

Machine Learning 2

by Yaocai Chen

Submission date: 28-Apr-2023 03:03PM (UTC+0100)

Submission ID: 203899465

File name: 40733_Yaocai_Chen_Machine_Learning_2_498640_729271630.docx (537.19K)

Word count: 3142

Character count: 17909

Machine Learning 2 1

Yaocai Chen

University of Wales Trinity Saint
David

2200435@student.uwtsd.ac.uk

Abstract— Facial expressions contain rich emotional information and are an important way for humans to transmit information. They can intuitively reflect human inner thoughts and play an indispensable role in interpersonal communication. With the vigorous development of computer vision and artificial intelligence, facial expression recognition technology has attracted widespread attention from scientists and gradually become a research hotspot. At present, facial expression recognition technology has been widely applied in fields such as intelligent transportation, medical treatment, human-computer interaction, criminal investigation, and online education, and has very important research value. Facial expression recognition technology extracts facial expression features, classifies them, and determines the category of facial expressions. The key to facial expression recognition technology is how to effectively extract facial expression features. Traditional facial expression recognition methods mainly rely on manual methods to extract facial expression features, which have disadvantages such as low feature extraction efficiency, complex feature extraction process, and low recognition accuracy. With the rapid development of deep learning and convolutional neural networks, facial expression recognition technology has also been rapidly improved. Convolutional neural networks can achieve automatic extraction of facial expression features, but how to design convolutional neural networks to extract more comprehensive and discriminative features, thereby improving the accuracy of facial expression recognition using convolutional neural networks, remains the focus of research on facial expression recognition. In response to the above issues, this article selects the DCNN model to improve the recognition accuracy of facial expressions.

Keywords— Deep Learning; DCNN; Convolutional Neural Network; Facial Expression Recognition

I. INTRODUCTION

Facial expressions are the most direct and effective emotional mode. The application of facial expressions is not based on the heart. However, the challenge of large-scale facial expression recognition mainly lies in its uniqueness, which limits the application of facial expression recognition due to the relatively acceptable facial expressions, low-quality facial images, and occluded facial expressions compared to dogs, as well as the immaturity of the 7-meter eyepiece. According to the excellent speech and action research conducted by scholars in the field of clothing, Disgust, Surprise, Anger, Sad, Fear, Happy, Neutral, AE.DP FFR 2013. CK+and JAFFE have established corresponding facial expression databases based on facial expression categories, such as RAF-DB , FER2013, CK+and JAFFE.

II. RELATED WORK IN LITERATURE

A. The Basic Principles of Convolutional Neural Networks

CNN evolved from Multi Layer perceptron (MLP). CNN adopts an end-to-end 2 approach, which can achieve automatic connection between input and output without manually adjusting parameters, making it suitable for feature extraction. When the original image is input into the convolutional layer, CNN can automatically learn the relevant features of the image from the input image, select the image features to be extracted, and finally output the classification results of the image. During the training process, CNN exhibits two characteristics: local connectivity and weight sharing. The following will briefly introduce these two characteristics.

(1) Local connection

In the early days, the main approach was to use fully connected sensing to connect the input layer and hidden layer of neural networks, which would greatly increase the number of parameters and complexity of the network, and also cause problems such as "gradient dispersion". The local connection of CNN can precisely alleviate the above problems. One of the advantages of the local connection method is that it can directly connect nodes in the hidden layer structure of CNN to other nodes in the input layer. This connection method can also be applied to layers within the hidden layer structure. From the figure, it can be seen that only three nodes in layer 1 are connected to nodes in layer 1, while nodes in layer 1+2 use the same connection pattern to connect nodes in layer 1+1.

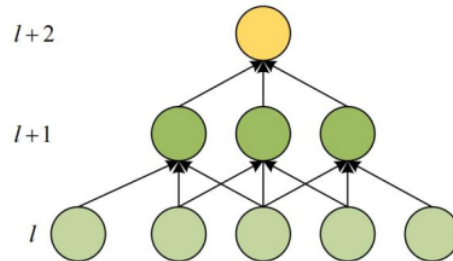


Fig. 1. Partial connection diagram

(2) Weight sharing

Weight sharing is different from local connections. The basic principle of weight sharing is that neurons with the same size of weight have the same characteristics. The central idea of weight sharing is that different neurons can achieve weight sharing on the feature plane.

In the figure, connecting lines with the same color indicate that neurons 3 have the same weight value. If the neurons in the

figure are fully connected, the parameter quantity of the neurons is $3 \times 5 = 15$. If the method of weight sharing is adopted, the parameter quantity of the neuron is 3, effectively reducing the parameter quantity.

B. The Basic Composition of Convolutional Neural Networks

The basic composition of CNN mainly includes convolution layer, pooling layer, activation function, full connection layer, etc. By alternately connecting several convolutional layers and pooling layers, CNN effectively reduces the number of parameters in the network model, resulting in less resource consumption and faster computation speed during the network operation process. CNN usually adds a activation function between the convolution layer and the pooling layer in order to introduce nonlinear capability. Finally, the extracted features are classified through a fully connected layer.

(1) Convolutional layer

Each convolutional layer contains multiple convolutional kernels (also known as filters), and convolutional operations are actually mathematical feature mapping problems. The specific operation process is to first multiply the values corresponding to the filter and the input feature map, and then add the resulting values as the output value. Then, the convolutional kernel is slid over the input feature map in a fixed step size, and the above steps are repeated until the corresponding values of the input feature map are fully traversed. Finally, combine all output values to obtain the corresponding output feature map.

Assuming the height and width of the input image are H and W , respectively, and the convolution kernel is $F \times F$, the sliding step (Stride) is S , P Representing an integer greater than or equal to zero used for zero filling operations, the height and width of the feature map after convolutional feature extraction can be calculated using formulas 1 and 2, respectively.

$$H_{output} = \frac{H - F + 2 \times P}{S} + 1$$

Fig. 2. formulas 1

$$W_{output} = \frac{W - F + 2 \times P}{S} + 1$$

Fig. 3. formulas 2

The schematic diagram of the feature map generated by the convolutional layer is shown in the figure.

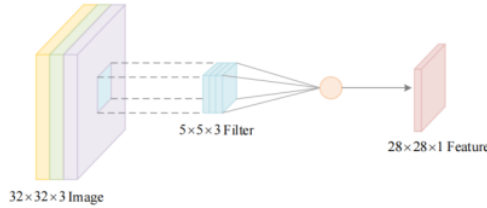


Fig. 4. Schematic diagram of feature maps generated by convolutional layers

(2) Pooling layer

The pooling layer is usually placed 30l after the convolutional layer, and its function includes the following two points. The first point is to reduce the size of the parameter matrix, compress data, and improve computer computing speed and reduce resource consumption. The second point is feature in variance in image processing. After pooling, the image size decreases, and some useless information is discarded, retaining only the most effective information that can transmit image features.

In image recognition and classification tasks, the most frequently used and widely used are average pooling and maximum pooling. The average pooling method divides the feature map into several regions based on a fixed sliding step size, and then takes the average of all feature values within the region, treating the obtained average value as a new feature value to represent the region. The maximum pooling method first divides the feature map into several regions based on a fixed sliding step size, and then represents the region with the maximum value within that region as a new feature value. Pooling and convolution are very similar in terms of operation. During the pooling process, the sampling window slides on the feature map from left to right and from top to bottom in a fixed step size. Then perform corresponding pooling within the corresponding sampling window, and the operation process is shown in the formula 3.

$$y_{i,j,k}^l = \text{pooling}(a_{i,j,k}^l)$$

Fig. 5. formulas 3

In formula 3, pooling (*) is the pooling function, $a_{i,j,k}^l$, which represents the output result of the previous layer on the network. Assuming the size of the feature map, the size of the sampling window, and the sliding step size are set to 4×4 , 2×2 , 2 respectively, the specific operations of average pooling and maximum pooling are shown in the figure.

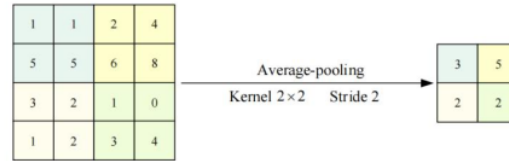


Fig. 6. Average pooling process

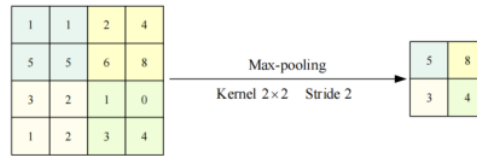


Fig. 7. Maximum pooling process

(3) Activation function

As a part of CNN, the activation function is used to introduce the nonlinear ability to CNN. If CNN lacks a activation function, then CNN can only carry out simple linear weighting on the weights of neurons. In the field of image classification, activation function with high application frequency include Sigmoid, Tanh, ReLU, among which ReLU activation function is the most widely used activation function. The three mentioned activation function will be introduced respectively below. The curves of the three activation function are shown in the following figure.

It can be seen from the figure that the output value of the Sigmoid activation function lies in the interval (0,1). The output value of Tanh activation function is in the interval (-1,1). When the input value of ReLU activation function is $x < 0$, the function value is $y=0$. When the input value is $x \geq 0$, the function value is $y=x$, which means that only some neurons of the neural network are activated, the network parameters become sparse, and the calculation is simpler. Therefore, in the process of network training, compared with the other two activation function, its convergence speed is faster.

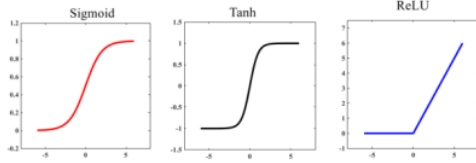


Fig. 8. Three commonly used activation function

The expression of Sigmoid activation function is:

$$f(x) = \frac{1}{1 + e^{-x}}$$

The expression of Tanh activation function is:

$$f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

The expression of ReLU activation function is:

$$f(x) = \begin{cases} 0, & x < 0 \\ x, & x \geq 0 \end{cases}$$

(4) Fully connected layer

The fully connected layer is usually adjacent to the pooling layer and located at the end of the CNN. Due to its fully connected approach, the computational load increases. When CNN performs image classification tasks, it is usually divided into two steps. The first step is feature extraction, and the second step is feature classification. The fully connected layer serves as the "classifier" of CNN to achieve correct classification of input images, as shown in the structural diagram.

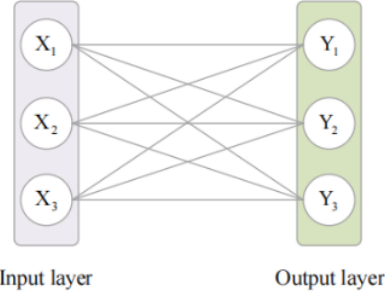


Fig. 9. Structure diagram of fully connected layer

The mathematical expression is shown in the equation.

$$Y_{w,b}(x) = f(W^T x + b)$$

$$Y_1 = W_{11} * x_1 + W_{12} * x_2 + W_{13} * x_3 + b_1$$

$$Y_2 = W_{21} * x_1 + W_{22} * x_2 + W_{23} * x_3 + b_2$$

$$Y_3 = W_{31} * x_1 + W_{32} * x_2 + W_{33} * x_3 + b_3$$

III. MAIN BODY

A. Facial Expression Recognition Method

Firstly, the input image is processed to crop out the facial region, and then the facial region is aligned to reduce the interference of noise on convolutional neural network feature extraction. Due to the large amount of data required for network training, some datasets with fewer images, such as the SFEW dataset, cannot support their training. Therefore, during the training process, it is necessary to enhance the input data, such as random cropping, rotation angle, flipping, etc. Data augmentation technology can effectively expand the number of images in the dataset and enrich the diversity of images.

The image obtained after preprocessing is sent to a convolutional neural network for feature extraction. During the training phase of the network, optimize and adjust the network parameters. The input image is first feature extracted through convolutional layers and the feature map is output. Then, the feature matrix of the output feature map is compressed through pooling layers to reduce computational complexity. After processing the input feature map through several convolutional and pooling layers, the network automatically sends the obtained output feature map to the fully connected layer for feature recognition and classification, and finally outputs the expression category and recognition accuracy corresponding to the input image. In the network testing phase, load the already trained network model and use the test set to test it. During the testing phase, data can be enhanced to improve the robustness of the network model. For facial expression recognition tasks, recognition accuracy is usually used as a measure of the quality of a network model. The higher the recognition accuracy, the better the performance of the network model and the better the recognition and classification effect.

Creating a model to recognize facial expressions can be divided into the following steps:

- Data collection: Collect a dataset containing facial expressions. You can use existing public datasets such as FER2013, CK+, etc., or collect data yourself, such as using a camera to capture your own facial expressions.
- Data preprocessing: Preprocessing the dataset, including data cleaning, data enhancement, etc. Data cleaning is to ensure data quality, and data enhancement is to increase the diversity of data sets, including rotation, cropping, translation and other operations.
- Feature extraction: Use feature extraction algorithms to extract facial expression features from images. The commonly used feature extraction algorithms include Haar features, LBP features, HOG features, etc.
- Model training: Train the dataset using machine learning algorithms or deep learning algorithms. Commonly used deep learning algorithms include convolutional neural networks (CNN), recurrent neural networks (RNN), etc.
- Model evaluation: Use a test set to evaluate the model, and evaluation indicators can include accuracy, recall, F1 score, etc.
- Model optimization: to optimize the model, you can try different algorithms, adjust hyperparameter and other methods.
- Deployment model: Deploy the model into practical applications, which can be deployed using mobile devices or cloud services.

The DCNN model is shown in the figure:

Layer (type)	Output Shape	Param #
conv2d_1 (Conv2D)	(None, 48, 48, 64)	1664
batchnorm_1 (BatchNormalizat	(None, 48, 48, 64)	256
conv2d_2 (Conv2D)	(None, 48, 48, 64)	102464
batchnorm_2 (BatchNormalizat	(None, 48, 48, 64)	256
maxpool2d_1 (MaxPooling2D)	(None, 24, 24, 64)	0
dropout_1 (Dropout)	(None, 24, 24, 64)	0
conv2d_3 (Conv2D)	(None, 24, 24, 128)	73856
batchnorm_3 (BatchNormalizat	(None, 24, 24, 128)	512

batchnorm_4 (BatchNormalizat	(None, 24, 24, 128)	512
maxpool2d_2 (MaxPooling2D)	(None, 12, 12, 128)	0
dropout_2 (Dropout)	(None, 12, 12, 128)	0
conv2d_5 (Conv2D)	(None, 12, 12, 256)	295168
batchnorm_5 (BatchNormalizat	(None, 12, 12, 256)	1024
conv2d_6 (Conv2D)	(None, 12, 12, 256)	590080
batchnorm_6 (BatchNormalizat	(None, 12, 12, 256)	1024
maxpool2d_3 (MaxPooling2D)	(None, 6, 6, 256)	0
dropout_3 (Dropout)	(None, 6, 6, 256)	0
flatten (Flatten)	(None, 9216)	0
dense_1 (Dense)	(None, 128)	1179776
batchnorm_7 (BatchNormalizat	(None, 128)	512
dropout_4 (Dropout)	(None, 128)	0
out_layer (Dense)	(None, 3)	387
=====		
Total params: 2,395,075		
Trainable params: 2,393,027		
Non-trainable params: 2,048		

Fig. 10. DCNN model

Here is the result of evaluation.

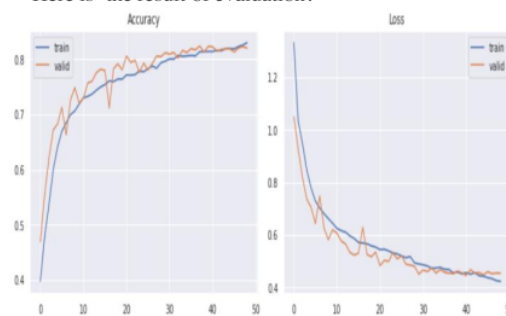


Fig. 11. Accuracy and Loss

	precision	recall	f1-score	support
0	0.93	0.91	0.92	899
1	0.77	0.76	0.77	608
2	0.73	0.76	0.75	620
accuracy			0.82	2127
macro avg	0.81	0.81	0.81	2127
weighted avg	0.83	0.82	0.82	2127

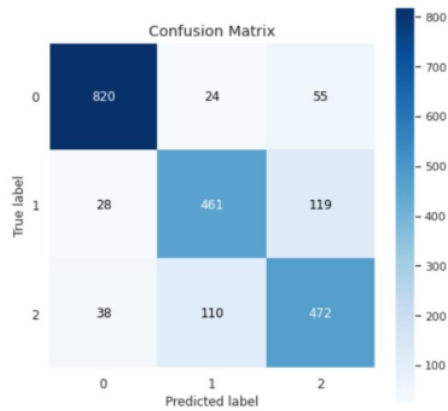


Fig. 12. confusion_matrix_dcnn

IV. CONCLUSION

Human facial expressions are an important way to express inner emotions and also an important means of social communication. With the development of artificial intelligence, if it is possible to infer a person's inner emotions at that time by recognizing human facial expressions, and predict possible behaviors based on their emotions at that time, corresponding services will be provided to meet their daily needs. Therefore, facial expression recognition has become an important research topic in computer vision. Facial expression recognition is widely used in fields such as human-computer interaction, online learning, assisted education, intelligent driving, and criminal investigation and case handling. Therefore, a faster and more accurate facial expression recognition model is needed.

Learn and research the relevant theoretical knowledge of deep learning, and separately study several classic neural networks commonly used in facial expression recognition. Explain the research background and significance of this article, and introduce the corresponding research progress of facial expression recognition at home and abroad, including application scenarios and the structural arrangement of this article.

Expounded the basic knowledge of the deep convolutional neural network studied in this topic, including the basic knowledge related to deep learning, the basic structure of the convolutional neural network, the characteristics of the convolutional neural network and various activation function, the characteristics and structure of various data sets used in this study, data preprocessing, and the activation function used in optimizing the network, laying a theoretical foundation for subsequent research.

Propose a facial expression recognition method combined with an improved DCNN network during the learning and research process. The most significant contribution of the network model is its residual structure. This design enables the deepening of the network layers to reduce the number of network parameters while also preventing overfitting. It also improves the design of the relevant activation function of the convolution layer in the training process of the convolutional neural network, and changes the original maximum pooling layer to the soft pooling layer after feature extraction. During

the experiment, ablation experiments were used for comparison. The effectiveness of the corrected network model has been verified, and it can greatly improve the performance of the network even when the number of network layers, i.e. the number of parameters, is small.

Although the facial expression recognition algorithm proposed in this article has achieved good recognition rate on the provided dataset, there are also some corresponding problems:

(1) Although the algorithm for manually extracting facial expression features has achieved a good expression recognition rate to some extent, the process of feature extraction is relatively complex and may show a relatively slow computational speed on large datasets. This may also be the reason why some researchers choose a simpler convolutional neural network as the feature extractor. In the future, they can focus on deep learning for further research.

(2) Shallow dense convolutional networks mainly increase the number of convolutional kernels in the convolutional module to improve the ability of feature extraction. They are a solution proposed for small datasets. When the dataset increases, the depth and width of the model need to be adjusted as needed. Meanwhile, dense convolutional networks require more periodic training due to the need to combine the characteristics of other layers during training, resulting in a relatively slow training speed.

(3) Parallel networks are an effective means to improve feature extraction capabilities and compensate for missing information in a single feature extraction algorithm. However, manual extraction algorithms may increase the overall parameter count of the model, resulting in a relatively slow training speed. In the future, research may need to be conducted in the direction of feature level fusion.

With the development of facial expression recognition research, manual expression feature extraction algorithms, various convolutional neural network expression feature extraction algorithms, and feature fusion algorithms have all achieved excellent results. However, in order to infer people's inner emotions, relying solely on facial expressions may be relatively shallow, while multimodal facial expression recognition (such as language, posture, etc.) may be more able to infer people's inner emotions. Therefore, Future research can invest more research in multimodal facial expression recognition.

REFERENCES

- [1] Mehrabian A, Ferris S R. Inference of attitudes from nonverbal communication in two channels[J]. *Journal of consulting psychology*, 1967,31(3): 248-252.
- [2] Darwin C. The expression of the emotions in man and animals[M]. University of Chicago press, 2015.
- [3] Ekman P, Friesen W V. Constants across cultures in the face and emotion[J]. *Journal of personality and social psychology*, 1971,17(2): 124.
- [4] Ekman P, Friesen W V. Facial Action Coding System (FACS): a Technique for the Measurement of Facial Actions[J]. *Rivista Di Psichiatria*, 1978,47(2): 126-138.
- [5] Sown M. A preliminary note on pattern recognition of facial emotional expression[C]. *The 4th International Joint Conferences on Pattern Recognition*.1978:408-410.
- [6] Jung H, Lee S, Yim J, et al. Joint fine-tuning in deep neural networks for facial expression recognition[C]. *Proceedings of the IEEE international conference on computer vision*.2015:2983-2991.
- [7] Mollahosseini A, Chan D, Mahoor M H. Going deeper in facial expression recognition using deep neural networks[C]. *2016 IEEE*

- Winter conference on applications of computer vision (WACV).IEEE,2016: 1-10.
- [8] Lopes A T,De Aguiar E, De Souza A F, et al. Facial expression recognition with convolutional neural networks: coping with few data and the training sample order[J]. Pattern recognition, 2017,61:610-628.
- [9] Verma M, KoboriH, Nakashima Y, et al.Facial expression recognition with skip-connection to leverage low-level features[C].2019 IEEE International Conference on Image Processing (ICIP).IEEE,2019:51-55.
- [10] Wang K, Peng X, Yang J, et al. Suppressing uncertainties for large-scale facial expression recognition[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 6897-6906.
- [11] Tang C,Zheng W, Yan J , et al. View-independent facial action unit detection[C].201712th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017).IEEE, 2017:878-882.

Machine Learning 2

GRADEMARK REPORT

FINAL GRADE

50/100

GENERAL COMMENTS

Instructor

Abstract could be more succinct and does not clearly summarize your paper. For us this means a summary of motivations, methods, results, conclusions.

Some odd phrases appear. "...7-meter eyepiece" Perhaps this is a translation issue.

Refer to figures using their number - eg figure 4, not "the figure"

A detailed review of relevant theory and concepts.

You need to review the dataset and choose, with justification 4 classes (at least).

Additionally, you need to explain you parameter choices for your network design - again, with justification.

References provided but you fail to adequately include inline citations.

In your related work you could have include ref to popular architectures and also papers that attempt a similar set of experiments using the same dataset.



Comment 1

title could be more helpful!



Comment 2

some prior feature extraction can also take place



Comment 3

which figure?

GRADING FORM: SMU GENERIC_GJ

YAOCAI CHEN

0

RESEARCH-INFORMED LITERATURE

KNOWLEDGE AND UNDERSTANDING

ANALYSIS

PRACTICAL APPLICATION AND DEP.

SKILLS FOR PROFESSIONAL PRAC.