

Multimodal Sentiment Analysis with Missing Modality: A Knowledge-Transfer Approach

Weide Liu* and Huijing Zhan†

* College of Computing and Data Science, Nanyang Technological University, Singapore

E-mail: weide001@e.ntu.edu.sg

† Singapore University of Social Sciences, Singapore

E-mail: hjzhan@suss.edu.sg

Abstract—Multimodal sentiment analysis aims to identify the emotions expressed by individuals through visual, language, and acoustic cues. However, most existing research assume that all modalities are available during both training and testing, which makes their algorithms susceptible to the missing-modality scenarios. In this paper, we propose a novel knowledge-transfer network to translate between different modalities to reconstruct the missing audio features. Moreover, we develop a cross-modality attention mechanism to maximize the information extracted from the reconstructed and observed modalities for sentiment prediction. Extensive experiments on three publicly available datasets demonstrate significant improvements over baseline methods and achieve comparable results to the previous methods with complete multi-modality supervision.

I. INTRODUCTION

With the rapid advancement of deep learning [1], [2], multimodal sentiment analysis (MSA) has emerged as a growing area of research in recent years, as it allows for a more comprehensive and effective understanding of an individual's emotions. Multiple sources of information are utilized, including visual, language, and acoustic modalities, which together provide a more complete picture of an individual's emotional state.

Recent works on MSA [3]–[12] have focused on developing effective methods for integrating and utilizing multi-modality information, under the assumption that all modalities are available during both training and testing. However, in real-world scenarios, missing modality is a common problem due to privacy concerns or technical difficulties. Especially in situations such as online meetings and network sharing, where data is frequently uploaded and downloaded, modalities can be missing during transmission. In these cases, it becomes essential to reconstruct missing modalities using the information from observed modalities.

Previous research studies [13], [14] have attempted to address the issue of missing modalities in multimodal sentiment analysis. In particular, Tsai *et al.* [14] proposed a joint generative-discriminative objective to obtain a robust multimodal representation and a surrogate inference model for missing modalities. Pham *et al.* [13] developed a multimodal translation network with a cyclic translation loss for forward adaptation between source and target modalities. However, the performance of their approaches degrades when complete modality information is not available during the training stage.

In this work, we propose a knowledge-transfer network to reconstruct missing acoustic modalities using transformer blocks and a consistency loss as a constraint during training. Additionally, we introduce a cross-modal attention network to effectively fuse representations from available modalities and the reconstructed features for a robust joint multimodal representation. This allows for more informative signals to be emphasized in the cross-modal attention blocks, leading to improved multimodal representation learning. Experiments on three multimodality sentiment analysis datasets indicate that our method can achieve comparable performance to those using complete modality supervision.

The main contributions of this work are concluded from those aspects:

- To the best of our knowledge, we are among the pioneering work to address the problem of missing modality imputation with the transformer framework.
- We propose a novel knowledge-transfer network to reconstruct the missing modality from available modalities.
- Extensive experiments validate the effectiveness of our proposed method on multimodal sentiment analysis and its robustness in the missing modality scenarios.

II. PROPOSED METHOD

As illustrated in Fig. 1, during training, given a video with visual, language, and acoustic modalities, denoted as X_v , X_L , and X_A , the modality-specific encoder independently maps each modality into its modality-aware feature. As we are addressing the problem of missing modality, we assume that the acoustic feature is not directly involved in multi-modality representation learning. More specifically, we propose a novel knowledge-transfer network to reconstruct the acoustic feature based on the available visual or language signals. The mutual relationships between the visual, language, and reconstructed acoustic features are modeled using the cross-modal attention network, which consists of a set of transformer blocks. Finally, the aggregated features are then used for the multimodal human language sentiment analysis.

A. Knowledge-Transfer Network

In this section, we will present our approach for reconstructing the missing audio modality information from the available visual and language modality features. We use a set

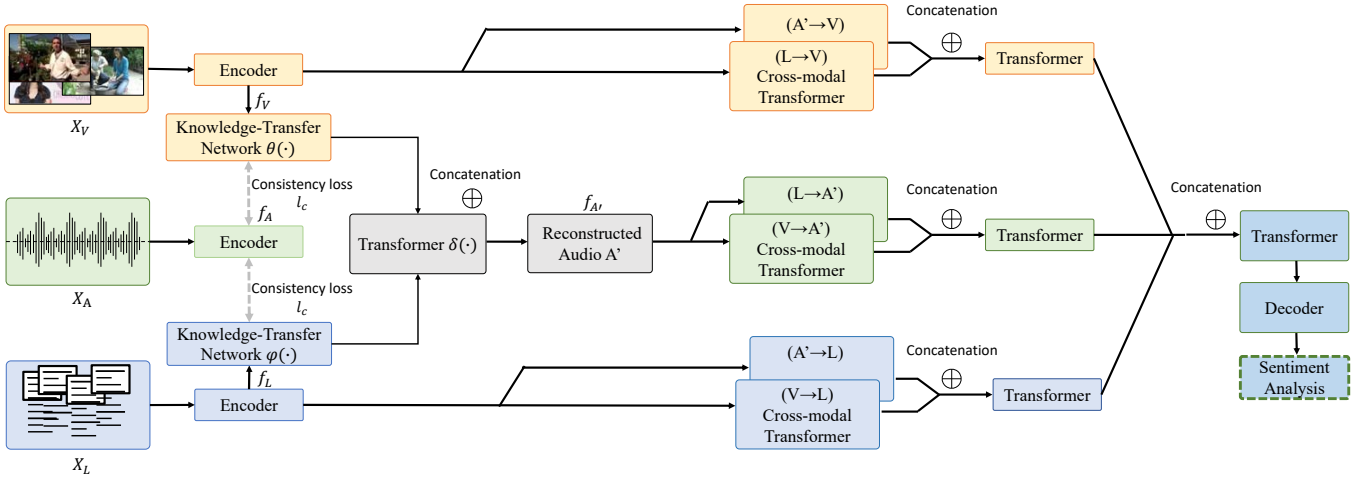


Fig. 1. The pipeline of our method. The A' denotes the reconstructed audio information.

of transformer-based encoders to convert the vision, language, and audio modalities into modality-specific features, denoted as \mathbf{f}_V , \mathbf{f}_L , and \mathbf{f}_A respectively. To ensure the reconstructed audio feature is close to the ground truth, we employ the consistency loss function ℓ_c , which minimize the Euclidean distance between the reconstructed and ground truth audio features, as defined below:

$$\mathcal{L}_{V \rightarrow A}^c = \ell_c(\theta(\mathbf{f}_V), \mathbf{f}_A), \quad \mathcal{L}_{L \rightarrow A}^c = \ell_c(\phi(\mathbf{f}_L), \mathbf{f}_A), \quad (1)$$

where the θ and the ϕ stand for the knowledge-transfer network for visual and language modality, consisting of a set of transformer blocks. Here $\theta(\mathbf{f}_V)$ and $\phi(\mathbf{f}_L)$ denote the acoustic features reconstructed from the visual and the language modality, respectively.

In order to reconstruct as much of the missing audio modality as possible, multiple available modalities are leveraged as supervision to train our network. Specifically, we combine the reconstructed acoustic features \mathbf{f}'_A from both the visual and language modalities, as shown below:

$$\mathbf{f}'_A = \delta([\theta(\mathbf{f}_V) || \phi(\mathbf{f}_L)]), \quad (2)$$

where $||$ denotes the concatenation operation which combines the reconstructed acoustic features from vision $\theta(\mathbf{f}_V)$ and language $\phi(\mathbf{f}_L)$. Instead of intuitive concatenation, we utilize a set of transformer blocks δ to encode the combined reconstructed acoustic information for effective representation.

B. Cross-modal Attention

To obtain a comprehensive joint multi-modal representation, it is essential to model the inter-dependency relationship between different modalities. Inspired by MulT [3], we also consider fusing cross-modal information by providing a latent adaptation across modalities. Consider a source modal feature m and a target modal feature m' , we map the target modal features m' into a latent space as the query $Q_{m'} = m'W_{Q_{m'}}$,

the key and value are obtained from the source modal features $K_m = mW_{K_m}$, and $V_m = mW_{V_m}$, where $W_{Q_{m'}}$, W_{K_m} and W_{V_m} are the weights. The latent adaptation from the source modality m to the target query feature m' is presented as the cross-modal attention (CM), $Z'_m := \text{CM}_{m \rightarrow m'}(m', m)$:

$$\begin{aligned} Z'_m &= \text{CM}_{m \rightarrow m'}(m', m) \\ &= \text{softmax} \left(\frac{Q_{m'} K_m^\top}{\sqrt{d_k}} \right) V_m \\ &= \text{softmax} \left(\frac{m' W_{Q_{m'}} W_{K_m}^\top m^\top}{\sqrt{d_k}} \right) m W_{V_m}, \end{aligned} \quad (3)$$

where $\sqrt{d_k}$ denotes the scaled parameter, and the d_k denotes the length of the target features. A residual connection is utilized to connect the original target feature m' and attended features Z'_m after the cross-modal attention computation as our final feature:

$$m' = m' + Z'_m. \quad (4)$$

Finally, the attended target feature with inter-modality correlation information are subsequently combined for the sentiment analysis prediction task, with the prediction denoted as y' . Note that the decoder consists of a set of self-attention transformer [8]. The standard cross entropy loss function ℓ_{ce} is utilized to compute the difference between the ground truth y and the prediction y' :

$$\mathcal{L}_e = \ell_{ce}(y, y'). \quad (5)$$

The overall loss function L to optimize is defined as below:

$$\mathcal{L} = \mathcal{L}_e + \lambda_1 \mathcal{L}_{V \rightarrow A}^c + \lambda_2 \mathcal{L}_{L \rightarrow A}^c, \quad (6)$$

where λ_1 and λ_2 refer to the trade-off parameters for the consistency loss from the visual and language modality, respectively.

TABLE I
RESULTS FOR MULTIMODAL SENTIMENT ANALYSIS ON THE CMU-MOSI DATASET WITH ALIGNED AND NON-ALIGNED MULTI-MODAL SEQUENCES. \uparrow : THE HIGHER, THE BETTER. \downarrow : THE LOWER, THE BETTER.

Metric	Acc $_7^\uparrow$	Acc $_2^\uparrow$	F1 $^\uparrow$	MAE $^\downarrow$	Corr $^\uparrow$
(Word Aligned) CMU-MOSI Sentiment					
RMFN [11]	38.3	78.4	78.0	0.922	0.681
MFM [14]	36.2	78.1	78.1	0.951	0.662
RAVEN [12]	33.2	78.0	76.6	0.915	0.691
MCTN [13]	35.6	79.3	79.1	0.909	0.676
Full modality (ours)	39.7	82.9	82.7	0.870	0.694
Vision only (ours)	34.1	76.6	76.4	0.721	0.421
Language only (ours)	36.4	78.1	77.3	0.751	0.619
Language and vision (ours)	38.7	81.3	81.2	0.849	0.688
Ours	39.1	82.3	82.2	0.858	0.691
(Unaligned) CMU-MOSI Sentiment					
CTC [15] + MCTN [13]	32.7	75.9	76.4	0.991	0.613
CTC [15] + RAVEN [12]	31.7	72.7	73.1	1.076	0.544
Full modality (ours)	39.3	82.3	82.1	0.861	0.690
Vision only(ours)	34.0	76.1	75.9	0.717	0.401
Language only (ours)	36.0	77.5	77.2	0.742	0.606
Language and vision (ours)	37.5	80.9	80.7	0.838	0.681
Ours	38.7	81.9	81.7	0.844	0.687

III. EXPERIMENT

Following the previous methods [11], [13], [14], we conducted experiments on three benchmark datasets for sentiment analysis and emotion recognition to validate the effectiveness of our proposed method.

A. Datasets and Experimental Settings

CMU-MOSI & MOSEI. CMU-MOSI [4] is a multimodal sentiment analysis dataset containing 2,199 short monologue video clips. CMU-MOSEI [16] consists of 23,454 video clips from YouTube, and each sample is assigned a sentiment score by human annotators, ranging from -3 (strongly negative) to 3 (strongly positive). Following previous methods [3], [13], [17], the performances are measured with a variety of evaluation metrics, including 7-class sentiment score classification (Acc $_7$), binary positive/negative sentiments prediction accuracy (Acc $_2$), F1 score, mean absolute error (MAE), and correlation of the model’s prediction with subjective study (Corr).

IEMOCAP IEMOCAP [18] is a multi-label emotion recognition dataset that contains around 10,000 videos. The dataset includes four classes: happy, sad, angry, and neutral. Unlike CMU-MOSI [4] and CMU-MOSEI [16], this dataset focuses on multi-label prediction, where a person can express multiple emotions simultaneously. Following the previous methods [3], [12], [19], the binary classification accuracy (Acc) and F1 score are reported in the experiments.

Implementation Details In this paper, we utilize the Multimodal Transformer [3] as our backbone and baseline. The transformer blocks within the model consist of three transformer layers [8]. We employ Adam optimization with 40 training epochs and maintain a constant learning rate of 1e-3 throughout the training process. We utilize the same training

TABLE II
RESULTS FOR MULTIMODAL SENTIMENT ANALYSIS ON CMU-MOSEI DATASET WITH ALIGNED AND NON-ALIGNED MULTIMODAL SEQUENCES. \uparrow : THE HIGHER, THE BETTER. \downarrow : THE LOWER, THE BETTER.

Metric	Acc $_7^\uparrow$	Acc $_2^\uparrow$	F1 $^\uparrow$	MAE $^\downarrow$	Corr $^\uparrow$
(Word Aligned) CMU-MOSEI Sentiment					
Graph-MFN [17]	45.0	76.9	77.0	0.71	0.54
RAVEN [12]	50.0	79.1	79.5	0.614	0.662
MCTN [13]	49.6	79.8	80.6	0.609	0.670
Full modality (ours)	50.7	80.6	80.8	0.623	0.700
Vision only (ours)	43.5	66.4	69.3	0.756	0.343
Language only (ours)	46.5	77.4	78.2	0.653	0.631
Language and vision (ours)	48.4	79.5	79.6	0.639	0.633
Ours	51.1	80.0	80.3	0.635	0.637
(Unaligned) CMU-MOSEI Sentiment					
CTC [15] + RAVEN [12]	45.5	75.4	75.7	0.664	0.599
CTC [15] + MCTN [13]	48.2	79.3	79.7	0.631	0.645
Full modality (ours)	49.7	79.8	80.1	0.641	0.681
Vision only (ours)	42.1	65.7	68.4	0.741	0.339
Language only (ours)	45.4	76.9	77.1	0.660	0.623
Language and vision (ours)	47.9	78.2	79.1	0.649	0.627
Ours	49.6	79.4	79.5	0.646	0.648

and testing split as that in [3] and CTC [15] is applied on the unaligned setting of baseline approaches.

B. Performance Comparison

We compare our results to state-of-the-art methods that utilize full modalities for supervision. Our method, however, is evaluated with certain missing modalities. The “Vision only” denotes that only the visual modal information (yellow sub-branch in Fig. 1) is utilized as input. “Language only” denotes that only the language modal information (blue sub-branch in Fig. 1) is utilized for training the model. “Language and vision” denotes that we fuse the visual and language modal information with the proposed cross-modal fusion module, but without the reconstructed acoustic modal information from our knowledge-transfer network. “Full modality” denotes the complete modality supervision.

CMU-MOSI. As shown in Table I, our method outperforms the lower bound results (including vision and language) for both the aligned and unaligned datasets. It’s worth noting that, with single-modal information, language performance is better than vision, indicating that language is more important than vision modal information in this dataset. When fusing vision and language without our reconstructed audio feature, the performance is significantly improved over methods with only one modal information, which demonstrates that more modal information can improve the performance. Our method achieves competitive results compared to the upper bound with the use of our reconstructed audio features. These results indicate that our reconstructed feature can effectively restore the missing audio modal information.

CMU-MOSEI. As depicted in Table II, the results for both aligned and unaligned settings for the MOSEI dataset demon-

TABLE III
RESULTS FOR MULTIMODAL EMOTIONS ANALYSIS ON IEMOCAP WITH ALIGNED AND NON-ALIGNED MULTIMODAL SEQUENCES.

Task Metric	Happy Acc [↑] F1 [↑]	Sad Acc [↑] F1 [↑]	Angry Acc [↑] F1 [↑]	Neutral Acc [↑] F1 [↑]
(Word Aligned) IEMOCAP Emotions				
RMFN [11]	87.5	85.8	83.8	82.9
MFM [14]	90.2	85.8	88.4	86.1
RAVEN [12]	87.3	85.8	83.4	83.1
MCTN [13]	84.9	83.1	80.5	79.6
Full modality (ours)	90.3	88.1	86.4	86.0
Vision only (ours)	83.7	81.6	81.5	81.2
Language only (ours)	85.3	85.9	85.7	84.2
Language and vision (ours)	89.1	86.8	85.9	85.0
Ours	90.1	87.6	87.5	85.5
(Unaligned) IEMOCAP Emotions				
CTC [15] + RAVEN [12]	77.0	76.8	67.6	65.0
CTC [15] + MCTN [13]	80.5	77.5	72.0	71.7
Full modality (ours)	84.8	81.9	77.7	74.1
Vision only (ours)	77.7	72.6	69.9	68.3
Language only (ours)	79.3	78.8	74.6	71.8
Language and vision (ours)	82.8	81.0	76.9	72.8
Ours	84.4	81.7	77.7	74.0

strate that our method attains comparable performance to the fully-supervised method and surpasses all previous methods. Interestingly, our method demonstrates even better performance than the full modality supervision on the evaluation metrics of Acc_7 and MAE on the aligned setting. This could be due to the fact that some of the audio information provided by the ground truth audio might negatively impact the prediction, such as background noise. For example, in a video of a boy who appears sad, his feelings can be accurately reflected by his facial expressions and speech, but if the audio includes background laughter, it may mislead the network’s ability to make an accurate judgement. However, our reconstructed audio modal information is obtained from vision and text, which is not affected by background noise, thus providing more accurate results.

Interestingly, our method occasionally surpasses the full-modality baseline, particularly in CMU-MOSEI. This outcome arises because real acoustic signals sometimes contain irrelevant or misleading noise (e.g., background laughter, music, or microphone distortion) that can bias the prediction. In contrast, the reconstructed audio representation is derived from visual and textual cues, effectively denoising the audio stream. As a result, our reconstructed features may be more informative for sentiment recognition than raw acoustic input, leading to performance exceeding the full-modality upper bound.

IEMOCAP. As evidenced by the results presented in Table III, our method demonstrates comparable performance to the full modality method and surpasses all baseline methods on both the aligned and unaligned settings. This conclusion is consistent with the results observed in the other two datasets, CMU-MOSI and CMU-MOSEI.

TABLE IV
THE EFFECTIVENESS OF EACH CROSS-MODAL TRANSFORMER. THE RESULTS ARE REPORTED ON THE CMU-MOSEI ALIGNED DATASET [16].

Target Modal	Acc [↑] ₇	Acc [↑] ₂	F1 [↑]	MAE [↓]	Corr [↑]
Language	49.0	79.7	80.2	0.636	0.632
Audio	48.2	79.6	80.0	0.639	0.627
Vision	48.4	79.5	79.6	0.641	0.633
Ours	51.1	80.0	80.3	0.635	0.630

TABLE V
THE EFFECTIVENESS OF DIFFERENT CONSISTENCY LOSS. THE RESULTS ARE REPORTED ON THE CMU-MOSEI ALIGNED DATASET [16].

Loss Function	Acc [↑] ₇	Acc [↑] ₂	F1 [↑]	MAE [↓]	Corr [↑]
L1	50.9	79.7	80.1	0.639	0.631
L2	51.1	80.0	80.3	0.635	0.637

C. Ablation Study

In this section, we evaluate the effectiveness of each individual cross-modal transformer on the CMU-MOSEI dataset. Specifically, we analyze the performance when only utilizing the cross-modal attention module for the language modality, represented as $A' \rightarrow L$ and $V \rightarrow L$, the blue sub-branch in Fig. 1. We also conduct similar evaluations for the visual and acoustic modalities. As shown in Table IV, the highest performance is achieved when the target modality is text (language). Combining all the cross-modal transformers further improves the performance. Additionally, Table V shows the results of using two different consistency losses to constrain the reconstructed features on the CMU-MOSEI dataset. We find that using the L2 loss leads to better performance.

IV. LIMITATION, FUTURE WORK AND CONCLUSION

While our method demonstrates strong robustness under missing-audio conditions, several limitations remain. First, extending the framework to scenarios where the visual or textual modalities are absent is non-trivial and warrants future investigation. Second, although the reconstruction enhances predictive performance, it does not guarantee interpretability or perceptual fidelity of the generated audio features. Finally, the reliance on transformer-based architectures introduces higher computational overhead compared to lightweight fusion models, which may restrict deployment in real-time or resource-constrained environments.

In conclusion, we present a knowledge-transfer network which utilizes a consistency loss for the task of multimodal learning with a missing modality for multimodal sentiment analysis. Our method specifically reconstructs the missing modal information and fuses it with the available modalities through a cross-modal attention network. Through extensive experiments on three sentiment analysis benchmarks, we demonstrate that the proposed method outperforms other baseline approaches and is capable of achieving comparable results to fully supervised multi-modality methods.

REFERENCES

- [1] W. Liu, C. Zhang, G. Lin, and F. Liu, "Crnet: Cross-reference networks for few-shot segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 4165–4173.
- [2] W. Liu, W. Zhou, J. Liu, *et al.*, "Modality-aware feature matching: A comprehensive review of single-and cross-modality techniques," *arXiv preprint arXiv:2507.22791*, 2025.
- [3] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the conference. Association for Computational Linguistics. Meeting*, NIH Public Access, vol. 2019, 2019, p. 6558.
- [4] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages," *IEEE Intelligent Systems*, vol. 31, no. 6, pp. 82–88, 2016.
- [5] Z. Zhang, J. M. Girard, Y. Wu, *et al.*, "Multimodal spontaneous emotion corpus for human behavior analysis," in *CVPP*, 2016, pp. 3438–3446.
- [6] Z. Zeng, J. Tu, B. Pianfetti, *et al.*, "Audio-visual affect recognition through multi-stream fused HMM for HCI," in *CVPR*, 2005, pp. 967–972.
- [7] Q. Gan, S. Wang, L. Hao, and Q. Ji, "A multimodal deep regression bayesian network for affective video content analyses," in *ICCV*, 2017, pp. 5123–5132.
- [8] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
- [9] M. X. Chen, O. Firat, A. Bapna, *et al.*, "The best of both worlds: Combining recent advances in neural machine translation," *arXiv preprint arXiv:1804.09849*, 2018.
- [10] J. D. M. C. K. Lee and K. Toutanova, "Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
- [11] P. P. Liang, Z. Liu, A. Zadeh, and L.-P. Morency, "Multimodal language analysis with recurrent multi-stage fusion," *arXiv preprint arXiv:1808.03920*, 2018.
- [12] Y. Wang, Y. Shen, Z. Liu, P. P. Liang, A. Zadeh, and L.-P. Morency, "Words can shift: Dynamically adjusting word representations using nonverbal behaviors," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 7216–7223.
- [13] H. Pham, P. P. Liang, T. Manzini, L.-P. Morency, and B. Póczos, "Found in translation: Learning robust joint representations by cyclic translations between modalities," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 6892–6899.
- [14] Y.-H. H. Tsai, P. P. Liang, A. Zadeh, L.-P. Morency, and R. Salakhutdinov, "Learning factorized multimodal representations," *arXiv preprint arXiv:1806.06176*, 2018.
- [15] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.
- [16] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2236–2246.
- [17] A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria, and L.-P. Morency, "Memory fusion network for multi-view sequential learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, 2018.
- [18] C. Busso, M. Bulut, C.-C. Lee, *et al.*, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.
- [19] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, "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)*, 2017, pp. 873–883.