network and tolerates a certain degree of distortion of the image. In this paper, the convolution neural network is used to identify the facial expression. The experimental results show that the method has high recognition rate and good generalization ability.

Acknowledgements

This work was supported by the National Science & Technology Pillar Program during the 12th Five-year Plan Period (Grant No. 2015BAD18B05).

References

[1]. HINTON G E, OSINDERO S, TEH Y W. A fast learning algorithm for deep belief nets [J]. Neural Computation, 2006, p. 1527 - 1554.
[2]. LECUN Y, BOTTOU L,BENGIO Y, et al.Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, p. 2278 - 2324.
[3]. HUBEL D H, WIESEL T N. Receptive fields,binocular interaction and functional architecture in the cat's visual cortex[J]. The Journal of Physiology,1962, p. 106 - 154.
[4]. TENENBAUM J B,SILVE V D,LANGFORD J C. A global geometric framework for nonlinear dimensionality reduction [J]. Science, 2000, 290: 2319 - 2323.
[5]. Maas A L, Hannun A Y, Ng A Y. Rectifier nonlinearities improve neural network acoustic models[C]//Proc. ICML, 2013, 30: 1.
[6]. VINCENT P,LAROCHELLE H,BENGIO Y, et al. Extracting and composing robust features with denoising autoencoders[C] Proceedings of the 25th International Conference on Machine Learning. New York,NY: ACM,2008:1096 - 1103.
[7]. LECUN Y,BOSER B,DENKER J, et al. Back propagation applied to handwritten zip code recognition [J]. Neural Computation, 1989, P. 541 - 551.
[8]. LUCEY P, COHN J F,KANADE T,et al. The Extended Cohn-Kanade Dataset (CK) : A complete dataset for action unit and emotion-specified expression[c] IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops(CVPRW). New York:IEEE, 2010: 94 - 101.