CLGSI: a multimodal sentiment analysis framework based on contrastive learning guided by sentiment intensity

Anonymous ACL submission

Abstract

Recently, contrastive learning has begun to gain popularity in multimodal sentiment analysis (MSA). However, most of existing MSA methods based on contrastive learning lacks more detailed learning of the distribution of sample pairs with different sentiment intensity differences in the contrastive learning representation space. In addition, limited research has been conducted on the fusion of each modality representation obtained by contrastive learning training. In this paper, we propose a novel framework for multimodal sentiment analysis based on **C**ontrastive **L**earning **G**uided by **S**entiment **I**ntensity (CLGSI). Firstly, the proposed contrastive learning guided by sentiment intensity selects positive and negative sample pairs based on the difference in sentiment intensity and assigns corresponding weights accordingly. Subsequently, we propose a new multimodal representation fusion mechanism, called **G**lobal-**L**ocal-**F**ine-**K**nowledge (GLFK), which extracts common features between different modalities' representations. At the same time, each unimodal encoder output is separately processed by a Multilayer Perceptron (MLP) to extract specific features of each modality. Finally, joint learning of the common and specific features is used to predict sentiment intensity. The effectiveness of CLGSI is assessed on two English datasets, MOSI and MOSEI, as well as one Chinese dataset, SIMS. We achieve competitive experimental results, which attest to the strong generalization performance of our approach. The code for our approach will be released in https://github.com/***/***

1 Introduction

Sentiment is one of the most important ways for human beings to perceive the world, and it can significantly affect human behavior and decision-making. MSA aims to comprehensively analyze human sentiments by integrating and examining information from diverse modalities (Cambria et al., 2014; Morency et al., 2011), such as text, video and audio. This integration and analysis enables machines to better understand and interpret human sentiments. Due to the rapid advancements in multimedia and computer technologies, MSA has garnered significant attention within the Natural Language Processing (NLP) community (Liu et al., 2022; Sun et al., 2020; Zadeh et al., 2017).

In recent times, contrastive learning has gained popularity in the field of MSA. The MSA approaches based on contrastive learning involve three significant issues: 1) the selection of positive and negative sample pairs, 2) the attention given to different positive and negative samples during the learning process, and 3) the integration of modality representations obtained after contrastive learning. Several researchers have proposed solutions to address these issues.

Mai et al. (Mai et al., 2022) first introduced contrastive learning in MSA and proposed Hycon. In their method, positive and negative sample pairs are first roughly divided using labels. During training, positive and negative sample pairs are dynamically selected based on the similarity across modalities. Another approach, ConFEDE, was proposed by Yang et al. (Yang et al., 2023), who argued that the text is generally more effective than audio and video in MSA. Thus, ConFEDE selects sample pairs to be trained during the learning process by considering text similarity, and only selects 2 positive samples and 4 negative samples for each anchor.

Although the aforementioned methods have yielded promising results, they do not account for differences in sentiment intensity between samples. Samples with sentiment intensity of -0.2 and -0.4 are likely to be treated as negative samples pair according to the pairs selection mechanism of HyCon and ConFEDE. However, they still have similarities in terms of labels and should not be pushed away in the representation space. In contrast, ConKI, proposed by Yu et al. (Yu et al., 2023), can alleviate this problem to some extent by selecting positive and negative sample pairs using predefined sentiment intervals (e.g., positive, weak positive, etc.) in the dataset.

Moreover, the majority of existing studies treat the learning of different sample pairs equally and lack a detailed learning of the distribution of sample pairs with varying sentiment intensity differences in the representation space. Nevertheless, in MSA, it is crucial to assign distinct attention to sample pairs with differing sentiment intensity differences during the optimization process of contrastive learning. For example, take the negative sample pair A: $\{y_1$=-0.4, y_2=+1$\}$ and B: $\{y_1$=-0.4, y_3=+0.6$\}$. It is obvious that A exhibits a larger sentiment intensity difference than B. Therefore, it is necessary to pay more attention to A, that is, letting the two samples in A have a greater relative distance in the representation space.

In addition, the modal representations obtained by contrastive learning training in the aforementioned studies are simply concatenated and fed into the MLP, which lacks further exploration of the integration of representation information, potentially restricting the model's generalization performance.

Considering the aforementioned limitations, we introduce a novel multimodal sentiment analysis framework based on **C**ontrastive **L**earning **G**uided by **S**entiment **I**ntensity (CLGSI). Our contributions are summarized as follows:

- We propose contrastive learning guided by sentiment intensity. The selection of positive and negative sample pairs in contrastive learning guided by the sentiment intensity difference, with corresponding weights being assigned accordingly. This enriches the contrastive learning process with fine-grained information.

- We propose a multi-modal representation fusion mechanism, **G**lobal-**L**ocal-**F**ine-**K**nowledge (GLFK), that mimics the human cognitive process. We use the GLFK mechanism to fuse the representations of each modality obtained by contrastive learning training to extract the common features across different modalities. At the same time, we use MLP to process the output of each modal encoder to extract the specific features of each modality. Finally, the joint learning of common features and specific features was used to predict the sentiment intensity.

- We conduct extensive experiments on public English and Chinese MSA datasets. Competitive experimental results show that CLGSI can better understand sentiment expressions under different cultural differences, which proves the good generalization performance and effectiveness of our model.

2 Related Work

2.1 Multimodal Sentiment Analysis

In the field of MSA, a major concern of researchers is the fusion and interaction between modalities. In earlier works, the main focus was on strategies for modality fusion. There are two common fusion strategies: early fusion and late fusion. Early fusion, constructs a joint feature representation by extracting the features of each modality and merging them at the input level (Morency et al., 2011; Park et al., 2016; Rosas et al., 2013; Zadeh et al., 2018b). Late fusion, firstly conducts sentiment analysis based on each modality, and then uses different mechanisms to incorporate the unimodal sentiment decision into the final decision. The common decision mechanism is weighted voting and majority voting (Alam and Riccardi, 2014; Cai and Xia, 2015; Kampman et al., 2018; Nojavanasghari et al., 2016).

Researchers have recently shifted their focus from solely modality fusion to also considering the interaction between modalities. For instance, Zadeh et al. (Zadeh et al., 2017) proposed a tensor fusion method that learns the intra-modal and inter-modal dynamics of three modalities in an end-to-end manner, aiming to improve MSA performance. Rahman et al. (Rahman et al., 2020) developed the Multimodal Adaptation Gate (MAG), which fine-tunes the BERT model (Devlin et al., 2018) to enhance MSA performance. Additionally, Han et al. (Han et al., 2021) proposed a method that simultaneously maximizes the mutual information (MI) between modalities and the MI between the multimodal fusion results and unimodal inputs, thus enhancing the model's capabilities.

Subsequently, researchers began to focus on the significance of simultaneously considering both the common and specific features across different modalities in the context of MSA. Hazarika et al. (Hazarika et al., 2020) proposed MISA, which di-

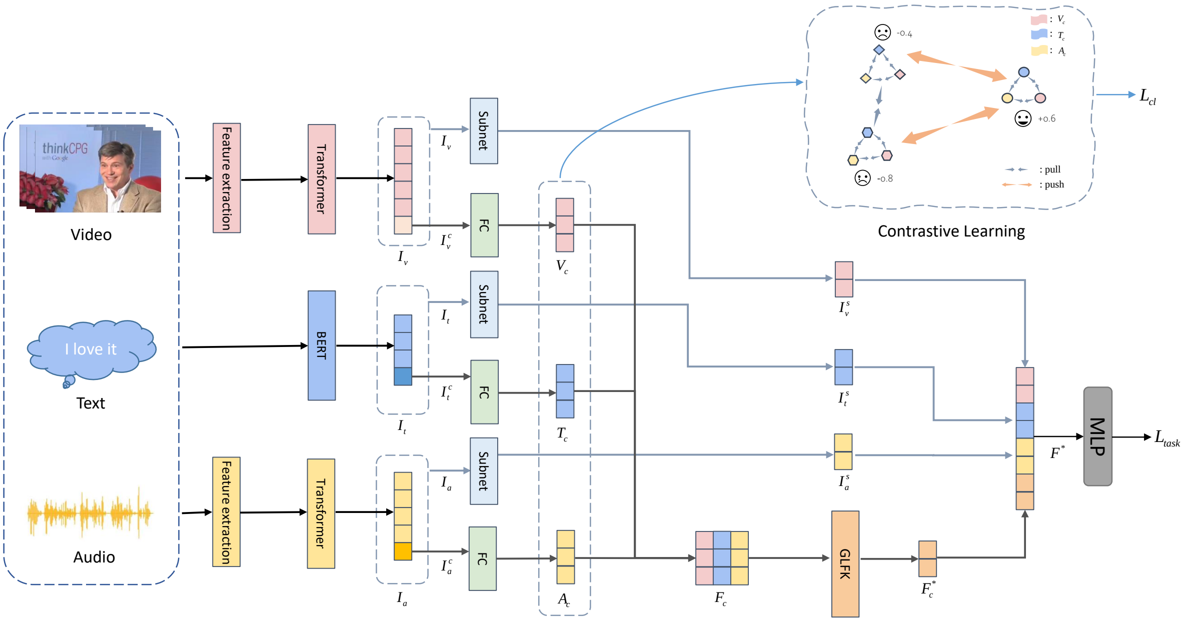


Figure 1: The overall architecture of CLGSI. F_c^* denotes the common features, while I_v^s, I_t^s and I_a^s represent the specific features associated with each modality.

vides each modality into modality-invariant subspaces and modality-specific subspaces, and then fuses them to predict sentiment. Similarly, Yang et al. (Yang et al., 2022) introduced FDMER, which decompose modalities into two subspaces, and introduce a modality discriminator to guide the parameter learning of the common and private encoder in an adversarial manner. In this study, we design two modules to extract the common features among diverse modalities and the specific features of each modality, and use these features to predict sentiment intensity.

2.2 Contrastive Learning

Contrastive learning, as an effective method for representation learning, has been widely explored in the community. Previous research on contrastive learning can be categorized into two main types: self-supervised contrastive learning (Akbari et al., 2021; Chen et al., 2020; Radford et al., 2021) and supervised contrastive learning (Hu et al., 2022; Khosla et al., 2020). The key distinction between these approaches lies in whether label information is employed to guide the selection of positive and negative sample pairs.

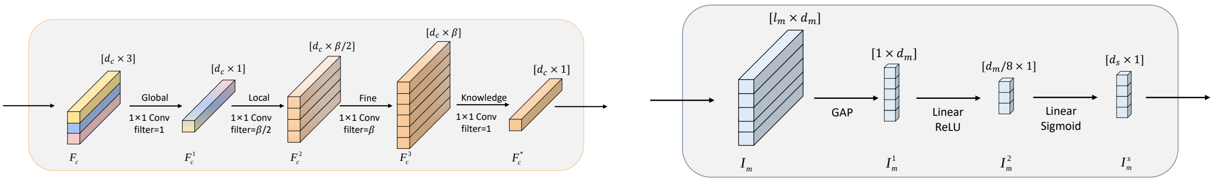
Recently, there has been a growing interest in supervised contrastive learning into MSA. For instance, Hycon, proposed by Mai et al. (Mai et al., 2022), is the first to leverage contrastive learning to enhance modal interactions in MSA. ConFEDE proposed by Yang et al. (Yang et al., 2023), used the similarity between texts to guide the joint execution of contrastive representation learning and contrastive feature decomposition. ConKI proposed by Yu et al. (Yu et al., 2023), utilizes contrastive knowledge injection so that the model can learn both specific and general knowledge representations for each modality. Although these works have achieved good results, they still have some limitations, as discussed in the introduction.

3 Methodology

3.1 Overall Architecture

The overall architecture of CLGSI is shown in Figure 1. Each input modality is encoded differently: Text uses the BERT, while video and audio use the pre-training toolkit for initial feature extraction (Yu et al., 2021), followed by the Transformer Encoder (Vaswani et al., 2017). The encoded representations of the sample's text, video, and audio modalities are denoted as $I_t \in \mathbb{R}^{l_t \times d_t}$, $I_v \in \mathbb{R}^{l_v \times d_v}$ and $I_a \in \mathbb{R}^{l_a \times d_a}$, respectively. Here, $l_{m \in \{t,v,a\}}$ represents the sequence length of each modality, and $d_{m \in \{t,v,a\}}$ represents the corresponding feature vector dimension.

Based on these representations, the common features between different modalities and the specific features of each modality are extracted separately. In the common feature extraction module, the contrastive learning guided by sentiment intensity is



(a) GLFK

(b) Subnet

Figure 2: The architecture of GLFK and Subnet

performed to enhance the representation ability of encoders. Finally, a 3-layer MLP was used to jointly learn the common features and specific features to predict the sentiment intensity.

3.2 Common Feature Extraction

In the common feature extraction module, the primary objective is to project the information from different modalities into the same representation space. For the text modality, the [CLS] vector from BERT $I_t^c \in \mathbb{R}^{1\times d_t}$ is used as the common vector representation. For the video and audio modalities, we use the last vector output from the last layer of the transformer encoder $I_v^c \in \mathbb{R}^{1\times d_v}$ and $I_a^c \in \mathbb{R}^{1\times d_a}$ as the common vector representation, respectively. Subsequently, these three vectors are transformed to the same dimension using a fully connected (FC) layer and a ReLU activation function, yielding $T_c \in \mathbb{R}^{d_c\times 1}$, $V_c \in \mathbb{R}^{d_c\times 1}$ and $A_c \in \mathbb{R}^{d_c\times 1}$. To enhance the representation capability of the encoders from different modalities, we employ the contrastive learning guided by sentiment intensity, enabling these information from different modalities to project onto the same representation space (see section 3.4 for details). Additionally, we stack V_c, T_c and A_c into a new matrix $F_c = [V_c; T_c; A_c] \in \mathbb{R}^{d_c\times 3}$ which serves as the input of GLFK, thereby facilitating the extraction of common features between different modalities.

The GLFK, a novel representation fusion mechanism inspired by human cognitive processes, comprises four components: Global, Local, Fine, and Knowledge (as illustrated in Figure 2(a)). To illustrate the mechanism, we draw an analogy between reading academic papers and our approach. Typically, individuals begin by reading the abstract to gain an overview of the research. This aligns with the Global component of GLFK, where we employ a 1×1 convolution (Conv) operation to globally compress the information. As a result, the $F_c \in \mathbb{R}^{d_c\times 3}$ is compressed to $F_c^1 \in \mathbb{R}^{d_c\times 1}$, facilitating an overall understanding of the content. Next, readers proceed to skim through the paper to grasp the main work, followed by in-depth reading to comprehend the technical details. This corresponds to the Local and Fine components of GLFK. Specifically, we utilize two 1×1 convolutions to expand $F_c^1 \in \mathbb{R}^{d_c\times 1}$ to $F_c^2 \in \mathbb{R}^{d_c\times \beta/2}$, and subsequently expand it to $F_c^3 \in \mathbb{R}^{d_c\times\beta}$, where β is a hyperparameter (set to 16 in this paper). Following these stages, readers possess a profound understanding and knowledge of the paper. Lastly, they summarize this knowledge, ultimately obtaining refined insights. This process aligns with the Knowledge component of GLFK, where a 1×1 convolution is employed to reduce the $F_c^3 \in \mathbb{R}^{d_c\times\beta/2}$ to $F_c^* \in \mathbb{R}^{d_c\times 1}$. Consequently, the common features across different modalities F_c^* are obtained.

3.3 Specific Feature Extraction

In the specific feature extraction module, our focus lies on efficiently capturing the comprehensive information within a modality. The sub-network (Subnet) structure for specific feature extraction is depicted in Figure 2(b). Given a modality $I_m \in \mathbb{R}^{l_m\times d_m}, m \in \{t, v, a\}$, we begin by utilizing global average pooling (GAP) along the sequence length to compressed I_m to $I_m^1 \in \mathbb{R}^{1\times d_m}$. Subsequently, a two-step nonlinear transformation is applied to project I_m^1 into a new lower-dimensional space:

$$I_m^s = \sigma_2(W_2\sigma_1(W_1 I_m^{1\,\mathrm{T}})), m \in \{t, v, a\}$$

where $W_1 \in \mathbb{R}^{(d_m/8)\times d_m}$, $W_2 \in \mathbb{R}^{d_s\times(d_m/8)}$, and $I_m^s \in \mathbb{R}^{d_s\times 1}$, the σ_1 represents the ReLU function, the σ_2 represents the Sigmoid function.

3.4 Contrastive Learning guided by Sentiment Intensity

3.4.1 Pair Selection

This section presents a two-step process to describe the selection of positive and negative sample pairs:

1) Initially, we determine the initial positive and negative pairs by calculating the difference between their corresponding sentiment intensities. Due to differing sentiment intensity ranges ([-3,3] in MOSI/MOSEI and [-1,1] in SIMS), we use uniform mapping to convert label values to [-1,1], only during contrastive learning. Given a batch B, we calculate the sentiment intensity difference between sample $i \in B$ and different samples using the following formula:

$$D_{(i,j)} = \left| y_i - y_j \right|, j \in B \ \ \& \ \ j \neq i \tag{1}$$

where y_i and y_j represent the sentiment intensity labels of samples i and j, respectively. Subsequently, we utilize a sentiment intensity difference threshold (κ), a hyperparameter set to 0.4 in this paper, to determine whether sample j is classified as an initial positive or negative sample of i. This process is illustrated in the subsequent equation:

$$\begin{cases} D_{(i,j)} > \kappa, \ (i,j) \in initial\ negative\ pairs \\ D_{(i,j)} \leq \kappa, \ (i,j) \in initial\ positive\ pairs \end{cases}$$

2) Based on the intra-modal and inter-modal cases, we provide a detailed division of positive and negative sample pairs. Given an set of initial positive and negative sample pairs, for a sample i, the intra-modal and inter-modal positive and negative sample pairs are chosen as follows:

- Intra-modal pairs:

$$\begin{aligned} P^i_{intra} = \{&(T^i_c, T^j_c), (V^i_c, V^j_c), (A^i_c, A^j_c) \\ &\mid (i,j) \in initial\ positive\ pairs\} \\ N^i_{intra} = \{&(T^i_c, T^k_c), (V^i_c, V^k_c), (A^i_c, A^k_c) \\ &\mid (i,k) \in initial\ negative\ pairs\} \end{aligned}$$

- Inter-modal pairs:

$$\begin{aligned} P^i_{inter} = \{&(V^i_c, T^i_c), (V^i_c, A^i_c), (T^i_c, A^i_c)\} \cup \\ \{&(V^i_c, T^j_c), (T^i_c, V^j_c), (V^i_c, A^j_c), \\ &(A^i_c, V^j_c), (T^i_c, A^j_c), (A^i_c, T^j_c) \\ &\mid (i,j) \in initial\ negative\ pairs\} \\ N^i_{inter} = \{&(V^i_c, T^k_c), (T^i_c, V^k_c), (V^i_c, A^k_c), \\ &(A^i_c, V^k_c), (T^i_c, A^k_c), (A^i_c, T^k_c) \\ &\mid (i,k) \in initial\ negative\ pairs\} \end{aligned}$$

where T^i_c, V^i_c, A^i_c correspond to the representations of three different modalities of sample i, while the rest of the symbols have the same meaning.

By combining the intra-modal pairs and inter-modal pairs of the sample i together, we obtain the positive and negative sample pairs P^i and N^i of sample i in contrastive learning process as follows:

$$\begin{aligned} P^i &= P^i_{intra} \cup P^i_{inter} \\ N^i &= N^i_{intra} \cup N^i_{inter} \end{aligned}$$

3.4.2 Contrastive Loss

After identifying the positive and negative sample pairs, we attempt to incorporate fine-grained information into the contrastive learning training process based on the sentiment intensity difference.

For instance, given samples i, j, and k, where the sentiment intensity difference from i to j and k are 0.5 and 1.6, respectively (as defined by (1)), both (i,j) and (i,k) are initial negative sample pairs of i. However, the sentiment intensity difference between sample i and k is noticeably greater. Thus, we assign a higher weight to (i,k) when calculating the contrastive loss to push samples i and k further apart in the representation space compared to samples i and j. In CLGSI, we design a weight function (as visualized in Figure 3) by using the non-linear function $|tanh(x)|$ as follows:

$$\omega_{(i,j)} = \begin{cases} \left|\tanh\left(D_{(i,j)} - 2\kappa\right)\right| \times 1.5, & (i,j) \in initial\ positive\ pairs \\ \left|\tanh\left(D_{(i,j)}\right)\right| \times 1.5, & (i,j) \in initial\ negative\ pairs \end{cases} \tag{2}$$

For ease of presentation, we integrate intra-modal and inter-modal contrastive learning into the same formula. Given a batch B, the contrastive loss is expressed as follows:

$$L_{cl} = -\mathbb{E}_{i \in B} \log \frac{\sum\limits_{(a,p) \in P^i} \delta(a,p)}{\sum\limits_{(a,q) \in P^i \cup N^i} \delta(a,q)} \tag{3}$$

where $\delta(a,p) = e^{[w_{(i,j)} * \frac{sim(a,p)}{\tau}]}$, and $w_{(i,j)}$ is the weight specified by the equation (2).

An illustrative example of the learning process is presented in the upper right corner of Figure 1.

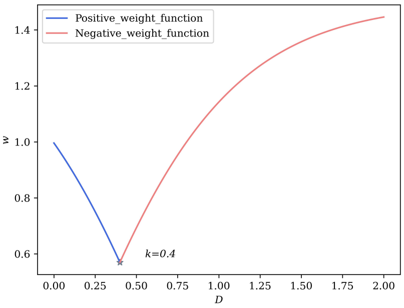


Figure 3: Weight function.

3.5 Overall Learning Objectives

After extracting the common and specific features, we concatenate the common feature vector F_c^* with the specific feature vectors I_v^s, I_t^s, I_a^s of the three modalities to obtain $F^* = [I_v^s, I_t^s, I_a^s, F_c^*] \in \mathbb{R}^{d_* \times 1}$, where $d_* = 3d_s + d_c$. We then feed F^* into a 3-layer MLP to predict the sentiment intensity value $\hat{y}_i$. Given the ground truth y_i, the mean absolute error is used to compute the MSA task loss, given by:

$$L_{task} = \frac{1}{N_b} \sum_{i}^{N_b} |y_i - \hat{y}_i|$$

where N_b is the number of samples in the batch B.

To combine both the task loss L_{task} and the contrastive loss L_{cl}, we define the overall learning objective of CLGSI as follows:

$$L_{overall} = L_{task} + \gamma L_{cl}$$

where γ is a hyperparameter.

4 Experiment

4.1 Dataset and Metrics

We conduct extensive experiments on three popular datasets: MOSI (Zadeh et al., 2016) and MOSEI(Zadeh et al., 2018c) in English, and SIMS (Yu et al., 2020) in Chinese. Appendix A provides further details on the dataset.

To ensure a fair comparison, we report our experimental results in both regression and classification. For regression, we report the mean absolute error (MAE) and Pearson correlation (Corr). For classification, we report the multi-class accuracy and F1 score. We calculate the accuracy of 2-class prediction (Acc-2) and 5-class (Acc-5) prediction for CH-SIMS, and the accuracy of 2-class prediction and 7-class prediction (Acc-7) for MOSI and MOSEI. In addition, the Acc-2 and F1 scores for SIMS are computed for positive/non-positive (including zero) classes. The Acc-2 and F1 scores for MOSI and MOSEI are reported for negative/positive (excluding zero) and negative/non-negative (including zero) classes. Higher values indicate better performance for all metrics except for MAE.

4.2 Baselines

We provide a comprehensive comparison between CLGSI and state-of-the-art baselines, which are summarized in Tables 1 and 2. These baselines include LF-DNN (Yu et al., 2020), MFN (Zadeh et al., 2018a), LMF (Liu et al., 2018), TFN (Zadeh et al., 2017), MulT (Tsai et al., 2019), MISA (Hazarika et al., 2020), MAG-BERT (Rahman et al., 2020), HyCon (Mai et al., 2022), Self-MM (Yu et al., 2021), and ConFEDE (Yang et al., 2023). For the sake of fair comparison, all the methods we selected have public code for easy replication. In Appendixes B and C, we provide comprehensive details of the models compared and the experimental setup, respectively.

4.3 Results

Tables 1 and 2 present the performance comparison results of each model on the SIMS, MOSI, and MOSEI datasets. Overall, CLGSI achieves competitive results compared to the baselines across all three datasets.

On the MOSI dataset, CLGSI outperforms all baselines in Acc-2, F1, Acc-7, and MAE. These results indicate that the newly introduced contrastive learning mechanism in CLGSI effectively learns the representations of different modalities, enabling the model to perform well even on small datasets. On the MOSEI dataset, CLGSI outperforms all baselines in Acc-2, F1 and Acc-7. Particularly, CLGSI improves by at least 0.5% over all baselines in Acc-2 and F1. For the SIMS dataset, CLGSI outperforms all baselines in Acc-2, while achieving comparable performance to the best baseline ConFEDE in the other four metrics.

Moreover, CLGSI demonstrates strong performance in multiclass classification metrics across all three datasets. On the MOSI dataset, Acc-7 surpasses the baselines by at least 1.36%. On the MOSEI dataset, Acc-7 outperforms the baselines by at least 1.1%. Although CLGSI falls slightly behind ConFEDE by 0.39% in Acc-5 on the SIMS dataset, it still outperforms the other baselines in Acc-5. This result shows that the contrastive learning mechanism in CLGSI can help the model correctly learn the sentiment information under different cultural differences, so as to enhance the fine-grained metric of multi-classification accuracy.

As a recently developed MSA method based on contrastive learning, ConFEDE exhibits superior overall performance among the baselines. Given its prominence, ConFEDE serves as the primary baseline for comparison with CLGSI. A combined analysis of Tables 1 and 2 reveals that CLGSI outperforms ConFEDE in terms of Acc-2 across all datasets. On the large dataset MOSEI and the small dataset MOSI, ConFEDE achieves Acc-7

Model	MOSI					MOSEI				
	Acc-2	F1	Acc-7	MAE	Corr	Acc-2	F1	Acc-7	MAE	Corr
LF-DNN	77.52/78.63	77.46/78.63	34.52	0.955	0.658	80.60/82.74	80.85/82.52	50.83	0.58	0.709
MFN	77.4/-	77.3/-	34.1	0.965	0.632	78.94/82.86	79.55/82.85	51.34	0.573	0.718
LMF	-/82.5	-/82.4	33.2	0.917	0.695	80.54/83.48	80.94/83.36	51.59	0.576	0.717
TFN	-/80.8	-/80.7	34.9	0.901	0.698	78.50/81.89	78.96/81.74	51.6	0.573	0.714
MulT	-/83.0	-/82.8	40	0.871	0.698	81.15/84.63	81.56/84.52	52.84	0.559	0.733
MISA	81.8/83.4	81.7/83.6	42.3	0.783	0.776	83.6/85.5	83.8/85.3	52.2	0.555	0.756
MAG-BERT	82.13/83.54	81.12/83.58	41.43	0.79	0.766	82.51/84.82	82.77/84.71	50.41	0.583	0.741
HyCon	-/85.2	-/85.1	46.6	0.713	0.79	-/85.4	-/85.6	52.8	0.601	**0.776**
Self-MM	83.44/85.46	83.36/85.43	46.67	0.708	0.796	83.76/85.15	83.82/84.90	53.87	0.531	0.765
ConFEDE	84.17/85.52	84.13/85.52	42.27	0.742	0.784	81.65/85.82	82.17/85.83	54.86	0.522	0.78
Self-MM*	82.54/84.77	82.68/84.91	45.79	0.712	**0.795**	82.68/84.96	82.95/84.93	53.46	**0.529**	0.767
ConFEDE*	83.24/84.76	83.23/84.8	41.98	0.755	0.779	82.36/84.78	82.45/84.55	52.99	0.55	0.757
CLGSI	**83.97/86.43**	**83.63/86.25**	**47.96**	**0.703**	0.79	**84.01/86.32**	**84.21/86.18**	**54.56**	0.532	0.763

Table 1: Results on MOSI and MOSEI. In Acc-2 and F1 score, the left and right sides of the slash ("/") represent "negative/non-negative" and "negative/positive", respectively. Models with * are reproduced under the same conditions, while other results are from (Yang et al., 2023).

Model	SIMS				
	Acc-2	F1	Acc-5	MAE	Corr
LF-DNN	78.87	79.87	41.62	0.42	0.612
MFN	77.9	77.88	39.47	0.435	0.582
LMF	77.77	77.88	40.53	0.441	0.576
TFN	78.38	78.62	39.3	0.432	0.591
MulT	78.56	79.66	37.94	0.453	0.561
Self-MM	80.04	80.44	41.53	0.425	0.595
ConFEDE	82.23	82.08	46.3	0.392	0.637
Self-MM*	78.71	78.76	42.94	0.411	0.601
ConFEDE*	81.05	**81.13**	**46.34**	**0.377**	**0.655**
CLGSI	**81.18**	80.59	45.95	0.408	0.634

Table 2: Results on SIMS. Models with * are reproduced under the same conditions, while other results are from (Yang et al., 2023).

of 52.99% and 41.98% respectively, showing a difference of 11.01%. On the other hand, CLGSI achieves Acc-7 of 54.56% and 47.96% respectively, showing a difference of 6.6%. This finding indicates that compared to ConFEDE, CLGSI demonstrates stronger generalization ability. On the SIMS dataset, CLGSI slightly underperforms ConFEDE. This is because ConFEDE utilizes additional unimodal labels provided in the dataset to pretrain unimodal encoders, leading to improved performance on the Chinese dataset. However, without additional unimodal labels in MOSI/MOSEI, ConFEDE performs worse than CLGSI overall. This indicates that ConFEDE relies on unimodal labels and pre-training. In contrast, CLGSI achieves competitive results on all three datasets without the need for additional pre-training.

4.4 Ablation Study

To evaluate the effectiveness of CLGSI's contribution, we conducted ablation studies on MOSI and SIMS. Specifically, for MOSI, we reported Acc-2 (excluding zero) and Acc-7, while for SIMS, we reported Acc-2 and Acc-5.

4.4.1 Effectiveness of the contrastive learning guided by sentiment intensity

To discuss the effect of the contrastive learning guided by sentiment intensity, we show the ablation result in Table 3, where "w/o CL" denoting the absence of the contrastive learning method, and "w/o Weight" indicating the utilization of sentiment labels to guide the selection of positive and negative sample pairs, without the incorporation of weights.

From the experimental results, we observe that the contrastive learning guided by sentiment intensity yields significant improvements for both MOSI and SIMS. However, the performance on MOSI is slightly degraded in the "w/o Weight" case. This can be attributed to the fact that MOSI includes more fine-grained sentiment intensity labels compared to SIMS. Consequently, without the incorporation of weights, contrastive learning may struggle to capture fine-grained information, thereby affecting the overall model performance. The proposed contrastive learning guided by sentiment intensity, which integrates sentiment intensity guidance-based weights, has yielded significant improvements in Acc-2 and Acc-7/Acc-5 for both MOSI and SIMS datasets. This highlights the effectiveness of the contrastive learning guided by sentiment intensity in enhancing the performance of the model.

Model	MOSI		SIMS	
	Acc-2	Acc-7	Acc-2	Acc-5
w/o CL	83.08	45.63	77.9	43.76
w/o Weight	82.77	44.75	79.21	44.2
CLGSI	**86.43**	**47.96**	**81.18**	**45.95**

Table 3: The ablation study results of the contrastive learning guided by sentiment intensity.

4.4.2 Effectiveness of GLFK

To demonstrate the efficacy of GLFK in CLGSI, we conducted a comparative analysis with the traditional “Add” and “Concatenate”. This means that V_c, T_c and A_c are directly added or concatenated into a one-dimensional vector and output as a common feature. Additionally, we devised two variants of GLFK for further evaluation. The first variant, GK, omits the local and fine components present in GLFK, while the second variant, referred to as LFK, excludes the global component. The results of our ablation studies (Table 4) reveal that the performance of the “Add” and “Concatenate” is inferior compared to GLFK. This can be attributed to their limited capacity for deeper and more comprehensive information interaction. On the coarse-grained metric (Acc-2), GK outperforms LFK by leveraging overall cognition of information. Conversely, LFK surpasses GK on the fine-grained metrics of Acc-7 and Acc-5, as it effectively captures detailed information. These findings underscore the importance of considering both global and detailed information in order to improve performance. GLFK facilitates complete information interaction across multiple modalities, enabling comprehensive and detailed information to be extracted for improved performance.

Model	MOSI		SIMS	
	Acc-2	Acc-7	Acc-2	Acc-5
Add	83.23	45.34	79.65	43.33
Concatenate	82.32	43.29	79.43	43.11
GK	82.77	45.04	79.87	43.76
LFK	82.32	47.96	79.43	45.3
CLGSI	**86.43**	**47.96**	**81.18**	**45.95**

Table 4: The ablation study results of GLFK.

4.4.3 Effectiveness of combination of common and specific features

In this subsection, we aim to evaluate the effectiveness of joint learning of common and specific features, the results of which are presented in Table 5. Specifically, “w/o Con” denotes the elimination of the common feature extraction module, while “w/o Spe” signifies the exclusion of the specific feature extraction module.

It can be seen from the results that the model’s performance is inferior when exclusively utilizing specific or common features compared to joint learning. In the case of “w/o Con”, the fusion of information between modalities solely relies on the final MLP. This shallow fusion approach leads to a certain level of performance degradation. In the case of “w/o Spe”, the model struggles to acquire effective common features for particularly complex sample scenarios, thereby negatively impacting performance. Nevertheless, when both common and specific features are jointly learned, we observe improved performance attributed to the complementarity between common and specific features.

Model	MOSI		SIMS	
	Acc-2	Acc-7	Acc-2	Acc-5
w/o Spe	84.6	39.36	78.34	44.86
w/o Con	83.38	43.59	78.56	44.2
CLGSI	**86.43**	**47.96**	**81.18**	**45.95**

Table 5: The ablation study results of the combination of common and specific features.

5 Conclusion

In this paper, we propose CLGSI, a novel MSA method. Firstly, CLGSI uses the contrastive learning guided by sentiment intensity to project different modalities into the same representation space. Then, by mimicking human cognitive process, GLFK is used to extract the common features between different modalities’ representations. At the same time, the output of each modal encoder was processed separately by MLP to extract the specific features of each modality. Finally, the joint learning of common and specific features was used to predict the sentiment intensity. We validate our model on both English and Chinese datasets, and the competitive results prove the good generalization performance and effectiveness of our model.

6 Limitation

While our model has shown impressive performance on MSA tasks, it is important to acknowledge the limitations that it faces. One notable limitation is that the proposed contrastive learning guide by sentiment intensity, cannot be directly applied to multimodal emotion recognition (MER) tasks. This is because the sample labels in MER tasks are different emotions (e.g., happy, angry, excited, etc.), and the sentiment intensity differences between them cannot be easily calculated. As a result, for MER tasks, we need to design an external mechanism that can transform the emotion labels and calculate their corresponding sentiment intensity differences. Our future work will aim to explore and develop effective mechanisms to address this limitation.

References

Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. 2021. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems*, 34:24206–24221.

Firoj Alam and Giuseppe Riccardi. 2014. Predicting personality traits using multimodal information. In *Proceedings of the 2014 ACM multi media on workshop on computational personality recognition*, pages 15–18.

Guoyong Cai and Binbin Xia. 2015. Convolutional neural networks for multimedia sentiment analysis. In *Natural Language Processing and Chinese Computing: 4th CCF Conference, NLPCC 2015, Nanchang, China, October 9-13, 2015, Proceedings 4*, pages 159–167. Springer.

Erik Cambria, Haixun Wang, and Bebo White. 2014. Guest editorial: Big social data analysis. *Knowledge-based systems*, (69):1–2.

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.

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

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

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.

Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu, Yuchuan Wu, and Yongbin Li. 2022. Unimse: Towards unified multimodal sentiment analysis and emotion recognition. *arXiv preprint arXiv:2211.11256*.

Onno Kampman, Elham J Barezi, Dario Bertero, and Pascale Fung. 2018. Investigating audio, video, and text fusion methods for end-to-end automatic personality prediction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 606–611.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673.

Yihe Liu, Ziqi Yuan, Huisheng Mao, Zhiyun Liang, Wanqiuyue Yang, Yuanzhe Qiu, Tie Cheng, Xiaoteng Li, Hua Xu, and Kai Gao. 2022. Make acoustic and visual cues matter: Ch-sims v2. 0 dataset and av-mixup consistent module. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, pages 247–258.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2018. Efficient low-rank multimodal fusion with modality-specific factors. *arXiv preprint arXiv:1806.00064*.

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*.

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.

Behnaz Nojavanasghari, Deepak Gopinath, Jayanth Koushik, Tadas Baltrušaitis, and Louis-Philippe Morency. 2016. Deep multimodal fusion for persuasiveness prediction. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 284–288.

Sunghyun Park, Han Suk Shim, Moitreya Chatterjee, Kenji Sagae, and Louis-Philippe Morency. 2016. Multimodal analysis and prediction of persuasiveness in online social multimedia. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 6(3):1–25.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. 2020. Integrating multimodal information in large pretrained transformers. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2020, page 2359. NIH Public Access.

Verónica Pérez Rosas, Rada Mihalcea, and Louis-Philippe Morency. 2013. Multimodal sentiment analysis of spanish online videos. *IEEE intelligent Systems*, 28(3):38–45.

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.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2019, page 6558. NIH Public Access.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Dingkang Yang, Shuai Huang, Haopeng Kuang, Yangtao Du, and Lihua Zhang. 2022. Disentangled representation learning for multimodal emotion recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1642–1651.

Jiuding Yang, Yakun Yu, Di Niu, Weidong Guo, and Yu Xu. 2023. Confede: Contrastive feature decomposition for multimodal sentiment analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7617–7630.

Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. 2020. Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 3718–3727.

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.

Yakun Yu, Mingjun Zhao, Shi-ang Qi, Feiran Sun, Baoxun Wang, Weidong Guo, Xiaoli Wang, Lei Yang, and Di Niu. 2023. Conki: Contrastive knowledge injection for multimodal sentiment analysis. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13610–13624.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*.

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 AAAI conference on artificial intelligence*, volume 32.

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 AAAI Conference on Artificial Intelligence*, volume 32.

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

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.

A Dateset

Table 6 shows the statistics of these datasets.

MOSI: The MOSI dataset is a popular dataset with three modalities (i.e. text, video, and audio). It was collected from 93 YouTube videos in which a speaker expressed an opinion on the film. MOSI contains 2199 speech video clips. Each segment is assigned a sentiment score ranging from -3 (strongly negative) to +3 (strongly positive).

MOSEI: The MOSEI dataset is a larger version of MOSI and contains 22856 annotated video clips over 250 different topics. As in MOSI, the sentiment score for each segment ranges from -3 (strongly negative) to +3 (strongly positive).

SIMS: The SIMS dataset is a Chinese multimodal dataset containing 2281 refined video clips. Each sample has a multimodal label and three unimodal labels with sentiment scores ranging from -1 (strongly negative) to +1 (strongly positive).

Dataset	Train	Valid	Test	Total
MOSI	1284	229	686	2199
MOSEI	16326	1871	4659	22856
SIMS	1368	456	457	2281

Table 6: The statistics of MOSI, MOSEI and SIMS.

B Baselines

LF-DNN: Late fusion DNN (LF-DNN) simply concatenates unimodal features extracted from unimodal features for sentiment inference (Yu et al., 2020)

MFN: Memory Fusion Network (MFN) (Zadeh et al., 2018a), which first learns view-specific interactions via LSTM, then learns cross-view interactions via attention network, and finally summarizes time via multi-view gated memory. The outputs of the MFN are concatenated as the final representation.

LMF: Low-Rank Multimodal Fusion (LMF) method (Liu et al., 2018) utilizes low-rank tensors to perform multimodal fusion efficiently.

TFN: The Tensor Fusion Network (TFN) (Zadeh et al., 2017) consists of 1) a modal embedding subnetwork to enrich the encoding of unimodal features as input and output after the neural network, 2) a tensor fusion layer to model unimodal, bimodal, and trimodal interactions using outer products, and 3) a sentiment inference subnetwork to perform sentiment inference.

MulT: The Multimodal Transformer (MulT) (Tsai et al., 2019) leverages directional pairwise cross-modal attention to learn the interactions between multimodal sequences and potentially adapt the flow from one modality to another.

MISA: MISA (Hazarika et al., 2020) is a multimodal framework that learns a modality-invariant and modality-specific representation for each modality. The learning process is optimized by including a combination of similarity loss, orthogonality loss, reconstruction loss, and task prediction loss.

MAG-BERT: Multimodal Adaptation Gates for Bert (MAG-BERT) (Rahman et al., 2020) are developed by applying multimodal adaptation gates at different layers of the BERT backbone.

HyCon: Hybrid Contrastive Learning for Trimodal Representations (HyCon) (Mai et al., 2022) is developed based on the contrastive learning method. It focuses on inter-sample and inter-class relationships, and reduce the modality gap.

Self-MM: Self-MM (Yu et al., 2021) first utilizes a self-supervised label generation module to obtain unimodal labels, and then jointly learns multimodal and unimodal representations based on multimodal labels and the generated unimodal labels.

ConFEDE: ConFEDE (Yang et al., 2023) is also a contrastive learning based framework. It performs contrastive representation learning and contrastive feature decomposition jointly to improve the representation of multimodal information. It decomposes each of the three modalities of video samples, including text, video frame and audio, into similarity features and dissimilarity features, and selects positive and negative sample pairs to learn with text as the center.

C Experimental Settings

Here, we briefly present the detailed setup of our experiments. All experiments were performed on a single NVIDIA RTX 4090 GPU. The trainable parameters of all implementations of CLGSI are under 120M. The training mode is full training, without additional pre-training. For Chinese text encoding, we use "bert-base-chinese"[1], and for English encoding, we use "bert-base-uncased"[2]. The number of layers of video transformer encoder and audio transformer encoder is 2. The optimizer is AdamW and the learning rate policy is warmup. Some of the key hyperparameters are shown in the table 7.

Para	MOSI	MOSEI	SIMS
Batch-size	64	128	64
Bert lr	5e-5	5e-5	5e-5
Visual Encoder lr	5e-3	5e-4	5e-4
Audio Encoder lr	1e-3	5e-4	5e-4
Others lr	1e-2	25e-4	5e-4
d_c	64	128	64
d_s	64	128	64

Table 7: Hyper-parameters of CLGSI for the multimodal sentiment analysis.

[1] https://huggingface.co/bert-base-chinese

[2] https://huggingface.co/bert-base-uncased