ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10446351

MULTI-GRAINED MULTIMODAL INTERACTION NETWORK FOR SENTIMENT ANALYSIS

Lingyong Fang[1], *Gongshen Liu*[1,*], *Ru Zhang*[2]

[1]Shanghai Jiao Tong University
[2]Beijing University of Posts and Telecommunications

ABSTRACT

Multimodal sentiment analysis aims to utilize different modalities including language, visual, and audio to identify human emotions in videos. Multimodal interaciton mechanism is the key challenge. Previous works lack modeling of multimodal interaction at different grain levels, and does not suppress redundant information in multimodal interaction. This leads to incomplete multimodal representation with noisy information. To address these issues, we propose Multi-grained Multimodal Interaction Network (MMIN) to provide a more complete view of multimodal representation. Coarse-grained Interaction Network (CIN) exploits the unique characteristics of different modalities at a coarse-grained level and adversarial learning is used to reduce redundancy. Fine-grained Interaction Network (FIN) employ sparse-attention mechanism to capture fine-grained interactions between multimodal sequences across distinct time steps and reduce irrelevant fine-grained multimodal interaction. Experimental results on two public datasets demonstrate the effectiveness of our model in multimodal sentiment analysis.

Index Terms— Multimodal Sentiment Analysis, Multimodal Fusion

1. INTRODUCTION

With a large amount of user-generated online content, such as videos, multimodal sentiment analysis (MSA) has received increasing attention in recent years and has important applications in human-computer interaction, video understanding, risk management, and other fields. Unlike unimodal sentiment analysis tasks, multimodal models can utilize different sources of information, such as language, visual, and audio, which facilitate the understanding of human emotions and intentions. How to model the interaction between this heterogeneous information is a major challenge.

Some previous works align different modality sequences based on word boundaries and then fused them based on the aligned sequences [1–3]. However, manual word-alignment process requires additional labor costs and time costs, and neglect long term dependencies across modalities. Therefore, recent studies [4–6] have focused on the fusion of unaligned sequence data. Tsai et al. [4] capture long-range dependencies between modalities through the attention mechanism. In addition to attention mechanism, some previous works [7–9] attempt to learn reliable cross-modal interactions over modality-invariant subspace where the distribution is bridged.

Although previous studies propose innovative multimodal interaction techniques, they only focus on the correlation between words, video frames, and audio frames, neglecting the importance of global semantics; Or they only employ a post-fusion approach for multimodal features, inevitably omitting some significant details. Considering the process of human multimodal sentiment analysis, humans initially analyze words, videos, and audio frames chronologically. Following this, they conduct an review for a holistic judgement. This is because certain emotional information can only be discerned in the global representation, and cannot be extracted from the fine-grained elements like individual words or transient expressions and tones. In addition, the excessive redundancy and noise in different modalities can adversely affect multimodal interaction and the accuracy of sentiment analysis tasks. For instance, video frames that lack emotional information or are unrelated to the text can pose difficulties for multimodal interaction.

To deal with these issues, in this paper, we propose Multi-grained Multimodal Interaction Network (MMIN) to improve the performance of multimodal sentiment analysis. The Coarse-grained Interaction Network (CIN) is aimed to extract modality-specific representations that capture distinctive characteristics of each modality from a coarse-grained perspective. We use an adversarial learning strategy to mitigate redundancy across modalities. The Fine-grained Interaction Network (FIN) is designed to extract cross-modal representations, which can capture the fine-grained semantic association between multimodal sequences across distinct time steps. We employ sparse-attention mechanism to alleviate the negative effects of redundant features during multimodal fine-grained interaction. Then we fuse representations of different grain

*Corresponing author. This research work has been funded by National Key RD Program of China(Grant No.2023YFC3303805), Joint Funds of the National Natural Science Foundation of China (Grant No. U21B2020), and Shanghai Science and Technology Plan (Grant No. 22511104400).

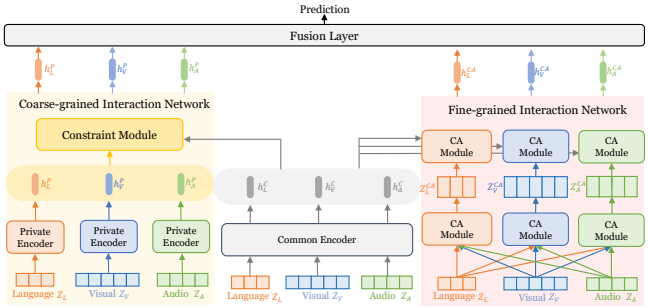
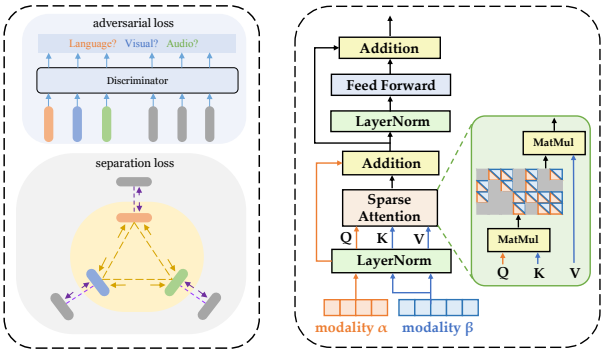


Fig. 1. Overview architecture of the Multi-grained Multimodal Interaction Network. The details of the constraint module and CA module are shown in Fig.2



(a) Constraint Module

(b) Cross-Attention Module (CA Module)

Fig. 2. The details of the modules.

levels to provide more comprehensive multi-grained information for multimodal sentiment analysis. Experiments on public multimodal sentiment benchmark datasets confirm the validity of our approach.

2. METHODOLOGY

2.1. Feature Extraction

For language modality, we feed the input text into BERT to obtain the language feature. For video and audio modalities, we use LSTM to capture the intra-modality interaction.

2.2. Coarse-grained Interaction Network

Encoder. We use the private encoder to extract modality-specific representation separately. For the language modality, we select the [CLS] vector in the BERT output as the input to the private encoder, and for the video and speech modalities, we select the hidden representations of the LSTM end state as the input to the private encoder. The inputs of each modality are denoted as h_L, h_V, and h_A.

$$h_m^P = \text{Private}_m(h_m; \theta_m^P), \quad m \in \{L, V, A\} \tag{1}$$

We employ common encoder to obtain modality-invariant representation. This encoder is also composed of fully-connected layers. In contrast to the private encoder, the parameters in the common encoder are shared for different modalities.

$$h_m^C = \text{Common}(h_m; \theta^C), \quad m \in \{L, V, A\} \tag{2}$$

These are coarse-grained interaction representations that express the overall information of the modalities.

Constraint Module. As shown in Fig.2(a), we constrain the different representations extracted by the encoder through the constraint module to extract purer modality-specific representations. Specifically, we constrain the outputs of the private encoder and common encoder by a series of loss functions.

The adversarial loss is used to make h_m^C reflect the common features of different modalities and to ensure that h_m^P can extract the unique information of each modality. We achieve this goal through a discriminator-based adversarial network. The discriminator $\mathcal{D}$ is a multi-class classifier, which identify which modality the input feature belongs to. For h_m^C, we hope that the discriminator cannot distinguish, which means that the modality-invariant representation belongs to the latent space shared by different modalities. According to the gradient reversal layer [10], we designed an adversarial loss for h_m^C.

$$\mathcal{L}_{ac} = -\frac{1}{n}\sum_{m}\sum_{i=1}^{n}\left(y_m \log\left(\mathcal{D}\left(h_m^C; \theta_{\mathcal{D}}\right)\right)\right) \tag{3}$$

where , $y_L = [1, 0, 0], y_V = [0, 1, 0], y_A = [0, 0, 1]$.

For h_m^P, we hope that the discriminator can accurately distinguish, so as to represent the personal information of the modality for the modality-specific representation. The adversarial loss of h_m^P is represented as:

$$\mathcal{L}_{ap} = -\frac{1}{n}\sum_{m}\sum_{i=1}^{n}\left(y_m \log\left(\mathcal{D}\left(h_m^P; \theta_{\mathcal{D}}\right)\right)\right) \tag{4}$$

The separation loss $\mathcal{L}_{sep}$ is used to make the modality-specific representation contain purer features specific to the different modalities. We reduce the redundancy between the modality-specific representation and the modality-invariant representation of the corresponding modality to ensure that the two representations extract different aspects of the modality while reducing the redundancy between modality-specific representations helps different private encoders to extract information unique to each modality. Previous studies [7, 11] have shown that non-redundancy effects can be achieved by applying orthogonality constraints to different representation vectors. When training a batch of multimodal input, $\mathbf{H}_m^P$ and $\mathbf{H}_m^C$ are two matrices, each row of which is a modality-specific representation h_m^P and a modality-invariant representation h_m^C of different modalities respectively.

$$\mathcal{L}_{sep} = \sum_{m} \left\| \mathbf{H}_m^{C^\top} \mathbf{H}_m^P \right\|_F^2 + \sum_{(m_1,m_2)} \left\| \mathbf{H}_{m_1}^{P^\top} \mathbf{H}_{m_2}^P \right\|_F^2 \quad (5)$$

where $\| \cdot \|_F^2$ is the squared Frobenius norm.

2.3. Fine-grained Interaction Network

To obtain fine-grained interaction, we designed the Interaction Network to exploit long term dependencies between elements across modalities.

Cross-Attention Module (CA Module). Following previous work [4, 5] about multimodal fusion, we employ a transformer variant for unaligned multimodal interaction. As shown in Fig. 2(b), CA Module includes multi-head layer normalization, cross-attention, and residual connection. We perform multi-modal fusion by passing fine-grained features of different modalities to N layers of stacked Cross-Attention Module (CA Module). For each layer of the CA Module, we use cross-modal sparse attention to perform fine-grained multi-modal interactions on different modal features X_α, X_β. We use sparsemax [12] for the normalization of attention weights, which leads to acquisition of sparse posterior attention weights. This causes the weights of redundant modality features to be assigned a value of zero.

Coarse-grained Representation Guided Interaction. There is heterogeneity between different modalities, and modality-invariant representation can guide the modality to focus on the most important interaction information between different modalities. Additionally, coarse-grained information can help fine-grained information reduce noise. We also implement the interaction process through CA Module. The modality-invariant representations of the different modalities are stacked together as a query, and the other modalities interact one by one as key and value. The corresponding vectors in the CA module output sequence are used as fine-grained interaction representations.

2.4. Fusion Layer

With the above CIN and FIN, we obtain the coarse-grained interaction representations and the fine-grained interaction representations. These representations are stacked into matrix and then fed into the transformer layer. The individual vectors in the output are concatenated together and fed into fully-connected layers for sentiment prediction. We use mean squared error loss to evaluate the quality of sentiment prediction and the loss function is denoted as $\mathcal{L}_{senti}$.

The final loss function is expressed as follows:

$$\mathcal{L} = \mathcal{L}_{senti} + \alpha \mathcal{L}_{sep} + \beta \mathcal{L}_{ac} + \gamma \mathcal{L}_{ap} \quad (6)$$

where α, β and γ are the trade-off parameters.

3. EXPERIMENTS

3.1. Datasets, Metrics and Settings

We evaluate our proposed model on CMU-MOSI [13] and CMU-MOSEI [14]. Following the previous works [4, 7, 15], we utilize four metrics to evaluate the performance of the proposed model. For the binary sentiment classification task, we report binary classification accuracy (Acc-2) and weighted F1 score (F1-Score). For the regression task, we report mean absolute error (MAE) and Pearson correlation (Corr).

To extract the features of visual modalities, video frames are processed by Facet [16] to generate a set of features consisting of 35 facial action units. To extract the features of audio modality, COVAREP [17] is utilized for generating features of acoustic signals. We use uncased BERT-Base as the pre-trained BERT in our model to extract the feature of language modality.

3.2. Comparison with Baselines

To evaluate the rationality and effectiveness of our methods, We compare proposed model with the following recent and competitive baselines: TFN [18], CIA [19], ICCN [20], PMR [5], MulT [4], MISA [7], MAG-BERT [21].

The results with baselines on the two datasets are shown in Table 1. We classify "Data Setting" into two categories: Unaligned and Aligned. The Aligned setting requires an additional step of manually aligning signals from different modalities based on word boundaries. In contrast, the Unaligned setting directly employs unaligned sequence data for multimodal fusion. Performance is generally better in aligned settings. It can be observed that the proposed MMIN model outperforms other models and obtains the best performance across the two datasets in all evaluation metrics.

Compared to MAG-BERT [21], which requires alignment settings, our model achieves better performance and does not require pre-alignment, thus reducing labor and time costs. MISA [7] emphasize the modality-specific representations, but ignore the fine-grained interactions of different modes in temporal order, and therefore some critical information may be lost, resulting in inferior performance to ours. In comparison to MulT [4] and PMR [5] which also use attention mechanism to fully exploit long-term dependencies across modalities, our approach validates the integration of modality-specific representations into the model and achieves superior results. The above observations suggest that it is beneficial to consider both fine-grained interaction information and modality-specific representations by CIN and FIN in multimodal sentiment analysis.

3.3. Ablation Study

To further explore the contributions of SIN, we conduct comprehensive ablation studies using the unaligned version of

Table 1. Comparison with baselines on CMU-MOSI and CMU-MOSEI benchmark

Model	MOSI				MOSEI				Data Setting
	MAE	Corr	Acc-2	F1-Score	MAE	Corr	Acc-2	F1-Score	
TFN*	0.901	0.698	-/80.8	-/80.7	0.593	0.700	-/82.5	-/82.1	Unaligned
CIA*	0.914	0.689	79.8/-	79.5/-	0.680	0.590	80.4/-	78.2/-	Aligned
ICCN*	0.860	0.710	-/83.0	-/83.0	0.565	0.713	-/84.2	-/84.2	Aligned
PMR†	-	-	-/83.6	-/83.4	-	-	-/82.4	-/82.1	Unaligned
MulT*	0.871	0.698	-/83.0	-/82.8	0.580	0.703	-/82.5	-/82.3	Unaligned
MISA*	0.783	0.761	81.8/83.4	81.7/83.6	0.555	0.756	83.6/85.5	83.8/85.3	Aligned
MAG-BERT‡	0.748	0.790	82.61/84.42	82.59/84.71	0.548	0.757	82.63/84.84	82.62/84.86	Aligned
MMIN (Ours)	**0.741**	**0.795**	**83.53/85.52**	**83.46/85.51**	**0.542**	**0.761**	**83.84/85.88**	**83.91/85.76**	Unaligned

*: from [7]; †: from [5]. Models with ‡ are reproduced under the same conditions. For Acc-2 and F1-Score, we use the segmentation marker -/- to report results, where the the left-side score is calculated as "negative/non-negative", while the right-side score is calculated as "negative/positive"

Table 2. Ablation Studies on CMU-MOSI Dataset.

Ablation	MAE	Corr	Acc-2	F1-Score
Role of Representations				
w/o CIN	0.773	0.777	82.07/83.84	82.03/83.86
w/o FIN	0.769	0.783	82.43/84.27	82.31/84.45
Role of Modality				
w/o Text	1.313	0.554	75.80/78.05	75.22/77.62
w/o Video	0.775	0.786	81.92/83.99	81.86/84.00
w/o Audio	0.801	0.789	81.63/83.54	81.68/83.63
Role of Regularization				
w/o $\mathcal{L}_{sep}$	0.755	0.785	82.49/84.38	82.50/84.45
w/o $\mathcal{L}_{ac}$	0.784	0.763	82.97/83.80	81.98/83.76
w/o $\mathcal{L}_{ap}$	0.754	0.789	82.94/84.45	82.94/84.50
MMIN(Full)	**0.741**	**0.795**	**83.53/85.52**	**83.46/85.51**

CMU-MOSI. The results are shown in Table 2.

Role of Representations. First, we removed CIN and FIN respectively to examine the validity of the two different types of representations in the proposed model. The process of representation learning is preserved when performing the experiments, and only parts of the representations are employed in the final prediction phase. Experimental results show that either removal of CIN or FIN leads to the degradation of model performance. This indicates that both networks are necessary and meaningful. CIN provides information on the individuality of the different modalities while FIN provides information on the fine-grained interactions of temporal order between the modalities. The two distinct representations complement each other to improve the model performance.

Role of Modality. We explore the effect of different modalities on our model performance. When performing the experiments, we remove the corresponding modality in both CIN and FIN. It can be observed that the removal of any modality leads to performance degradation, indicating that each modality contributes to the model and our model is able to fully exploit the value of each modality to the network. The model performance decreases significantly after moving out of the text modality, probably because the text modality contains more information and is more critical for multimodal sentiment analysis tasks.

Role of Regularization. In order to explore the role of different regularizations, we removed each loss function separately to perform the experiment. We observe that all three loss functions improve model performance. When $\mathcal{L}_{sep}$ is removed, it is difficult for the model to extract modality-specific representation from the modalities. The representation contains redundant noise which leads to the decrease of model performance. $\mathcal{L}_{ac}$ can help extract common information from the model and thus assist the model in identifying individual information, therefore, the lack of $\mathcal{L}_{ac}$ has a negative impact on the network training. When $\mathcal{L}_{ap}$ is moved out, the modality-specific representation learned by the model may be trivial, causing bad performance.

4. CONCLUSION

In this paper, we introduce Multi-grained Multimodal Interaction Network (MMIN) for multimodal sentiment analysis. Coarse-grained Interaction Network extracts modality-specific representation that capture the distinctive characteristics between modalities while Fine-grained Interaction Network extracts cross-modal representation that learns correlations between elements from different modalities. Our comprehensive experiments demonstrated the superiority of MMIN. In addition, our approach can be extended to other multimodal applications.

5. REFERENCES

[1] Yue Gu, Kangning Yang, Shiyu Fu, Shuhong Chen, Xinyu Li, and Ivan Marsic, "Multimodal affective analysis using hierarchical attention strategy with word-level alignment," in *Proc. of ACL*, 2018, vol. 2018, p. 2225.

[2] Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency, "Words can shift: Dynamically adjusting word representations using nonverbal behaviors," in *Proc. of AAAI*, 2019, vol. 33, pp. 7216–7223.

[3] Wei Han, Hui Chen, and Soujanya Poria, "Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis," in *EMNLP*. Nov. 2021, pp. 9180–9192, Association for Computational Linguistics.

[4] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proc. of ACL*, 2019, pp. 6558–6569.

[5] Fengmao Lv, Xiang Chen, Yanyong Huang, Lixin Duan, and Guosheng Lin, "Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences," in *Proc. of CVPR*, 2021, pp. 2554–2562.

[6] Lingyong Fang, Gongshen Liu, and Ru Zhang, "Sense-aware bert and multi-task fine-tuning for multimodal sentiment analysis," in *2022 International Joint Conference on Neural Networks (IJCNN)*, July 2022, pp. 1–8.

[7] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria, "Misa: Modality-invariant and -specific representations for multimodal sentiment analysis," in *Proc. of ACM MM*, 2020, pp. 1122–1131.

[8] Dingkang Yang, Haopeng Kuang, Shuai Huang, and Lihua Zhang, "Learning modality-specific and -agnostic representations for asynchronous multimodal language sequences," in *Proc. of ACM MM*, 2022, pp. 1708–1717.

[9] Dingkang Yang, Shuai Huang, Haopeng Kuang, Yangtao Du, and Lihua Zhang, "Disentangled representation learning for multimodal emotion recognition," in *Proc. of ACM MM*, 2022, pp. 1642–1651.

[10] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky, "Domain-adversarial training of neural networks," *Journal of Machine Learning Research*, vol. 17, no. 59, pp. 1–35, 2016.

[11] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan, "Domain separation networks," in *Proc. of NeurIPS*, 2016, vol. 29.

[12] Andre Martins and Ramon Astudillo, "From softmax to sparsemax: A sparse model of attention and multi-label classification," in *Proceedings of The 33rd International Conference on Machine Learning*. June 2016, pp. 1614–1623, PMLR.

[13] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency, "Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages," *IEEE Intelligent Systems*, vol. 31, no. 6, pp. 82–88, 2016.

[14] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency, "Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph," in *Proc. of ACL*, 2018, pp. 2236–2246.

[15] Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu, "Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis," *Proc. of AAAI*, vol. 35, no. 12, pp. 10790–10797, 2021.

[16] iMotions, "Facial expression analysis," 2017.

[17] Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer, "Covarep—a collaborative voice analysis repository for speech technologies," in *Proc. of ICASSP*, 2014, pp. 960–964.

[18] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proc. of EMNLP*, 2017, pp. 1103–1114.

[19] Dushyant Singh Chauhan, Md Shad Akhtar, Asif Ekbal, and Pushpak Bhattacharyya, "Context-aware interactive attention for multi-modal sentiment and emotion analysis," in *Proc. of EMNLP*, 2019, pp. 5647–5657.

[20] Zhongkai Sun, Prathusha Sarma, William Sethares, and Yingyu Liang, "Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis," *Proc. of AAAI*, pp. 8992–8999, 2020.

[21] Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, AmirAli Bagher Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque, "Integrating multimodal information in large pretrained transformers," in *Proc. of ACL*, 2020, pp. 2359–2369.