MULTI-CHANNEL ATTENTIVE GRAPH CONVOLUTIONAL NETWORK WITH SENTIMENT FUSION FOR MULTIMODAL SENTIMENT ANALYSIS

Luwei Xiao[1,†], *Xingjiao Wu*[1,†], *Wen Wu*[1,⋆], *Jing Yang*[1], *Liang He*[1,⋆]

[1] School of Computer Science and Technology, East China Normal University, Shanghai, China

{louisshaw, Xingjiao.Wu}@stu.ecnu.edu.cn, {wwu}@cc.ecnu.edu.cn, {jyang, lhe}@cs.ecnu.edu.cn

ABSTRACT

Nowadays, with the explosive growth of multimodal reviews on social media platforms, multimodal sentiment analysis has recently gained popularity because of its high relevance to these social media posts. Although most previous studies design various fusion frameworks for learning an interactive representation of multiple modalities, they fail to incorporate sentimental knowledge into inter-modality learning. This paper proposes a Multi-channel Attentive Graph Convolutional Network (MAGCN), consisting of two main components: cross-modality interactive learning and sentimental feature fusion. For cross-modality interactive learning, we exploit the self-attention mechanism combined with densely connected graph convolutional networks to learn inter-modality dynamics. For sentimental feature fusion, we utilize multi-head self-attention to merge sentimental knowledge into inter-modality feature representations. Extensive experiments are conducted on three widely-used datasets. The experimental results demonstrate that the proposed model achieves competitive performance on accuracy and F1 scores compared to several state-of-the-art approaches.

Index Terms— multimodal sentiment analysis, sentimental knowledge, Multi-channel Attentive Graph Convolutional Network

1. INTRODUCTION

Sentiment analysis (SA) is one of the widely studied fields in the Natural Language Process (NLP), which aims to predict the individual's attitude towards a product or event [1]. With the rapid development of Internet technology, the Multimodal Sentiment Analysis (MSA) increasingly attracts more attention in recent years [2]. The goal of MSA is to determine the sentiment of a video, an image or a text-based on multiple modal features (see Fig.1)[3, 4]. Understanding an individual's attitude towards a specific entity is helpful in a wide range of application domains [5] (e.g., business intelligence, recommendation system, government intelligence, etc.).

According to different fusion strategies, existing methods can be divided into the following two categories: a) Approaches that make an advance on the LSTM architecture
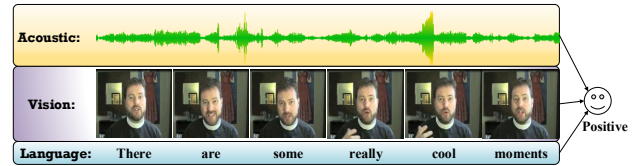
---

† Equal contribution

⋆ Co-corresponding author

Fig. 1: An example of multimodal sentiment analysis task. This task aims to predict the sentiment of a given utterance.

[6, 7], b) Frameworks that emphasize applying the expressiveness of tensors for multimodal representation [8, 9]. Although previous studies have made salient improvements on benchmark datasets, there are still challenges for multimodal sentiment analysis. For one thing, most previous studies exploit simple operations (e.g., concatenation, addition) to learn inter-modality dynamics and conduct fusion within one step, which might weaken the overall performance of the multimodal framework since the cross-modal joint representation learned is less expressive. For another, few previous publications pay attention to incorporating the sentimental knowledge accompanying the language text into the multimodal system. Moreover, the importance of sentimental knowledge has been verified by sentence-level [10] and aspect-level [11, 12] sentiment analysis tasks, which can be adopted to assist the fusion of different modalities to be more sentiment-specific based on the rich sentimental information contained in the language. In order to solve the above problems, this paper proposes a novel Multi-channel Attentive Graph Convolutional Network (MAGCN) to fuse intra-/inter-modality with sentimental knowledge hierarchically. The main contribution of this paper is as follows:

- We introduce sentiment embedding to make the use of sentimental knowledge for learning sentiment-specific inter-/intra-modality fusion representation.

- We introduce densely connected graph convolutional network from relation extraction task to multimodal sentiment analysis, which successfully models the inter-/intra-modality feature representations.

- We define a novel consistency constraint loss to further enhance the commonality between modalities.

- The proposed model yields new state-of-the-art results on three MSA benchmark datasets.

2. RELATED WORK

For approaches that make an advance on the LSTM architecture, Zadeh *et al.* [6] designed a novel Memory Fusion Network which is composed of a System of LSTMs and Delta-memory Attention Network to learn both the view-specific interactive representations and the cross-view interactive representations. Liang *et al.* [7] put forward a Recurrent Multistage Fusion Network to learn cross-modal interactive information by exploiting a specialized multistage fusion approach. For methods that lay emphasis on applying the expressiveness of tensors for multimodal representation, Zadeh *et al.* [8] proposed a Tensor Fusion Network to explicitly model the intra-modality and inter-modality dynamics end-to-end via a 3-fold Cartesian product from the modality embedding. Liu *et al.* [9] exploited a Low-rank Multimodal Fusion network that first obtains the unimodal representation and leverage low-rank weight tensors to finish multimodal fusion.

Recently, some other approaches have achieved new state-of-the-art results. Pham *et al.* [13] designed a framework of learning joint representations, a cycle consistency loss is proposed to ensure that the joint representations retain maximal information from all modalities. Mai *et al.* [14] devised a divide, conquer and combine strategy to fuse cross-modality dynamically and hierarchically. Tsai *et al.* [15] utilized a Multimodal Factorization Model which factorizes features into multimodal discriminative and modality-specific generative factors and designs a common generative-discriminative objective to optimize across multimodal data and labels. Chen *et al.* [16] introduced a method consists of a shallow fusion and aggregation module, and an auxiliary sentimental words classification is added to assist and guide the profound fusion of the three modalities.

3. APPROACH

The overall architecture of MAGCN is illustrated in Fig.2. This section will describe our approach in more detail, including the contextual utterance representation, the inter-modality representation learning, the unimodal language representation learning, and the inference module.

3.1. Contextual Utterance Representation

The sentiment knowledge in each word should be incorporated into the learning process of intra-modality and inter-modality to generate richer multimodal representation. Inspired by Chen *et al.* [16], we propose to utilize Bing Liu's Opinion Lexicon [17] to train a sentiment embedding for each word as the external knowledge. The function of sentiment embedding is to indicate whether a word is a sentimental word (If a word is in the Opinion Lexicon, then it is). The weight of each dimension for sentiment embedding is trainable during the whole process. $S = \{s_1, s_2, \cdots, s_n\}$ refers to sentiment embedding for input sequence. Given the original representation of text $E = \{e_1, e_2, \cdots, e_n\}$, acoustic $X_A = \{a_1, a_2, \cdots, a_n\}$ and vision $X_V = \{v_1, v_2, \cdots, v_n\}$,
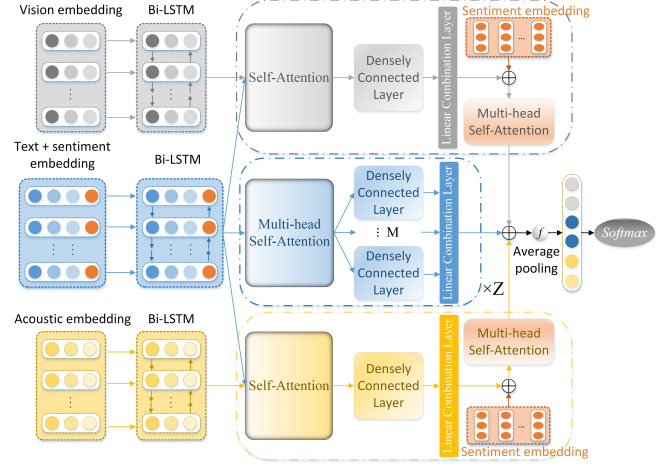

Fig. 2: The overall architecture of proposed MAGCN.

we concatenate the text embedding and sentiment embedding. The representation for a word is $x_i = [e_i; s_i]$, the language modality is denoted as $X_L = \{x_1, x_2, \cdots, x_n\}$. We first exploit the Bi-LSTM to model the representation of the above three modalities and obtain the feature encoding of them, denoted as H_L, H_A and H_V, respectively.

3.2. Inter-Modality Representation Learning

In this part, we combine self-attention, densely connected graph convolutional networks with multi-head self-attention to learn the cross-modal feature representation between language modality and other modalities. The self-attention mechanism [18] and densely connected graph convolutional networks (DCGCN) [19] are exploited to learn the bimodal fusion between language modality and other modalities (i.e., vision or acoustics) simultaneously. Firstly, an affinity matrix is generated by self-attention mechanism:

$$A^{LV} = soft \max \left(\frac{H_L W_Q \times (H_V W_K)^T}{\sqrt{d}} \right) \quad (1)$$

which can be treated as a fully connected edge-weighted graph. These weights can be viewed as the strength of relatedness between language modality and vision (or acoustics) modality. "d" denotes the dimension of H_L. $W_Q \in R^{d \times d}$ and $W_K \in R^{d \times d}$ are both parameter matrices. Next, the DCGCN [19] is introduced to capture rich local and non-local dependency information between different modalities. $g_j^l = [x_j; h_j^1; \cdots; h_j^{l-1}]$ is denoted as the concatenation of original node feature embedding and other nodes embedding generated in layers $1, 2, \cdots, l - 1$, where x_j refers to the initial input for DCGCN, produced from H_V (or H_A). In DCGCN, each densely connected layer has L sublayers. The dimension of each sublayer is determined by the input dimension d and L. For instance, if the input dimension is 60 and $L = 3$, the hidden dimension of each sub-layer will be $d_{sub} = d/L = 60/3 = 20$. An example of DCGCN is shown in Fig.3. Each layer of DCGCN is computed as follows:
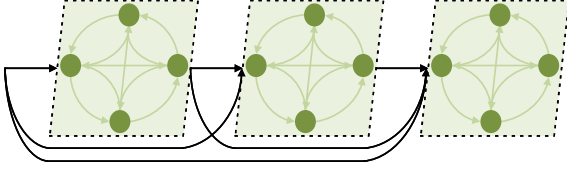
Fig. 3: The architecture of densely connected graph convolution networks(DCGCN).

$$h_i^l = \sigma(\sum_{j=1}^n A_{ij}^{LV} W^l g_j^l + b^l) \qquad (2)$$

where $W^l \in R^{d_{sub} \times d^l}$ and $b^l \in R^{d_{sub}}$ are trainable matrix and bias, respectively. $d^l = d + d_{sub} \times (l-1)$. After we feed the fully connected edge-weighted graph A^{LV} (or A^{LA}) and H_V (or H_A) into DCGCN, we obtain the bimodal fusion representation H_{LV} (or H_{LA}). Then, a linear transformation is followed to obatin the output of DCGCN H_{LV}^{gcn}. Finally, we concatenate the bimodal fusion representation H_{LV} (or H_{LA}) with sentiment embedding S as the input HS_{LV} (or HS_{LA}) for multi-head attention, learning the sentiment-awareness bimodal representation HS_{LV}^{out} (or HS_{LA}^{out}) :

$$HS_{LV}^{out} = MHA(HS_{LV}, HS_{LV}) \qquad (3)$$

where MHA is multi-head self-attention [20]. HS_{LV}^{out} is the final inter-modality feature representation of the language modality and visual modality. The final inter-modality feature representation of the language modality and acoustic modality HS_{LA}^{out} can be generated in the same way.

3.3. Unimodal Language Representation Learning
This module comprises Z identical blocks. Each block is composed of multi-head self-attention and DCGCN for extracting richer semantic information from context representations. Given the language modality representation H_L, we first generate several fully connected graphs by exploiting multi-head self-attention:

$$\widetilde{B}^{(t)} = \text{softmax}\left(\frac{H_L \mathbf{W}_t^Q \times (H_L \mathbf{W}_t^K)^T}{\sqrt{d/M}} \right) \qquad (4)$$

where $\widetilde{B}^{(t)} \in R^{n \times n}$ is the t-th matrix constructed corresponding to the t-th head. Up to M matrices are constructed in total. For M attention graphs, M independent DCGCN are required. Each layer of DCGCN is computed as follows:

$$h_{t_i}^l = \sigma(\sum_{j=1}^n \widetilde{B}_{ij}^{(t)} W_t^l u_j^l + b_t^l) \qquad (5)$$

where $u_j^l = [h_j^0; h_j^1; \ldots; h_j^{l-1}]$. h_j^0 is obtained from H_L. $t \in [1, 2, \ldots, M]$. With different matrices, different parameter matrix $W_t^l \in R^{d_{sub} \times d^l}$ and bias term $b_t^l \in R^{d_{sub}}$ are picked up for training. Then, the context feature representation from M different DCGCN are concatenated, a linear

transformation is exploited to obtain the final feature representation of language modality:

$$H_L^{out} = W_{out} [h^1; h^2; \cdots ; h^M] + b_{out} \qquad (6)$$

where $W_{out} \in R^{(d \times M) \times d}$ and $b_{out} \in R^d$ are trainable matrix and bias, respectively.

3.4. Decision Inference Module
In order to improve the commonality between the inter-modality representation, we design a consistency loss function for achieving this goal. L_2-normalization is used to normalize the embedding matrices H_L^{out}, HS_{LV}^{out} and HS_{LA}^{out} and denoted them as C_{Lnor}, C_{LVnor} and C_{LAnor}, respectively. Then, they can be used to guide the inter-modality representations to enhance their commonality further as follows:

$$L_c = \left\| C_{Lnor} \cdot C_{LVnor}^T - C_{Lnor} \cdot C_{LAnor}^T \right\|_F^2 \qquad (7)$$

Finally, we concatenate H_L^{out}, HS_{LV}^{out} and HS_{LA}^{out}, and an average pooling function is applied, the final feature representation for our model is obtained, which is then feed into $softmax$ to produce the sentiment prediction y_s. The overall loss function of our model is shown as follows:

$$Loss = \alpha \left(\frac{1}{N} \sum_{i=1}^N |y_s^i - \hat{y}_s^i| \right) + \beta L_c \qquad (8)$$

where α and β are hyper-parameters. N denotes all training samples in the dataset. y_s^i and $\hat{y}_s^i$ are predicted sentiment and true sentiment for i-th sample, respectively.

4. EXPERIMENTS
4.1. Datasets
CMU-MOSI [21],CMU-MOSEI [22] and YouTube [23] are selected as our experimental datasets. GloVe [24] is utilized to generate the language embedding, the visual representations are obtained by Facet library[1] and acoustic features are extracted by applying COVAREP [25].

4.2. Baseline Models
Some classical studies [8, 9] and recently state-of-the-art approaches [13, 14, 16, 15] which are introduced in Section 2 have been chosen as baseline models for experiments. Accuracy and F1 scores are picked up as our evaluation metrics.

4.3. Experimental Results and Analysis
The experimental results on YouTube, MOSEI and MOSI datasets are shown in Table 1. From Table 1, there are a certain number of issues worth noting. Firstly, For the YouTube dataset, compared with the best baseline model SWAFN, the MAGCN gives a higher overall performance, which outperforms SWAFN by 1.2% on accuracy and 1.3 % on F1 score. Since the YouTube dataset contains relatively few samples, most baselines may be overfitting on its training set. However, our model achieves the best performance, demonstrating the better generalization capability of the proposed model. Secondly, for MOSEI and MOSI datasets, MAGCN still achieves the best performance over all evaluations. One

---

[1]https://imotions.com/biosensor/fea-facial-expression-analysis/

Table 1: Experimental results (%). "N/A" refers to unreported results. Highest results are in bold.

Model	YouTube		CMU-MOSEI		CMU-MOSI	
	Acc	F1	Acc	F1	Acc	F1
LMF [9]	N/A	N/A	60.27	53.87	76.4	75.7
TFN [8]	45.0	41.0	59.4	57.33	74.6	74.5
MFM [15]	53.3	52.4	N/A	N/A	78.1	78.1
MCTN [13]	51.7	52.4	N/A	N/A	79.3	79.1
HFFN [14]	N/A	N/A	60.37	59.07	80.2	80.3
SWAFN [16]	55.0	53.3	61.03	59.32	80.2	80.1
MAGCN	**56.2**	**54.6**	**61.95**	**59.70**	**80.6**	**80.3**

Table 2: Experimental results (%) of proposed model using unimodal, bimodal and multimodal features. "V" means Video, "A" means Audio, "L" means Language.

	Unimodal			Bimodal			Multimodal
	V	A	L	A+V	L+V	L+A	L+V+A
Acc	59.1	58.6	78.5	59.4	78.8	79.1	**80.6**
F1	59.2	58.4	78.7	59.3	78.5	78.9	**80.3**

characteristic these baselines have in common is that they treat language modality equally with other modalities. In contrast, we specifically implement a language feature learning module in a more straightforward and powerful way and design a consistency loss function to partially guide other modalities to focus on features from language modality. All in all, the experimental results on three datasets demonstrate the effectiveness and robustness of the proposed model.

4.4. Ablation Study

Contribution of Different Modalities We conduct a series of different experiments on the MOSI dataset to investigate the influence of each modality. Table 2 displays the experimental results of using unimodal, bimodal and multimodal features. Firstly, from the unimodal perspective, the use of language features outperforms acoustic features or vision features significantly, which verifies that language modality contains much more crucial information than other modalities. Secondly, from a bimodal perspective, we notice that although the combination of language features and acoustic features or vision features improves the accuracy compared with only utilizing language features, not all F1 scores can be improved. Besides combining acoustic features with video features, the model still gives frustrating results, indicating that the language modality should be dominant for this task. Thirdly, for incorporating three modalities, with the assistance of sentimental knowledge and consistency loss as well as DCGCN, MAGCN achieves state-of-the-art performance.

Contributions of Sentiment Embedding and Consistency Loss To evaluate the contributions of sentiment embedding and consistency loss and DCGCN, ablation studies are per-

Table 3: Ablation study results (%). Highest results are in bold. "w/o SE" means no sentiment embedding is exploited. "w/o CL" means no consistency loss is applied. "w/o DCGC-N" indicates that DCGCN is replaced by Vanilla GCN.

Model	CMU-MOSEI		CMU-MOSI	
	Acc	F1	Acc	F1
MAGCN	**61.95**	**59.70**	**80.61**	**80.32**
MAGCN w/o SE	60.88	58.41	79.43	79.13
MAGCN w/o CL	61.29	59.25	78.82	78.61
MAGCN w/o DCGCN	61.08	59.15	79.27	79.05

formed. In Table 3, the removal of sentiment embedding leads to a drop in both the accuracy and F1 score over two datasets, which indicates that sentiment knowledge is beneficial to this task. We find that for MOSEI, the removal of sentiment knowledge leads to worse performance than the lack of consistency loss. In contrast to the MOSEI, the MOSI dataset tends to highlight the information from language modality since removing the consistency loss leads to the dramatic fall of about 2% in accuracy and F1 score. We suppose that samples from MOSI are more context-aware while data in MOSEI contain more useful sentimental information. The results demonstrate the effectiveness of consistency loss in guiding the acoustic and vision modality to learn more details from language representation, highlighting the necessity of carefully analyzing human language in MSA.

Contributions of Densely Connected Graph Convolutional Neural Nwtworks In Table 3, we also use vanilla GCN to replace DCGCN. Without dense connection, the model is deficient in accurately capturing latent semantic correlations between nodes, especially those connected by in-direct, multi-hop paths cross-modality relations, which causes disappointing performance on both datasets. The results demonstrate that DCGCN is quite crucial for our framework.

5. CONCLUSION

In this paper, we propose a framework to improve the MSA by using self-attention combined with DCGCN to learn inter-modality feature representation, then fuse sentiment knowledge via multi-head self-attention and exploiting a module consists of multi-head self-attention and DCGCN to extract intra-language feature representation. A consistency loss is designed to enhance the commonality between inter-modality feature representations. Experiments on three well-known datasets (YouTube, CMU-MOSI and CMU-MOSEI) show the effectiveness of the proposed model.

6. ACKNOWLEDGEMENTS

7. REFERENCES

[1] Jie Zhou, Jimmy Xiangji Huang, Qinmin Vivian Hu, and Liang He, "Sk-gcn: modeling syntax and knowledge via graph convolutional network for aspect-level sentiment classification," *KBS*, vol. 205, pp. 106292, 2020.

[2] Xingjiao Wu, Luwei Xiao, Yixuan Sun, Junhang Zhang, Tianlong Ma, and Liang He, "A survey of human-in-the-loop for machine learning," *arXiv preprint arXiv:2108.00941*, 2021.

[3] Sijie Mai, Songlong Xing, and Haifeng Hu, "Analyzing multimodal sentiment via acoustic-and visual-lstm with channel-aware temporal convolution network," *TASLP*, vol. 29, pp. 1424–1437, 2021.

[4] Ayush Kumar and Jithendra Vepa, "Gated mechanism for attention based multi modal sentiment analysis," in *ICASSP*, 2020, pp. 4477–4481.

[5] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

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

[7] Paul Pu Liang, Ziyin Liu, AmirAli Bagher Zadeh, and Louis-Philippe Morency, "Multimodal language analysis with recurrent multistage fusion," in *EMNLP*, 2018, pp. 150–161.

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

[9] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency, "Efficient low-rank multimodal fusion with modality-specific factors," in *ACL*, 2018, pp. 2247–2256.

[10] Zeyang Lei, Yujiu Yang, Min Yang, and Yi Liu, "A multi-sentiment-resource enhanced attention network for sentiment classification," in *ACL*, 2018, pp. 758–763.

[11] Hao Tian, Can Gao, Xinyan Xiao, Hao Liu, Bolei He, Hua Wu, Haifeng Wang, and Feng Wu, "Skep: Sentiment knowledge enhanced pre-training for sentiment analysis," in *ACL*, 2020, pp. 4067–4076.

[12] Luwei Xiao, Yun Xue, Hua Wang, Xiaohui Hu, Donghong Gu, and Yongsheng Zhu, "Exploring fine-grained syntactic information for aspect-based sentiment classification with dual graph neural networks," *Neurocomputing*, vol. 471, pp. 48–59, 2022.

[13] Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos, "Found in translation: Learning robust joint representations by cyclic translations between modalities," in *AAAI*, 2019, vol. 33, pp. 6892–6899.

[14] Sijie Mai, Haifeng Hu, and Songlong Xing, "Divide, conquer and combine: Hierarchical feature fusion network with local and global perspectives for multimodal affective computing," in *ACL*, 2019, pp. 481–492.

[15] Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov, "Learning factorized multimodal representations," in *ICLR*, 2018.

[16] Minping Chen and Xia Li, "Swafn: Sentimental words aware fusion network for multimodal sentiment analysis," in *International Conference on Computational Linguistics*, 2020, pp. 1067–1077.

[17] Xiaowen Ding, Bing Liu, and Philip S Yu, "A holistic lexicon-based approach to opinion mining," in *Proceedings of the 2008 international conference on web search and data mining*, 2008, pp. 231–240.

[18] Jianpeng Cheng, Li Dong, and Mirella Lapata, "Long short-term memory-networks for machine reading," in *EMNLP*. Association for Computational Linguistics, 2016, pp. 551–561.

[19] Zhijiang Guo, Yan Zhang, and Wei Lu, "Attention guided graph convolutional networks for relation extraction," in *ACL*, 2019, pp. 241–251.

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *NeurIPS*, 2017, pp. 5998–6008.

[21] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency, "Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos," *arXiv preprint arXiv:1606.06259*, 2016.

[22] 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*, 2018, pp. 2236–2246.

[23] Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi, "Towards multimodal sentiment analysis: Harvesting opinions from the web," in *Proceedings of the 13th international conference on multimodal interfaces*, 2011, pp. 169–176.

[24] Jeffrey Pennington, Richard Socher, and Christopher D Manning, "Glove: Global vectors for word representation," in *EMNLP*, 2014, pp. 1532–1543.

[25] Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer, "Covarepa collaborative voice analysis repository for speech technologies," in *ICASSP*, 2014, pp. 960–964.