MECG: Modality-Enhanced Convolutional Graph for Unbalanced Multimodal Representations

Anonymous ACL submission

Abstract

In multimodal sentiment analysis tasks, it is very challenging to model the relationships between different modalities and fusing them. The problem in this area is the unbalance of sentiment representation and distribution across the different modalities, resulting in a fusion process that deviates from the multimodal sentiment semantic space. We propose a novel fusion framework, MECG, which is based on graph convolutional neural networks and provides an efficient approach for fusing unaligned multimodal sequences. With the help of text modalities, we first use the multimodal enhancement module to enhance visual and acoustic modalities for obtaining more discriminative modalities, thus assisting the subsequent aggregation process. In addition, we construct text-driven multimodal feature graphs for modality fusion, which can effectively deal with the unbalance issue among modalities in the graph convolution aggregation process. Finally, we integrate the fused information extracted by MECG into the verbal representation, thus dynamically transforming the original word representations toward the most accurate multimodal sentiment-semantic space. Our model proves its effectiveness and superiority on two publicly available datasets, CMU-MOSI and CMU-MOSEI.

1 Introduction

With the rapid development of multimedia technology, multimodal sentiment analysis (MSA) has become a popular topic, and how to perform efficient sentiment analysis on data with different modalities is a great challenge for artificial intelligence(Ngiam et al., 2011). Compared with individual modal sentiment analysis, multimodal sentiment analysis can help us understand the sentiment behind the data more effectively and precisely, and is therefore widely used in sentiment analysis tasks. In general, the different modalities serve as a complement to each other in order to better bridge the semantic and sentiment divergence (Zhang et al., 2018).

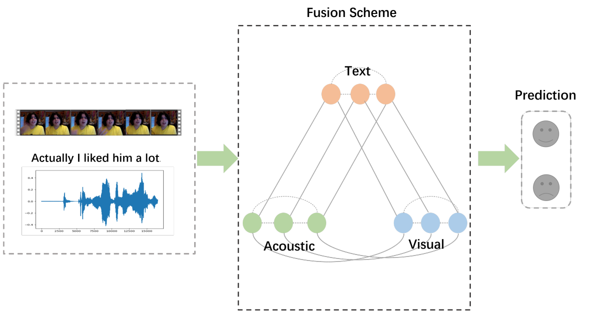

Figure 1: Previous fusion schemes developed by graph construction for the task of multimodal sentiment analysis. The displayed data were taken from the CMU-MOSI dataset.

A key part of MSA is multimodal fusion, where a model aims to extract and integrate information from all input modalities in order to discriminate sentiment. For instance, Zadeh et al. proposed tensor fusion networks to extract unimodal, bimodal, and trimodal interactions (Zadeh et al., 2017), and MISA firstly projected each modality into two different subspaces and investigated the commonalities and specific features from the corresponding subspace (Hazarika et al., 2020). Tsai et al. introduced multimodal converters (MulT) to capture intermodal messages using a cross-modal attention mechanism (Tsai et al., 2019a). Tang et al proposed a Coupled-Translation Fusion Network(Tang et al., 2021) and Multi-Way Multi-Modal Transformer for multimodal learning(Tang et al.). Hu et al. proposed a multimodal GCN fusion scheme to capture the long distance information in emotion recognition (Hu et al., 2021). However, as show in Figure 1, previous fusion structure obtained in multimodal learning task is a ternary symmetric structure, where the bidirectional cross-modalities are modeled in a somewhat identical manner. Notably, it was found in many previous studies (Chen et al., 2017; Sun et al., 2020) that the critical information distributed in the three modalities is unbalanced, where the text modality containing more sentimental information compared to complementary modalities(visual and acoustic)(Guo et al., 2022). Therefore, the architecture that does not consider the relative importance of these three modalities cannot properly integrate them. In order to deal with the above issue, this paper will complete the intermodal information fusion under the dominance of text modality.

We propose an enhancement module for the acoustic and visual modalities in the preprocessing stage, with the aim of being able to compute the most relevant cross-modality sentiment context between the visual and acoustic modalities with respect to the text modality. Actually, the text modality plays the most critical role among the three modalities and contains the richest information (Sun et al., 2020; Chen et al., 2017). Hence, the above enhancement module can be used to allow the acoustic and visual modalities to remove the redundant contextual information within the modality that is less relevant to the sentiment analysis task with the help of the text. Additionally, the above procedure is able to improve the sentiment analysis capabilities of visual and acoustic. Thus helping to further fuse and extract the data among the three modalities in the next step. The proposed graph convolution module MECG is able to extract the complementary information from the other two modalities and the text modality. We believe that the graph structure can well represent the correspondence between the modalities, and the weights between the nodes can not only reflect the important information within the modalities but also the correlation between the modalities. We enter the three modalities into the graph convolution module, and get the neighboring edge values by calculating the node similarity. Therefore, the effective information of visual and acoustic modalities to text modality can be extracted by aggregation calculation respectively. The extracted information is fed into BERT (Devlin et al., 2018a) Encoder as an auxiliary, combined with the text information, to extract the final sentiment contexts via the self-attention mechanism. Our contribution can be summarized as follows:

- We propose a multimodal enhancement module that allows acoustic and visual modalities to eliminate their own redundant information with the help of text modalities to improve emotion discrimination.
- The designed graph fusion framework MECG can learn complementary information of visual and acoustic to text modalities through inter-modal association, thus helping the transformation of text representaions in multimodal semantic space.
- We validate the effectiveness and superiority of the proposed multimodal fusion framework in two publicly available multimodal sentiment analysis databases.

2 Related Work

In this section, we give a brief overview of some related work on multimodal sentiment analysis and fusion.

Due to the ambiguity of single language modality, sentiments are limited in their expressiveness (Williams et al., 2018). This ambiguity typically occurs in scenarios such as slang and sarcasm. To overcome the limitations of single language modality, additional information about nonverbal behavior can be an important complement.

Multimodal sentiment analysis (MSA) is a branch of multimodal language analysis (Zadeh et al., 2018), which predicts the final human emotional expression through the joint analysis of text, acoustic and visual information (Morency et al., 2011). In this field, people focus on the construction of new multimodal neural network and the fusion method between modalities so as to obtain superior effect in sentiment analysis (Liang et al., 2018; Tsai et al., 2019b). In the past few years, deep neural networks have been used to learn multimodal representations in sentiment analysis, such as Long Short-Term Memory (LSTM), which is used to model long-term dependencies from low-level multimodal features. Most previous work in this area has focused on early or late fusion. For example, Zadeh et al. (Zadeh et al., 2017) proposed a tensor fusion network (TFN) that fuses different modal representations at a deeper level. As the attentional mechanism became more popular, Zadeh et al. (Zadeh et al., 2019) modified the LSTM with a new multi-attentional block to capture the interaction between different modalities over time. Recursive Memory Fusion Network (RMFN) (Liang et al., 2018) captures the subtle interactions between modalities in a multi-stage manner, enabling each stage to focus on a subset of signals. The novel training scheme proposed by Yu et al. generates additional unimodal labels for each modality simultaneously with its main task and trains simultaneously with it. Developing modal-specific representations is facilitated by unimodal subtasks (Yu et al., 2021).

Most of the aforementioned research in this area is based on the assumption that multimodal data sources are already aligned. However, more general approaches for sentiment analysis or emotion recognition tasks should be investigated on unaligned multimodal data sources. Multimodal Transformer (MulT) (Tsai et al., 2019a) deploy three transformers for one modality to capture interactions with the other two modalities in a self-attention manner. Yang et al. first converted the unaligned multimodal sequence data into a graph. Then, they designed a method called MTAG to capture the various interactions between modalities (Yang et al., 2021). Hu et al. designed a multimodal fusion convolutional network, MMGCN, which provides a more efficient way to utilize multimodal and remote context information(Hu et al., 2021).

Graph Convolutional Networks have been widely used in the field of multimodal sentiment in the last few years, especially for Emotion Recognition in Conversation (ERC), and have achieved competitive performance due to their ability to handle non-Euclidean numbers and intuitive structure. DialogueGCN(Ghosal et al., 2019) uses a graph-based structure to capture conversational dependencies between corpora. MMGCN(Hu et al., 2021) further proposes a GCN-based multimodal fusion approach for multimodal ERC tasks to improve recognition performance.

However, previous work focused on the correspondence between different modalities, and we emphasized the role of text modality in MECG instead of directly aggregating the information of the three modalities as in previous studies. Additionally, each GCN layer filters node features based on the original neighbor matrix, which limits the effectiveness of the filtering operation. Thus, our multimodal enhancement module processes the raw data first. As stated above, our proposed model is an end-to-end multimodal sentiment analysis method using unaligned multimodal data sources and deal with the issue of unbalance of different modalities. This fusion approach simultaneously establishes the dominance of text modality in multimodal sentiment analysis. To the best of our knowledge, there is little research on modeling unaligned multimodal data sources with text modality as the dominant modality.

3 Methodology

In this section, we first briefly define the problem and then describe our model.

The task of the MSA is to predict the emotional intensity, polarity,or emotional labeling of a given multimodal input (video clip). The video includes three modalities: t(text), a(acoustic) and v(visual). The above modalities are represented as: $M_t \in R^{T_t \times d_t}$, $M_a \in R^{T_a \times d_a}$, $M_v \in R^{T_v \times d_v}$. T_m and d_m represent sequence length (e.g., number of frames) and feature vector size of modality m.

3.1 Modality Encoding

We use $BERT$ (Preo Tiuc Pietro and Devlin Marier. 2019) to encode input sentences. The raw sentence $M_t = (w_1, \ldots, w_n)$, n here means the length of every utterance. We first add [CLS] and [SEP] at the beginning and end of the sentence respectively, and then embed the sentence, the resulting text mode is $X_t = (m_0, m_1, \ldots, m_{n+1})$, For visual and acoustic, we first convolve them to the same time dimension to obtain the processed visual M'_v and acoustic M'_a modalities for subsequent processing. Compared to previous work (Hazarika et al., 2020; Yu et al., 2021), we use two bidirectional LSTM (Hochreiter and Schmidhuber, 1997) to capture the time characteristics of these modes instead of one-way LSTM. Specifically, we use bidirectional Long Short-Term Memory (LSTM) networks to encode sentiment context within acoustic and visual modalities. The above procedure is formulated as follows:

$$X_s = \overrightarrow{LSTM}\left(M'_s\right)\overleftarrow{LSTM} \quad s \in \{a, v\} \qquad (1)$$

3.2 Modality Enhance Module

As shown in the Figure 3, the proposed multimodal enhancement module can compute the complex sentiment contexts within the acoustic and visual modalities that are most relevant to the text modality. First, in the multimodal enhancement module, the input text modality and acoustic modality (or visual modality) are mapped together into a cross-modal sentiment interaction space using a dot product operation. That is, a joint cross-modal sentiment representation space is constructed.The information of the acoustic modality (or visual modality)

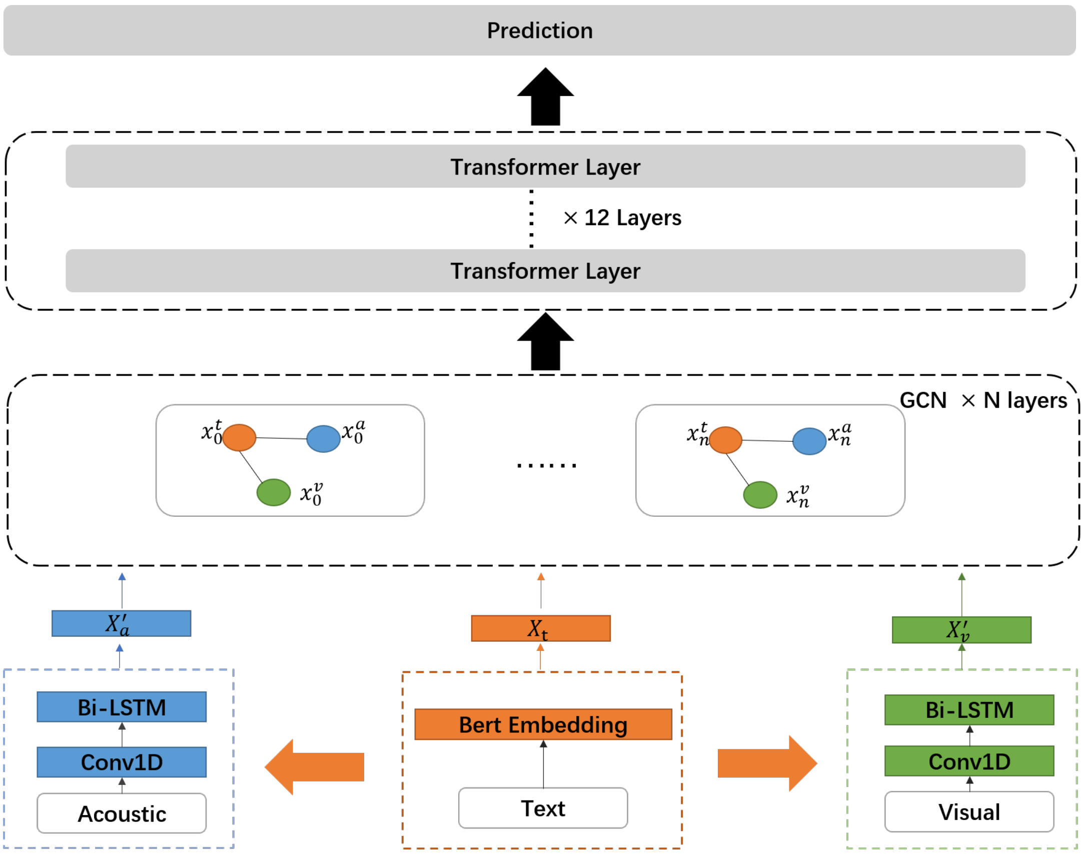

Figure 2: Overall architecture of the model.The model has four components: (1) modality encoding, (2)modality enhance module (3) multimodal GCN interaction module, and (4) sentiment classify. The raw multimodal signals are processed by the modality encoding to obtain numerical sequential vectors. Then, in the modality enhance module, it helps the acoustic and visual modalities to obtain richer sentiment analysis capabilities. And the GCN module extracts complementary information between the three modalities. Finally, theshifted word representations is fed into the BERT Encoder layers to predict the sentiment.

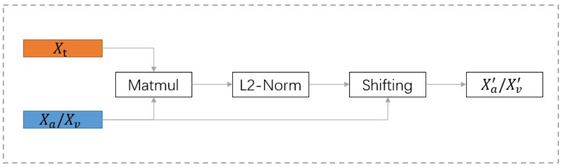

Figure 3: Multimodal enhancement module, where acoustic and visual modalities can obtain richer sentiment analysis capabilities with the help of text modalities

is further enriched in the cross-modal joint representation space with the guidance of text modality. Then, L2 normalization is used to normalize the joint representation data, with the aim of computing the guidance of the text modality on the other two modalities. The above procedure is formulated as follows:

$$X_m = ||X_t \cdot X_s||_2 \quad s \in \{a, v\} \tag{2}$$

After normalizing the joint representation, we use it to change the position of the acoustic modality (or visual modality) in its original semantic space, and finally obtain the acoustic modality X'_a(or visual modality X'_v), which follows:

$$X'_s = X_m + X_s \quad s \in \{a, v\} \tag{3}$$

The above operation can enrich the sentiment context inside the acoustic modality and visual modality to some extent, i.e., the more expressive and discriminative modalities.

3.3 GCN Module

To capture cross-modal sentiment contexts, we construct a spectral domain graph convolution network to encode multimodal contextual information inspired by (Chen et al., 2020; Li et al., 2019; Hu et al., 2021). In the graph convolution module, we will perform text-driven cross-modal fusion, and it is not necessary to stack many layers of graph convolution to obtain great results. In the GCN module, we construct 2N graphs(N is the number of text/viusal/acoustic modalities in the training set), which incudes N text-acoustic graphs $G_{ta} = (V, E)$, and N text-visual graphs $G_{tv} = (V, E)$. V denotes the utterance nodes of two modalities, and the number of nodes in a graph can be freely divided according to the temporal dimension of the modalities. $E \subset V \times V$, is a set containing different modal relationships, which denote the sentiment contexts in the temporal and feature domain. We construct each multimodal graph as follows.

3.3.1 Nodes

The nodes of the three modalities are represented as x_i^t, x_i^a,x_i^v, initialized by the input modal information X_t, X'_a , X'_v. That is, the modalities information corresponding to each sentence is used as nodes. We can divide the number of nodes according to the time dimension and the number of nodes in each graph can be, initialized with the number of modalities.

3.3.2 Edges

Calculating the weights of the edges in each graph is crucial. The traditional graph convolutional neural networks (GCN) attempts to use 1 and 0 to represent weights, which fails to reflect the connection importance between each node. We consider that if two nodes are similar, which means that the correlation between these two nodes is higher. To capture the correlation between different nodes, we use cosine distance to calculate the weights between each graph node (Skianis et al., 2018). The specific representation is formulated as follows:

$$Weight_i = 1 - \frac{\arccos\left((x_i^t, x_i^s)\right)}{\pi} \quad s \in \{a, v\} \tag{4}$$

Based on the above basic information of node and edge weights, we construct multiple shallow bimodal undirected graphs to compute the correlation between the text modality and the other two modalities. Specifically, the reformulated graph Laplacian matrix (Kipf and Welling, 2016) of the undirected graph $G = (V, E)$ is $\widetilde{\mathcal{P}}$:

$$\widetilde{\mathcal{P}} = \mathcal{D}^{-1/2}\tilde{\mathcal{A}}\mathcal{D}^{-1/2} = (\mathcal{D}+\mathcal{I})^{-1/2}(\mathcal{A}+\mathcal{I})(\mathcal{D}+\mathcal{I})^{-1/2} \tag{5}$$

where $\mathcal{A}$ denotes the adjacency matrix, $\mathcal{D}$ denotes the diagonal matrix of the graph G, and $\mathcal{I}$ denotes the unit matrix. The GCN iterations of different layers can be expressed as:

$$\mathcal{H}^{(l+1)} = \sigma\left(\left((1-\alpha)\widetilde{\mathcal{P}}\mathcal{H}^{(l)} + \alpha\mathcal{H}^{(0)}\right)\left((1-\beta)\,\mathcal{I} + \beta\mathcal{W}^{(l)}\right)\right) \tag{6}$$

where α and β are two hyperparameters, σ denotes the activation function. $\mathcal{W}^{(l)}$ is the learnable weight matrix. $\beta = \log\left(\frac{\eta}{l} + 1\right)$, where η is also a hyperparameter. The remaining connections to the first layer $\mathcal{H}^{(0)}$ are added to the representation $\widetilde{\mathcal{P}}\mathcal{H}^{(l)}$ and the constant mapping $\mathcal{I}$ is added to the weight matrix $\mathcal{W}^{(l)}$.

We will concatenate together the features obtained from each graph for the subsequent classification:

$$h_s = [h_{ta}, h_{tv}] \tag{7}$$

3.4 Sentiment Classification

We then utilize the the linear transformation layer to shift h_s to the semantic space of text modality X_t. And, the output of linear transformation layer is further transmitted to the BertEncoder that is initialized in huggingface (Devlin et al., 2018b) and associated with 12 layers. Note that, the first token of the output vector of the last layer refers to [CLS] that comprised the information needed for classification task. Then, we use a linear layer to analyze the obtained $h_s^{''}$, leading to the y utilized to reach the final sentiment prediction.

$$\begin{aligned} h_s^{'} &= Dropout\left(LayerNorm\left(h_s + X_t\right)\right) \\ h_s^{''} &= BERT(h_s^{'}) \\ y &= Wh_s^{''} + b \end{aligned} \tag{8}$$

We use mean square error (MSE) as the loss function in this sentiment analysis task because it is a regression task.

$$\mathcal{L}_{task} = MSE(y, \hat{y}) \tag{9}$$

4 Experiment

4.1 Datasets

We evaluated our model on two public datasets.

- **CMU-MOSI** dataset (Zadeh et al., 2016) is a universal benchmark for evaluating the performance of fusion networks in emotional intensity prediction tasks. This data set contains many YouTube videos. It contains 2,199 speech videos, edited from 93 videos played by 89 different narrators. Each segment was manually labeled with a real score, ranging from -3 to +3, indicating the relative strength of negative (score below zero) or positive (score above zero) emotions

- **CMU-MOSEI** dataset (Zadeh et al., 2018) is an upgraded version of CMU-MOSI on sample number. It also enriches the versatility of the speaker, covering a wider range of topics. This dataset contains 23,453 video clips, which are annotated in the same way as CMU-MOSI. The clips were pulled from 5,000 videos, involving 1,000 different speakers and 250 different topics.

4.2 Preprocessing

The multimodal signals in our experiments were unaligned (Yu et al., 2021).

Text Modality The current work is in favor of advanced pre-trained language models. Therefore, the raw text input is encoded by BERT in all experiments.

Acoustic Modality P2FA(Yuan et al., 2008) is used to extract visual features. It is a common toolkit and have been used frequently.

Visual Modality We used COVAREP(Degottex et al., 2014), which is a professional acoustic analysis framework for feature extraction.

4.3 Baselines and Metrics

We compare our results with several advanced multimodal fusion frameworks. We consider pure learning-based models such as TFN (Zadeh et al., 2017), LMF (Liu et al., 2018), MFM (Tsai et al., 2019b) and MulT (Tsai et al., 2019a). And methods involving feature space manipulation, such as ICCN(Sun et al., 2020) and MISA(Hazarika et al., 2020). We also compared recent competitive baselines for our models, such as MAG-Bert (Rahman et al., 2020), Self-MM (Yu et al., 2021), MMIM (Han et al., 2021),and MMGCN (Hu et al., 2021).

- TFN (Zadeh et al., 2017): Tensor Fusion network extracts the local features of the three modes and uses the 3-fold Cartesian product to decompose each mode into tensors for internal and external fusion of the modalities

- LMF (Liu et al., 2018): The approach is to decompose the high-order tensors of modal into many low-order factors, and then fuse them based on these factors.

- MFM (Tsai et al., 2019b): It is a generative discriminant model, which connects inference network and generative network with specific modal factors for fusion.

- MULT (Tsai et al., 2019a): It leverages the cross-modal attention mechanism, allowing stacked Transformer networks to mitigate the time alignment problems of multimodal signals.

- ICCN (Sun et al., 2020): Interactive canonical correlation networks rely on mathematical metrics to minimize canonical losses between modal pairs to achieve fusion.

- MISA (Hazarika et al., 2020): In this paper, different modes are mapped to two independent feature spaces to complete the fusion.

- MAG-BERT (Rahman et al., 2020): Multimodal Adaptation Gate designed a multimodal adaptive gate, which was added to different pre-training models, such as BERT and XLnet.

- SELF-MM (Yu et al., 2021): Self-supervised Multi-Task Learning assigns each modal unimodal training task to find the optimal, so that the back propagation of gradient can be adjusted in the multi-modal task

- MMIM (Han et al., 2021): MultiModal InfoMax preserves mission-critical information, it layered to maximize mutual information in the multimodal fusion pipeline.

- MMGCN (Hu et al., 2021): The framework uses deep graph convolutional aggregation of data from three modal inputs in parallel in the ERC task to obtain long distance multimodal information.

4.3.1 Metrics

A set of measures we used are: mean absolute error (MAE), the average absolute difference between the predicted value and true value. Pearson correlation (Corr), which measures the degree of prediction bias, represents the proportion of predictions that correctly fall into the same range of the seven ranges between -3 and +3, And binary classification accuracy (Acc-2) and F1 scores calculated for positive/negative and non-negative/negative classification results.

4.3.2 Basic Settings

The hyperparameter settings are as follows: The number of GCN layers of MOSI and MOSEI is 1 layer. Dropout is 0.5. The learning rate is 0.00001.

item	MOSI	MOSEI
learning rate	1e-5	1e-5
α	0.2	0.5
η	0.9	0.4
LSTM hidden dim	15	15
dropout	0.5	0.5

Table 1: Hyperparameters for best performance

α and η were set to [0.2, 0.9] and [0.5,0.4] in MOSI and MOSEI. The batch sizes for MOSI and MOSEI are [48,128,128] and [64,128,128] for train, valid as well as test.

5 Results and Discussions

Table 3 shows the results of MECG on the CMU-MOSI and CMU-MOSEI datasets compared with the baseline models. Based on the data alignment requirements, we classify the baseline models into two categories: Unaligned and Word Aligned. Compared with the adopted baseline, the proposed models yielded competitive results, and MECG still outperformed the models using aligned multimoal data without aligning the mltimodal data in advance. We compared our proposed model with all baseline models of the CMU-MOSI and CMU-MOSEI datasets in multimodal settings described above. In more detail, MECG is superior to SOTA in some metrics (Acc-2, F1) of CMU-MOSI and CMU-MOSEI. For the other metrics, we achieved close performance to SOTA.

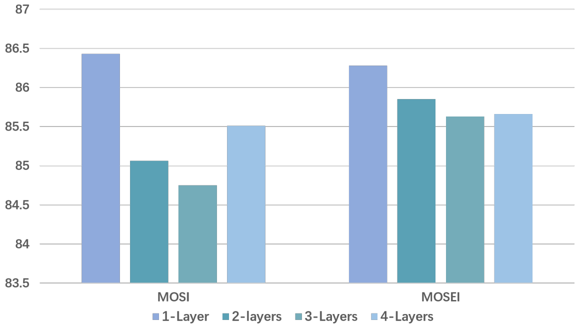

Figure 4: We study the effect of different GCN aggregation layers on CMU-MOSI and CMU-MOSEI dataset.

5.1 Ablation Study

We conducted an ablation study on the CMU-MOSEI dataset to test the functionality of the overall architecture and components presented in this paper.

First, as shown in Figure 4. The results showed that it is not necessary to stack too many layers of the GCN, but only to perform one aggregation and residual concatenation for each multimodal graph to obtain the best results. In addition, as shown in Table 2. In seven experiments, we validate the effectiveness of the bimodal fusion architecture in GCN module. Specifically, we (1) used only one pair as input; (2) freely combined three multimodal bimodal pairs, (TV/TA/VA); and (3) added a visual-acoustic(VA) complementary module to build a ternary symmetric model.

In terms of results, type (2) outperforms type (1), indicating the importance of all three modalities. Moreover, the performance of the acoustic-focused input (TA+VA) is close to that of the text-focused input (our TV+TA), i.e., our architecture can operate on these modalities pairs as well. In contrast, when visual-acoustic (VA) input pairs are added, the performance of type (3) decreases. This means that, even if all modalities are included in the input, the redundant structure can introduce malicious noise, corrupt useful information and clutter the model. The largest improvement over the baseline shown in Table 2 comes from TV+TA pairs.

To investigate the effect of the fine-grained granularity of time on the graph convolution module, as shown in Table 4, we cut the nodes in each multimodal graph according to the time dimension, yielding a different number of nodes for aggregation. The results show that in the MOSI dataset, we obtained optimal results by dividing each modality of each graph into two nodes by time dimension for aggregation. In contrast, in the MOSEI dataset, the optimal results are obtained without division.

5.2 Further Analysis

We use the visualization method t-SNE to visualize and analyze the experimental results, which

Node Num(each modality)	MOSI	MOSEI
1	86.28	**86.24**
2	**86.43**	86.21
3	85.67	86.13
4	85.67	86.10
5	85.97	86.18
10	85.82	85.99

Table 2: We divide each modality to different numbers of nodes in each graph and compute the Acc-2 on the CMU-MOSI and CMU-MOSEI.

	(Word Aligned) MOSI				(Word Aligned) MOSEI			
Models	MAE	Corr	Acc-2	F1	MAE	Corr	Acc-2	F1
LMF	0.917	0.695	-/82.5	-/82.4	0.623	0.677	-/82.0	-/82.1
MFM	0.877	0.706	-/81.7	-/81.6	0.568	0.717	-/84.4	-/84.3
ICCN	0.862	0.714	-/83.0	-/83.0	0.565	0.713	-/84.2	-/84.2
MISA	0.804	0.764	80.79/82.10	80.77/82.03	0.568	0.724	82.59/84.23	82.67/83.97
MAG-BERT	0.731	0.789	82.5/84.3	82.54/84.43	0.543	0.755	82.51/84.82	82.77/84.71
	(Unaligned) MOSI				(Unaligned) MOSEI			
Models	MAE	Corr	Acc-2	F1	MAE	Corr	Acc-2	F1
TFN	0.901	0.698	-/80.8	-/80.7	0.593	0.700	-/82.5	-/82.1
MULT	0.861	0.711	81.5/84.0	80.6/83.9	0.580	0.703	-/82.5	-/82.3
MMGCN	0.757	0.775	83.40/85.23	83.38/85.22	0.580	0.770	82.84/85.10	82.91/84.99
Self-MM	0.712	0.795	82.54/84.77	82.68/84.91	0.529	0.767	82.68/84.96	82.95/84.93
MMIM	0.700	0.800	84.14/86.06	84.00/85.98	0.526	0.772	82.24/85.97	82.66/85.94
Ours(No enhance)	0.750	0.783	83.53/85.67	83.47/85.55	0.575	0.798	83.00/85.77	83.24/85.69
Ours	0.737	0.795	**84.44/86.43**	**84.37/86.31**	0.565	**0.799**	**83.32/86.24**	**83.42/86.18**

Table 3: Results on CMU-MOSI and CMU-MOSEI. All models use BERT as the text encoder. For Acc-2 and F1, we have two sets of non-negative/negative (left) and positive/negative (right) evaluation results. NOTE: - means the results are not given in the paper.

Description(in GCN module)	MAE	Corr	Acc-2	F1
TV only	0.577	0.797	85.80	85.66
TA only	0.576	0.797	85.88	85.82
VA only	0.577	0.796	85.69	85.63
TA+VA (Acoustic-driven GCN)	0.569	0.798	85.91	85.85
TV+VA (Visual-driven GCN)	0.575	0.798	85.55	85.49
TV+TA (Text-driven GCN)	**0.565**	**0.799**	**86.24**	**86.18**
TV+TA+VA	0.566	0.798	85.96	85.91

Table 4: The effect of modalities on the CMU-MOSEI dataset

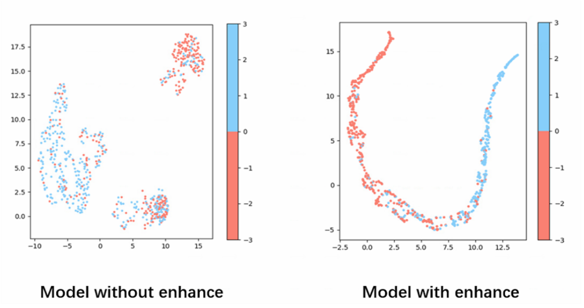

Figure 5: The t-SNE visualization of the effect of multimodal enhance module

can present the data distribution of the multimodal sentiment fusion information obtained from model learning more intuitively. Figure 5 shows what data information characterizes both positive and negative emotions for the sentiment classification task. Blue dots correspond to positive emotion data, red dots refers to negative emotion data. After learning our model, we obtain more discriminative multimodal sentiment fusion information.Figure 5 shows a further visual analysis of the multimodal enhancement module, and the subplot on the left is the result without the modality enhance module. The classification effect is obviously not as good as the result of the right side subplot which includes the multimodal enhance module.

6 Conclusion

In this paper, we propose a text-driven multimodal fusion framework, MECG, for multimodal sentiment analysis tasks. We propose multimodal enahnce module, which helps to enrich the modalities reperesentations and remove redundant information before constructing multimodal graphs and aggregating multimodal information, thus improving multimodal information imbalance during graph convolution. MECG constructs multimodal transformed word representations that dynamically capture changes in different nonverbal contexts. We conducted comprehensive experiments on two datasets (CMU-MOSI, CMU-MOSEI) followed by an ablation study, and the results validated the effectiveness and necessity of our fusion process.

References

Ming Chen, Zhewei Wei, Zengfeng Huang, Bolin Ding, and Yaliang Li. 2020. Simple and deep graph convolutional networks. In *International Conference on Machine Learning*, pages 1725–1735. PMLR.

Minghai Chen, Sen Wang, Paul Pu Liang, Tadas Baltrušaitis, Amir Zadeh, and Louis-Philippe Morency. 2017. Multimodal sentiment analysis with word-level fusion and reinforcement learning. In *Proceedings of the 19th ACM international conference on multimodal interaction*, pages 163–171.

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

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018b. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 154–164.

Jiwei Guo, Jiajia Tang, Weichen Dai, Yu Ding, and Wanzeng Kong. 2022. Dynamically adjust word representations using unaligned multimodal information. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3394–3402.

Wei Han, Hui Chen, and Soujanya Poria. 2021. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9180–9192.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM international conference on multimedia*, pages 1122–1131.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Jingwen Hu, Yuchen Liu, Jinming Zhao, and Qin Jin. 2021. Mmgcn: Multimodal fusion via deep graph convolution network for emotion recognition in conversation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5666–5675.

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Guohao Li, Matthias Muller, Ali Thabet, and Bernard Ghanem. 2019. Deepgcns: Can gcns go as deep as cnns? In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9267–9276.

Paul Pu Liang, Ziyin Liu, AmirAli Bagher Zadeh, and Louis-Philippe Morency. 2018. Multimodal language analysis with recurrent multistage fusion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 150–161.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. 2018. Efficient low-rank multimodal fusion with modality-specific factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2247–2256.

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

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. 2011. Multimodal deep learning. In *ICML*.

Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, AmirAli Bagher Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. 2020. Integrating multimodal information in large pretrained transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2359–2369.

Konstantinos Skianis, Fragkiskos Malliaros, and Michalis Vazirgiannis. 2018. Fusing document, collection and label graph-based representations with word embeddings for text classification. In *Proceedings of the Twelfth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-12)*, pages 49–58.

Zhongkai Sun, Prathusha Sarma, William Sethares, and Yingyu Liang. 2020. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8992–8999.

Jiajia Tang, Kang Li, Ming Hou, Xuanyu Jin, Wanzeng Kong, Yu Ding, and Qibin Zhao. Mmt: Multi-way multi-modal transformer for multimodal learning.

Jiajia Tang, Kang Li, Xuanyu Jin, Andrzej Cichocki, Qibin Zhao, and Wanzeng Kong. 2021. Ctfn: Hierarchical learning for multimodal sentiment analysis using coupled-translation fusion network. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5301–5311.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019a. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569.

Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019b. Learning factorized multimodal representations. In *International Conference on Representation Learning*.

Jennifer Williams, Steven Kleinegesse, Ramona Comanescu, and Oana Radu. 2018. Recognizing emotions in video using multimodal dnn feature fusion. In *Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML)*, pages 11–19.

Jianing Yang, Yongxin Wang, Ruitao Yi, Yuying Zhu, Azaan Rehman, Amir Zadeh, Soujanya Poria, and Louis-Philippe Morency. 2021. Mtag: Modal-temporal attention graph for unaligned human multimodal language sequences. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1009–1021.

Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. 2021. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10790–10797.

Jiahong Yuan, Mark Liberman, et al. 2008. Speaker identification on the scotus corpus. *Journal of the Acoustical Society of America*, 123(5):3878.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114.

Amir Zadeh, Chengfeng Mao, Kelly Shi, Yiwei Zhang, Paul Pu Liang, Soujanya Poria, and Louis-Philippe Morency. 2019. Factorized multimodal transformer for multimodal sequential learning. *arXiv preprint arXiv:1911.09826*.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6):82–88.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. 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)*, pages 2236–2246.

Yazhou Zhang, Dawei Song, Peng Zhang, Panpan Wang, Jingfei Li, Xiang Li, and Benyou Wang. 2018. A quantum-inspired multimodal sentiment analysis framework. *Theoretical Computer Science*, 752:21–40.