Video emotion recognition based on Convolutional Neural Networks

Chen Li*, Yuliang Shi¹ and Xianjin Yi¹

¹Faculty of Information Technology, Beijing University of Technology, Beijing, 100123, China

*Corresponding author’s e-mail: lichenatrrox@emails.bjut.edu.cn

Abstract. The existing video sentiment analysis methods only obtain features from the spatial and temporal signals of the video for sentiment classification, and cannot solve the difficulty of not knowing which emotion contributes the most to the entire video sentiment analysis in the video sentiment analysis. To solve this problem, a neural network with video frame weight vector is proposed. First, the video frame feature is obtained through the reel neural network, and then the weight vector layer is used to calculate the weight of the feature, and finally the frame feature with weight is put into the LSTM Training to obtain a video sentiment analysis model. We verified on the BAUM-1s data set. The results show that this method is better than existing methods in accuracy.

1. Introduction

Nowadays, the explosive growth of user video puts forward high requirements for the computational understanding of visual media data. Video content can convey strong emotional information to viewers, but to a large extent, the emotional ability of computer to understand video has not been well solved[1][5]. There are many applications for computer to understand the emotional content of video. For example, video recommendation service can match the interest of users with the emotion of video content.

There are two problems in video emotion analysis. First of all, in a video, only a few small segments express the content of the video, while other parts are not related to the video. Secondly, there may be multiple emotions in a video. It is difficult to answer which emotion contributes the most to the overall emotional analysis of the video.

Video is composed of images. Due to the large-scale data set, CNN improves the accuracy of image classification, and is better than the manual feature method. However, the spatial and temporal signals of video are ignored only for single image feature extraction[2]. The spatial and temporal signals in video are very important for emotional video modeling. At present, the popular emotion analysis method is to extract features from video with spatial and temporal signals. For example, 3dnn with spatial and temporal streams can extract spatial and temporal signals from video[3]. However, because 3dnn is too dependent on the sequence of image frames, combined with the difficulties of emotional analysis, it can not be well used in video emotion analysis. Xian Zhang et al. Proposed an end-to-end winnder neural network framework[9]. The framework extracts image features by CNN, and then classifies these image features into LSTM. This method can extract the spatial and temporal signals of video. But it does not solve the difficulties of emotional analysis.

One of the difficulties in sentiment analysis is that we do not know which emotions in the video contribute greatly to the overall emotion of the video. For example, we often see two scenes: one person laughs when crying, and the other one laughs and laughs, which are totally different classification
results⁶⁸. The main reason is that the focus of emotion classification in these two scenes is in the bit behind the video frame. However, the weight of these time series is not taken into account in the current video classification when extracting images. Therefore, this paper trains a weight set in the latitude of the image frame through the video data set, so as to express the weight of different frames on the video sentiment analysis results, and then train the video CNN + LSTM network to analyze the video emotion⁴¹. In this paper, it is found that the correct rate of video emotion recognition is improved by first extracting image frames according to weight and then analyzing video emotion.

2. Experimental methods

Based on the characteristics of video emotion, a new neural network structure is proposed:

![Figure 1 Neural network structure.](image)

The network is divided into three parts: Convolution neural network, Video frame weight and LSTM.

2.1. Convolution neural network

Convolution neural network is a deep artificial neural network, which is widely used in the field of computer vision. It has been applied to many task scenarios, including image classification, semantic segmentation, and achieved good results⁷¹⁰.

Convolutional neural network is a feature extractor composed of multiple convolution layers and pooling layers. Convolution formula:

\[
s(i, j) = (X * W) = \sum_m \sum_n x(i + m, j + n)w(m, n)
\]  

Where \( W \) is the convolution kernel and \( X \) is our input.

In this paper, a three-layer convolution and a two-layer full connection are used to extract the image features of each frame in the video. Image data is used as input.

In convolution neural network, in order to reduce the size of the parameter matrix and prevent over fitting to a certain extent, we often add a pooling layer between the convolution layers. Our common pooling layer: 1) maximum pooling: select the maximum value of the image area as the pooled value of the region; 2) average pooling: calculate the average value of the image area as the pooled value of the region. The pool layer selected in the standard is the largest pool layer.
2.2. Video frame weight

In order to obtain the weight of video frames, a vector with the same length as the number of video frames is constructed to represent the weight:

\[(x_1, x_2, ..., x_n)\]  \hspace{1cm} (2)

In order to fix the value of this vector between 0 and 1, let's first pass through a softmax:

\[S_i = \frac{e^{x_i}}{\sum_k^n e^{x_k}}\]  \hspace{1cm} (3)

\(S_i\) is the weight processed by softmax.

After that, we combine each frame feature vector with the video frame weight vector dot multiplication:

\[a_t = a_t \cdot S_i\]  \hspace{1cm} (4)

The video frame feature vector with frame weight is obtained.

2.3. LSTM

Long short term memory (LSTM) is a special cyclic neural network. The traditional neural network can't adapt to the changing time series and can't extract the continuous features in the time series. The standard recurrent neural network solves this problem. Because the input is processed in time sequence, the output of the recurrent neural network is mainly based on the previous context, so it can better extract continuous time features. However, the ordinary cyclic network has the problem of long-term dependence in learning long-term sequences, and can not get the best results. LSTM is specially designed to solve the long-term dependence problem of general RNN. The neurons in LSTM are shown in Figure 2:

\[z_t = \sigma(W_z \cdot [h_{t-1}, x_t])\]  \hspace{1cm} (5)

\[r_t = \sigma(W_r \cdot [h_{t-1}, x_t])\]  \hspace{1cm} (6)

\[\tilde{h}_t = \tanh(W \cdot [r_t \cdot h_{t-1}, x_t])\]  \hspace{1cm} (7)

\[h_t = (1 - z_t) \cdot h_{t-1} + z_t \cdot \tilde{h}_t\]  \hspace{1cm} (8)

Where \(x_t\) is the eigenvector input at time \(t\), \(h_t\) is the neuron output at time \(t\). The structure of the recurrent neural network in this paper is as follows:
Figure 3 Structure of recurrent neural network.

Since 2d conv convolution only considers the features of the image in the video, and does not consider the temporal and spatial features of the video, so we choose to use LSTM to extract the spatiotemporal features of the image features.

After inputting the frame eigenvectors from the previous network into LSTM in turn, a feature vector is finally obtained. The feature vector is classified by softmax.

3. Experimental steps

3.1. Video preprocessing
First of all, the video data is cleaned. The length of the video data in the data set is different, which will affect the verification of our experimental results. We make a comparison between 50-150 frames and find that 96 frames are the best for the experimental accuracy, so we unify the video data to 96 frames. For videos longer than 96 frames, we lose the extra frames at the beginning and end (l-96) / 2. For video data shorter than 96 frames, we insert the first and last (l-96) / 2 video frames into the tail. For the image size, we use bilinear interpolation algorithm to cut the image size to 150 × 110 × 3. After our video data cleaning, the final video matrix size is 150 × 110 × 3 × 96.

3.2. Training neural network
We use cross entropy to train our network, and the probability of network output is obtained by softmax.

In the part of convolution neural network, we have three winding layers, two maximization layers and two full connections. The input image size is 150 × 110 × 3.

4. Analysis of experimental results
We have done experiments on the open dataset baum-1s. Baum-1s database contains 1222 video samples. The video shows five emotions: joy, anger, sadness, depression and calm.

In order to verify the influence of our network design on the experimental accuracy, we first verify the influence of removing the weight vector layer on the experimental results. As shown in Table 1, the unauthorized duplicate layer removes the second part of our network, that is, the frame weight part, so that the FC5 in the first part of the winder network layer is directly put into the third part of LSTM. We use the verification set in the data set to verify. As can be seen in Table 1, after adding the weight vector layer to the data set baum-1s, the recognition accuracy of the network has been significantly improved, which proves that our weight vector layer is very effective.

| Method                  | Accuracy |
|-------------------------|----------|
| OURS                    | 49.83%   |
| Remove the weight layer | 46.51%   |

We also test the influence of placing our weight vector layer in different positions on the experimental results. The experiment compares the influence of placing our weight vector layer in FC4 and FC5
respectively. Train together in each case. As can be seen in Table 2, the best effect is to put our weight vector layer after FC5. The reason may be that after the FC5 layer, the video image features are more abundant.

| Method       | Accuracy |
|--------------|----------|
| After fc4    | 48.27%   |
| After fc5    | 49.83%   |

Finally, we compare with the most popular deep network methods (such as CNN, hog 3D, 3dcnn and CNN + LSTM). As can be seen from table 3, the recognition rate of 3dcnn, hog 3D and CNN + LSTM is higher than that of single CNN, because they all take advantage of the spatial characteristics of video. Compared with the above method, our network has higher accuracy.

| Method    | Accuracy |
|-----------|----------|
| CNN       | 41.09%   |
| HOG 3D    | 45.32%   |
| 3DCNN     | 46.23%   |
| CNN+LSTM  | 46.51%   |
| OURS      | 49.83%   |

The results show that our method can be used in small video sentiment analysis. In the network, there are two parts for extracting video features, namely, the winner neural network part for extracting static image features, the LSTM part for extracting video spatial features, and a weight vector layer for adding weight to the video frame features. Because the contribution of each video frame to the final video emotion classification is different, we introduce this weight vector layer to add weight to different frame features. Good results have been achieved.

5. Expectation

In this paper, a new neural network is proposed in the field of video emotion analysis. According to the characteristics of video emotion analysis, the network structure is studied and designed to classify video emotion. This method can effectively train the features which contribute a lot to the video emotion classification, so as to carry out effective emotion classification. The experimental results show that the accuracy of the proposed method is improved from 46.51% to 49.83% in baum-1s database. Therefore, it is proved that the proposed method is helpful for video emotion classification. However, the network needs a long training time because it contains a large number of parameters. In the future, it is necessary to further study how to reduce the network parameters and improve the calculation time.

References
[1] ZHANG S.Q., HUANG T.J. (2018) Learning Affective Features with a Hybrid Deep Model for Audio-Visual Emotion Recognition. IEEE Trans., 25:455-488.
[2] JI S.W., XU W., YANG M. (2013) 3D Convolutional Neural Networks for Human Action Recognition. IEEE., 25:221-231.
[3] Zhao, S., Liu, Y., Han, Y., Hong, R., Hu, Q., Tian, Q. (2018) Pooling the Convolutional Layers in Deep ConvNets for Video Action Recognition. IEEE Trans. Circuits Syst. Video Technol., 28:1839–1849.
[4] Wang, Z., Ruan, Q., An, G. (2016) Facial expression recognition using sparse local Fisher discriminant analysis. Neurocomputing., 174:756–766.
[5] Huibin, L.I., Sun, J., Zongben, X.U., Chen, L. (2017) Multimodal 2D+3D Facial Expression Recognition with Deep Fusion Convolutional Neural Network. IEEE Trans. Multimed., 19:2816-2831.
[6] Haryanto, I., Ariyanto, M., Caesarendra, W., Dewoto, H.K. (2017) Development of Speech Control for Robotic Hand Using Neural Network and Stream Processing Method. Internetworking Indones. J., 9:59–64.

[7] Gajewski, J., Vališ, D. (2017) The determination of combustion engine condition and reliability using oil analysis by MLP and RBF neural networks. Tribol. Int., 115:557–572.

[8] YING L., LI B., YUE B. (1998) Gradient-based learning applied to document recognition. In Proceedings of the IEEE., 125:2278-2324.

[9] Tang, Y., Zhang, X.M., Wang, H. (2018) Geometric-Convolutional Feature Fusion Based on Learning Propagation for Facial Expression Recognition. IEEE Access., 6:42532–42540.

[10] Deng, W., Zhang, S.J., Zhao, H.M., Yang, X.H. (2018) A novel fault diagnosis method based on integrating empirical wavelet transform and fuzzy entropy for motor bearing. IEEE Access., 6: 35042–35056.