An Approach to Decrease Interference among Modalities in Emotion Detection

Yu Dai, Jiancheng Weng, Yi Liu, Jian Liu and Boyu Chang
College of Software, Northeastern University, P.R.China
Email: daiy@swc.neu.edu.cn

Abstract. Emotion detection of video clips through multimodal information helps in a number of applications, including view mining, recommending system, public opinion and public sentiment monitoring. Current multimodal fusion techniques are applied to identify relevant information among modalities, which can introduce the interference of different modalities when detecting the emotion in the video. In this paper, we introduce a method which uses Q-Learning algorithm in attention-based network for decreasing interference among different emotion modalities. Experimental results prove that our method can improve the performance of basic attention-based network when detecting the emotion in the video.

1. Introduction
With the popularity of online social media, more and more people tend to share their life with others through a video clip, which always contains information of different modalities such as visual and audio information. Emotion analysis with the aid of these multimodal information helps in different areas such as recommending systems and public opinion monitoring. For example, government can collect public opinions on a new policy.

An utterance is a segment of speech bounded by breaths and pauses. A video clip usually contains multiple utterances. The goal of considering an utterance as a processing unit is to label sentiment at a more detailed level. An utterance often contains information of three modalities: audio, visual and text.

Recently, a number of approaches has been proposed to detect sentiment at an utterance level [1][2][3]. However, most of the current methods always neglect the relationship among different modalities which may result in low performance of the model. That is to say, there exists interferences among different modalities which may make the classifier difficult to find out the true emotion. For this problem, currently many works has been done[4][5]. Most of the studies focus on extracting better representation of related features. Besides, decision-level fusion always lays on the assumption that modalities have no or little relevance with each other while neglecting the fact that in the decision-level fusing, it needs to consider more on the interference among different modalities.

To this end, we propose a method, which combines attention mechanism, used as a feature fusion method, and Q-Learning algorithm, used as a feature interference mechanism, to improve the performance of emotion classifier. The experiment result shows our method can improve the performance of basic attention-based models, which reveals that the interference among different modalities are decreased.

The following sections in this paper can be summarized as follows: Section 2 introduces relevant studies in multimodal sentiment analysis. Section 3 describes the methodology in detail. Section IV shows the experiment result and corresponding analysis. Section 4 is conclusion.
2. Related Works

Various fusion techniques have been proposed in recent studies. Feature-level fusion, is to fuse the features before training. In early stages of relevant studies, researchers simply jointed features of different modalities [6] as feature-level fusion. Jin et al [7], proposed a kernel-level method to fuse the features, which takes feature-level fusion to a subtler layer. The other one is decision-level fusion. To fuse modality information at a decision level, researchers train unimodal features respectively with different models and fuse the classifying result with different weight-based mechanism. Decision-level fusion lays on the assumption that modalities have no or little relevance with each other. Some traditional methods of ensemble learning such as bagging and boosting are applied at a decision level. In addition, Wang et al. [8] proposed an EGG-based decision-level method, combined with unimodal SVM classifiers. Gera et al. [9] combined SVM and HMM with a F-score-based decision-level approach. Song et al. [10] fuse video prediction by an EdNet and audio prediction by an independent classifier with ANN and KNN. Recently, combining the feature-level fusion and decision-level fusion has become a tendency. Ortega et al. [11] used a DNN to combine both fusion techniques. Jenifer et al. [12] proposed an intermediate-level fusion to combine the two levels of fusion and studied on the impact of PCA on the intermediate-level fusion.

With the advancement of relevant areas such as NLP, more and more mechanism has been applied in more complicated models. Poria et al. [13] propose an attention-based feature-level fusion method to get better performance on extracting the sequence message of the context. Zadeh et al. [4] propose a Memory Fusion Network, which realized the synchronization of multimodal information by multi-view g. Majumder et al. [14] presents a DialogRNN, which takes the state of speakers into consideration while training and utilizes three GRU gates to extract better representation. The studies reveal that the techniques of fusing modality information become more mature.

However, studies mentioned above ignore the fact that interference can be introduced while fusing these modalities. From this perspective, we introduce Q-Learning algorithm to deduce the interference among modalities.

3. Method

In the following sections, we describe our method in detail.

3.1. Overview of the Method

Framework of our approach is shown in Figure 1.

![Figure 1. Structure of our framework](image-url)
In the following, we will give a detailed description of the main components in the framework.

3.2. Unimodal Feature Extraction and Pre-training

We follow the approaches of Poria [1] to extract unimodal features of audio, visual and text. A CNN is used to extract the text features. Each utterance is represented by a matrix of Google word2vec vectors [15]. Audio features are extracted by openSMILE toolkit, involving low-level descriptors and their statistics. Video features are extracted using a 3D-CNN [16], which can pick out sequential information of one clip of utterance.

In this step, some basic statistics such as accuracy, loss, f1 score of each training epoch are recorded. These statistics can help measure the performance of feature fusion and basic models. More importantly, these statistics are needed when initializing the feature filtering mechanism. Utterances in a clip are not independent, thus emotion detection of one specific utterance should be based on the context of the whole clip. We use a model similar with CAT-LSTM [13] for pre-training, which is proved to have ability to extract better sequence message representation of the context. There are four models, which trains features as bimodal combinations, that is, audio-visual model (signified as AV), visual-text model (signified as VT) and audio-text model (signified as AT), as well as a tri-model of all the three modalities (signified as AVT).

3.3. Fusion in the Feature-Level

Alignment of features of different features leads to better performance [17]. Then, this paper uses a self-attention-based network to align the features due to the proved performance in feature alignment [18]. Then, global feature after fusion in self-attention-based Network can be filtered by Q-Leaning filter mechanism.

We design four different private networks, involving different bimodal networks audio-visual, audio-text, visual-text networks and trimodal networks. One of these networks is chosen to train the global feature in one training epoch. The modality which interferes with other modalities can be filtered by feeding the global feature into different private training network which does not contain parameters of that modality. To get the appropriate training network, we introduce the idea of reinforcement learning. Let global feature to be an agent. The state of the agent is the training epoch it is in. Actions of agent include the four networks, that is {AV, TV, AT, AVT}.

Then, Q-Learning algorithm is employed which is to improve the decision policy by incremental improvement of the reward. Here, let $S$ represent the current state of global feature. $S \in \{AV, TV, AT, AVT\}$, with $\{AV, TV, AT\}$ represents the different combinations of two modalities and $\{AVT\}$ represents the tri-modality. Let $A$ represents the action the global feature will take in next training episode. Let $Q$ be the possibility of each action to be taken. A Q-table here is used to record the relation of $Q, S$ and $A$. Let $n$ is the maximum episode of training, $C_n$ is to signify the pretraining classifier, $A^*$ is used to signify the action the agent takes based on estimation in state $S'$. $\max_{A'} Q(S', A')$ is to signify the maximum estimation of $Q(S')$. $\gamma \in [0, 1)$ is the discount factor, $\alpha$ is the learning rate, and $\alpha > 0$. $R$ is the reward the agency gets in each training epoch, which is represented by classifying accuracy. Then, the Q-table can be updated as below.

Algorithm 1. Update Q-table

Repeat (for each episode):
  Initialize $S$
  Repeat (for each step of episode):
    Choose $A$ from $S$ using policy derived from $Q$
    Take action $A$, observe $R$, $S'$
    $Q(S, A) \leftarrow Q(S, A) + \alpha [ R + \gamma \max_{A'} Q(S', A') - Q(S, A) ]$
  $S \leftarrow S'$
until $S$ is terminal.
3.4. Private Training

According to the list output by Q-Learning filter mechanism, the global feature will be trained in networks shown in figure 2, each represents a saved action, with a Bi-GRU layer for its concise structure and better performance in extracting relevant information of an utterance in both ends.

![Figure 2. A private training network](image)

4. Experiments

4.1. Dataset

The dataset used in our experiment is CMU-MOSI dataset [19]. The Multimodal Corpus of Sentiment Intensity (CMU-MOSI) dataset is a collection of 2199 opinion video clips from 93 videos by 89 speakers. Each opinion clip is annotated with sentiment in the range [-3,3]. The dataset is rigorously annotated with labels for subjectivity, i.e., positive and negative. We divided the dataset to a train set with 62 videos and test set with 31 videos. That is, 1447 utterances in the training set and 752 utterances in the testing set.

4.2. Results and Analysis

We apply our method to bimodal emotion classification, including audio-visual, text-visual and audio-text. Take audio-visual modality as an example, Actions and States in the Q-table are A, V and AV, as well as the private training networks.

Figure 3 shows loss of each training epoch. Compared to the audio-text model, the trimodal training has lower rate of convergence, which proves that video features may have interference with these two modalities.
Figure 3. Loss of each training epoch

Figure 4 shows accuracy of each training epoch. Trimodal training, represented in black dot line, doesn’t outperform other bimodal training, which also proves the existence of interference among modalities. As figure 4 shows, our model gains the best performance among these models in multimodal emotion detection, by a narrow margin though. We use Q-Learning algorithm to improve the attention mechanism, so that samples can be filtered dynamically while training. We consider this feature fusion mechanism combined with a feature filter mechanism can decrease possible interference of different modalities.

Figure 4. Accuracy of each training epoch

5. Conclusions

With the development of fusion technologies of multimodal information in emotion detection, it becomes a trend on eliminating interference among modalities while fusing multimodal information. In this paper, we proposed an approach which uses Q-Learning algorithm to improve the attention-based feature fusion mechanism. Presented model gains a better performance on CMU-MOSI dataset. In addition, we appeal to presentation of one specific type of datasets, which emotion representation is
less evident due to the incompliance among multimodal information. We hope that hidden emotion can be detected by eliminating interference among different modalities.

6. Acknowledgement

This paper is supported by the National Key Research and Development Plan under Grant No. 2017YFB1401803.

7. References

[1] Poria S, Cambria E, Hazarika D, et al. Context-dependent sentiment analysis in user-generated videos[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2017: 873-883.

[2] Poria S, Cambria E, Gelbukh A. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis[C]//Proceedings of the 2015 conference on empirical methods in natural language processing. 2015: 2539-2544.

[3] Zadeh A, Chen M, Poria S, et al. Tensor fusion network for multimodal sentiment analysis[J]. arXiv preprint arXiv:1707.07250, 2017.

[4] Zadeh A, Liang P P, Mazumder N, et al. Memory fusion network for multi-view sequential learning[C]//Thirty-Second AAAI Conference on Artificial Intelligence. 2018.

[5] Liu P, Qiu X, Huang X. Adversarial multi-task learning for text classification[J]. arXiv preprint arXiv:1704.05742, 2017.

[6] Chen L S, Huang T S, Miyasato T, et al. Multimodal human emotion/expression recognition[C]//Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition. IEEE, 1998: 366-371.

[7] Jin Y, Song P, Zheng W, et al. A feature selection and feature fusion combination method for speaker-independent speech emotion recognition[C]//2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2014: 4808-4812.

[8] Wang S, Du J, Xu R. Decision fusion for EEG-based emotion recognition[C]//2015 International Conference on Machine Learning and Cybernetics (ICMLC). IEEE, 2015, 2: 883-889.

[9] Gera A, Bhattacharya A. Emotion recognition from audio and visual data using f-score based fusion[C]//Proceedings of the 1st IKDD Conference on Data Sciences. ACM, 2014: 1-10.

[10] Song K S, Nho Y H, Seo J H, et al. Decision-Level Fusion Method for Emotion Recognition using Multimodal Emotion Recognition Information[C]//2018 15th International Conference on Ubiquitous Robots (UR). IEEE, 2018: 472-476.

[11] Ortega J D S, Senoussaoui M, Granger E, et al. Multimodal fusion with deep neural networks for audio-video emotion recognition[J]. arXiv preprint arXiv:1907.03196, 2019.

[12] Williams J, Comanescu R, Radu O, et al. Dnn multimodal fusion techniques for predicting video sentiment[C]//Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML). 2018: 64-72.

[13] Poria S, Cambria E, Hazarika D, et al. Multi-level multiple attentions for contextual multimodal sentiment analysis[C]//2017 IEEE International Conference on Data Mining (ICDM). IEEE, 2017: 1033-1038.

[14] Majumder N, Poria S, Hazarika D, et al. Dialoguernn: An attentive rnn for emotion detection in conversations[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2019, 33: 6818-6825.

[15] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013.

[16] Tran D, Bourdev L, Fergus R, et al. Learning spatiotemporal features with 3d convolutional networks[C]//Proceedings of the IEEE international conference on computer vision. 2015: 4489-4497.

[17] Zadeh A, Liang P P, Poria S, et al. Multi-attention recurrent network for human communication comprehension[C]//Thirty-Second AAAI Conference on Artificial Intelligence. 2018.

[18] Xu H, Zhang H, Han K, et al. Learning alignment for multimodal emotion recognition from speech[J]. arXiv preprint arXiv:1909.05645, 2019.
[19] Zadeh A. Micro-opinion sentiment intensity analysis and summarization in online videos[C]/Proceedings of the 2015 ACM on International Conference on Multimodal Interaction. ACM, 2015: 587-591.