Project Plan Group 21

1. Problem

Emotion recognition has been increasingly playing an essential role in human-computer interactions. Unimodal emotion classifiers, systems that rely on a single aspect of emotional expression, have not shown outstanding performance. To address this, multi-modal recognizers, which are trained on multiple aspects of emotional expression, have been developed. Another issue is that data collected for training is lab-made and usually spoken by a single person, which means emotional intensity and arousal could be misleading and not reflecting real-world human interactions.

As we have several sources of data, the way we combine them could affect the performance of recognition system. which feature classes should be kept, which classes should be ignored. Should we combine all feature classes then do classification or do prediction on each feature class then combine classification result into the final output? Those questions could be partially answered in the following sections.

2. Data

MELD contains 1400 dialogues and 13000 utterances from Friends TV series. All dialogues has multiple speakers involved and simulate daily conversations. Although these conversations were scripted in advance and acted in a controlled environment, the authenticity is sufficiently high as speakers are professional actors/actresses.

There are 7 emotions: Anger, Disgust, Sadness, Joy, Neutral, Surprise, and Fear, which are labeled to each utterance in a dialogue. Additionally, the dataset has 3 sentiment, positive, negative, and neutral, annotated to each utterance.

3. Methods and Experiments

3.1. Methods

A tri-modal recognition system which combine three data classes: audio, text, and video, is utilized in this project. In addition, the way we combine these three modals should not only preserve the characteristics of each modal but also reflect their inter-modal relationship. Hence, the hybrid fusion model is selected in this project. The hybrid fusion model firstly combines extracted features from two classes then do classification on the newly-merged features. Given that bimodal classification result, the prediction result from the third feature class is combined to output the final emotion recognition result.

To evaluate how good our system is, we will compare the F-Score obtained for each emotion and sentiment to the result generated by the baseline model of MELD data set. The baseline models have 3 several variants: unimodal of audio , unimodal of text, bimodal of text and audio.

3.2. Experiments

Based on the bi-modal baseline model, relying on audio and text, we would develop a tri-modal model which utilizes information from audio, text, and video. All data source has already-extracted feature sets which we could take advantage of without extracting from scratch. However, the referenced audio features were extracted by openSMILE which is a bit out-of-date. We plan to re-extract the audio features using

Word2Vec, which is a more cutting-edge model, to see if there is any improvement in the performance of emotion recognition system. However, this task is not the top-priority in the case that we have a extremely limited amount of time.

In terms of bi-modal approach, we would like mix the combination of data sources, e.g. audio-video and text-video. As the baseline model has already utilized text-audio, facial expression extracted from video could be a top candidate to see if additional video-based information would enhance the recognition precision. Moreover, when deal with bi-modality, we have to make a wise choice on which fusion model should be implemented. While the feature-level approach is able to propagate the inter-modal relationship between different data sources, the decision-level method preserve the characteristics of each single data class.

Regarding unimodal models, we would try to harden them as efficient as possible based on the baseline unimodal model. There are rooms for improvement such as accuracy, computation complexity, and resource consumption. The most interesting we would like to know is that if a hardened unimodal model could outperform a multi-modal model in terms of classification accuracy. Another interesting thing needs to be explored is that which single data source dominates the overall accuracy of classification.

4. Division of Work

- Cuong: implement tri-modal emotion recognition system
- Daniel: implement uni-modal emotion recognition system
- Raihan: implement bi-modal emotion recognition system

Schedule:

- 8/11 - 22/11: Implmentation and Experiment
- 23/11 - 29/11: Result and Conclusion
- 30/11 - 4/12: Presentation and Final Report