ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10096644

A MULTI-STAGE HIERARCHICAL RELATIONAL GRAPH NEURAL NETWORK FOR MULTIMODAL SENTIMENT ANALYSIS

Peizhu Gong *Jin Liu** *Xiliang Zhang* *Xingye Li*

School of Information Engineering, Shanghai Maritime University, China

ABSTRACT

Multimodal sentiment analysis targets at accurately perceiving the emotional states by incorporating related information from multiple sources. However, existing methods mostly neglect the unbalanced contributions and inherent relational interactions across distinct modalities. In this paper, we propose a multi-stage hierarchical relational graph neural network (MHRG), catering to intra- and inter-modal dynamics learning with modality calibration. In the first stage, modality-specific graph convolution modules are introduced to learn the intra-modal sequential semantics. In the second, we design a modality-adaptive modification module to determine the contribution of each modality based on the prediction confidence. Finally, diverse inter-modal dynamics are considered respectively by a novel hierarchical relational graph fusion method for further aggregation according to the type of interactions. Extensive experiments on benchmark datasets demonstrate that MHRG outperforms the existing methods and achieves the state-of-the-art performance.

Index Terms— multimodal sentiment analysis, graph neural network, relational interactions, inter-modal dynamics

1. INTRODUCTION

Multimodal sentiment analysis (MSA), as a critical way to determine an individual's attitude towards a specific entity, has attracted increasing research attention. Earlier works mostly assumed that multimodal sequences are aligned in the resolution of words and modeled the cross-modal interactions on the aligned steps by recurrent neural networks (RNNs). However, the problems with these RNN-based approaches [1, 2] are reflected in two aspects. For one, it is impractical to consider only short-term interactions at the word level. For another, RNNs are prone to slow inferring speed due to its recurrent nature and have limited capacity of learning long-term dependencies. To address such challenges, transformer-based models [3] which learn representations directly from unaligned multimodal streams have been proposed. Tsai et al. [4] proposed a cross-modal Transformer to learn the directional attention between elements across modalities. Wu et al. [5] designed a multi-head attention-based fusion network that considers the interactions between any two pair-wise modalities differently. However, the cross-modal Transformer is a bi-modal operation that only account for two modalities' input at a time, resulting in a large number of parameters needed to retain original modality information.

Recent remarkable works [6] use graph neural networks (GNN) to model intra-modal and inter-modal dynamics. However, previous GNN-based approaches have the following limitations: Modal-temporal attention graph (MTAG) [7] captures the rich interactions across modalities, while ignoring the unbalanced contribution of unimodal representation. Although Huang and Liu's studies [8, 9] take unimodal graph into account, they commonly treat inter-modal dynamics without investigating inherent multi-relational interactions. In sum, few works have paid attention to the unbalanced contributions and inherent relational interactions across distinct modalities. In this paper, we propose a multi-stage hierarchical relational graph neural network (MHRG), catering to intra- and inter-modal dynamics learning with modality calibration. In the first stage, modality-specific graph convolution modules are introduced to learn sequential semantics. Then, we design a modality-adaptive modification module to determine the contribution of each modality according to the prediction confidence. Finally, relation-specific interactions are considered respectively by a novel hierarchical relational graph fusion method for further aggregation. Extensive experiments demonstrate that MHRG outperforms the existing methods and achieves the state-of-the-art performance.

The main contributions are summarized as follows: (1) we propose MHRG that models both intra- and inter-modal dynamics in graph structure. (2) A modality-adaptive modification module is designed to identify the contribution of each modality according to the prediction confidence. (3) A hierarchical relational graph fusion method is proposed to aggregate relation-specific interactions respectively across modalities.

2. METHODOLOGY

The overall architecture of MHRG is depicted in Fig. 1, which is composed of modality-specific graph convolution module, modality-adaptive modification module (MAMM) and hierarchical relational graph fusion module (HRGF). More details are introduced in the following subsections.

2.1. Modality-specific Graph Convolution Module

The input to MHRG consists of three sequences of features from textual (t), visual (v) and acoustic (a) modalities, denoted as $X_m \in \mathbb{R}^{N \times d_m}$, $m \in \{a, v, t\}$. N and $d_{(.)}$ represent the number of time steps and feature dimension. For each

*Correspongding author. Email: jinliu@shmtu.edu.cn

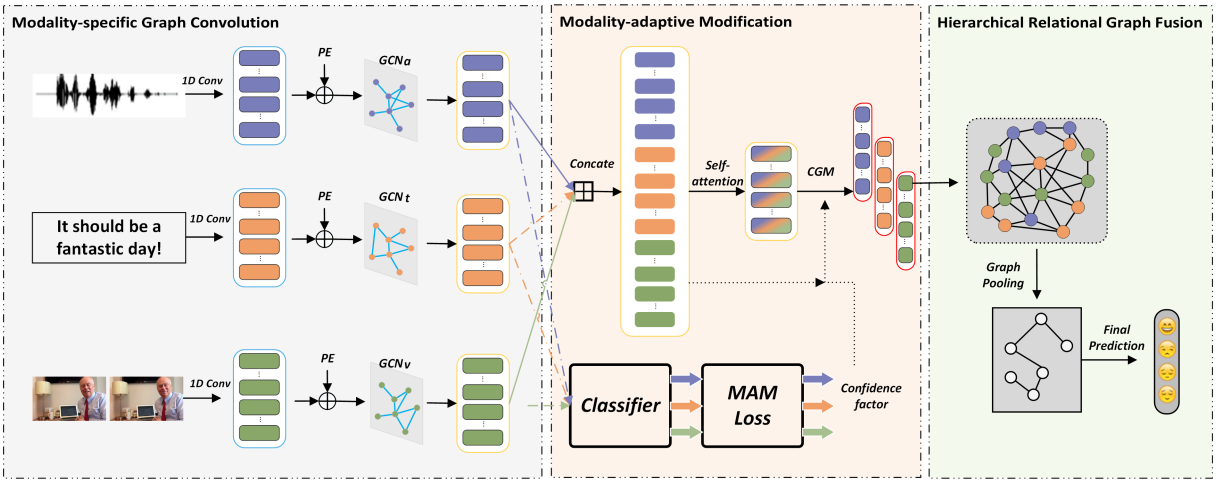


Fig. 1. The overall architecture of the proposed MHRG, which is composed of three main components: a modality-specific graph convolution module, a modality-adaptive modification module and a hierarchical relational graph fusion module. PE denotes the positional encoding, GCN_m denotes the graph convolution for modality m.

modality, we build an undirected graph $G(\mathcal{V}, \mathcal{E})$ to capture sequential semantics, composed of nodes $\mathcal{V}$ and edges $\mathcal{E}$.

Node Generator. Since information density varies between modalities, we pass the input sequence through a 1D temporal convolution layer to overcome this difference. This operation also allows each element of the input sequence to have sufficient awareness of its neighbors.

$$\hat{X}_m = \text{Conv}\,1D\left(X_m, k_m\right) \in \mathbb{R}^{N\times d} \tag{1}$$

where $\hat{X}_m$ represents the node embedding in the unimodal graph, k is the size of the convolutional kernel and d is the common dimension which is the same for all modalities. To avoid losing positional information, we add a positional encoding to each node's feature vector.

Edge Generator. In this paper, the edge set is represented by a similarity matrix. We perform a simple dot-product attention mechanism to measure the similarity between pairwise nodes and define the similarity matrix $A_m \in \mathbb{R}^{N\times N}$.

$$A_m = Q \cdot K^T = \sigma\left(\left(W_Q\hat{X}_m\right)\cdot\left(W_K\hat{X}_m\right)^T\right) \tag{2}$$

where σ is the RELU activation function, W_Q and W_k are learnable parameters. An edge will have a high similarity score if they have a strong semantic relationship.

Modality-specific Graph Convolution. After obtaining the node embeddings and the similarity matrix, we introduce graph convolution to explore the sequential semantics in each modality. To ensure the stability of graph topology structure, we add a residual connection to learn the feature shift from original graph:

$$\begin{gathered}\tilde{X}_m^{l+1} = W_m^r\left(D_m^{-1}A_m\tilde{X}_m^l W_m^l\right) + \tilde{X}_m^l \\ \tilde{X}_m^0 = \hat{X}_m\end{gathered} \tag{3}$$

where $\tilde{X}_m^l$ denotes the node embedding in the l^{th} iteration and $D_m \in \mathbb{R}^{N\times N}$ is the diagonal degree matrix of A_m. $W_m^r \in \mathbb{R}^{N\times N}$ and $W_m^l \in \mathbb{R}^{N\times N}$ are weight matrix acting on the residual structure and node representation respectively.

2.2. Modality-adaptive Modification Module

The core idea of the modality-adaptive modification module (MAMM) is to identify the importance of each modality, which introduces a modality-adaptive modification loss (MAM Loss) and a confidence gated mechanism (CGM). Leveraging the prediction confidence of all the modalities, MAM Loss is designed to dynamically adjust each modal representation from a global perspective. Based on the notion that the smaller the value of loss, the higher the prediction confidence, MAM Loss can be formulated as:

$$l_m^{MAM} = \theta_m \cdot l_m = \left(\prod_{\rho}^{\{a,v,t\}-m} \mu \cdot l_\rho\right)\cdot l_m \tag{4}$$

where θ_m is a confidence factor for modality m, calculated from the rest unimodal loss values and a scaling factor μ. In our method, Mean Absolute Error (MAE) is used for unimodal loss and the scaling factor is set to the arithmetic mean of all unimodal losses.

In addition, we propose a CGM to further minimize the negative impact of noisy unimodal representation on the final prediction. The pipeline of CGM can be summarized as: we concatenate the feature embeddings of all the modalities and perform a self-attention mechanism to calculate a shift vector. This shift vector has the same dimension as the unimodal representation and it will be weighted summed with each modal vector separately to obtain the modified representation. The

Table 1. The comparison experiments on CMU-MOSI and IEMOCAP datasets.

Model\Metric	CMU-MOSI					IEMOCAP							
						Happy		Sad		Angry		Neutral	
	Acc^7	Acc^2	F-score	MAE	Corr	Acc	F-score	Acc	F-score	Acc	F-score	Acc	F-score
RAVEN [1]	33.2	78	76.6	0.915	0.691	77	76.8	67.6	65.6	65	64.1	62	59.5
MCTN [2]	35.6	79.3	79.1	0.909	0.676	80.5	77.5	72	71.7	64.9	65.6	49.4	49.3
MULT [4]	35.3	80.6	79.3	0.972	0.681	84.8	81.9	77.7	74.1	73.9	70.2	62.5	59.7
BIMHA [5]	36.4	80.3	80	0.925	0.671	86.3	85.3	83.6	82.9	74.2	74.4	70.1	71.2
MTAG [7]	38.9	82.3	82.1	0.866	0.722	91.7	88.4	89.3	86.1	89.6	87.2	**79.1**	**72.3**
TGCN [9]	36.9	82.6	82.5	0.871	0.722	92.5	88.9	**90.4**	**87.8**	88.9	86.3	78.9	70.4
Multimodal Graph [10]	34.1	80.6	80.5	0.913	0.698	84.3	81.3	83.7	82.1	75.9	72.7	70.9	69.7
MAGCN [11]	36.8	80.6	80.3	0.882	0.706	85.6	82.3	81.6	78.5	74.3	72.1	67.4	64.8
MHRG	**39.4**	**83.9**	**83.7**	**0.857**	**0.741**	**92.7**	**89.4**	89.7	86.3	**89.8**	**87.5**	76.3	70.7

equations of CGM are given as follows:

$$\begin{aligned} X' &= F^{\alpha}\left(\delta\left(\tilde{X}_a \oplus \tilde{X}_v \oplus \tilde{X}_t\right)\right) \\ X_m^{\text{final}} &= S^{\gamma}\left(\theta_m\right)\tilde{X}_m + \left(1 - S^{\gamma}\left(\theta_m\right)\right)X' \end{aligned} \tag{5}$$

where $\oplus$ represents the concatenation operation, δ is self-attention mechanism, F^{α} is a fully connected layer and S^{γ} is a scoring function which takes confidence factor θ_m as input to determine the extent of modifications.

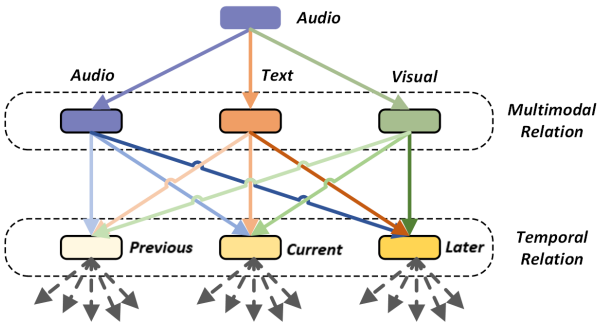


Fig. 2. The details of HRGF module. Multimodal and temporal relations are gradually considered by HRGF.

2.3. Hierarchical Relational Graph Fusion Module

The nodes of the multimodal graph are composed of the modified representations from all modalities. In terms of edges, following [7], we consider two types of inter-modal interactions as directed edges, i.e., multimodal edges and temporal edges, denoted as $\mathcal{E}_M$ and $\mathcal{E}_T$ respectively. Specifically, the multimodal edges are classified into nine types according to diverse subjects and objects, while the temporal edges are divided into "previous", "current" and "later" based on sequence order. As a result, there are a total of 27 interaction combinations. As depicted in Fig. 2, the HRGF incrementally models the inter-modal interactions in multimodal graph for further aggregation. The equation for the l^{th} iteration of HRGF is shown below:

$$h_i^{l+1} = \sigma\left(\sum_{e' \in \mathcal{E}_M} W_{e'}^{l}\left(\sum_{e \in \varepsilon_T}\sum_{j \in \mathcal{N}_i} W_e^{l} h_j^{l}\right) + W_i^{l} h_i^{l}\right) \tag{6}$$

where h_i^l is the hidden state of node v_i in the l^{th} iteration, and $\mathcal{N}_i$ is the set of adjacent nodes of v_i. W_e and $W_{e'}$ represent weight matrix of temporal and multimodal edges, while W_i is the weight for self-loop operation. σ is the RELU nonlinear activation function. Through such progressive relation-specific aggregation, our model can enhance the interactions across distinct modalities and implement deep fusion. After that, graph pooling is introduced to fuse related nodes and simplify the graph structure. In addition, considering the problem that the diversity of inter-modal interactions may cause computational difficulties and disrupt prediction accuracy, we provide an alternative graph pruning strategy. The main idea is to select the *top-k* largest relational values and keep them unchanged, while the rest are reset to zero.

3. EXPERIMENTS

3.1. Datasets

IEMOCAP [12] contains a total of five sets of conversation videos between ten exclusive speakers. It includes 7,433 utterances and each utterance is annotated with a category label. To be consistent with previous studies, we select four of the most common labels, namely "angry", "happy", "sad", and "neutral" for experiments.

CMU-MOSI [13] is a typical multimodal sentiment analysis dataset that contains 2,199 short video clips. Each sample is labeled with a sentiment score from -3 (strongly negative) to 3 (strongly positive). We use the segmentation methods provided in the CMU-SDK [14] to launch our experiments.

3.2. Implementation Details and Evaluation metrics

We develop our model on Pytorch with RTX2080Ti as GPU. GloVe word embeddings, COVEREP, and Facet [15, 16] are applied for extracting the utterance-level features of textual, acoustic and visual modalities respectively. The embedding size of tri-modality is set to 32 for CMU-MOSI and 128 for

Table 2. Ablation studies on CMU-MOSI dataset, which can be divided into five groups.

Ablation	Acc^2	F-score	MAE
Modalities			
T Only	80.2	79.6	0.927
V Only	58.2	58.2	1.317
A Only	59.7	61.3	1.227
A+V	62.3	62.4	1.194
A+T	81.9	80.8	0.905
V+T	82	81.7	0.894
Modification			
No MAM Loss	82.6	81.3	0.868
No CGM	82.8	81.7	0.863
Edge types			
No edge type	80.7	79.7	0.926
mutlimodal edges only	83	82.4	0.864
temporal edges only	82.5	81.3	0.872
Multimodal Fusion			
Concate+FC	80.3	79.6	0.92
cross-modal Transformer [17]	81.7	80.8	0.883
Vanilla GCN [18]	83.2	83.1	0.86
Pruning			
Pruning 50%	71.1	69.3	1.113
Pruning 20%	**86.6**	**86.3**	**0.828**
MHRG	**83.9**	**83.7**	**0.857**

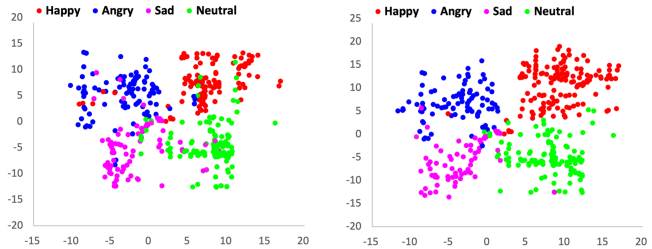
IEMOCAP. The initial learning rate is set to 0.01 for CMU-MOSI and 0.0001 for IEMOCAP. The experimental results are evaluated in two forms: classification and regression. For classification, we report the F-score, 7-class and binary accuracy (Acc^7 and Acc^2). For regression, we report mean absolute error (MAE) and Pearson correlation (Corr). Except for MAE, higher values denote better performance for all metrics.

3.3. Experimental Results and Analysis

The overall performance of the approaches on CMU-MOSI and IEMOCAP datasets is listed in Table 1. Table 1 shows that MHRG outperforms previous methods on CMU-MOSI in all metrics, which demonstrates the superiority of our method. It can also be observed that MHRG achieves the state-of-the-art performance on both Acc^2 and F-score of "happy" and "angry" emotion, competitive performance on "sad", and the worst on "neutral". We infer that one of the major causes is that 'happy' and 'angry' have a relatively more pronounced emotional tendency, whereas 'neutral' does not.

3.4. Ablation Study

We conduct five groups of ablation experiments on the CMU-MOSI dataset to investigate the contribution of each component, and the results are listed in Table 2. Firstly, we study the influence of different modalities. It is observed that all modalities are beneficial for learning better multimodal representations. Among them, textual modality brings the most significant performance gain presumably due to its high information density. Secondly, we verify the role of modification module. For a vivid demonstration, we employ t-SNE [19] to visualize the original and modified audio embedding distributions as an example in Fig. 3. The results show that MAM Loss as well as CGM help to learn more meaningful intra-modal information. Thirdly, we investigate the effect of edge types on our model performance. It highlights the need to take various types of inter-modal interactions into account. And the model performance keeps growing, as we gradually consider multimodal and temporal relations. Then, the multimodal fusion methods are discussed. Benefiting from the structure of hierarchical aggregating, HRGF obtain a better performance compared to vanilla GCN. It also proves that straightforward fusion methods like simple concatenation will burden multimodal learning. Finally, we compare the MHRG with different pruning rates set. It is believed that appropriate pruning can make MHRG focus on the truly important interactions, avoiding over-smoothing issue of GNNs to some extent.



(a) original (b) modified

Fig. 3. Visualization of audio embedding distributions.

4. CONCLUSION

In this paper, we present a multi-stage hierarchical relational graph neural network (MHRG) to model intra and inter-modal dynamics using graph structures. The method concentrates on solving the issues of imbalanced contributions and heterogeneous relations across multiple modalities. For optimizing intra-modal representation, MAMM is designed to dynamically adjust the contribution of each modality according to the prediction confidence. Furthermore, we propose HRGF to aggregate relation-specific dynamics respectively. Extensive experiments on benchmark datasets demonstrate the effectiveness and superiority of MHRG.

5. ACKNOWLEDGMENTS

This work was supported by the National Key Technologies Research and Development Program by the National Key Technologies Research and Development Program of China under Grant 2021YFC2801001, in part by the National Social Science Foundation of China under Grant 20&ZD130.

6. REFERENCES

[1] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, vol. 33, pp. 7216–7223.

[2] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, vol. 33, pp. 6892–6899.

[3] Peizhu Gong, Jin Liu, Yihe Yang, and Huihua He, "Towards knowledge enhanced language model for machine reading comprehension," *IEEE Access*, vol. 8, pp. 224837–224851, 2020.

[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 *Proceedings of the conference. Association for Computational Linguistics. Meeting*. NIH Public Access, 2019, vol. 2019, p. 6558.

[5] Ting Wu, Junjie Peng, Wenqiang Zhang, Huiran Zhang, Shuhua Tan, Fen Yi, Chuanshuai Ma, and Yansong Huang, "Video sentiment analysis with bimodal information-augmented multi-head attention," *Knowledge-Based Systems*, vol. 235, pp. 107676, 2022.

[6] Dou Hu, Xiaolong Hou, Lingwei Wei, Lianxin Jiang, and Yang Mo, "Mm-dfn: Multimodal dynamic fusion network for emotion recognition in conversations," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7037–7041.

[7] 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.

[8] Jiaxing Liu, Sen Chen, Longbiao Wang, Zhilei Liu, Yahui Fu, Lili Guo, and Jianwu Dang, "Multimodal emotion recognition with capsule graph convolutional based representation fusion," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6339–6343.

[9] Jian Huang, Zehang Lin, Zhenguo Yang, and Wenyin Liu, "Temporal graph convolutional network for multimodal sentiment analysis," in *Proceedings of the 2021 International Conference on Multimodal Interaction*, 2021, pp. 239–247.

[10] Luwei Xiao, Xingjiao Wu, Wen Wu, Jing Yang, and Liang He, "Multi-channel attentive graph convolutional network with sentiment fusion for multimodal sentiment analysis," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 4578–4582.

[11] Sijie Mai, Songlong Xing, Jiaxuan He, Ying Zeng, and Haifeng Hu, "Analyzing unaligned multimodal sequence via graph convolution and graph pooling fusion," *arXiv preprint arXiv:2011.13572*, 2020.

[12] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[13] 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.

[14] Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency, "Multi-attention recurrent network for human communication comprehension," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018, vol. 32.

[15] Jeffrey Pennington, Richard Socher, and Christopher D Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[16] Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer, "Covarep—a collaborative voice analysis repository for speech technologies," in *2014 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 2014, pp. 960–964.

[17] Weidong Chen, Xiaofeng Xing, Xiangmin Xu, Jichen Yang, and Jianxin Pang, "Key-sparse transformer for multimodal speech emotion recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6897–6901.

[18] Will Hamilton, Zhitao Ying, and Jure Leskovec, "Inductive representation learning on large graphs," *Advances in neural information processing systems*, vol. 30, 2017.

[19] Laurens Van Der Maaten, "Learning a parametric embedding by preserving local structure," in *Artificial intelligence and statistics*. PMLR, 2009, pp. 384–391.