Research on Video Emotion Recognition Based on Attention Mechanism LSTM Model

Xiaobin Zheng [a)]

Faculty of Computer, Guangdong University of Technology, Guangzhou 510006, China.

[a)] Corresponding author: gdutzxb@hotmail.com

Abstract. The study of video-based emotion recognition is still a difficult and hot research issue in the area of computer vision and human-computer interaction. How to effectively extract key frames from videos, construct spatiotemporal feature space of videos, and build model based on the mapping relationship between temporal feature space and video emotion type space, has become an important issue of video-based emotion recognition study. To solve this problem, this paper proposes a video emotion recognition method based on the attention mechanism LSTM model. Based on the features extracted by CNN, this method uses the LSTM model based on attention mechanism to model video temporal feature space and construct video emotion recognition model. The experimental results on CHEAVD data set show that this method can effectively improve the recognition rate of video emotion recognition task.

Key words: human-computer; LSTM; CHEAVD; CNN; videos.

INTRODUCTION

With the development of artificial intelligence, the research on human-machine emotional interaction has attracted extensive attention in the field of artificial intelligence. The research of human-machine emotional interaction needs the technology and method of emotion analysis [1]. In addition, the advent of the big data era has made people increasingly aware of the importance of data, while video is the largest part of mass data. Based on emotional analysis technology and massive video data, Research on video emotion recognition is of great importance and applying prospect.

Many researchers at home and abroad are engaged in the research of video emotion recognition. The basic idea of video emotion recognition method is to construct the mapping relationship model between the underlying feature space and the emotional type space of video. Traditional methods adopt features of manual design as underlying features and learn the mapping relationship between the underlying feature space and the emotional type space by means of pattern classifier and rule inference. Silva et al. [2] proposed extracting features of video and audio, and respectively constructing the facial expression recognition system based on video temporal features and the hidden Markov model based on audio feature. Schuller et al. [3] proposed using a tree extension naive Bayesian classifier, and a multiple hidden Markov model to realize the automatic segmentation of video and emotion recognition. The problem of the traditional method is that the low-level features of manual design largely determine the results of video emotion recognition model. The experimental results have higher requirements for the design of the underlying features.

In recent years, with the breakthrough of deep learning, video emotion recognition research based on deep learning method has become the mainstream. Most of the research uses convolutional neural networks and recursive neural networks. The literature [4] predicted the emotion type of video by using the multimodal characteristics of video as the input of the SVM classifier, and no further research was done on the modeling of temporal characteristics. In the follow-up study, the research on the modeling of temporal characteristics by deep learning method is more and more concerned. For example, the recursive neural network is introduced in the literature [5] to learn the temporal characteristics of video. However, there is too much redundancy in video. How to extract the key video frame information while learning the temporal characteristics is the direction to improve the recognition rate of video emotion recognition.

According to above problems, as shown in Fig.1, this paper puts forward the video emotion recognition framework on attention mechanism LSTM model.

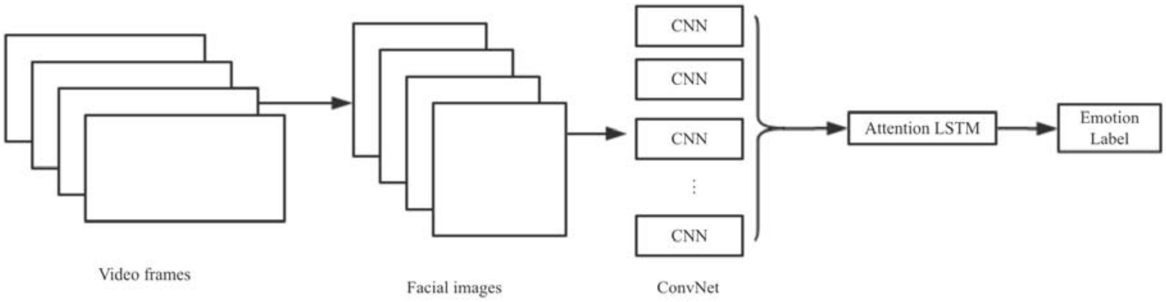

FIGURE 1. Video emotion recognition framework

VIDEO EMOTION RECOGNITION FRAMEWORK BASED ON ATTENTION MECHANISM LSTM MODEL

The main goal of video emotion recognition is to describe the relation of the video information and video emotions. In this paper, we study the relationship between video visual information and video emotions. There are two main tasks: to construct video visual information characteristics of space and time space; and to construct the relationship model between video spatiotemporal feature space and emotion type space.

Focusing on above tasks, this paper designs a framework shown in Fig.1. The framework can be broadly divided into the following stages: the data preprocessing stage; the spatial feature modeling stage; and the temporal feature modeling stage.

Preprocessing

Video consists of several video frames, each of which is an image, but the overall image has too much interference. In this paper, a face image extracted from video frame is used as input. The face image is extracted by IntraFace toolkit, where the OpenCV's Viola & Jones face detector is applied for face detection and initialization of the Intraface tracking library. The face size is set to 100*100. Since the face images in some video cannot be extracted using IntraFace toolkit, this paper adopts the open source MTCNN model to re-examine some video for face detection and face matching pretreatment.

Spatial Feature Modeling

The extraction of the underlying features of video is the basis of video emotion recognition. This paper uses the convolution neural network to extract features of face images. At present, the network structures widely used in computer vision tasks include common network structures such as AlexNet, GoogLeNet and VGG-16. The data volume of CHEAVD data sets is small. In order to prevent overfitting, this paper adopts the convolutional neural network structure similar to AlexNet. This paper has modified the AlexNet network. In order to better illustrate, this article will name the modified convolution neural network structure ConvNet. The input layer of the input image size is 100*100. The convolution kernel size is 3*3. The last full connection layer output dimension is set to the emotional type space dimension.

There are too few training samples in the data set. In this paper, the public face expression image data set FER2013 is adopted to pre-train the modified convolutional neural network. Then, in the CHEAVD data set, the weight of the fixed convolutional layer is retrained and the weight of the whole connection layer is retrained until the training result converges or reaches the expected number of iterations.

Most literatures use the convolution neural network as feature extraction method. The output of the full connection layer is usually used as the representation of the image. This paper uses the output of the last fully connected layer as input.

Temporal Feature Modeling

The LSTM model can learn the temporal characteristics of video from sequence information. This paper combines CNN model and RNN model. The output of CNN model is the input of RNN model.

Compared with the standard RNN network, the LSTM network uses memory cells to store and output information, which is beneficial to the discovery of a longer range of temporal information. The LSTM network consists of a memory unit containing a gate structure. Its calculation formula is:

$$\begin{cases} i_t = \sigma\left(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i\right) \\ f_t = \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f\right) \\ c_t = f_t c_{t-1} + i_t tanh\left(W_{xc}x_t + W_{hc}h_{t-1} + W_{ci}h_{t-1} + b_c\right) \\ o_t = \sigma\left(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o\right) \\ h_t = o_t \tanh\left(c_t\right) \end{cases} \tag{1}$$

Where: σ is the sigmoid activation function, i, f, o and c respectively represent the input gate, the forget gate, the output gate and cell activation vector, W represents weight matrix (for example, w_{hi} represents the weight matrix between the hidden layer and the input gate), b represents the bias (for example, b_i represents the bias vector of input gate).

The LSTM model can solve the problem of gradient extinction and gradient explosion in the standard RNN network. However, video has a lot of redundant video frames. Effectively extracting the key frame information of video will help to improve the recognition rate of video emotion recognition task. In order to solve the above problems, this paper draws on the idea of attention mechanism and adopts the LSTM model based on attention mechanism on video emotion recognition task.

$\mathrm{H} \in R^{d \times N}$ is defined as the matrix of the hidden layer output vector $\left[h_1, \ldots, h_N\right]$ generated by the LSTM model, where d is the dimension of the hidden layer output vector, and N is the number of hidden layer output vectors. The attention mechanism will generate an attention weight vector α and the weighted hidden layer representation r.

$$\begin{cases} M = tanh\left(W_h H\right) \\ \alpha = \mathrm{softmax}\left(\omega^T M\right) \\ r = H\alpha^T \end{cases} \tag{2}$$

Where: $M \in R^{d \times N}, \alpha \in R^N, r \in R^d, W_h \in R^{d \times d}, \omega \in R^d$ are the corresponding weight matrixes.

The calculation formula of the hidden layer eigenvector output is:

$$h^* = \tanh\left(W_p r + W_x h_N\right) \tag{3}$$

Where: $h^* \in R^d, W_p \in R^{d \times d}, W_x \in R^{d \times d}$ are the corresponding weight matrixes.

h^* can be regarded as a video temporal feature vector, and the video temporal feature vector is used as input of the Softmax layer. The probability distribution of emotion types is calculated as follows:

$$y = \text{softmax}\left(W_s h^* + b_s\right) \tag{4}$$

Where: W_s and b_s are the weights and biases of the SoftMax layer respectively.

EXPERIMENT AND ANALYSIS

In order to validate and analyze the effectiveness of video emotion recognition method proposed in this paper, the benchmark results of CHEAVD data sets are selected as the comparison. Among them, the benchmark method uses random forest method to predict video emotion type. The data set in the Baseline-1 experiment is not sampled. The data set in the Baseline-2 experiment was conducted to reduce the sampling operation, and 100 samples were selected from the samples of all emotion categories in the training set to form a new training set.

On the CHEAVD data set, the default parameters of all models are: the learning rate is 0.005, the learning rate attenuation ratio is 0.99, the number of iterations is 100,000, and the coefficient of L1 paradigm is 0. 0001.The Early Stopping method is adopted in the experiment, and the iteration is stopped according to the result of the validation data set before the model is convergent to the training data set.

Datasets

This paper adopts the CHEAVD data set, the data set is from the institute of automation, Chinese academy of sciences. The data from interception by the film and television play video clips, each video clips were marked for some common emotions (happy, sad, angry, surprise, disgust, fear, anxiety) or the neutral emotion. Video has a total length of 141 minutes, including video fragments of 238 speakers from movies, TV shows and talk shows. The length of each video fragment is about 1 to 19 seconds. The entire emotional dataset contains 2,852 video fragments, of which the training set is 1981 and the test set is 243.

Evaluation

Video emotion recognition task belongs to the multiple classification problems. The different emotional categories of data is very uneven, so this paper adopts the MAP (Macro Average Precision) as a measure of the prediction results, and uses Accuracy as a measure of the prediction results. The calculation method of two evaluation ways as follows:

$$\text{ACC} = \frac{\sum_{i=1}^{s} TP_i}{\sum_{i=1}^{s} TP_i + FN_i} \tag{5}$$

$$\begin{cases} P_i = \dfrac{TP_i}{TP_i + FN_i} \\ \text{MAP} = \dfrac{1}{s} \times \sum\limits_{i=1}^{s} P_i \end{cases} \tag{6}$$

Where: represents the emotion category. represents the number of samples which is in the category and predicted as category . represents the number of samples which is in the category and predicted as other categories. represents the precision of category .

Results and Analysis

In view of the above-mentioned methods, this paper conducted experiments on CHEAVD data sets and obtained the prediction results under different models through experiments. Table 1 shows the different model predictive results of the test set, you can see two different temporal feature modeling method is compared with the Baseline method has large improvement, the fusion model the best prediction results have been achieved.

LSTM model based on attention mechanism increased by 2.1% compared with basic LSTM model, the mechanism of attention by paying attention to important information, video frame model helps learning to the key information, video frames to avoid the interference of information redundancy of video frame, thus improve the recognition rate of video emotion recognition task.

TABLE 1. The results of different methods

Methods	ACC (%)	MAP (%)
Baseline-1	28.81	11.46
Baseline-2	12.35	10.71
ConvNet+LSTM	37.85	20.95
ConvNet+Attention LSTM	38.49	23.04

CONCLUSION

Video emotion recognition research is to build the space-time characteristics of video information and video emotion category mapping relationship of the space, among them, how to effectively to modeling the space-time characteristics of video and reduce the influence of the redundancy of video frame information will help to improve the recognition rate video emotion recognition. To this end, this paper proposes a video emotion recognition method based on time series multi-model fusion modeling. The validity of this method is verified by comparing with the benchmark method. In addition, the video recognition framework proposed in this paper can further improve the recognition rate of the model by extending different time series feature models and adding video information with multimode state.

ACKNOWLEDGMENTS

Author brief introduction: Zheng Xiaobin (1992-), male, master graduate, the main research direction includes machine learning and computer vision.

REFERENCES

1. Zeng Z, Pantic M, Roisman G I, et al. A survey of affect recognition methods: Audio, visual, and spontaneous expressions[J]. IEEE transactions on pattern analysis and machine intelligence, 2009, 31(1): 39-58.
2. De Silva L C, Ng P C. Bimodal emotion recognition[C]//Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on. IEEE, 2000: 332-335.
3. Schuller B, Rigoll G, Lang M. Hidden Markov model-based speech emotion recognition[C]// International Conference on Multimedia and Expo, 2003. ICME '03. Proceedings. IEEE, 2003: I-401-4 vol.1.
4. Kahou S E, Pal C, Bouthillier X, et al. Combining modality specific deep neural networks for emotion recognition in video[C]//Proceedings of the 15th ACM on International conference on multimodal interaction. ACM, 2013: 543-550.
5. Ebrahimi Kahou S, Michalski V, Konda K, et al. Recurrent neural networks for emotion recognition in video[C]//Proceedings of the 2015 ACM on International Conference on Multimodal Interaction. ACM, 2015: 467-474.