Dynamic Facial Expression Recognition Based on ResNet and LSTM

Hong Guo* and Jiayou Chen
School of Computer Science and Technology. Wuhan University of Science and Technology. Wuhan, China

*Corresponding author

Abstract. The face image data collected by ordinary face recognition system is usually static. In real life, to fully capture the state information of human emotion change, we need to collect dynamic face expression change data, which needs to capture a video sequence to reflect the process of emotion change. In this paper, we combine ResNet residual neural network and long-term memory network (LSTM) to extract features from video sequences and then recognize them. In the residual neural network, the residual element has a very good effect on the training and optimization process of deep CNN model, and shortens the convergence time of the model. Combined with the LSTM network, the captured dynamic sequence data is incorporated into the CNN model. Compared with the static single frame expression recognition system, the research of dynamic expression recognition has more application value.

Keywords: Convolutional neural network; ResNet; LSTM; Expression recognition.

1. Introduction
In the age of artificial intelligence, it is the desire of many engineers to let machines understand human behavior. Facial expression is a more intuitive way to convey emotional information in the process of human communication. More and more AI engineers begin to devote themselves to the important research topic of facial expression recognition[1]. Recently, the application of facial expression recognition in people's life appears more and more scenes, such as setting different lights and music according to the change of the homeowners' expression in the context of smart home; detecting whether the driver has fatigue driving in the intelligent driving car; the service industry automatically judges the customer's satisfaction degree of entering the store according to the customer's expression, etc.

In the traditional expression recognition system, the feature extraction part and classifier design part of facial expression are carried out separately, so for an excellent recognition system, it is necessary to extract very accurate features first and design a proper classifier at the same time. Scholars at home and abroad have done a lot of research on these two aspects in the past. They have proposed the methods of feature extraction based on Gabor wavelet transform, local binary pattern (LBP), histogram of oriented gradients (HOG), popular learning, and support vector machine (SVM), etc[2]. Although these methods have achieved a very good recognition effect for most of the normal data, the recognition accuracy of the algorithm still needs to be improved when the data are greatly affected by external interference factors such as light, angle and other human interference factors[3].

In recent years, with the increasing attention of AI, deep learning, especially CNN( convolutional neural network), has made a breakthrough in the field of image recognition. The main principle is that by changing the depth of the hidden layer and increasing the number of neurons between each network
layer, enables the neural network model to naturally integrate the extracted features of high, medium and low levels. Especially through the combination of cross-channel information, the ability of image recognition is significantly improved compared with the traditional way of feature extraction and classifier. The earliest convolutional neural network, LeNet, was proposed by LeCun in 1998, which effectively solved the visual task of handwritten digit recognition, and established the basic structure of convolutional neural network, that is, convolution layer, pooling layer and full connection layer[4]. At the same time, LeNet is also considered as the originator of CNN, because many subsequent neural network models are based on it. For example, AlexNet, the winner of ImageNet in 2012, proposed using relu instead of sigmoid as the activation function to speed up the convergence of SGD, using dropout to avoid over fitting, overlapping and maximizing pooling, avoiding the fuzzy effect of average pooling in CNN before, and proposed LRN layer to normalize and accelerate training using the adjacent data. In 2014, the VGGNet model proposed by the visual geometry group (VGG) of Oxford University reduced the top-5 error rate to 7.32% in the ilsvrc competition. Its essence is actually to reproduce AlexNet on a finer granularity, which uses a smaller convolution core to realize a deeper convolution neural network. In the same year, GoogleNet appeared in ilsvrc2014 for the first time, and won the championship with greater advantages[5]. It introduced a new network structure, perception, which regarded the original node as a network, a network in the network structure. Therefore, GoogleNet was also called perception V1, followed by v2-v4 and other improved versions, reducing the top-5 error rate of Imagenet dataset to 4.8%. In 2015, ResNet, a deep residual network proposed by the Kaiming and others from China, won the championship in islvrc and coco. ResNet has a very deep layer, which uses a large number of residual modules as the basic component of the network. Compared with the traditional network, the biggest change is to add an identity mapping layer, which can effectively prevent the neural network from disappearing or exploding gradient when the depth is increasing. The network has a milestone innovation in the history of deep learning. It won the ilsvrc competition in 2015 and reduced the top-5 error rate of Imagenet data set to 3.57%.

In this paper, the residual network ResNet, a deep convolution neural network and memory network LSTM are used to recognize facial expressions. ResNet mainly extracts facial expression features, and LSTM is responsible for memory function. In the experiment, if the intercepted single frame data is put into the LSTM, a lot of processing data and redundant information will be generated, which will bring great difficulties to the training. However, if all pixels of the whole video sequence are directly put into the network, long-term memory cannot be realized[6]. Therefore, ResNet network is used to reduce the dimension of data, that is to extract the features of graphics, and then LSTM is used for long-term memory. In data set, this paper uses the hybrid data set composed of two kinds of expression data sets, CK + and fer2013, to train the network. Compared with a single data set, using the hybrid data set is very helpful to improve the generalization ability of the network[7].

2. Introduction to Residual Network

In the research of image recognition using convolutional neural network, with the increasing complexity of the content to be recognized, more complex models are needed to complete the recognition work, but the more complex models mean deeper network structure and larger neurons, and such complex network structure is too much for the training set data due to the increase of neurons in the training. Because of the sensitivity, it is easy to cause over fitting, and the appearance of ResNet is to solve the performance degradation problem of convolutional neural network because of the depth deepening[8].

The structure of the residual network is shown in Figure 1-1, which is a simple example model of the residual element with only two layers of network. The core way of thinking is to add a jumping shortcut connection branch in the construction of convolution neural network, so as to form a basic residual unit. Because ResNet uses such a jump structure to build neural network, it also has a method called "shortcut connection", which is to find a shortcut connection network. So why use this way to build the network and such a network can effectively solve the problem of performance degradation of deep neural network? As shown in Figure I, first of all, X is the input signal, H (x) is the expected final mapping output of the network layer, and the residual mapping is the difference between the output and the output, that is, f (x) = H (x) -X, then H (x) = f (x) + X, and the final mapping at sight is the sum of the input mapping and the residual mapping, that is to say, through this structure, the optimization
target can be transformed from $H(x)$ to $f(x)$, which needs to be supplemented. In the past research, we hope to achieve the effect of shallow network by making only one equivalent mapping on the upper layers of the deep network structure on the basis of the shallow layer. This idea is often a failure, because it is difficult for the algorithm to train it to the ideal effect, resulting in the deep network structure not only not being optimized, but also achieving a lower effect than the original. As for the residual network, it turns the original training target $H(x)$ into residual $f(x)$. At this time, instead of training the above layers into an equivalent mapping, it approaches to zero infinitely, so the training effect is actually much less difficult than training to an equivalent mapping. That is to say, if a four-layer network has reached an optimal function in the first three layers, then there is no need for further training in the last layer. At this time, with a jump structure like ResNet, the optimization goal will become close to 0. This process does not introduce new parameters, nor will it increase the computational complexity, which effectively relieves the deep convolution neural network. The gradient disappears in the back propagation when the model is trained, which solves a series of problems such as performance degradation.

![Figure 1. Residual learning unit module.](image)

In addition to the above two-level residual learning unit, there are three-level structure. As shown in Figure II, ResNet34 structure is shown on the left and ResNet50 / 101 / 152 structure is shown on the right, with the main purpose of reducing the number of parameters. The left image is two convolutions of $3 \times 3 \times 256$, the right image is the first convolution of $1 \times 1$ to reduce the 256 dimension channel to 64 dimension, and then through $1 \times 1$ convolution recovery, that is to say, the process of dimension reduction and dimension increase is carried out.

![Figure 2. Comparison of two-layer and three-layer residual element modules.](image)

3. Basic Structure of LSTM

Long and short term memory network (LSTM) is a special kind of cyclic neural network. The main way to memorize information is to introduce memory unit and control grid. LSTM can control all kinds of information flow in the network by changing the corresponding gate structure, so as to achieve the purpose of preserving all kinds of complex network element information for a long time. Its main contribution is to update the existing memory of the information network, and to set the parameters of the hidden layer of the new input network, so as to successfully solve the problem of gradient disappearance in BPTT training. It divides the first layer recurrent neural network into four layers, and then interacts. The basic LSTM unit consists of a storage unit and three control gates. The structural unit is shown in Figure III.
The following describes the three control gates and a memory unit structure respectively.

3.1. Input gate
The input of LSTM includes $H_{t-1}$ and $X_t$, where $H_{t-1}$ is the hidden state of the previous time, and $X_t$ is the newly received data at the current time. The specific formula is as follows, in which tanh function combines $H_{t-1}$ and $X_t$, and takes them as the final input. And the $\sigma$ is the activation function.

$$I_t = \sigma(X_tW_{ix} + H_{t-1}W_{ih} + b_I)$$

3.2. Oblivion gate
Forgetting gate is to control the forgetting of the current node's historical information. It can select the information remembered by the network and activate the function through sigmoid. In this way, redundant and duplicate historical records can be removed, and management system capacity information and associated information can be refined. If the result of the function is close to 1, it means that the information value of the memory is higher than before, and it will try to keep the current information value for the next stage. If the function value is close to 0, it is proved that the memory information value is lower than the previous value, then it will discard most of the memory information.

$$F_t = \sigma(X_tW_{xf} + H_{t-1}W_{hf} + b_f)$$

3.3. Output gate
The gate mainly controls the output information of nodes. If the node information represents the main feature, the output effect will be improved; if it does not represent the main feature, the output information will be reduced. Meanwhile, it can also determine the output of the previous memory update, so as to control the size of the next information output.

$$O_t = \sigma(X_tW_{io} + H_{t-1}W_{ho} + b_o)$$

3.4. Memory cells
The purpose of this part is to store state information, that is, to retain long-term historical information. In this part, candidate memory cells need to be calculated first. The calculation method is similar to the above three gates. The difference is that tanh function in the range of [-1,1] is used for activation function. The following formula is the calculation method of candidate memory cells in time step $t$.

$$C_t = \tanh(X_tW_{ic} + H_{t-1}W_{hc} + b_c)$$

Then through the input gate, forget gate, and output gate, the flow of information in the hidden state is controlled, which is generally realized by using multiplication by elements (symbol $\odot$). The calculation of current time step memory cell $H_t \in \mathbb{R}^{n_h}$ combines the information of upper time step memory cell and current time step candidate memory cell, and controls the flow of information by forgetting door and input door, the formula is as follows.
4. Combination of CNN and LSTM

The built-in LSTM unit has played an important role in solving the gradient problem of RNN network, and has successfully improved the important ability of long-term storage of RNN network structure. This paper mainly combines two important tools, LSTM and CNN, to complete the work of feature extraction of dynamic facial expression data, and through in-depth study of the internal law of expression time series data to truly realize the recognition of dynamic facial expression. In the process, the first step is to obtain the image data of people in the database, then extract them into N frames when extracting the data, and then extract the corresponding CNN features of each image to form CNN feature sampling layer. The single layer of LSTM is connected with each layer of CNN, and then the upper and lower layers of LSTM are connected to form a feature learning layer of LSTM. Finally, the feature classification layer is formed by SVM (support vector machine). Figure IV shows the overall structure of the eventually formed DEBN network.

![Overall network structure](image)

5. Experimental Results and Analysis

The experimental hardware platform of this paper is Intel (R) Core (TM) i5-6500 CPU, the main frequency is 3.2GHz, the memory is 16GB, the operating system is Ubuntu 18.04 LTS version, and the processing is accelerated by NVIDIA gtx1060 GPU. The software platform is based on the Google deep learning framework tensorflow + Python 3.6 on PyCharm for training.

The data sets are CK+ and fer2013. Among them, CK+ contains 593 video sequences of 123 subjects, among which 327 video sequences are labeled with facial expressions at the last frame. There are seven emoticons in this database: contempt, happiness, sadness, surprise, anger, disgust and fear[9]. The fer2013 data set was published by FER in the ICML 2013 competition. There are 28709 training gray images in the database, the size of which is 48 × 48 pixels. There are 3589 public test examples represented as fer2013-pub. The dataset contains 3589 private test samples, represented as fer2013. There are seven tags in this database: neutral, happy, sad, surprised, angry, disgusted and scared[10]. Because there are different expression tags in the two datasets, the disdainful tags are discarded from the expression datasets of CK+. 

\[ C_t = F_t \otimes C_{t-1} + I_t \otimes \hat{C}_t \]
In the experiment, all the images are unified in pixel size, and the pixel values are normalized. The training time is 74 hours, the learning rate is initialized to 0.8 ~ 1.2, and batch normalization is added after each convolution layer and pooling layer. The best classification accuracy of the selected model is obtained after 300 time periods. Each epoch randomly selects 640000 data as training data. Then we use the triple network to extract the features of the complete dataset, and train a simple 1-layer model to complete the classification task. Figure V shows the confusion matrix evaluation after training. When batch standardization is applied, the network will converge faster. The proposed model converges after about 100000 steps of batch standardization.

In the experiment, with the deepening of the number of network layers, the accuracy of expression recognition is not degraded by the performance in the rising process, and the convergence is also good, which shows the advantage of using ResNet as a convolutional neural network model.

6. Conclusion

This paper mainly introduces an algorithm of expression classification and recognition based on deep learning. The algorithm first introduces the advantages of convolution neural network and cyclic neural network in the prediction of expression dynamic sequence, then introduces the related concepts of ResNet and LSTM. Using the correlation and advantages of the two, an algorithm model of "CNN + LSTM" is proposed. The model combines CNN's efficient feature extraction and LSTM's dynamic sequence prediction. Compared with the single convolution neural network algorithm, it has a great improvement.

But in this paper, the research work of facial expression recognition can not meet the latest and most advanced vertex position that can continue to develop, so in the end, the possible direction of future work improvement is proposed: first, data processing, the first part is the research on the method of facial position estimation and detection, first of all, the related processing of the input pictures, so that the detection method of facial position can reduce the illumination, human The sensitivity of face inclination, pixel features of original size and size of photos is reduced, so that the next step of recognition can be more simple. Secondly, in the aspect of network optimization and improvement, we can deeply study the principle of convolutional neural network, so that it is no longer just a black box training mode, explore the reasons for its good performance, and then strengthen the treatment for the good places, and discard the bad places. This may not only improve the effectiveness of the training model, but also improve the training and operation time.

References

[1] Yin, L., Wei, X., Sun, Y., et al.: A 3D facial expression database for facial behaviour research. In: 7th International Conference on Automatic Face and Gesture Recognition, FGR 2006, IEEE, pp. 211–216, 2006.

[2] Guanming Lu, hailui Zhu, Qiang Hao, et al. Facial expression recognition based on deep residual network [J]. Data collection and processing, 2019, 34 (01): 54-61

[3] Zhao, G., Huang, X., Taini, M., et al.: Facial expression recognition from near-infrared videos. Image Vis. Comput. 29(9), 607–619, 2011.

[4] Fengping An, Zhiwen Liu. "Facial expression recognition algorithm based on parameter adaptive initialization of CNN and LSTM", The Visual Computer, 2019

[5] Maninchedda, F., Oswald, M.R., Pollefeys, M.: Fast 3d reconstruction of faces with glasses. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, pp. 4608–4617, 2017.
[6] Yongfeng Huang, Yujuan Zhang. "A Two-Phase Model for Sleep Staging Using Single Channel EEG", 2018 International Conference on Big Data and Artificial Intelligence (BDAI), 2018.

[7] Zhao, X., Liang, X., Liu, L., et al.: Peak-piloted deep network for facial expression recognition. In: European Conference on Computer Vision, Springer, Cham, pp. 425–442, 2016.

[8] Rima Tri Wahyuningrum, Lilik Anifah, I Ketut Eddy Purnama, Mauridhi Hery Purnomo. "A New Approach to Classify Knee Osteoarthritis Severity from Radiographic Images based on CNN-LSTM Method", 2019 IEEE 10th International Conference on Awareness Science and Technology (iCAST), 2019.

[9] Liu, M., Shan, S., Wang, R. et al.: Learning expression lets on spatio-temporal manifold for dynamic facial expression recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1749–1756, 2014.

[10] Guihua Wen, Zhi Hou, Huihui Li, Danyang Li, Lijun Jiang, Eryang Xun. "Ensemble of Deep Neural Networks with Probability-Based Fusion for Facial Expression Recognition", Cognitive Computation, 2017.