Modeling Intra- and Inter-Modal Relations: Hierarchical Graph Contrastive Learning for Multimodal Sentiment Analysis

Zijie Lin [1], Bin Liang [1]*, Yunfei Long [2], Yixue Dang [3], Min Yang [4], Min Zhang [1], Ruifeng Xu [1,5,6]*

[1] Harbin Institute of Technology, Shenzhen, China
[2] University of Essex, Colchester, UK [3] China Merchants Securities Co ., Ltd.
[4] Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences
[5] Peng Cheng Laboratory, Shenzhen, China
[6] Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies
lzjjeffery@163.com, 18B951033@stu.hit.edu.cn
xuruifeng@hit.edu.cn

Abstract

The existing research efforts in Multimodal Sentiment Analysis (MSA) have focused on developing the expressive ability of neural networks to fuse information from different modalities. However, these approaches lack a mechanism to understand the complex relations within and across different modalities, since some sentiments may be scattered in different modalities. To this end, in this paper, we propose a novel hierarchical graph contrastive learning (HGraph-CL) framework for MSA, aiming to explore the intricate relations of intra- and inter-modal representations for sentiment extraction. Specifically, regarding the intra-modal level, we build a unimodal graph for each modality representation to account for the modality-specific sentiment implications. Based on it, a graph contrastive learning strategy is adopted to explore the potential relations based on unimodal graph augmentations. Furthermore, we construct a multimodal graph of each instance based on the unimodal graphs to grasp the sentiment relations between different modalities. Then, in light of the multimodal augmentation graphs, a graph contrastive learning strategy over the inter-modal level is proposed to ulteriorly seek the possible graph structures for precisely learning sentiment relations. This essentially allows the framework to understand the appropriate graph structures for learning intricate relations among different modalities. Experimental results on two benchmark datasets show that the proposed framework outperforms the state-of-the-art baselines in MSA.

1 Introduction

Multimodal Sentiment Analysis (MSA) has received increasing research attention in recent years. Different from textual sentiment analysis, MSA generally contains three modalities: text (caption), visual and acoustic, in which the visual and acoustic information can complement the text for identifying the sentiment score and thus preferably detecting the sentiment polarity (e.g. *positive*, *negative*, etc.). As shown in Figure 1, although the text modality expresses the positive sentiment, we can infer that the correct sentiment of this example is *negative* in the light of the study of visual and acoustic information. Therefore, dealing with MSA needs to consider learning and fusing the information from different modalities.

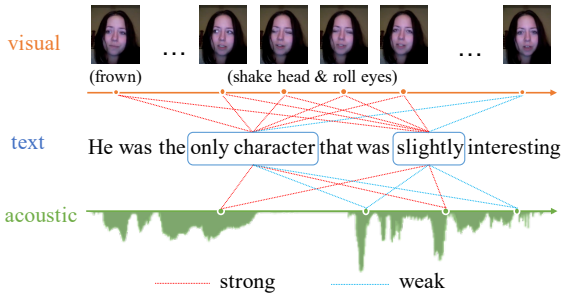

Figure 1: An example of intricate relations among different modalities in MSA.

Early MSA work attempted to fuse the information from different modalities by tensor-based features fusion (Snoek et al., 2005; Zadeh et al., 2017; Liu et al., 2018) or attention-based features fusion (Zadeh et al., 2018b,a; Tsai et al., 2019a). Furthermore, some representation learning-based approaches (Tsai et al., 2019b; Hazarika et al., 2020) aim to model the consistency and the variability between modalities for extracting the sentiment cues among modalities or consider both fusion and alignment of multimodal sequential data with a graph model (Yang et al., 2021). Despite the promising progress made by current work, they generally focus on fusing multimodal representations via class-driven supervised learning or multi-task learning, which fails to understand the intricate

* Corresponding author.

Proceedings of the 29th International Conference on Computational Linguistics, pages 7124–7135
October 12–17, 2022.

relations within and across modalities for better sentiment extraction. As shown in Figure 1, the sentiment is scattered within each modality and across different modalities.

In this paper, we study how to understand representations within and across modalities, enabling the highly correlated modal representations to be explicitly linked for learning the multimodal sentiment information. To reach this goal, in the light of developing the merit of graph structure for modeling intricate relations of representations, we first build a unimodal graph for each modality and further build a multimodal graph for each instance based on the unimodal graphs. Concretely, for the intra-modal graphs, to account for the underlying relations within each modality, we construct a syntax-aware graph for the text modality based on the dependency tree of the sentence and build sequential connection graphs for visual and acoustic modality. For the inter-modal graph, we build a fully-connected inter-modal graph based on the modality-specific graphs to capture the potential relations across different modalities. Then, we apply a graph attention networks (Veličković et al., 2018) architecture to model the semantic relations by means of specifying different weights to different nodes in a neighborhood, without requiring any costly matrix operation (such as inversion) or depending on knowing the graph structure upfront.

Following that, we propose a hierarchical graph contrastive learning (HGraph-CL) framework to model the correlation and difference of graph information within a specific modality and further across different modalities. Specifically, for the intra-modal level, inspired by You et al. (2020), we first devise a self-supervised graph contrastive learning strategy based on the graph augmentations, aiming to explore more appropriate graph structures and derive robust graph representations for each modality. In addition, inspired by Khosla et al. (2020); Gunel et al. (2021), we employ a supervised contrastive learning strategy to make sense of the correlation and difference between different classes, so as to capture the similarity between examples in one class and contrast them with examples in other classes. Moreover, for the inter-modal level, we also perform these two contrastive learning strategies to learn the graph representations for better generalizability, transferability, and robustness in learning sentiment cues compared with pure class-driven methods.

The main contributions of our work can be summarized as follows:

(1) The MSA task is approached from a novel perspective that explores intra- and inter-modality graph construction to leverage the potential sentiment relations within and across modalities.

(2) A novel hierarchical graph contrastive learning (HGraph-CL) framework is devised for better sentiment relations extraction at an intra-modal level and further at an inter-modal level.

(3) Performance evaluation on two benchmark datasets shows the superiority and robustness of the proposed framework compared to several competitive baselines.

2 Related Work

2.1 Multimodal Sentiment Analysis

The MSA task aims to predict sentiment polarity by aiding text with visual and acoustic information. Since the raw visual and acoustic data is in frames and the text is in words, MSA can be broadly classified into word-level and utterance-level depending on the granularity of the data used. Among them, utterance-level approaches (Zadeh et al., 2017; Liu et al., 2018; Yu et al., 2021) perform modal fusion in global representation, while word-level approaches (Tsai et al., 2019b; Wang et al., 2019; Tsai et al., 2019a; Rahman et al., 2020; Hazarika et al., 2020; Wu et al., 2021) are more concerned with local modal interactions. Furthermore, Rahman et al. (2020) proposes a Multimodal Adaptation Gate (MAG) mechanism to perform modal fusion at word-level, which does not rely on complex structure and can be embedded in pre-trained attention models. Hazarika et al. (2020) proposes a multimodal representation learning framework to model modality-invariant and modality-specific information within the example by projecting each modality to two distinct subspaces. Besides, Yu et al. (2021) trains MSA together with three unimodal sentiment analysis tasks, and proposes a heuristic approach to generate unimodal labels. Yang et al. (2021) proposes a novel graph-based neural network to analyze multimodal sequential data for MSA. We propose a novel graph-based approach to address MSA's modal interactions and fusion problem. In contrast to the existing graph works in MSA, our proposed method constructs intra-modal graphs based on prior knowledge of the modalities (e.g., textual dependencies). Furthermore, we create edges between any two nodes from different

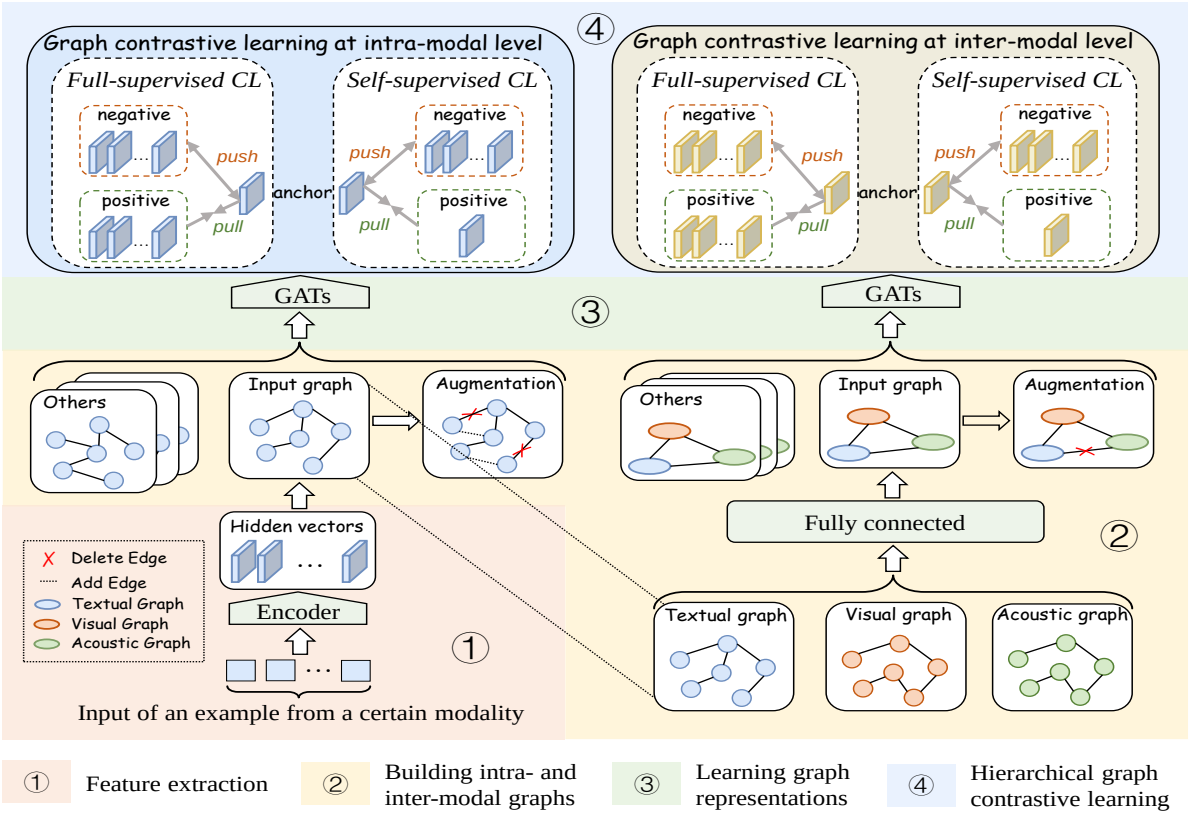

Figure 2: The architecture of the proposed HGraph-CL framework.

intra-model graphs for more complex inter-modal relationships to preserve any possible association.

2.2 Contrastive Learning

Our work also relates to contrastive learning. Contrastive learning (CL) is originally proposed as a self-supervised learning method for solving the lack of supervised signals (Chen et al., 2020; Liu et al., 2021). However, CL often requires effective data augmentation as a foundation. Recent work (Khosla et al., 2020) proposes supervised contrastive learning methods in combination with class information, which is capable of learning the class distribution of examples and without data augmentation. Mai et al. (2022) proposes a hybrid contrastive learning strategy for MSA, but lacks exploring the potential relationship within and among modalities. The combination with graph networks is another new application of contrastive learning (You et al., 2020; Zhu et al., 2020). The graph networks can model the association between nodes, and data augmentation on graph structures is feasible and operable. Common augmentation methods include additions and deletions of nodes or edges, masking of the representations of nodes or edges, etc. Therefore, to explore more appropriate graph structures, inspired by You et al. (2020), we apply the graph augmentations by deleting and adding edges in graphs, and thus derive multifarious but similar graph structures with respect to the source.

3 Method

In this section, we describe the proposed HGraph-CL framework in detail. As illustrated in Figure 2, the framework mainly consists of four components:

1) **Feature extraction**, which applies BERT (Devlin et al., 2019) and BiLSTMs (Hochreiter and Schmidhuber, 1997) to extract features from the three modalities of text, images and audio.

2) **Building intra- and inter-modal graphs**, which constructs intra- and inter-modal graphs based on the hidden vectors learned from the text, images, and audio.

3) **Learning graph representations**, which learns intra- and inter-modal graph representations and leverages the potential sentiment relations within and across modalities. We believe a graph network can model the complex relationship between different modalities while updating the node representations.

4) **Hierarchical graph contrastive learning**, which performs contrastive learning based on heterogeneous graphs at the intra-modal level and inter-modal level. Contrastive learning can help the model understand the similarity and differences of the data across different modalities. Moreover, subtle differences in the graphs may also affect the learning of models on samples. Therefore, we propose the hierarchical graph contrastive learning strategy to augment the learning of the graph representations at both the data level and label level.

3.1 Task Definition

Formally, supposing there is an example consisting of a text t and the corresponding image frames v and audio a from a video, the goal of multimodal sentiment analysis (MSA) is to predict a sentiment score y, which is a constant from -3.0 to 3.0, for each example. Additionally, according to the sentiment score y, we thus identify the sentiment polarity (i.e. *positive* if $y > 0$ or *negative* if $y < 0$).

3.2 Feature Extraction

Given an input example with L tokens for each modality $\boldsymbol{x} = (\boldsymbol{x}^t, \boldsymbol{x}^v, \boldsymbol{x}^a)$, where t,v, and a denote the text, visual, and acoustic modalities respectively. Then, three encoders are used to extract features from the three modality data. Among them, the encoder of text modality is BERT (Devlin et al., 2019), which takes the text representation $\boldsymbol{x}^t$ as input to derive the hidden representations $\boldsymbol{e}^t \in \mathbb{R}^{L \times d_t}$ of all the tokens of the text modality:

$$\boldsymbol{e}^t = \text{BERT}([CLS]\boldsymbol{x}^t[SEP])_{1:L} \quad (1)$$

where d_t is the dimension of text hidden vectors.

We use Facet (Zhu et al., 2006) to extract a set of visual features, including facial markers, facial action units, head pose, visual trajectory, and HOG features. And we use COVAREP (Degottex et al., 2014) to extract a set of low-level acoustic features, including 12 mel cepstral coefficients (MFCCs), pitch tracking and turbid/clear segmentation features, gating source parameters, peak slope parameters, and maximum dispersion quotient. The visual/acoustic features are aligned with the text at token level by averaging the frames of video/audio recording over the time interval align to a token. The lengths of obtained sequences $\boldsymbol{x}^v$ and $\boldsymbol{x}^a$ are the same as the text sequences $\boldsymbol{x}^t$.

Owing to the sequential structure of visual and acoustic modalities, we adopt BiLSTM (Hochreiter and Schmidhuber, 1997) as visual and acoustic encoders to embed each token into a d_v-dimensional vector and a d_a-dimensional vector respectively. Here, the encoded visual representations $\boldsymbol{e}^v \in \mathbb{R}^{L \times d_v}$ and acoustic representations $\boldsymbol{e}^a \in \mathbb{R}^{L \times d_a}$ are computed as follow:

$$\boldsymbol{e}^v = \text{BiLSTM}(\boldsymbol{x}^v) \quad (2)$$

$$\boldsymbol{e}^a = \text{BiLSTM}(\boldsymbol{x}^a) \quad (3)$$

Then, the representations of the three modalities are mapped to the same d_h-dimensional hidden vector space using three projection layers $\text{p}^t(\cdot), \text{p}^v(\cdot), \text{p}^a(\cdot)$, which are presented as $\boldsymbol{h}^t, \boldsymbol{h}^v, \boldsymbol{h}^a \in \mathbb{R}^{L \times d_h}$. A projection layer is a dense layer with a ReLU activation function.

3.3 Building Intra- and Inter-Modal Graphs

This section describes how to construct the intra-modal and inter-modal graphs for each multimodal instance. Inspired by Kipf and Welling (2017), for both intra- and inter-modal graphs, we build the graphs to be undirected and set a self-loop for each node, to make use of more sufficient relations of the sentiment expression of nodes.

Intra-Modal Graph To leverage the intricate sentiment implications within each modality, we first build three intra-modal graphs to explicitly account for the modality-specific relations of the representations towards the three modalities of a multimodal instance.

To be specific, for the text-modality graph, to leverage the syntax-aware relations of the textual information, inspired by Zhang et al. (2019); Liang et al. (2020, 2022), we construct the text modality graph $\boldsymbol{G}^t \in \mathbb{R}^{L \times L}$ based on the dependency tree of the sentence*. That is, we link the context tokens if there is a relation between these two tokens in the dependency tree. For visual and acoustic modalities, the nodes are averaged video/audio tokens alighed to textual tokens, described in Section 3.2. Since the input representations of these two modalities are sequential, we connect the adjacent nodes in the sequences for the visual modality graph $\boldsymbol{G}^v \in \mathbb{R}^{L \times L}$ and the acoustic modality graph $\boldsymbol{G}^a \in \mathbb{R}^{L \times L}$ to capture the sequence relations of these two modalities.

Inter-Modal Graph To fuse the multimodal representations for extracting the sentiment implica-

*We use the spaCy toolkit to obtain the dependency tree of a sentence: https://spacy.io/.

tions produced by multiple modalities, we construct an inter-modal graph $\boldsymbol{G}^m \in \mathbb{R}^{3L\times 3L}$ for each multimodal instance based on the derived intra-modal graphs. Specifically, we first combine the three intra-modal graphs and then employ a fully connected solution to link the cross-modal tokens among the intra-modal graphs of the three modalities, to capture the potentially scattered relations of the multimodal instance.

3.4 Learning Graph Representations

Based on the intra- and inter-modal graphs derived in Section 3.3, we employ Graph Attention Network (Veličković et al., 2018) to update the nodes in the graphs by aggregating the information from the neighborhoods with varying weights. Specifically, in a GAT layer, for each neighbor, the representations of the current node *i* and the neighbor *j* are concatenated and then mapped to a scalar s_{ij} as the attention coefficient. Then normalizing the attention coefficients of all neighbors by softmax.

$$s_{ij} = \text{LeakyReLU}(\boldsymbol{a}[\boldsymbol{W}h_i \| \boldsymbol{W}h_j]) \quad (4)$$

$$\alpha_{ij} = \text{softmax}_j(s_{ij}) = \frac{\exp(s_{ij})}{\sum_{k\in\mathcal{N}_i} s_{ik}} \quad (5)$$

where $\boldsymbol{a}$ is a weight vector, W is a weight matrix, and $||$ is the concatenation operation. $\mathcal{N}_i$ denotes the set of node *i* and its neighbors. Finally, the representation of node *i* is updated with a weighted sum of the representations of neighbors and itself, and multi-head attention mechanism is applied to stabilize the learning process of self-attention.

$$\tilde{h}_i = \|_{k=1}^{K} \sigma\left(\sum_{j\in\mathcal{N}_i} \alpha_{ij}^k \boldsymbol{W}^k h_j\right) \quad (6)$$

where *k* denotes the *k*-th attention head and σ is a sigmoid function to provide non-linearity.

This mechanism essentially allows the model to make sense of the intra- and inter-modal sentiment associations by modeling the relations in the graphs with the GAT operation. For example, the edge weight between a "smile"(visual) node and a "happy"(textual) node should be higher than the edge weight between a "frown"(visual) node and the "happy"(textual) node.

For intra-modal graphs, the hidden representations $\boldsymbol{h}^t$, $\boldsymbol{h}^v$, and $\boldsymbol{h}^a$ of the three modalities and the corresponding graphs $\boldsymbol{G}^t$, $\boldsymbol{G}^v$, and $\boldsymbol{G}^a$ are fed as inputs into the GAT layers (GATs) to derive the unimodal graph representations $\boldsymbol{r}^t, \boldsymbol{r}^v, \boldsymbol{r}^a \in \mathbb{R}^{d_h}$:

$$\boldsymbol{r}^t = \text{READOUT}^t(\text{GATs}(\boldsymbol{h}^t, \boldsymbol{G}^t)) \quad (7)$$

$$\boldsymbol{r}^v = \text{READOUT}^v(\text{GATs}(\boldsymbol{h}^v, \boldsymbol{G}^v)) \quad (8)$$

$$\boldsymbol{r}^a = \text{READOUT}^a(\text{GATs}(\boldsymbol{h}^a, \boldsymbol{G}^a)) \quad (9)$$

where $\text{GATs}(\cdot)$ denotes the operation of GAT layers. Note that following You et al. (2020), we use a $\text{READOUT}(\cdot)$ function to aggregate the node representations to derive the graph representation.

On the other hand, for an inter-modal graph, we also apply GATs to model the relations in the graph, aiming to extract the relations between modalities for better learning sentiment cues. Given the inter-model graph $\boldsymbol{G}^m$ of a multimodal instance, the corresponding nodes are represented as all nodes from $\boldsymbol{h}^t$, $\boldsymbol{h}^v$ and $\boldsymbol{h}^a$, and the multimodal graph representation $\boldsymbol{r}^m \in \mathbb{R}^{d_h}$ is derived as follow:

$$\boldsymbol{r}^m = \text{READOUT}^m(\text{GAT}([\boldsymbol{h}^t \| \boldsymbol{h}^v \| \boldsymbol{h}^a], \boldsymbol{G}^m)) \quad (10)$$

To this end, owing to the merit of graph attention networks that the weights of associated edges can be adjusted according to the attention mechanism during the training process, the degree of association of nodes on a fully connected graph can be quantified by the weights of the edges, and thus deriving appropriate graph representation.

3.5 Hierarchical Graph Contrastive Learning

In this section, we detail the proposed hierarchical graph contrastive learning strategy in our HGraph-CL framework. As shown in Figure 2, our hierarchical graph contrastive learning strategy first performs at the intra-modal level, and further performs at the inter-modal level. Here for each level, we devise a fully-supervised contrastive loss based on the sentiment labels to improve the graph representation for better sentiment learning and a self-supervised contrastive loss based on the graph augmentations to explore more appropriate graph structures for deriving precise graph representation.

3.5.1 Graph Contrastive Learning at Intra-Modal Level

Fully-Supervised Contrastive Loss Based on Sentiment Labels Inspired by the work on fully-supervised contrastive learning (Khosla et al., 2020; Gunel et al., 2021), for a mini-batch, we adopt sentiment labels as the supervised signal to perform fully-supervised loss for capturing the similarity between examples in one class and contrasting them

with examples in other classes. Specifically, given N examples in a mini-batch, the examples can be divided into $S_1, S_2, ...$ according to sentiment polarity. Considering the binary classification, then $|S_1| = M, |S_2| = N - M, |\cdot|$ is the cardinality of the set. For the *anchor* example $s_i \in S_1$, a *positive* pair can be represented as (s_i, s_j), here $s_j \in S_1, j \neq i$. While the rest $N - M$ samples are regarded as *negative* examples. The pairwise objective $\ell_1(\boldsymbol{r}_i^{\mathcal{M}}, \boldsymbol{r}_j^{\mathcal{M}})$ between the unimodal graph representation $\boldsymbol{r}_i^{\mathcal{M}}$ of s_i and the graph representation $\boldsymbol{r}_j^{\mathcal{M}}$ of $s_j \in S_1$ are defined as:

$$\ell_1(\boldsymbol{r}_i^{\mathcal{M}}, \boldsymbol{r}_j^{\mathcal{M}}) = -\log \frac{e^{\text{sim}(\boldsymbol{r}_i^{\mathcal{M}}, \boldsymbol{r}_j^{\mathcal{M}})/\tau}}{\Sigma^{sup}} \tag{11}$$

$$\begin{aligned}\Sigma^{sup} = &\sum_{k, s_k \in S_1, k \neq i} e^{\text{sim}(\boldsymbol{r}_i^{\mathcal{M}}, \boldsymbol{r}_k^{\mathcal{M}})/\tau} \\ &+ \sum_{l, s_l \in S_2} e^{\text{sim}(\boldsymbol{r}_i^{\mathcal{M}}, \boldsymbol{r}_l^{\mathcal{M}})/\tau}\end{aligned} \tag{12}$$

where $\mathcal{M} \in \{t, v, a\}$, $\text{sim}(\cdot)$ is the similarity function, $\text{sim}(\boldsymbol{u}, \boldsymbol{r}) = \boldsymbol{u}^{\mathrm{T}}\boldsymbol{r}/\|\boldsymbol{u}\|\|\boldsymbol{v}\|$, and τ is the temperature parameter. We use the multimodal sentiment polarity labels as the unimodal sentiment polarity labels, and calculate the supervised contrastive loss on the unimodal graph representations and the multimodal graph representation, respectively. Finally, we sum the two losses to obtain the overall supervised contrastive loss. Following Khosla et al. (2020), the computation of intra-modal supervised contrastive loss $\mathcal{L}_{sup}^{intra}$ is:

$$\begin{aligned}\mathcal{L}_{sup}^{intra} = \sum_{\mathcal{M}} [&\sum_{s_i \in S_1} \frac{1}{|S_1 - 1|} \sum_{j, s_j \in S_1, j \neq i} \ell_1(\boldsymbol{r}_i^{\mathcal{M}}, \boldsymbol{r}_j^{\mathcal{M}}) \\ &+ \sum_{s_k \in S_2} \frac{1}{|S_2 - 1|} \sum_{l, s_l \in S_2, l \neq k} \ell_1(\boldsymbol{r}_k^{\mathcal{M}}, \boldsymbol{r}_l^{\mathcal{M}})]\end{aligned} \tag{13}$$

If more labels are correctly predicted, the value of the loss function will be lower until it converges. On the other hand, if the model cannot predict most of the labels, the loss function will fail to converge.

Self-Supervised Contrastive Loss Based on Graph Augmentations To apply the self-supervised contrastive learning, we implement augmentation data by supplementary and corrupting graphs. The augmented graphs U, V are obtained by randomly deleting or adding a certain ratio of edges, aiming at the exploration of more appropriate graph structures.

Based on the graph augmentation, for an *anchor* graph representation $\boldsymbol{r}_i^{\mathcal{M}}$, we regard the representation derived by the corresponding augmented graph $\boldsymbol{u}_i^{\mathcal{M}}$ as the *positive* sample, while others are regarded as *negative* samples. For the N examples in a mini-batch, the pairwise objective $\ell_2(\boldsymbol{r}_i^{\mathcal{M}}, \boldsymbol{u}_i^{\mathcal{M}})$ between $\boldsymbol{r}_i^{\mathcal{M}}$ and $\boldsymbol{u}_i^{\mathcal{M}}$ is defined as:

$$\ell_2(\boldsymbol{r}_i^{\mathcal{M}}, \boldsymbol{u}_i^{\mathcal{M}}) = -\log \frac{e^{\text{sim}(\boldsymbol{r}_i^{\mathcal{M}}, \boldsymbol{u}_i^{\mathcal{M}})/\tau}}{\Sigma^{self}} \tag{14}$$

$$\begin{aligned}\Sigma^{self} = \; &e^{\text{sim}(\boldsymbol{r}_i^{\mathcal{M}}, \boldsymbol{u}_i^{\mathcal{M}})/\tau} \\ &+ \sum_{j=1}^{N} \mathbb{1}_{[j \neq i]} [e^{\text{sim}(\boldsymbol{r}_i^{\mathcal{M}}, \boldsymbol{r}_j^{\mathcal{M}})/\tau} \\ &+ e^{\text{sim}(\boldsymbol{r}_i^{\mathcal{M}}, \boldsymbol{u}_j^{\mathcal{M}})/\tau}]\end{aligned} \tag{15}$$

where $\mathbb{1}_{[j \neq i]} \in \{0, 1\}$ is the indicator function and equals 1 iff $j \neq i$. The intra-modal self-supervised contrastive loss $\mathcal{L}_{self}^{intra}$ is as follow:

$$\mathcal{L}_{self}^{intra} = \frac{1}{2N} \sum_{\mathcal{M}} \sum_{i=1}^{N} \left[\ell_2(\boldsymbol{u}_i^{\mathcal{M}}, \boldsymbol{r}_i^{\mathcal{M}}) + \ell_2(\boldsymbol{r}_i^{\mathcal{M}}, \boldsymbol{u}_i^{\mathcal{M}})\right] \tag{16}$$

3.5.2 Graph Contrastive Learning at Inter-Modal Level

Corresponding to Section 3.5.1, we perform the fully-supervised contrastive loss and the self-supervised contrastive loss on the multimodal graph representations r^m, and obtain the inter-modal supervised and self-supervised contrastive loss $\mathcal{L}_{sup}^{inter}$ and $\mathcal{L}_{self}^{inter}$, which are defined as:

$$\begin{aligned}\mathcal{L}_{sup}^{inter} = &\sum_{s_i \in S_1} \frac{1}{|S_1 - 1|} \sum_{j \in S_1, j \neq i} \ell_1(\boldsymbol{r}_i^m, \boldsymbol{r}_j^m) \\ &+ \sum_{s_k \in S_2} \frac{1}{|S_2 - 1|} \sum_{l \in S_2, l \neq k} \ell_1(\boldsymbol{r}_k^m, \boldsymbol{r}_l^m)\end{aligned} \tag{17}$$

$$\mathcal{L}_{self}^{inter} = \frac{1}{2N} \sum_{i=1}^{N} [\ell_2(\boldsymbol{u}_i^m, \boldsymbol{r}_i^m) + \ell_2(\boldsymbol{r}_i^m, \boldsymbol{u}_i^m)] \tag{18}$$

3.6 Sentiment Prediction

The multimodal graph representation $\boldsymbol{r}^m$ is fed into a fully-connected layer to predict the sentiment score y:

$$y = \boldsymbol{W}^p \cdot \boldsymbol{r}^m + b^p \tag{19}$$

where $\boldsymbol{W}^p$ and b^p are the weight matrix and bias. Then, the output y_i of the prediction layer for the i-th example is compared with the ground truth y_i^* to calculate the loss of the MSA task $\mathcal{L}^{msa}$:

$$\mathcal{L}^{msa} = \frac{1}{N} \sum_{i}^{N} |y_i - y_i^*| \tag{20}$$

The overall loss of our framework is defined as:

$$\begin{aligned}\mathcal{L} =& \mathcal{L}^{msa} + w_1 * (\mathcal{L}_{sup}^{intra} + \mathcal{L}_{sup}^{inter}) \\ &+ w_2 * (\mathcal{L}_{self}^{intra} + \mathcal{L}_{self}^{inter})\end{aligned} \tag{21}$$

where w_1, w_2 are hyperparameters, controlling the effect of different losses.

4 Experiments

4.1 Datasets and Metrics

We evaluate our approach on two benchmarks, MOSI (Zadeh et al., 2016) and MOSEI (Zadeh et al., 2018c). Table 1 shows the basic statistics of the two datasets. Each example of MOSI or MOSEI contains a continuous sentiment score in the interval $[-3, 3]$ and three modal data. Image frames and audio frames are aligned to text content at the word level.

Following Zadeh et al. (2018b), binary accuracy (Acc-2) and weighted F1 score (F1-Score) are selected as classification metrics, mean absolute error (MAE) and Pearson correlation coefficient (Corr) are selected as regression metrics.

4.2 Training Setting

We take *mean absolute error* as the loss function. For contrastive learning, *N-pairs loss* (Sohn, 2016) and *SupCon loss* (Khosla et al., 2020) are naturally suitable for this scenario. For the textual encoder, we use a pre-trained BERT (bert-base-uncased) and finetune it when training. For visual and acoustic encoders, we train BiLSTMs from scratch. The optimizer chosen is Adam (Kingma and Ba, 2015) and the parameters of BERT and other model parameters are optimized separately. We use a lower learning rate {5e-6, 1e-5} and warm-up strategy for the BERT and a larger learning rate {1e-3, 1e-3} for the other parts. The hyperparameters w_1 and w_2 are selected from {0.1, 0.2, 0.3, 0.4, 0.5}. The results of our model and the reproduced models take the average results obtained from five runs with different random seeds for obtaining stable results. More training settings are presented in Table 2. In addition, we use a learning rate adjustment strategy to update the learning rate when training.

4.3 Baselines and Performance

To verify the effectiveness of our approach, we compare it with the following BERT-based methods: TFN (Zadeh et al., 2017), LMF (Liu et al., 2018), MFM (Tsai et al., 2019b), MulT (Tsai et al., 2019a), MAG-BERT (Rahman et al., 2020), MISA (Hazarika et al., 2020), Self-MM (Yu et al., 2021), HyCon-BERT (Mai et al., 2022). The details of the introduction are presented in Appendix A.

	train	valid	test	total
MOSI	1283	229	686	2198
MOSEI	16326	1871	4659	22856

Table 1: The example size of MOSI and MOSEI.

Parameter	MOSI	MOSEI
epoch	20	6
batch size	4	24
max length	50	128
hidden size	128	128
LSTM layers	1	1
GAT layers	2	1
GAT attention heads	1	1
dropout	0.2	0.1
augmentation ratio	0.2	0.2
BERT learning rate	5e-6	1e-5
other learning rate	1e-3	1e-3
$\mathcal{L}^{sup}$ weight w_1	0.1	0.1
$\mathcal{L}^{self}$ weight w_2	0.1	0.1

Table 2: Training setting details

The comparison results of our HGraph-CL framework and the baseline models are presented in Table 3. We observe that our proposed HGraph-CL consistently outperforms all the baseline models on the two datasets, which verifies the effectiveness of our approach in the MSA task. Moreover, compared with the intra-example representation learning approaches (MFM, MISA), our HGraph-CL achieves outstanding improvement on the two datasets. This indicates that exploring the sentiment implications from both intra- and inter-modal levels is significant for improving the performance of MSA. Furthermore, the significance tests of our HGraph-CL over Self-MM‡ and MAG-BERT‡ present a statistically significant improvement in Acc-2 and F1-Score on MOSI and Acc-2, F1-Score and Corr on MOSEI (with $p < 0.05$).

5 Analysis

5.1 Ablation Study

To verify the impact of the hierarchical graph contrastive learning on performance, we conduct ablation experiments on the two datasets and show the results in Table 4. We can observe that the class distribution is useful for the classification of MOSI and MOSEI datasets, and slightly improve regression. In contrast, the representation distribution learning improves regression significantly

Model	MOSI				MOSEI			
	Acc-2 ↑	F1-Score ↑	MAE ↓	Corr ↑	Acc-2 ↑	F1-Score ↑	MAE ↓	Corr ↑
TFN†	-/80.8	-/80.7	0.901	0.698	-/82.5	-/82.1	0.593	0.700
LMF†	-/82.5	-/82.4	0.917	0.695	-/82.0	-/82.1	0.623	0.677
MFM†	-/81.7	-/81.6	0.877	0.706	-/84.4	-/84.3	0.568	0.717
MulT†	81.5/84.1	80.6/83.9	0.861	0.711	-/82.5	-/82.3	0.580	0.703
MISA†	81.8/83.4	81.7/83.6	0.783	0.761	83.6/85.5	83.8/85.3	0.555	0.756
MAG-BERT	84.2/86.1	84.1/86.0	**0.712**	0.796	84.7/-	84.5/-	-	-
Self-MM	84.0/86.0	84.4/86.0	0.713	0.798	82.8/85.2	82.5/85.3	0.530	0.765
HyCon-BERT	-/85.2	-/85.1	0.713	0.790	-/85.4	-/85.6	0.601	**0.776**
MAG-BERT‡	81.5/83.1	81.5/83.1	0.808	0.761	81.4/84.6	81.9/84.6	0.552	0.756
Self-MM‡	83.1/84.9	83.1/84.9	0.736	0.791	80.5/84.2	80.0/84.2	0.531	0.764
HGraph-CL (ours)	**84.3/86.2** *	**84.6/86.2** *	0.717	**0.799**	**84.5/85.9** *	**84.5/85.8** *	**0.527**	0.769 *

Table 3: Main results on MOSI and MOSEI. ↑ denotes the higher the evaluation metric the better, and ↓ denotes the lower the evaluation metric the better. Results with † are retrieved from (Hazarika et al., 2020) , with ‡ are reproduced using the source code released by the authors, and with * indicate the significance test over Self-MM‡ and MAG-BERT‡ presents a statistically significant improvement. For Acc-2 and F1-Score, the left side of / is the result of dividing examples by *positive*/*non-positive* following (Zadeh et al., 2018b), and the right side is the result of dividing examples by *positive*/*negative* following (Tsai et al., 2019a).

Graph	CL	MOSI				MOSEI			
		Acc-2 ↑	F1-Score ↑	MAE ↓	Corr ↑	Acc-2 ↑	F1-Score ↑	MAE ↓	Corr ↑
Intra, Inter	*Sup, Self*	**84.3/86.2**	**84.6/86.2**	**0.717**	**0.799**	**84.5**/85.9	**84.5**/85.8	**0.527**	**0.769**
	Sup	84.0/85.9	84.2/85.8	0.733	0.788	84.3/**86.0**	84.3/**86.0**	0.535	0.766
	Self	83.8/85.8	84.0/85.7	0.718	0.793	84.1/85.4	84.1/85.3	0.533	0.766
Intra	*Sup, Self*	83.9/85.9	84.1/85.8	0.729	0.790	84.2/85.6	84.2/85.4	0.531	0.765
	Sup	83.6/85.8	83.8/85.7	0.731	0.788	83.9/85.9	83.9/85.6	0.539	0.767
	Self	83.6/85.5	83.8/85.4	0.726	0.789	83.9/85.3	84.0/85.1	0.533	0.764
Inter	*Sup, Self*	84.1/86.2	84.3/86.2	0.723	0.792	84.0/85.7	84.0/85.6	0.529	0.767
	Sup	83.7/86.0	83.8/85.9	0.733	0.789	84.1/85.7	84.1/85.3	0.539	0.763
	Self	83.6/85.9	83.8/85.8	0.722	0.793	83.9/85.3	83.9/85.0	0.531	0.766
None	*None*	83.0/85.1	83.1/85.0	0.756	0.784	82.8/85.1	83.1/85.0	0.539	0.763

Table 4: The performance with different contrastive learning strategies on MOSI and MOSEI. We conduct an ablation study to analyze the impact of graph structure and contrastive learning strategy. {*Intra, Inter*} denotes performing graph contrastive learning at both intra- and inter-modal levels, {*Intra*} denotes at the intra-modal level only, while {*Inter*} denotes at the inter-modal level only. {*Sup, Self*} denotes the result with class distribution learning and representation distribution learning, {*Sup*} denotes the result with class distribution learning only, {*Self*} denotes the result with representation distribution learning only. {*None*} denotes without any strategy.

on MOSI and MOSEI, and makes sense for classification. On the other hand, performing graph contrastive learning at the intra-modal level only or at the inter-modal level only can improve both classification and regression on two datasets. Additionally, our model with complete hierarchical graph contrastive learning can achieve the best overall performance, with a significant improvement over the model without it. The results suggest that hierarchical graph contrastive learning has a great positive impact on the classification and the regression of HGraph-CL.

5.2 Effect of GAT Layers

We convert the measurement of the relations between different modalities into learning the edges of a multimodal graph by GATs. Furthermore, we want to explore the effect of the number of GAT layers on performance. Thus we evaluate our model with different layers of GATs from 1 to 5 on both two datasets to quantify the effect. The experimental results are shown in Figure 3. We can observe that our model achieves the best performance with a small number of layers. Another observation is that the volatility of the classification performance is greater when choosing a big number of layers. The possible reason is that the deep GAT layer will learn similar representations of different nodes, which is called over-smoothing (Li et al., 2018). Over-smoothing may result in the modality-specific information being discarded, and the results suggest that too many layers make a negative effect.

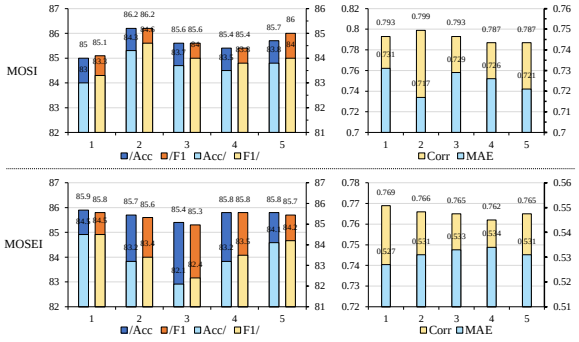

Figure 3: Results with different numbers of GAT layers

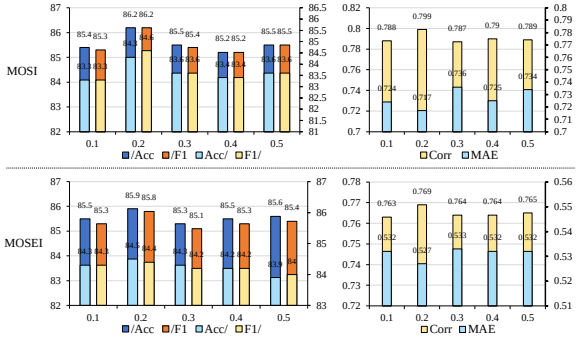

Figure 4: Results of different deleting/adding ratios

5.3 Effect of Deleting/Adding Ratio

To investigate the effect of different ratios of deleting/adding edges in deriving graph augmentations on the performance, we conduct experiments with different values of ratio and report the results in Figure 4. From the experimental results, we can observe that different ratios introduce a considerable impact on performance. When the delete ratio is small (0.1), the possibility of exploration of graph structures is limited, thus leading to a poorer performance. In addition, noticeable performance degradation is also shown when the ratio is greater than 0.2. This indicates that excessively exploring the possible relations may weaken the learning ability of graph contrastive learning. Therefore, we set the ratio to 0.2 in our experiments.

5.4 Case Study

To show the relationship strength between different representations, we select an example from the MOSI dataset and visualize the weights of edges between text and visual nodes and present them in Figure 5. We can observe that the negative word nodes *jack ass* have a stronger relationship with the visual nodes representing *frowning faces*, and are weakly related to these visual nodes representing *normal faces*. It shows that the proposed model can understand the relations of representations across modalities, enabling the highly correlated modal representations to be explicitly linked for learning the multimodal sentiment information.

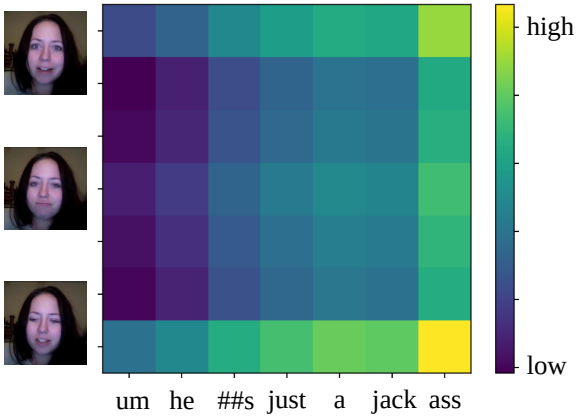

Figure 5: A case of the weights of edges between text and visual nodes

6 Conclusion

This paper proposes a novel hierarchical graph contrastive learning (HGraph-CL) framework for multimodal sentiment analysis (MSA), in which graph contrastive learning is performed at intra-modal and inter-modal levels. For the graph contrastive learning strategy performed at each level, we devise a fully-supervised contrastive loss and a self-supervised contrastive loss. The fully-supervised contrastive loss is devised to improve the learning of sentiment cues by capturing the similarity between examples in one class and the contrast among different classes. And the self-supervised contrastive loss is devised to explore a more appropriate graph structure based on the graph augmentations for making use of sentiment relations within each modality and across different modalities. Experimental results on two benchmark datasets show that our method outperforms the state-of-the-art baselines in MSA.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China (62006062, 62176076), Shenzhen Foundational Research Funding (JCYJ20200109113441941 and JCYJ2021032411 5614039), The Major Key Project of PCL2021A06, Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies (2022B1212010005).

References

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

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, Florence. IEEE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Veselin Stoyanov. 2021. Supervised contrastive learning for pre-trained language model fine-tuning. In *International Conference on Learning Representations*.

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*, MM '20, page 1122–1131, New York, NY, USA. Association for Computing Machinery.

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

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673. Curran Associates, Inc.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*, San Diego.

Thomas N Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *Proceedings of the 28th International Conference on Computational Linguistics*.

Qimai Li, Zhichao Han, and Xiao-Ming Wu. 2018. Deeper insights into graph convolutional networks for semi-supervised learning. In *Thirty-Second AAAI conference on artificial intelligence*.

Bin Liang, Hang Su, Lin Gui, Erik Cambria, and Ruifeng Xu. 2022. Aspect-based sentiment analysis via affective knowledge enhanced graph convolutional networks. *Knowledge-Based Systems*, 235:107643.

Bin Liang, Rongdi Yin, Lin Gui, Jiachen Du, and Ruifeng Xu. 2020. Jointly learning aspect-focused and inter-aspect relations with graph convolutional networks for aspect sentiment analysis. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 150–161, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Che Liu, Rui Wang, Jinghua Liu, Jian Sun, Fei Huang, and Luo Si. 2021. Dialoguecse: Dialogue-based contrastive learning of sentence embeddings. *EMNLP*.

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, Melbourne, Australia. Association for Computational Linguistics.

Sijie Mai, Ying Zeng, Shuangjia Zheng, and Haifeng Hu. 2022. Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, pages 1–1.

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, Online. Association for Computational Linguistics.

Cees GM Snoek, Marcel Worring, and Arnold WM Smeulders. 2005. Early versus late fusion in semantic video analysis. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 399–402, New York.

Kihyuk Sohn. 2016. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in neural information processing systems*, pages 1857–1865.

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, Florence, Italy. Association for Computational Linguistics.

Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov.

2019b. Learning factorized multimodal representations. In *International Conference on Learning Representations*.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *International Conference on Learning Representations*.

Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2019. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):7216–7223.

Yang Wu, Zijie Lin, Yanyan Zhao, Bing Qin, and Li-Nan Zhu. 2021. A text-centered shared-private framework via cross-modal prediction for multimodal sentiment analysis. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4730–4738, Online. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. 2020. Graph contrastive learning with augmentations. In *Advances in Neural Information Processing Systems*, volume 33, pages 5812–5823. Curran Associates, Inc.

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

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, Copenhagen, Denmark. Association for Computational Linguistics.

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018a. Memory fusion network for multi-view sequential learning. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 5634–5641, Palo Alto.

Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. 2018b. Multi-attention recurrent network for human communication comprehension. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 5642–5649, Palo Alto.

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. 2018c. 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, Melbourne.

Chen Zhang, Qiuchi Li, and Dawei Song. 2019. Aspect-based sentiment classification with aspect-specific graph convolutional networks. 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 4568–4578, Hong Kong, China. Association for Computational Linguistics.

Qiang Zhu, Mei-Chen Yeh, Kwang-Ting Cheng, and Shai Avidan. 2006. Fast human detection using a cascade of histograms of oriented gradients. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1491–1498, New York. IEEE.

Yanqiao Zhu, Yichen Xu, Feng Yu, Qiang Liu, Shu Wu, and Liang Wang. 2020. Deep Graph Contrastive Representation Learning. In *ICML Workshop on Graph Representation Learning and Beyond*.

A Introduction of Baselines

Since our HGraph-CL framework is designed based on BERT, to verify the validity of our approach, we select the following state-of-the-art models with a BERT-based version for comparison:

TFN (Zadeh et al., 2017): The tensor fusion network performs outer product on three modal representations, then the multimodal representation vector fused with uni-, bi-, and tri- modalities will be obtained.

LMF (Liu et al., 2018): The low-rank multimodal fusion network is based on TFN and gets some improvement. It uses tensor decomposition to decompose the parameter tensor of the outer product layer.

MFM (Tsai et al., 2019b): The multimodal factorization model factorizes representations into two sets of independent factors, and optimizes for a joint generative-discriminative objective across multimodal data and labels.

MulT (Tsai et al., 2019a): With a slight modification in the structure of the transformer encoder, MulT proposes a cross-modal transformer network to align information from one mode to another.

MAG-BERT (Rahman et al., 2020): A variant of BERT, adding a multimodal shifting gate unit behind the input layer. By fusing with visual and acoustic information, the word embeddings can be shifted in a direction that can express sentiment polarity better in the feature space.

MISA (Hazarika et al., 2020): proposes a multitask framework for intra-example representation learning. It projects each modality to two distinct subspaces to model modality-specific and -invariant information.

Self-MM (Yu et al., 2021): designs a label generation module based on the self-supervised learning strategy to acquire independent unimodal supervision. Jointly training uni- and multimodal sentiment analysis tasks have got the state-of-the-art performance on MOSI and MOSEI.

HyCon-BERT (Mai et al., 2022): proposes hybrid contrastive learning of tri-modal representations to explore cross-modal interaction and reduce the gap among modal representations.