Amplifying a Sense of Emotion toward Drama-
Long Short-Term Memory Recurrent Neural Network for dynamic emotion recognition

105062575 何元通　　104061701 周惶振

1. Introduction

Emotion is a core fundamental internal attribute of humans that governs our behaviors and decision-makings. There has already been a tremendous research effort in modeling humans using a variety of measurable signals, whichs aims at enabling machines to sense emotional states automatically [1, 2].

One key components in advancing such research is the availability of databases for researchers to develop robust recognition algorithms and carry out meaningful analyses. Human interaction often involves complex processes of communicative goals and emotional behaviors, which not only are expressed verbally and nonverbally but also are reflected in the inner responses of humans [3].

We want to use the NNIME database to study the emotion behavior (such as arousal and valence state) in small duration (like in real time), and to augment a sense of emotional feeling with visual demonstration.

After all of this, we just wonder how the emotion application can be. Therefore, we think of the amplification of the emotion in video. There are a lot of video that you would feel awkward watching it because of its boring and no effect. So we want to amplify the context in the video to make the video better.

2. Dataset Description

2.1 NNIME-Emotion Corpus

The increasing availability of large-scale emotion corpus with advancement in emotion recognition algorithms have enabled the emergence of next-generation human-machine interfaces. The database is a result of the collaborative work between engineers and drama experts. This database includes recordings of 44 subjects engaged in spontaneous dyadic spoken interactions.

The multimodal data includes approximately 11-hour worth of audio, video, and electrocardiogram data recorded continuously and synchronously. The database is also completed with a rich set of emotion annotations of discrete and continuous-in-time annotation from a total of 50 annotators per subject.

The emotion annotation further includes a diverse perspectives: peer-report, director-report, self-report, and observer-report. This carefully-engineered data collection and annotation processes provide an additional valuable resource to quantify and investigate various aspects of affective phenomenon and human communication. To our best knowledge, the NNIME is one of the few large-scale Chinese affective dyadic inter-action database that have been systematic-ally collected, organized, and to be publicly released to the research community.

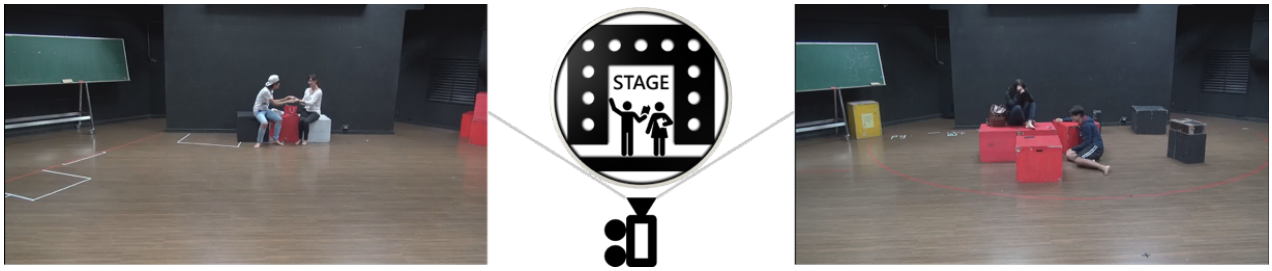


Figure 1. It depicts an actual snapshot of two different recording sessions extracted from the stage front-facing video camcorder (left and right). The middle depicts the camera setup in relation to the stage

2.1 Post-Processing

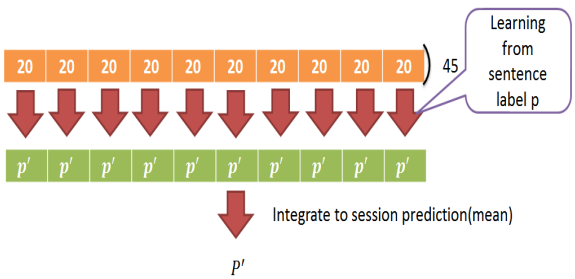
Each audio file corresponds to data collected from one of the micro-phones for each session. We manually segmented all audio files (two in every session with each lasted approximately 3 minutes long) into utterances. This resulted in a total of 6701 utterances.

Further, we marked each utterance as speech, laugh, sigh, sobbing, or dience background noise in order to enable further studies in understanding the role of non-verbal vocalization in affective interactions. We also manually completed the transcripts for all of the sessions.

3. Methodology &Experiment setup

To increase more visual effect on video, we train audio and lexical LSTM model together. Hope that higher accuracy improved through text feature if we can get video transcript in the future.

3.1 Audio



We use OpenSmile[4] to extract 45 LLDs(pitch、intensity、Mfcc(s) and their delta and delta delta) in the audio feature. Due to the difference of each sentence length, splitting sentence to the same step size is essential, so we first need to select a good step size to get more better emotion expression. Accordingly, a complete sentence may be split into a lot of frames, each frame representing a chunk of voice with the same step size. By this way, machine can see more detail in audio, also learn more large of amount of data. And we take sentence label being frame ground truth.

3.2 Text

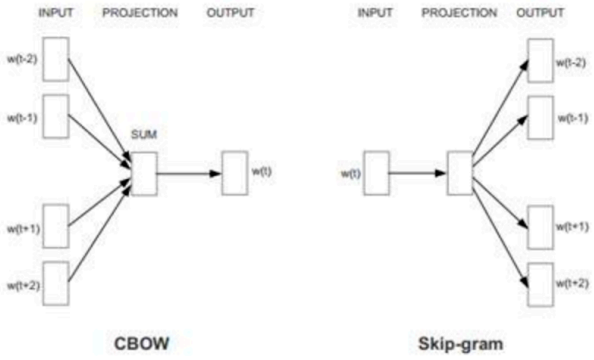
In Chinese, if we want to make machine understand the meaning in the context, a good word segmentation is necessary. In our project, we use CKIP[5] developed by Sinica to do that. For the first step of sentence segmentation, two algorithms are adopted.

The first one is “jieba”, which we have used in the course. We simply change its dictionary to traditional Chinese and call its API for segmentation. The second one is the segmenter from the Lab of the instructor; for the convenience, we call it JJ here. Since we call this segmenter with requests function

and it need to implement through Internet, it runs a little lower than jieba. In addition, according to the results, jieba runs better than JJ. To my observation, it may due to the fact that JJ sometimes segments a Chinese vocabulary into several individual characters. In this segmentation approach, it might not be able to show the meanings in this sentence as a character may have different meanings when it meets other characters. As a consequence, jieba is eventually chosen to perform for segmentation before feature transformation.

Word representation is a difficult issue in AI field, because it has the more complicate structure compared to that more intuitive and lower level behavior signal like audio, and it needs more higher level Cognitive ability to learn text for machine, this field is often called semantic analysis. For past few years, rapid development in Deep learning has a major breakthrough in word embedding.

Word2vec is a word embedding method developed by Tomas Mikolov Google[6], it can maps words into high dimension semantic space so that form more compact relation between word and word. After using Facebook FastText[7] to train Word2vec, we generate size 300 word vector and adopt the same method in the audio to split word to the same step size.



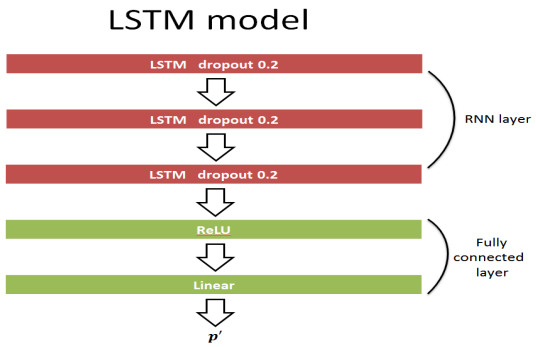
3.3 Model

3.3.1 SVR

Support vector regression is used to predict the final results. We can get approximately 30 to 40 percentage of correlation or sometimes even 70 percentage of correlation in the better situation without fine-tuning. Furthermore, for the purpose of making the result better, we have once performed feature selection. Nevertheless, perhaps the feature are tuned the best in fast-text, we found that using all features can perform better than simply using merely some part. Eventually, we have consider to use other data mining algorithms such as binary tree… etc. in the future works to show the advantage of our main purpose of LSTM-RNN.

3.3.2LSTM-RNN

By RNNLM[8], we train two LSTM model for audio and text respectively, the structure showed in figure 2, and parameter showed below,



after training LSTM model, we integrate the prediction of activation and valence of each frame to sentence by mean

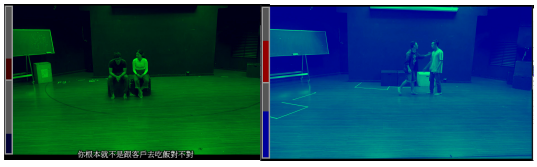
Audio parameter	Text parameter
Node : 45 35 25 10 1	Node : 300 200 100 50 1
Epoch : 20	Epoch : 30
Optimizer : Adam	Optimizer : Adam
Learning rate : 0.001	Learning rate : 0.001
Batch size : 200	Batch size : 50
Cost func. : MSE	Cost func. : MSE

3.4 The processing of the visualized effect

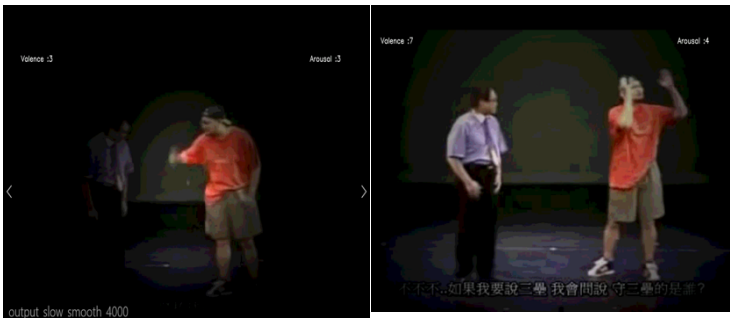
Our purpose is to enhance the feeling while the audience are watching the video, including audio, video and some other feelings. So we make some visualizing effect onto the video for example. In the future, we could make some more audio effects onto it.

Here we make two versions of the effect. One is RGB version, and the other is Vignette version.

1. RGB version is to change the RGB value according to the value of Valence and Arousal. As the Valence and Arousal increase, the Red and blue elements would enhance in the video accordingly. Two bars are placed at the left-most position of the video. The upper bar shows the value of arousal, and the lower one shows the value of valence. The two bars can make people more convenient to observe the two value. Moreover, to make the observation much simpler, we also filter the video through a color filter. The higher the arousal is, the higher the red value of RGB will be given. So as to the valence, but it operates on the blue value. The green value does not change for the purpose of maintain the original image. As a consequently, theoretically, when the arousal gets higher, the tone of the frame color tends to be orange. And, the frame tone tends to be blue, when the valence gets higher. But, in practical, with the mutual influence of the three colors, it tends to be green when red is higher than blue, and be blue when blue is higher than red. Also, it is predictable, when the two values are both high, the lightness will become higher. Considering that it is not obvious of the color change, we decide to modify the influence of the two value. Fortunately, for the analysis of the value distribution, almost values are distributed at the range between 0.4 to 0.8. We raise the impact on the values between 0.4 to 0.8, and reduce the influence on the values out of this range. The adjustment makes it more apparent on visualization that people can observe the change of the two value simply.



2.Vignette version is to change the parameter of the Vignette filter. As the Valence increases, the bright area of the Vignette would expand to make the frame look brighter. As for the value of Arousal is more than 5, we make a twinkling effect, which would make the video switching between the black-and-white frame and the colorful frame. The implement of Vignette filter is making two dimensional Gaussian distribution. By adjusting the mean and variance of x-axis and y-axis we could make the central spot light effect with different size of bright area. While the value of valence is higher, we let the variance of the distribution increase to make the bright area larger. Which is as shown below



4. Result and analysis

The final result is showed in below .audio result is better than baseline about 0.11 in activation and 0.04 in valence, but text is much worse. We think that maybe when training Word2vec which need to join more emotion corpus to improve model performance or this database is not appropriate for text LSTM model. Therefore, in the future, we may do some improvement in our algorithm or collect more corpus. First of all, we can combine word feature with audio feature generated from LSTM middle layer by more advanced algorithm. Secondly, by using python crawler to collect a large amount of corpus about psychology or drama. Finally, we can introduce more complicated algorithm like bidirectional RNN to try to capture more time series information in text and audio.

	Activation	Valence
Baseline(Audio)	0.32	0.09
Audio	0.43	0.13
Baseline(Text)	0.43	0.32
Text	0.1	0.04

5. Conclusion

We probably know that it's almost impossible to manifest without emotions. Emotions give passion to your thoughts, making them "loud" enough to leave an impression. Some people however have trouble adding emotion – it either isn't very strong, or there's no emotion at all. Therefore, we want to amplify emotions on movie scenes by "Fisheye effect" or "color filters". Not only enhancing emotion experience but also providing directly emotional clues for people who have trouble sensing emotions. However, our model doesn't perform well. Audio LSTM-RNN model only increases 11% (0.32 to 0.43) on activation evaluation and 4 % (0.09 to 0.13) on valence evaluation than SVR model. Text is not better than SVR model, and it might be caused by too small Chinese sentence data. We will do more effort on Chinese emotion recognition, and hoping we can do the robust model and finding interesting knowledge about emotions in the future.

6. Reference

[1] Picard, Rosalind W., and Roalind Picard. *Affective computing*. Vol. 252. Cambridge: MIT press, 1997.

[2] Narayanan, Shrikanth, and Panayiotis G. Georgiou. "Behavioral signal processing: Deriving human behavioral informatics from speech and language." *Proceedings of the IEEE* 101.5 (2013): 1203-1233.

[3] Quan, Changqin, and Fuji Ren. "A blog emotion corpus for emotional expression analysis in Chinese." Computer Speech & Language 24.4 (2010): 726-749.

[4] Eyben, Florian, Martin Wöllmer, and Björn Schuller. "Opensmile: the munich versatile and fast open-source audio feature extractor." Proceedings of the 18th ACM international conference on Multimedia. ACM, 2010.

[5] Ma, Wei-Yun, and Keh-Jiann Chen. "Introduction to CKIP Chinese word segmentation system for the first international Chinese Word Segmentation Bakeoff." *Proceedings of the second SIGHAN workshop on Chinese language processing-Volume 17*. Association for Computational Linguistics, 2003.

[6] Efficient Estimation of Word Representations in Vector Space Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean
(Submitted on 16 Jan 2013 (v1), last revised 7 Sep 2013 (this version, v3))
https://arxiv.org/abs/1607.04606

[7] Enriching Word Vectors with Subword Information Piotr Bojanowski, Edouard Grave, Armand Joulin, Tomas Mikolov (Submitted on 15 Jul 2016) https://arxiv.org/abs/1607.04606

[8] RNNLM - Recurrent Neural Network Language Modeling Toolkit Tomas Mikolov, Stefan Kombrink, Anoop Deoras , Lukas Burget, Jan “Honza” Cernocky
http://www.fit.vutbr.cz/~imikolov/rnnlm/rnnlm-demo.pdf