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/ICASSP48485.2024.10446572

MULTIMODAL SENTIMENT ANALYSIS BASED ON 3D STEREOSCOPIC ATTENTION

Jian Huang[⋆]*, Yuanyuan Pu*[⋆,∗]*, Dongming Zhou*[⋆]*, Hang Shi*[⋆]*, Zhengpeng Zhao*[⋆]*, Dan Xu*[⋆]*, Jinde Cao*[†]

[⋆] School of Information Science and Engineering, Yunnan University, China
[⋆] University Key Laboratory of Internet of Things Technology and ApplicationYunnan Province
[†]Southeast University, China and Yonsei Frontier Lab, Yonsei University, South Korea

ABSTRACT

In the multimodal (text, audio, and visual) sentiment analysis, the current methods generally consider the bi-modal sentiment interaction, resulting in inadequate mining and fusion of relations between modalities. In this paper, we propose the concept of multimodal 3D (3-Dimensional) stereoscopic attention for the first time, which constructs the tri-modal stereoscopic attention with temporal sequences simultaneously to adequately structure the sentiment interaction. To solve the problems of stereoscopic attention construction such as the increased complexity of algorithms caused by rising dimensions, we propose a progressive construction method with 2D attention as an intermediate process. To implement sentiment relations based on stereoscopic attention to integrating modal information sufficiently, a forward propagation mechanism is proposed, which optimizes the representations of each modality with multimodal modulation. The results on two public datasets confirm the superiority of the proposed method in all metrics to the baselines.

***Index Terms*—** Multimodal sentiment analysis; Sentiment interaction; Stereoscopic attention

1. INTRODUCTION

Given a video, multimodal sentiment analysis (MSA) targets learning the polarity of sentiment the speaker expressed through different modalities such as verbal (text), audio, and visual (facial expression). Previous researchers[1] argued that complementary information between modalities is an advantage over unimodal sentiment analysis. However, due to the asynchronous nature of multimodal sequences, the behaviors that express the same sentiment may not occur at the same time, e.g., a smile may be associated with a positive word said in the past. To address this problem, previous works [2] utilized Recurrent Neural Networks (RNNs) and Transformers [3, 4, 5]. However, RNNs only integrate unimodal information and miss interactions between modalities. To ameliorate these drawbacks, Transformer-based models are designed. However, the attention mechanism of the Transformer only explores the bi-modal interactions. With the popularity of graph neural networks (GNNs), GNNs-based models [6] have been introduced to MSA. To faithfully characterize the connections of modalities, pairwise modeling in graphs is inadequate. For example, a smile coupled with a positive word is positive, while audio represents sarcasm and ultimately leads to an opposite shift to negative. Therefore, considering only the bi-modal relationship is not sufficient.

To construct the sentiment interactions among the three sequence-structured modalities simultaneously, we propose a 3D stereoscopic attention mechanism for the first time, which helps to understand how multimodal information affects the final sentiment polarity. However, there are two main obstacles in model design: 1) Different from 2D attention, there is dimension rising in the 3D attention generation; 2) Based on the generated 3D attention, how to integrate the information between each 2D sequence modality. As for the former, we propose a progressive stereoscopic attention construction, which first constructs the attention relations between two modalities and combines them to obtain a unified representation to generate the stereoscopic attention with the last modality. For the latter, we design a forward propagation for stereoscopic attention, which optimizes the representation of each modality with the other two modalities and maintains the transfer ability of our model. The experimental results demonstrate our method performs better in all metrics on CMU-MOSI and CMU-MOSEI datasets.

The main contributions of this paper are summarized as follows: (1) The concept of 3D stereoscopic attention is proposed for the first time to fully explore sentiment interactions between modalities. (2) A progressive stereoscopic attention construction method is designed to construct and optimize the attention weights between modalities in time step levels. (3) To adequately integrate the individual information of each modality, a forward propagation is proposed.

2. METHODOLOGY

The multimodal stereoscopic attention framework is shown in Fig. 1. The framework is mainly composed of three parts: unimodal encoder, progressive stereoscopic attention, and modalities fusion. We elaborate on them in the following.

∗ Correspongding author. Email: yuanyuanpu@ynu.edu.cn

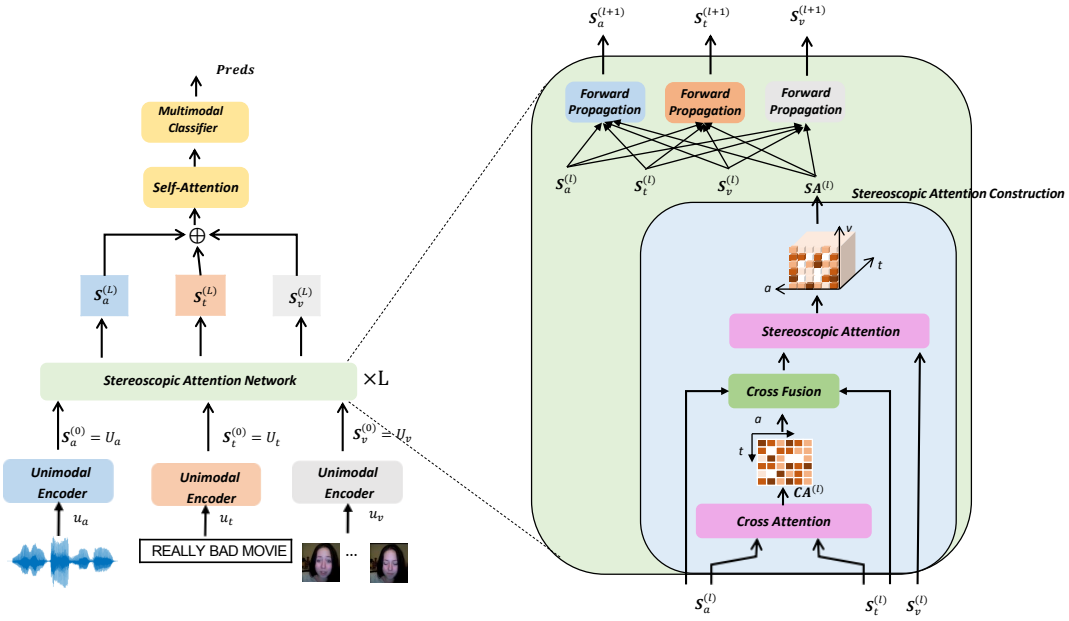

Fig. 1. The overall architecture of the proposed method.

2.1. Unimodal Encoder

The sentence-level tri-modal features are input to the model, containing three modalities: text (t), audio (a), and visual (v), which is denoted as $\boldsymbol{u}_m \in \mathrm{R}^{T_m \times d_m}$, $m \in \{t, a, v\}$, where T_m and d_m are denoted as the temporal length and the feature dimension of each time step. We utilize Bi-LSTM [7] (Bidirectional Long Short Term Memory) to encode the sequences, and then the features of each modality are mapped to the same dimension d through the fully connected layer as follows:

$$\boldsymbol{U}_m = BiLSTM\left(\boldsymbol{u}_m\right), m \in \{t, a, v\} \tag{1}$$

2.2. Progressive Stereoscopic Attention

To effectively construct tri-modal stereoscopic attention, we adopt a progressive construction method, as shown in Figure 1. Firstly, 2D cross-modal attention[3] is generated between the features of text and audio, which reflects the strength of the relationship between the temporal steps of the two modalities. Then, the temporal features of the two modalities are fused with their attention weights. In the paper, the text and audio modalities are utilized to generate the 2D cross-modal attention in the first stage as an example, and the rest of the cases are the same. The text and audio features are mapped to calculate the 2D cross-modal attention as follows:

$$\boldsymbol{CA}^{(l)} = Softmax\left(\frac{\boldsymbol{W}_t^{(l)}\boldsymbol{S}_t^{(l)}\boldsymbol{W}_a^{(l)\mathrm{T}}\boldsymbol{S}_a^{(l)\mathrm{T}}}{\sqrt{d}}\right) \tag{2}$$

where $\boldsymbol{CA}^{(l)}$ denotes Cross-modal Attention at l-th layer. $\boldsymbol{W}_t^{(l)}$ and $\boldsymbol{W}_a^{(l)}$ represent the learnable parameters of text and audio features. d is the dimension of encoded features. Noting that $\boldsymbol{S}_m^{(0)} = \boldsymbol{U}_m$. Then, the 2D cross-modal attention weights are implemented to fuse features of the two modalities.

$$\boldsymbol{F}_{i,j}^{(l)} = \boldsymbol{S}_{t,i}^{(l)} + \boldsymbol{CA}_{i,j}^{(l)} \odot \boldsymbol{S}_{a,j}^{(l)}, \{i \in T_t, j \in T_a\} \tag{3}$$

where $\boldsymbol{F}^{(l)} \in \mathrm{R}^{T_t \times T_a \times d}$ represents the weighted fusion of the temporal features of $\boldsymbol{S}_t^{(l)} \in \mathrm{R}^{T_t \times d}$ and $\boldsymbol{S}_a^{(l)} \in \mathrm{R}^{T_a \times d}$. According to the matrix dimension formula, the rising dimension of the fused features is proved as follows:

$$\begin{aligned} &dim\left(\boldsymbol{F}^{(l)}\right) = dim\left(\boldsymbol{S}_t^{(l)} + \boldsymbol{S}_a^{(l)}\right) \\ &= dim\left(\boldsymbol{S}_t^{(l)}\right) + dim\left(\boldsymbol{S}_a^{(l)}\right) - dim\left(\boldsymbol{S}_t^{(l)} \cap \boldsymbol{S}_a^{(l)}\right) \end{aligned} \tag{4}$$

where $dim\left(\cdot\right)$ denotes the dimension of matrices.The fused features of text and audio are then combined with visual features to generate the final trimodal stereoscopic attention matrix as follows:

$$\boldsymbol{SA}^{(l)} = Softmax\left(\frac{\boldsymbol{F}^{(l)}\boldsymbol{W}_v^{(l)\mathrm{T}}\boldsymbol{S}_v^{(l)\mathrm{T}}}{\sqrt{d}}\right) \tag{5}$$

where $\boldsymbol{SA}^{(l)}$ denotes stereoscopic attention. The sentiment interaction between text, audio, and visual can be represented by the attention weights $\boldsymbol{SA}^{(l)} \in \mathrm{R}^{T_t \times T_a \times T_v}$. Moreover, in order to improve the ability to capture semantics, $\boldsymbol{CA}^{(l)}$ and $\boldsymbol{SA}^{(l)}$ are both in multi-head [14].

$$\boldsymbol{CA}^{(l)} = [\boldsymbol{CA}_{head,1}^{(l)}; \cdots; \boldsymbol{CA}_{head,n}^{(l)}] \tag{6}$$

$$\boldsymbol{SA}^{(l)} = [\boldsymbol{SA}_{head,1}^{(l)}; \cdots; \boldsymbol{SA}_{head,m}^{(l)}] \tag{7}$$

where $[;]$ denotes the concatenation. n and m indicate the number of attention heads.

Table 1. Results of our method and baselines on CMU-MOSI and CMU-MOSEI. * indicates results from previous papers.

Model	CMU-MOSI				CMU-MOSEI			
	Acc2↑	F1↑	MAE↓	Corr↑	Acc2↑	F1↑	MAE↓	Corr↑
MFN[2]	77.26	77.38	0.9534	0.6672	80.23	80.77	0.5693	0.7202
MulT[3]	81.47	81.4	0.7892	0.7763	80.90	80.87	0.5690	0.7240
MISA[8]	81.95	81.91	0.7596	0.7771	81.10	81.18	0.5710	0.7310
BERT-MAG[9]	82.42	82.38	0.7313	0.7836	81.90	82.32	0.5640	0.7597
BBFN[5]	80.33	80.32	0.8216	0.6277	82.29	82.21	0.5820	0.7270
Self-MM[10]	82.51	82.47	0.7251	0.7896	82.17	82.46	0.5351	0.7605
MMIM[11]	82.33	82.28	0.7514	0.7718	80.04	80.47	0.5794	0.7352
UEGD*[12]	79.90	79.90	0.8860	0.6910	81.20	81.70	0.5430	0.7480
DMD[13]	82.07	82.08	0.7470	0.7859	82.81	83.06	0.5473	0.7518
Ours	**83.68**	**83.71**	**0.7124**	**0.7941**	**83.87**	**83.91**	**0.5210**	**0.7680**

2.3. Modalities Fusion

Inspired by Transformer [15], we propose a forward propagation mechanism based on stereoscopic attention to fuse the information of each modality. Taking the text modality as an example, the text features are modulated by the audio and visual information, and the modulation factor α and β ensure the offset within a reasonable range. Firstly, stereoscopic attention is used to get important information about audio and visual modalities to text modality.

$$\boldsymbol{S}_{v\to t}^{(l)} = \sum_{i=0}^{T_a} \frac{\left(\boldsymbol{SA}^{(l)}\boldsymbol{S}_v^{(l)}\right)_{T_t\times i\times d}}{T_a} \tag{8}$$

$$\boldsymbol{S}_{a\to t}^{(l)} = \sum_{j=0}^{T_v} \frac{\left(\boldsymbol{SA}^{(l)}\boldsymbol{S}_a^{(l)}\right)_{T_t\times j\times d}}{T_v} \tag{9}$$

where $\boldsymbol{S}_{v\to t}^{(l)} \in \mathrm{R}^{T_t\times d}$ represents the important mapping features of the visual modality for the textual modality. The information from the two modalities is then used to modulate the textual features.

$$\boldsymbol{S}_t^{(l+1)} = \boldsymbol{S}_t^{(l)} + \alpha * \boldsymbol{S}_{v\to t}^{(l)} + \beta * \boldsymbol{S}_{a\to t}^{(l)} \tag{10}$$

Finally, to improve the model's robustness and learning ability, the modulated text is followed by the fully connected and normalized layer to obtain the updated features.

$$\boldsymbol{S}_t^{(l+1)} = LayerNorm\left(FC\left(\boldsymbol{S}_t^{(l+1)} + \boldsymbol{S}_t^{(l)}\right)\right) \tag{11}$$

where $LayerNorm()$ denotes layer normalization. In addition, the features of audio and visual modalities are fused in the same pattern.

3. EXPERIMENTS

3.1. Datasets

CMU-MOSI: The CMU-MOSI [16] dataset consists of 93 videos of 89 distinct speakers, divided into 2,199 utterance-level video clips, which is manually annotated with a real-valued sentiment intensity score in the range [-3, +3].

CMU-MOSEI: The CMU-MOSEI [17], which is the extended version of the CMU-MOSI dataset, consists of 22,856 annotated video clips from more than 1000 online speakers and 250 different topics.

3.2. Experiment Details

Developed on the PyTorch framework, our model implements Mean Absolute Error (MAE) as the loss function for the sentiment prediction task. For a fair comparison with the baselines, we apply Bert [18], Facet[1], and COVAREP [19] to extract text, visual, and audio features respectively. The experiments consist of two tasks: regression and classification. Binary accuracy (Acc2, negative/non-negative) and F1 scores are reported to evaluate the results of the classification task, while MAE and Pearson correlation (Corr) are reported to evaluate the results of the regression task. Except for MAE, higher values reflect better performance for all metrics. Meanwhile, the results and the reproducible baselines are the average performance of five experiments.

3.3. Experiments

Results and Discussion: The results of our method and baselines are shown in Table 1. Compared with RNNs-based and Transformer-based models such as MFN and BBFN, our method achieves remarkable improvement in all metrics. Moreover, our method also outperforms state-of-the-art baselines such as Self-MM and DMA, which proves that our method can effectively improve the performance of tasks by modeling sentiment interactions between three modalities simultaneously.

Ablation Study: The effectiveness of each component in our method is verified by ablation experiments on the CMU-MOSI dataset, and the results are shown in Table 2. First of

[1]https://imotions.com/platform/

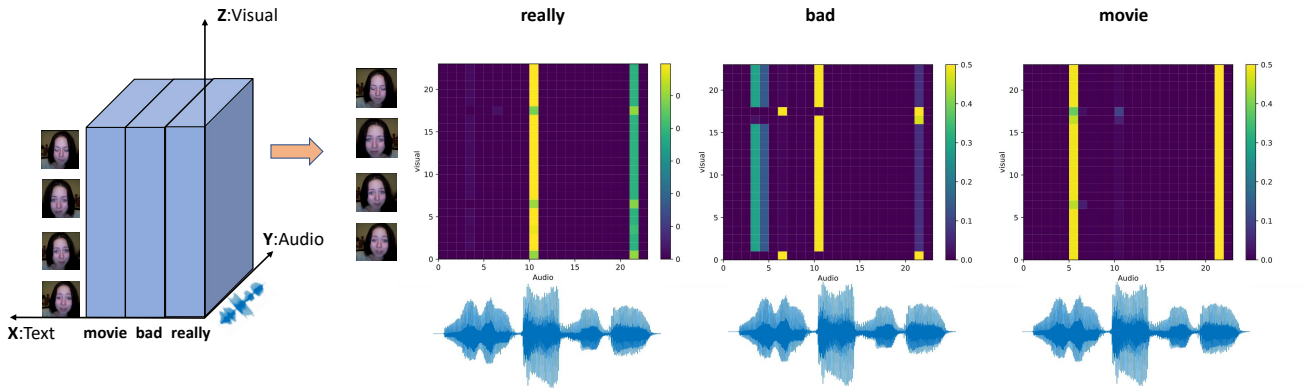

Fig. 2. Visualization of the sample "really bad movie", representing the attention matrices of different words corresponding to audio and visual.

Table 2. Ablation study of our proposed method on CMU-MOSI. "w/o" denotes removing the component.

Model	Acc2↑	F1↑	MAE↓	Corr↑
w/o SA	80.32	80.21	0.9012	0.7351
UniAtten	81.25	81.29	0.8797	0.7418
CrossAtten	82.93	82.96	0.8452	0.7492
TV,A	83.23	82.87	0.7194	0.7585
AV,T	83.54	83.28	0.7185	0.7691
1 Layer	82.05	82.02	0.7430	0.7679
3 Layer	82.16	82.14	0.7342	0.7705
4 Layer	81.40	81.47	0.7496	0.7660
Ours	**83.68**	**83.71**	**0.7124**	**0.7941**

all, when the stereoscopic attention module is removed, all metrics decrease, which indicates that the module is able to optimize the integration of the features and thus improve performance. To demonstrate the effectiveness of stereoscopic attention, it is replaced with UniAtten and CrossAtten respectively and the results show the superiority of the stereoscopic attention. In addition, the experiments verify the effect of the generation order of stereoscopic attention. Such as "TV, A", text, and visual modalities are first generated into a 2D attention matrix and then the features are fused with the audio to generate the final stereoscopic attention. From the results, the generation order shows little impact on the task performance, which shows the strong robustness of our method.

Case Study: In this section, we select a sample from the CMU-MOSI dataset and visualize its stereoscopic attention matrix as shown in Fig. 2. The text of the sample is "really bad movie" and the stereoscopic matrix is sliced at the word level. As shown in the figure, the negative word "bad" and the phonetic "bad" have the highest attentional weights in relation to the visual modality of head shaking. In addition, the negative word "bad" and the degree word "really" are phonetically second only to itself. The above results are consistent with human sentiment understanding mechanisms. Therefore the proposed 3D stereoscopic attention can effectively align the sentiment information between different modalities and and improve the interpretability of the multimodal sentiment.

Effect of Attention Layers: To explore the effect of the number of layers of stereoscopic attention, the number of layers is analyzed on the CMU-MOSI dataset shown in Table 2. As the number of layers increases, the performance increases and peaks at 2 layers. Subsequently, the performance starts decreasing with the increase in the number of layers, which is attributed to the fact that the model parameters incrementally increase with the number of layers, resulting in overfitting due to insufficient learning.

4. CONCLUSION

In this paper, the concept of multimodal stereoscopic attention is proposed for the first to explore the sentiment interactions among text, audio, and visual modalities. Meanwhile, a progressive method to construct stereoscopic attention and a fusion strategy based on forward propagation is proposed to improve the performance of the MSA task. Experimental results on two public datasets indicate that our method outperforms the baselines in all metrics. In future work, we consider optimizing the method of constructing stereoscopic attention and applying it to different multimodal tasks.

5. ACKNOWLEDGMENTS

The work is supported by the National Natural Science Foundation of China(61271361,61761046,62162068,62362070), Key Project of Applied Basic Research Program of Yunnan Provincial Department of Science and Technology (202001BB050043), the Major Science and Technology Special Project in Yunnan Province (202302AF080006), Postgraduate Science Foundation of Yunnan University under Grant KC-22221992.

6. REFERENCES

[1] Sarah A Abdu, Ahmed H Yousef, and Ashraf Salem, "Multimodal video sentiment analysis using deep learning approaches, a survey," *Inf. Fusion*, vol. 76, pp. 204–226, 2021.

[2] Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency, "Memory fusion network for multi-view sequential learning," in *Proc. AAAI Conf. Artif. Intell., AAAI*, 2018, vol. 32.

[3] 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. Annu. Meet. Assoc. Comput Linguist.* NIH Public Access, 2019, vol. 2019, p. 6558.

[4] Saurav Sahay, Eda Okur, Shachi H Kumar, and Lama Nachman, "Low rank fusion based transformers for multimodal sequences," *arXiv preprint arXiv:2007.02038*, 2020.

[5] Wei Han, Hui Chen, Alexander Gelbukh, Amir Zadeh, Louis-philippe Morency, and Soujanya Poria, "Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis," in *ICMI - Proc. Int. Conf. Multimodal Interact.*, 2021, pp. 6–15.

[6] Jianing Yang, Yongxin Wang, Ruitao Yi, Yuying Zhu, Azaan Rehman, Amir Zadeh, Soujanya Poria, and Louis-Philippe Morency, "Mtag: Modal-temporal attention graph for unaligned human multimodal language sequences," *arXiv preprint arXiv:2010.11985*, 2020.

[7] Kazuya Kawakami, *Supervised sequence labelling with recurrent neural networks*, Ph.D. thesis, Technical University of Munich, 2008.

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

[9] Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque, "Integrating multimodal information in large pretrained transformers," in *Proc. Annu. Meet. Assoc. Comput Linguist.* NIH Public Access, 2020, vol. 2020, p. 2359.

[10] Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu, "Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis," in *Proc. AAAI Conf. Artif. Intell., AAAI*, 2021, vol. 35, pp. 10790–10797.

[11] Wei Han, Hui Chen, and Soujanya Poria, "Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis," *arXiv preprint arXiv:2109.00412*, 2021.

[12] Atsushi Ando, Ryo Masumura, Akihiko Takashima, Satoshi Suzuki, Naoki Makishima, Keita Suzuki, Takafumi Moriya, Takanori Ashihara, and Hiroshi Sato, "On the use of modality-specific large-scale pre-trained encoders for multimodal sentiment analysis," in *IEEE Spok. Lang. Technol. Workshop, SLT - Proc.* IEEE, 2023, pp. 739–746.

[13] Yong Li, Yuanzhi Wang, and Zhen Cui, "Decoupled multimodal distilling for emotion recognition," in *Proc IEEE Comput Soc Conf Comput Vision Pattern Recognit*, 2023, pp. 6631–6640.

[14] Chen Xi, Guanming Lu, and Jingjie Yan, "Multimodal sentiment analysis based on multi-head attention mechanism," in *ACM Int. Conf. Proc. Ser.*, 2020, pp. 34–39.

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Adv. neural inf. proces. syst.*, vol. 30, 2017.

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

[17] 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 *ACL - Annu. Meet. Assoc. Comput. Linguist., Proc. Stud. Res. Workshop*, 2018, pp. 2236–2246.

[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[19] Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer, "Covarep—a collaborative voice analysis repository for speech technologies," in *ICASSP IEEE Int Conf Acoust Speech Signal Process Proc*. IEEE, 2014, pp. 960–964.