Sentiment Word Aware Multimodal Refinement for Multimodal Sentiment Analysis with ASR Errors

Yang Wu Yanyan Zhao Hao Yang Song Chen Bing Qin
Xiaohuan Cao Wenting Zhao

Abstract

Multimodal sentiment analysis is an increasingly popular research area and lots of models have been proposed. However, the performance of current state-of-the-art models degrade significantly when they are deployed in the real world. the author find that the main reason is real-world applications usually access text output by using automatic speech recognition(ASR) models, which may include sentiment substitution errors due to the limited performance of modern ASR models. To address this issue, the author propose sentiment aware multimodal refinement model(SWRM[1]), which can dynamics refine the sentiment substitution errors in the text output by using multimodal sentiment clue. Our improvement is to replace early fusion with tensor fusion[2], which makes multimodal feature retain both unimodal and multimodal information. Experiment results on the real-world datasets incluing MOSI-Speechbrain, MOSI-IBM, and MOSI-iFlytek, show that our method improve the performance of SWRM model.

Keywords：Multimodal sentiment analysis, ASR errors, Sentiment word, Tensor fusion.

1 Introduction

Multimodal sentiment analysis(MSA) has been an emerging research field for its potential applications in human-computer interaction. How to effectively fuse multimodal information including textual, acoustic, and visual to predict the sentiment is a very challenging problem and has been addressed by many previous studies. Some works focus on introducing additional information into the fusing model, such as the alignment information between different modal features and unimodal sentiment labels[3]. And other works consider the semantic gaps between multimodal data and adopt the adversarial learning[4] and multi-task learning[5]to map different modal features into a shared subspace.

Despite the apparent success of the current state-of-the-art models, their performance decreases sharply, when deployed in the real world, The main reason is that the input texts are provided by the ASR models, which usually are with errors because of the limitation of model capacity. To address this issue, We propose the sentiment word aware multimodal refinement (SWRM) model, which can detect the positions of the sentiment words in the text and dynamically refine the word embeddings in the detected positions by incorporating multimodal clues. We consider leveraging the multimodal sentiment information, namely the negative sentiment conveyed by the low voice and sad face, and textual context information to help the model reconstruct the sentiment semantics for the input embeddings. Specifically, we first use the sentiment word location module to detect the positions of sentiment words and meanwhile utilize the strong language model, BERT[6], to generate

the candidate sentiment words. Then we propose the multimodal sentiment word refinement module to refine the word embeddings based on the multimodal context information. The refinement process consists of two parts, filtering and adding. We apply the multimodal gating network to filter out useless information from the input word embeddings in the filtering process and use the multimodal sentiment word attention network to leverage the useful information from candidate sentiment words as the supplement to the filtered word embeddings in the adding process. Finally, the refined sentiment word embeddings are used for multimodal feature fusion.

We build three real-world multimodal sentiment analysis datasets based on the existing dataset, CMU-MOSI[7]. Specifically, we adopt three widely used ASR APIs including SpeechBrain, IBM, and iFlytek to process the original audios and obtain the recognized texts. Then, we replace the gold texts in CMU-MOSI with the ASR results and get three real-world datasets, namely MOSI-Speech, MOSI-IBM, MOSI-iFlytek.

Early fusion consists in simply concatenating multimodal features mostly at input level. This fusion approach does not allow intra-modality dynamics to be efficiently modeled. This is due to the fact that inter-modality dynamics can be more complex at input level and can dominate the learning process or result in overfitting.

To tackle this problem, We use another method to fuse three modality, named tensor fusion, which explicitly aggregates unimodal, bimodal and trimodal interactions. Experiment results on three real-world datasets and CMU-MOSI dataset show that tensor fusion method improve the performance of the SWRM model.

2 Related works

Performing the cross-modal alignment is helpful for multimodal feature fusion. Chen[8] considered that the holistic features mainly contain global information, which may fail to capture local information. Therefore, they applied the force-alignment to align the visual and acoustic features with the words and further obtained the word-level features. To effectively fuse them, they proposed the GME-LSTM(A) model, which consists of two modules, the gated multimodal embedding and the LSTM with the temporal attention. However, obtaining the word-level features needs to perform the force-alignment, which is time-consuming. To address it, Tsai[9] proposed the MulT model, which uses the crossmodal attention to align different modal features implicitly. Instead of performing the alignment in the time dimension, some works focusing on semantic alignment. Hazarika[5] considered that the semantic gaps between heterogeneous data could hurt the model performance and proposed the MISA model, which maps the different modal data into a shared space before multimodal feature fusion. Wu[10] first utilized the cross-modal prediction task to distinguish the shared and private semantics of non-textual modalities compared to the textual modality and then fuse them. The above works show that performing the cross-modal alignment is helpful for multimodal feature fusion.

Training the MSA models in an end-to-end manner is more effective.Most of the previous studies adopt a two-phase pipeline, first extracting unimodal features and then fusing them. Dai[11] considered that it may lead to suboptimal performance since the extracted unimodal features are fixed and cannot be further

improved benefiting from the downstream supervisory signals. Therefore, they proposed the multimodal endto-end sparse model, which can optimize the unimodal feature extraction and multimodal feature fusion jointly. The experimental results on the multimodal emotion detection task show that training the models in an end-to-end manner can obtain better results than the pipeline models.

Leveraging the unimodal sentiment labels to learn more informative unimodal representations is useful for multimodal feature fusion.Yu[12] considered that introducing the unimodal sentiment labels can help the model capture the unimodal sentiment information and model the difference between modalities. Motivated by it, they built the CH-SIMS dataset, which contains not only the multimodal sentiment labels but also unimodal sentiment labels. And based on it, they proposed a multi-task learning framework to leverage two types of sentiment labels simultaneously. However, this method needs unimodal labels, which is absent for most of the existing datasets. To address it, Yu[3] proposed the Self-MM model, which first generates the unimodal labels by utilizing the relationship between the unimodal and multimodal labels and then uses the multi-task learning to train the model. These two works both address the usefulness of introducing unimodal labels.

Comparing to the above works, we evaluate the SOTA MSA models on the real-world datasets and observe that the performance of models decreases sharply because of the erroneous ASR texts. Through in-depth analysis of the ASR outputs, we find the sentiment word substitution error in the ASR texts could hurt the MSA models directly. To address it, we propose the sentiment word aware multimodal refinement model, which only uses the ASR texts in the training and testing phrases.

3 Method

Our model consists of three modules including the sentiment word location module, multimodal sentiment word refinement module, and multimodal feature fusion module. We first use the sentiment word location module to detect the possible positions of sentiment words and then utilize the multimodal sentiment word refinement module to dynamically refine the word embeddings in the detected positions. Finally, the refined word embeddings are fed into the multimodal feature fusion module to predict the final sentiment labels.

3.1 Overview

SWRM model Figure1：

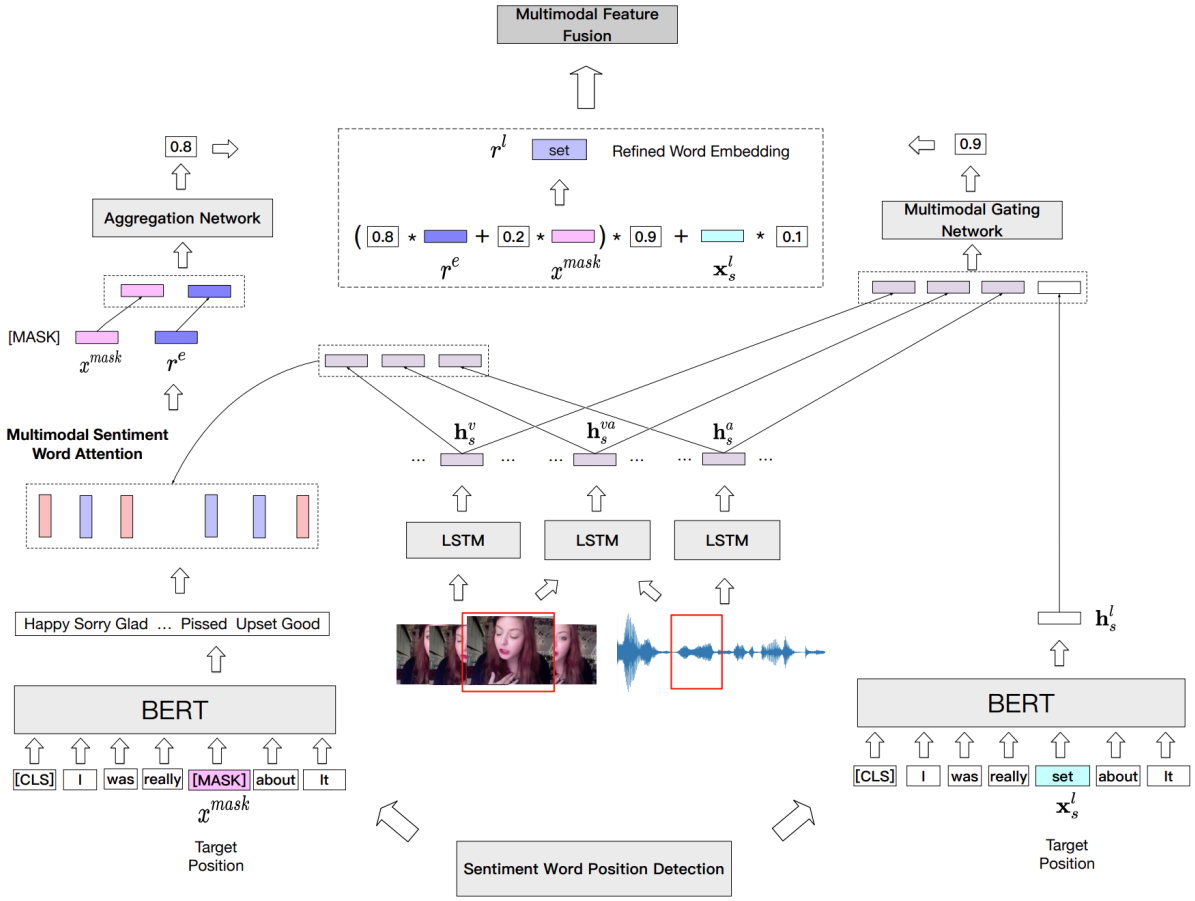

Figure 1: Overview of the model

3.2 Sentiment Word Position Detection

The core idea of the sentiment word position detection module is to find out the possible positions of sentiment words in the ASR texts. To achieve it, we consider adopting a powerful language model, since language model can model the context information of the sentiment words such as syntactic and grammatical information and predict the appropriate words for the target position. Specifically, we choose the BERT model as our language model since the masked language modeling pretraining objective meets our needs perfectly.

Given a sentence, we first mask every word in the sentence sequentially, and then we use the BERT pre-train model to predict the possible words in the position of the masked word. we will get the Top-K candidate words. Next, we sort the candidate words by prediction probabilities and get the most possible position of the candidate word by using sentiment lexicons. Considering that in some cases there is not a sentiment word in the candidate word, we use a sentiment threshold to filter out the impossible position of the sentence.

3.3 Multimodal Sentiment Word Refinement

To refine the ASR error texts, we propose the multimodal sentiment word refinement module, which refine the word embeddings of sentiment words from two aspects. One is that we use multimodal gating network to filter out the useless information from the input word embeddings. The another one is that we use multimodal sentiment attention network to incorporate the useful information from candidate sentiment word produced by BERT model.

Given an utterance, which includes three modal unaligned features, word embeddings, acoustic features, and visual features. Firstly, we utilize the pseudo-alignment method to make acoustic and visual features

corresponding to word embeddings. Then, we apply the BERT model and LSTM networks to encode the features and get context-aware representations. In order to capture high level sentiment semantics, we also use an LSTM networks to fuse acoustic and visual features to get acoustic-visual bimodal features.

Subsequently, we propose the multimodal gating network to filter the word embeddings, which implemented by a non-linear layer. Specifically, we cancatenate the unimodal context-aware representations and bimodal representation in the same position and feed them into a non-linear neural network, producing the gate value, which is used to filter out the useless information from the word embedding. To make the model ignore the impossible one, we use the gate mask to achieve it.

Furthermore, we propose a novel multimodal sentiment attention network to leverage the sentiment-related information from the candidate words to complement the word embeddings. Specifically, we use a linear layer to implement the multimodal sentiment word attention network. In addition, there may not be suitable sentiment words in the candidate words, we incorporate the embedding of the special word [mask] to let the BERT model handle this problem based on the context. Then, we design an aggregation network to balance the contributions of the special word embedding mask and the sentiment embedding. Finally, we obtain the refined word embedding.

3.4 Multimodal Feature Fusion

It is noted that our proposed refinement approach only modifies the textual input token embeddings, which makes it easy to be adapted for other multimodal feature fusion models based on BERT.

The multimodal feature fusion method described in the original paper is early fusion. The most common early fusion operation is feature concatenate, Specifically, concatenate textual, acoustic, and visual features to obtain trimodal features and then feed into a non-linear layer to capture the interactions between them. This fusion approach does not allow the intra-modality dynamics to be effectively modeled. Hence, we use another fusion method, named tensor fusion, which explicitly aggregates unimodal, bimodal, and trimodal interactions. This fusion approach learns both inter-modality and intra-modality dynamics end-to-end.

4 Implementation details

4.1 Comparing with released source codes

This section shows the difference between the source code and our implemented method.

```
fusion_h = torch.cat([text, audio, video], dim=-1)
fusion_h = self.post_fusion_dropout(fusion_h)
fusion_h = F.relu(self.post_fusion_layer_1(fusion_h), inplace=False)
```

Figure 2: Early fusion in source code

```
# implement tensor-fusion method for feature-fusion
add_one = torch.ones(size=[audio.shape[0], 1], requires_grad=False).type_as(audio).cuda()
_audio = torch.cat((add_one, audio), dim=1)
_video = torch.cat((add_one, video), dim=1)
_text = torch.cat((add_one, text), dim=1)
fusion_h = torch.bmm(_audio.unsqueeze(2), _video.unsqueeze(1))
fusion_h = fusion_h.view(-1, (audio.shape[1] + 1) * (video.shape[1] + 1), 1)
fusion_h = torch.bmm(fusion_h, _text.unsqueeze(1)).view(audio.shape[0], -1)
fusion_h = self.post_fusion_dropout(fusion_h)
fusion_h = F.relu(self.post_fusion_layer_1(fusion_h), inplace=False)
```

Figure 3: Implement tensor fusion

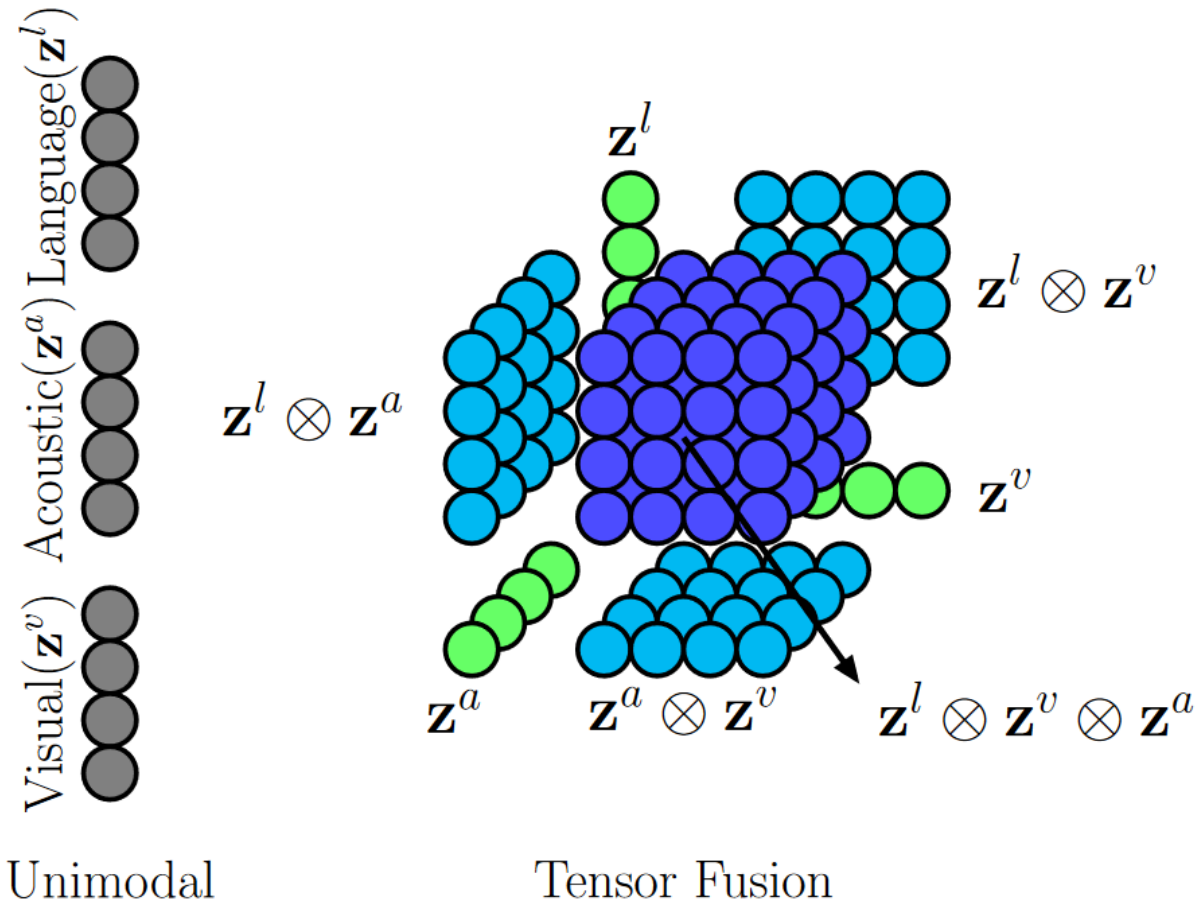

Figure 4: Tensor fusion with three types of subtensors: unimodal, bimodal and trimodal

We use tensor fusion approach for multimodal feature fusion. Specifically, we concatenate the textual, visual, acoustic representation with one vector. After using three-fold Cartesian product, the output features contain both unimodal, bimodal, and trimodal dynamics.

In Figure 4, The green subregions represent unimodal embeddings, the light blue subregions represent bimodal interactions, and the deep blue subregions represent trimodal interactions.

Procedure 1 Multimodal Sentiment Word Refinement.

Input: acoustic x^a, visual x^v, language x^l

Output: acoustic representation v^a, visual representation v^v, language representation v^l

for *i **in** batch data* **do**

$$u^v = \text{Pseudo}(x^v) \qquad u^a = \text{Pseudo}(x^a)$$

$$h^l = \text{BERT}(x^l) \qquad h^v = \text{LSTM}_\text{v}(u^v) \qquad h^a = \text{LSTM}_\text{a}(u^a) \qquad h^{va} = \text{LSTM}_\text{va}([u^v; u^a])$$

$$g^v = \text{Sigmoid}\left(W_1\left(\left[\mathbf{h}_s^l; \mathbf{h}_s^v; \mathbf{h}_s^a; \mathbf{h}_s^{va}\right]\right) + b_1\right) \qquad r^v = (1 - g^v p)\,\mathbf{x}_s^l$$

$$g_t^e = W_2\left(\left[x^{c_t^s}; \mathbf{h}_s^v; \mathbf{h}_s^a; \mathbf{h}_s^{va}\right]\right) + b_2 \qquad w_t^e = \frac{e^{g_t^e}}{\sum_{t=1}^k e^{g_t^e}} \qquad r^e = \textstyle\sum_{t=1}^k w_t^e x^{c_t^s}$$

$$g^{\text{mask}} = \text{Sigmoid}\left(W_3\left(\left[r^e; x^{\text{mask}}\right]\right) + b_3\right) \qquad r^{add} = g^{mask} r^e + \left(1 - g^{mask}\right) x^{\text{mask}}$$

$$r^l = (g^v p)\, r^{add} + r^v$$

$$\mathbf{z^l} = \left\{x_1^l, x_2^l, \ldots, r^l, \ldots, x_{n_l}^l\right\}, \text{ where } \mathbf{x}^l = \left\{x_t^l : 1 \leqslant t \leqslant n_l, x_t^l \in \mathbb{R}^{d_x^l}\right\}$$

end

for *i **in** refined word embeddings* **do**

$$v^l = \text{BERT}_{\text{textual}}\left(\mathbf{z}^{\text{l}}\right) \qquad v^v = \text{LSTM}_{\text{visual}}\left(\mathbf{x}^{\text{v}}\right) \qquad v^a = \text{LSTM}_{\text{acoustic}}\left(\mathbf{x}^{\text{a}}\right)$$

end

Procedure 2 Tensor Fusion for Multimodal Feature Fusion.

for *i **in** unimodal representation* **do**

Input: acoustic representation v^a, visual representation v^v, language representation v^l

Output: Probability p^f

$$\mathbf{v}^m = \begin{bmatrix} \mathbf{v}^l \\ 1 \end{bmatrix} \otimes \begin{bmatrix} \mathbf{v}^v \\ 1 \end{bmatrix} \otimes \begin{bmatrix} \mathbf{v}^a \\ 1 \end{bmatrix}$$

$$v^f = \text{Relu}\left(W_4 \mathbf{v}^m + b_4\right)$$

$$p_f = W_5 v^f + b_5$$

end

4.2 Experimental environment setup

We use Adam as the optimizer and the learning rate is 5e-5. The batch size is 64. The sentiment threshold is set to 0.5 while detecting the sentiment word position. The number of the candidate words is 50. We run five times and report the average performance. The random seeds we used are 1111, 1112, 1113, 1114, and 1115.

4.3 Case study

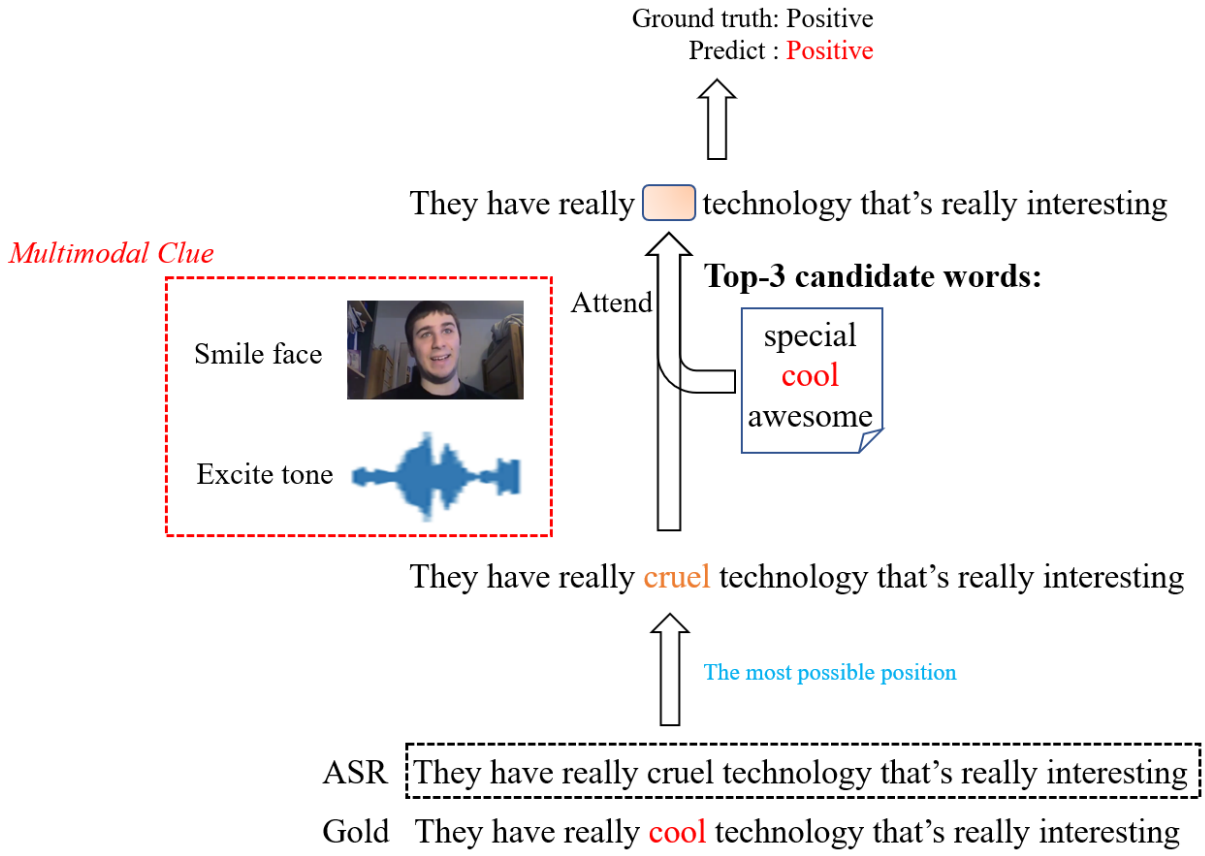

Figure 5: Case study for the SWRM

4.4 Main contributions

The main contributions of this work are as follows:

(1) We reproduce the main structure of the sentiment word aware refinement module for multimodal sentiment analysis and evaluate the reproduced model on three real-world datasets and CMU-MOSI dataset.

(2) To improve the performance of the reproduced model, We use tensor fusion approach to fuse three modality for multimodal feature fusion and evaluate the improved model on three real-world datasets and CMU-MOSI dataset. Experiment results show that the improved model outperform the origin model and reproduced model in some evaluation metrics.

5 Results and analysis

Datasets	Models	Evaluation Metrics					
		Has0-Acc ↑	Has0-F1 ↑	Non0-Acc ↑	Non0-F1 ↑	MAE ↓	Corr ↑
MOSI-SpeechBrain	论文	74.58	74.62	75.70	75.82	90.56	67.47
	复现	74.49	74.55	75.70	75.84	91.84	67.2
	改进	**74.96**	**75.0**	**76.31**	**76.44**	**90.56**	**67.86**
MOSI-IBM	论文	**78.43**	**78.47**	79.70	79.80	82.91	73.91
	复现	78.22	78.25	79.66	79.76	83.47	73.35
	改进	78.37	78.34	**79.91**	**79.95**	**82.12**	**74.0**
MOSI-iFlytek	论文	**80.47**	**80.47**	**81.28**	**81.34**	**78.39**	**75.97**
	复现	80.26	80.23	81.31	81.33	77.19	75.78
	改进	80.38	80.42	81.16	81.25	80.05	75.58
MOSI-Gold	复现	83.2	83.15	84.85	84.84	**72.08**	78.82
	改进	**83.56**	**83.53**	**85.03**	**85.05**	72.52	**79.04**

Figure 6: Experimental results

(1) In MOSI-SpeechBrain, MOSI-IBM, and MOSI-Gold datasets, our improved model outperform the origin SWRM model and the reproduced model in most evaluation metrics.

(2) In MOSI-iFlytek dataset, our improved model outperform the reproduced model in Has0-Acc, Has0-F1 metrics.

6 Conclusion and future work

In this work, we reproduce the sentiment word aware refinement module of the SWRM model, this module refine sentiment word ASR errors by introducing multimodal sentiment clues. Further more, for multimodal sentiment analysis, multimodal feature fusion is essential, we use tensor fusion method to fuse three modality(acoustic, visual, and refined textual) for multimodal feature fusion. Experiment results show that this fusion method improve the performance of the SWRM model. For future work, we will continue to explore the potential of this fusion approach.

References

[1] WU Y, ZHAO Y, YANG H, et al. Sentiment Word Aware Multimodal Refinement for Multimodal Sentiment Analysis with ASR Errors[C]//Findings of the Association for Computational Linguistics: ACL 2022. Dublin, Ireland: Association for Computational Linguistics, 2022: 1397-1406.

[2] ZADEH A, CHEN M, PORIA S, et al. Tensor Fusion Network for Multimodal Sentiment Analysis[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark: Association for Computational Linguistics, 2017: 1103-1114.

[3] YU W, XU H, YUAN Z, et al. Learning Modality-Specific Representations with Self-Supervised Multi-Task Learning for Multimodal Sentiment Analysis[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(12): 10790-10797.

[4] MAI S, HU H, XING S. Modality to Modality Translation: An Adversarial Representation Learning and Graph Fusion Network for Multimodal Fusion[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(01): 164-172.

[5] HAZARIKA D, ZIMMERMANN R, PORIA S. MISA: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis[C]//MM '20: Proceedings of the 28th ACM International Conference on Multimedia. Seattle, WA, USA: Association for Computing Machinery, 2020: 1122-1131.

[6] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics, 2019: 4171-4186.

[7] ZADEH A, ZELLERS R, PINCUS E, et al. MOSI: Multimodal Corpus of Sentiment Intensity and Subjectivity Analysis in Online Opinion Videos[J]. CoRR, 2016, abs/1606.06259. arXiv: 1606.06259.

[8] CHEN M, WANG S, LIANG P P, et al. Multimodal Sentiment Analysis with Word-Level Fusion and Reinforcement Learning[C]//ICMI '17: Proceedings of the 19th ACM International Conference on Multimodal Interaction. Glasgow, UK: Association for Computing Machinery, 2017: 163-171.

[9] TSAI Y H H, BAI S, LIANG P P, et al. Multimodal Transformer for Unaligned Multimodal Language Sequences[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, 2019: 6558-6569.

[10] WU Y, LIN Z, ZHAO Y, et al. A Text-Centered Shared-Private Framework via Cross-Modal Prediction for Multimodal Sentiment Analysis[C]//Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Online: Association for Computational Linguistics, 2021: 4730-4738.

[11] DAI W, CAHYAWIJAYA S, LIU Z, et al. Multimodal End-to-End Sparse Model for Emotion Recognition[C]//Proceedings of the 2021 Conference of the North American Chapter of the Association for

Computational Linguistics: Human Language Technologies. Online: Association for Computational Linguistics, 2021: 5305-5316.

[12] YU W, XU H, MENG F, et al. CH-SIMS: A Chinese Multimodal Sentiment Analysis Dataset with Fine-grained Annotation of Modality[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, 2020: 3718-3727.