Contextual multimodal sentiment analysis with information enhancement

Dandan Jiang, Dongsheng Zou, Zhipeng Deng and Jingpei Dan

College of Computer Science, Chongqing University, Chongqing, 400030, China

*Corresponding author’s e-mail: dszou@cqu.edu.cn

Abstract. The results of sentiment analysis research can be applied to various social situations, including human communication understanding and dialogue system. Multimodal sentiment analysis involves the identification of sentiment in videos. This paper proposes a multimodal sentiment analysis framework, in which the Attention-based Bi-directional Gated Recurrent Unit (AT-BiGRU) model is used to obtain the contextual relationship among utterances, and the Information Enhancement Fusion (IE-Fusion) model is used to fuse the multimodal features of each utterance. Experimental results show that our framework achieves 82.85% and 65.77% accuracies on two different benchmark datasets, CMU-MOSI and IEMOCAP, respectively. Our method outperforms the state-of-the-art models on the datasets.

1. Introduction
Sentiment analysis and emotional recognition can be applied to various social situations. They provide more opportunities for social media in understanding user preferences, habits and contents [1]. The research has been widely used in human communication understanding, dialogue system and so on. Emotional recognition and sentiment analysis have aroused the growing interest in the scientific and business field. Recently, in the field of multimodal sentiment analysis, the methods proposed by [2] regards each utterance as a separate entity, ignoring relationships and dependence among utterances. An Attention-based Bi-directional Gated Recurrent Unit model (AT-BiGRU) is proposed by us to solve the problem. The model considers the contextual relationship among utterances, and prioritizes the utterances with higher relevance to the target utterance.

This paper proposed a novel framework, Information Enhancement Fusion AT-BiGRU (IEFAT-BiGRU). It takes into account the factors that different modalities have the varying degree influence on target utterance and considers the dependence among utterances. Our proposed methodology is evaluated on two benchmark datasets, viz. CMU Multimodal Opinion-Level Sentiment Intensity (CMU-MOSI) [3] and Interactive Emotional Dyadic Motion Capture (IEMOCAP) [4].

2. Proposed methodology

2.1. Problem definition
Suppose that there are $P$ videos, represented as $V = v_1, v_2, ..., v_i ... v_p$. And each utterance in the $i^{th}$ video is expressed as $v_i = v_{i,1}, v_{i,2}, ..., v_{i,j} ... v_{i,L_i}$, where $v_{i,j}$ is the $j^{th}$ utterance in video $v_i$ and $L_i$ is the number of utterances in the video. Our goal is to classify the sentiment of each utterance $v_{i,j}$ in the video. The rest of the utterances in the video except $v_{i,j}$ are considered to be context of $v_{i,j}$ and provide important information. Below, we outline the proposed methodology through five main steps,
Figure 1 shows the overall framework.

**2.2. IE-Fusion—multimodal feature fusion with information enhancement**

In multimodal sentiment classification tasks, not all modalities are equally relevant. In order to give priority to the modalities that have greater impacts on the target utterance, we propose an IE-Fusion model. The model can enhance the information of the modality that has a greater influence on the correct classification result of the target utterance. Then it fuses the multiple modalities after the information enhancement. The model is shown in figure 2.

Through the CT-BiLSTM, we have obtained the context-dependent unimodal features of text, audio and visual, and they have the same dimension size $d$ of feature vectors. $Z$ denotes the features of multimodal information enhanced fusion, and it is calculated as follows:

$$Z = n \cdot Z_a + p \cdot Z_v + m \cdot Z_t$$  \hspace{1cm} (1)

Here, $Z_a$ = acoustic features, $Z_v$ = visual features, $Z_t$ = textual features, $Z_a, Z_v, Z_t \in \mathbb{R}^d$, and $n, p, m \in \mathbb{N}$. In these experiments, the value of $n, p, m$ is set to an integer between [1, 10]. Next, output $Z$ is input to the AT-BiGRU (Section 2.3, figure 2) to obtain the contextual information.

Figure 2. Proposed IEFAT-BiGRU model consists of IE-Fusion and AT-BiGRU. A = Acoustic; V = Visual; T = Text.
2.3. AT-BiGRU

In order to consider the influence of the surrounding utterances on the target utterance, i.e. the inter-utterance relationship. We make use of an AT-BiGRU model to magnify the contextual information with high relevance to the target utterance. The model is shown in figure 2. An attention mechanism is introduced by us to expand the influence of important utterances on target utterance. The attention network can calculate the degree of correlation among all utterances around the target utterance, and assign different weights.

3. Experiments and analysis

We appraise proposed methodology on two benchmark datasets, CMU-MOSI corpus and IEMOCAP database. To prove the effectiveness of IEFAT-BiGRU, the proposed model will be compared to multiple advanced networks in multimodal utterance-level sentiment detection. They are bc-LSTM [5], CATF-LSTM [6], GME-LSTM(A) [7], TFN [2], MFN [8], CMN [9], ICON [10].

The importance of IE-Fusion in the process of multimodal fusion is demonstrated, as shown in Table 1. Tables 2 (CMU-MOSI dataset) and 3 (IEMOCAP dataset) display the comparison results of IEFAT-BiGRU and its variants with baseline and the state-of-the-art models. As expected, IEFAT-BiGRU outperforms all methods on both datasets. At the same time, we compare IE-Fusion with simple fusion technology (Simple-Fusion).

### Table 1. Comparison between Simple-Fusion and IE-Fusion. A=Acoustic; V=Visual; T=Text.

| Modality | CMU-MOSI | IEMOCAP |
|----------|----------|---------|
|          | Simple-Fusion | IE-Fusion | Simple-Fusion | IE-Fusion |
| A+V      | 60.81     | 62.23    | 45.82         | 52.93     |
| A+T      | 80.32     | 81.25    | 63.34         | 64.97     |
| T+V      | 79.71     | 81.78    | 63.22         | 63.97     |
| A+V+T    | 81.01     | 82.85    | 63.90         | 65.77     |

### Table 2. Comparison of IEFAT-BiGRU and its variant with state-of-the-art models on CMU-MOSI dataset. A=Acoustic; V=Visual; T=Text.

| Modality | IEFAT-BiLSTM | IEFAT-BiGRU | IEF-BiGRU | bc-LSTM | CATF-LSTM | GME-LSTM(A) | MFN | IEFAT-BiGRU |
|----------|--------------|-------------|-----------|---------|-----------|-------------|-----|-------------|
| A        | 61.84        | 62.10       | 58.78     | 60.3    | -         | -           | -   | 62.45       |
| V        | 61.30        | 62.77       | 57.31     | 55.8    | -         | -           | -   | 61.18       |
| T        | 78.86        | 78.72       | 79.79     | 78.1    | -         | -           | -   | 79.65       |
| A+V      | 60.64        | 61.30       | 57.58     | 62.1    | 62.9      | -           | -   | 62.23       |
| A+T      | 80.72        | 79.52       | 80.85     | 80.2    | 80.1      | -           | -   | 81.25       |
| T+V      | 81.25        | 80.32       | 80.45     | 79.3    | 79.9      | -           | -   | 81.78       |
| A+V+T    | 80.59        | 81.38       | 81.11     | 80.3    | 81.3      | 76.5        | 77.4| 82.85       |

“-” indicates that the method does not apply to input in this case.

### 3.1. Comparison with the state of the art

CMU-MOSI: As shown in Table 2, for CMU-MOSI dataset, IEFAT-BiGRU is superior to all comparison models. We can see that compared with GME-LSTM (A), our model accuracy increases significantly (about 6.35%). As evidenced by experimental results, proposed model has a positive effect on the performance increase. IEMOCAP: As reported in Table 3, our model’s classification results are significantly better than the state-of-the-art methods. Compared to the advanced models, the accuracies of our model have been improved by 0.93%-2.57% and 1.77%-6.97% in bimodal and trimodal experiments, respectively. We believe that enhancement is due to our model considers the different effects of each modality on the target utterance.
Table 3. Comparison of IEFAT-BiGRU and its variants with state-of-the-art models on IEMOCAP dataset. A=Acoustic; V=Visual; T=Text.

| Modality | IEFAT-BiLSTM | IEFAT-GRU | bc-LSTM | TFN | MFN | CMN | ICON | IEFAT-BiGRU |
|----------|--------------|-----------|---------|-----|-----|-----|------|-------------|
| A        | 49.06        | 48.75     | 50.31   | -   | -   | -   | 50.7 | 50.94       |
| V        | 38.29        | 37.48     | 39.41   | -   | -   | -   | 41.2 | 37.79       |
| T        | 62.47        | 63.53     | 61.41   | -   | -   | -   | 58.3 | 62.97       |
| A+V      | 54.49        | 51.06     | 52.62   | -   | -   | -   | 52.0 | 52.93       |
| A+T      | 64.59        | 64.15     | 63.84   | -   | -   | -   | 63.8 | 64.97       |
| T+V      | 63.03        | 63.34     | 60.04   | -   | -   | -   | 61.4 | 63.97       |
| A+V+T    | 65.02        | 64.40     | 63.97   | 59.8| 58.8| 60.1| 61.9 | 64.0        | 65.77       |

“-” indicates that the method does not apply to input in this case.

4. Conclusion
We have presented IEFAT-BiGRU, a multimodal framework for sentiment analysis. It adopts the IE-Fusion model to carry out multimodal feature fusion for utterances. The model considers the factor that different modalities have varying degree influence in the process of multimodal fusion. Meanwhile, IEFAT-BiGRU includes an AT-BiGRU model, which uses contextual information of the utterances to predict sentiment at the utterance-level. Experiments show that the proposed framework improves significantly compared with the baseline and is superior to the state-of-the-art methods.

Acknowledgments
This work was supported by the National Nature Science Foundation of Chin [grant number 51608070] and the Fundamental Research Funds for the Central Universities [grant number 2019CDYGYB012].

References
[1] Poria, S., E. Cambria, R. Bajpai, and A. Hussain. (2017) A review of affective computing: From unimodal analysis to multimodal fusion. Information Fusion, 37:98-125.
[2] Zadeh, A., M. Chen, S. Poria, E. Cambria, and L.P. Morency. (2017) Tensor fusion network for multimodal sentiment analysis. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen. pp. 1103-1114.
[3] Zadeh, A., R. Zellers, E. Pincus, and L. P. Morency. (2016) Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. IEEE Intelligent Systems, 31: 82–88.
[4] Busso, C., M. Bulut, C. C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan. (2008) IEMOCAP: interactive emotional dyadicmotion capture database. Language Resources & Evaluation, 42 (4):335-359.
[5] Poria, S., E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.P. Morency. (2017) Context-dependent sentiment analysis in user-generated videos. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vancouver. pp. 873–883.
[6] Poria, S., E. Cambria, D. Hazarika, N. Mazumder, A. Zadeh, and L. P. Morency. (2017) Multi-level Multiple Attentions for Contextual Multimodal Sentiment Analysis. In: ICDM. New Orleans. pp. 1033–1038.
[7] Chen, M., S. Wang, P. P. Liang, T. Baltrušaitis, and L. P. Morency. (2017) Multimodal sentiment analysis with word-level fusion and reinforcement learning. In: Proceedings of the 19th ACM International Conference on Multimodal Interaction. New York. pp. 163–171.
[8] Zadeh, A., P. P. Liang, N. Mazumder, S. Poria, and L. P. Morency. (2018) Memory Fusion Network for Multi-view Sequential Learning. In: The Thirty-Second AAAI Conference on Artificial Intelligence. New Orleans. pp. 5634-5641.
[9] Hazarika, D., S. Poria, A. Zadeh, E. Cambria, L.P. Morency, and R. Zimmermann. (2018) Conversational memory network for emotion recognition in dyadic dialogue videos. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. New Orleans. pp. 2122–2132.

[10] Hazarika, D., S. Poria, R. Mihalcea, E. Cambria, and R. Zimmermann. (2018) Icon: Interactive conversational memory network for multimodal emotion detection. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels. pp. 2594–2604.