1 Related works

Many people are doing emotional calculations in the text field. For example, [1] is doing microblogging sentiment calculations. However, relatively little work has been done in the field of image sentiment calculation. In [1], Erik Cambria et al. performed an emotional calculation on the video. In the thesis, the author divides the video into three parts: image, audio, and text, and then performs the fusion. Here we only perform emotional calculations on the images in the video. In this paper, we propose CNN-based framework for the image sentiment computing, which gives the sentiment attitude of different videos.

2 Method

The first step to get the feature of the videos is to get the key frame extraction. One simple method is selecting frames at regular intervals. In this passage, we select the head and tail frames as the key frames. The tail plus the head is used as the feature information. Compared with the subduction of both frames, this method collects the mean features of videos. However, the subduction only reduces the common features. As the video data is very large, we extracted three frames per second from the training videos. We used OpenCV to crop faces in the frames and save them as new frames. In this way, we could reduce the amount of training data. The new frames were then passed through a CNN architecture similar to Fig.1.

The first convolution layer contains 32 kernels of size 5*5 and the first pooling layer is of dimension 2*2 ; then the second convolution layer contains 32 kernels of size 3*3 and the second pooling layer is of dimension 2*2. This layer is followed by a logistic layer of fully connected 300 neurons and a softmax layer.

3 Experiment and Observations

In this paper we choose the MOSI dataset as the training dataset and test dataset. The MOSI dataset is a dataset rich in sentimental expressions where 93 people review topics in English [2]. The videos are segmented with each segment's sentiment label scored between +3 to -3 by 5 annotators. We took the average of these labels as the sentiment polarity and, hence, considered only two classes (positive and negative). We take k-fold cross validation on the dataset. In this experiment, we set k as 5. The accuracy recall and f1-score are shown in Table 1.

accuracy	recall	f1-score
0.57	0.46	0.54

Table 1: The result in the 5-fold cross validation.

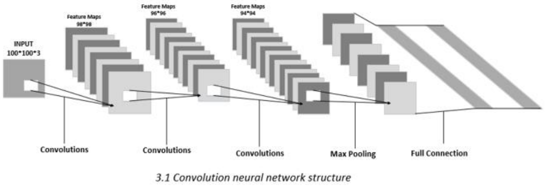

Figure 1: Convolusion Neural Network Structure.

4 future work

From the result we can see that the accuracy and f1-score are not high. Since our experiment wants to divide the videos into two classes, so the result seems not ideal. I think the result should be improved by doing some optimization. For example, we should use some algrithm to extract the frames. We can also subtract the adjacent images to reflect changes in the contours of the face.

References

[1] Bai Xue et al., "A Study on Sentiment Computing and Classification of Sina Weibo with Word2vec," IEEE International Congress on Big Data, 2014, pp.358-363.

[2] Erik Cambria et al., “Benchmarking Multimodal Sentiment Analysis,” arxiv.org, 29. Jul. 2017.

[3] Amir Zadeh et al., “Multimodal Sentiment Intensity Analysis in Videos: Facial Gestures and Verbal Messages,” IEEE Intelligent Systems, 2016, 31(6), 82-88.