Semi-IIN: Semi-supervised Intra-Inter Modal Interaction Learning Network for Multimodal Sentiment Analysis

Jinhao Lin, Yifei Wang, Yanwu Xu, Qi Liu[*]

South China University of Technology
ftlinjinhao@mail.scut.edu.cn, ywang634@Outlook.com, ywxu@ieee.org, drliuqi@scut.edu.cn

Abstract

Despite multimodal sentiment analysis being a fertile research ground that merits further investigation, current approaches take up high annotation cost and suffer from label ambiguity, non-amicable to high-quality labeled data acquisition. Furthermore, choosing the right interactions is essential because the significance of intra- or inter-modal interactions can differ among various samples. To this end, we propose Semi-IIN, a Semi-supervised Intra-inter modal Interaction learning Network for multimodal sentiment analysis. Semi-IIN integrates masked attention and gating mechanisms, enabling effective dynamic selection after independently capturing intra- and inter-modal interactive information. Combined with the self-training approach, Semi-IIN fully utilizes the knowledge learned from unlabeled data. Experimental results on two public datasets, MOSI and MOSEI, demonstrate the effectiveness of Semi-IIN, establishing a new state-of-the-art on several metrics.

Code — https://github.com/flow-ljh/Semi-IIN
Extended version — https://arxiv.org/abs/2412.09784

Introduction

Multimodal sentiment analysis (MSA) has attracted increasing attention in recent years due to the rapid development of online social media platforms (Poria et al. 2020). Multimodal data offers more emotional cues than unimodal sentiment analysis, allowing machines to interpret human behaviors better and make more precise sentiment predictions (Zhang, Xu, and Lin 2021; Hu, Lu, and Zhao 2021). However, utilizing various modalities for analyzing human emotions continues to be a significant obstacle, particularly in the context of multimodal interactions with unlabeled data. Existing methods can be categorized into supervised learning and semi-supervised learning approaches. The former focuses on multimodal fusion and alignment, where the goal is to extract complementary information from different modalities and better understand human emotions. For multimodal fusion, current approaches acquire joint representations by imposing constraints (Hazarika, Zimmermann, and Poria 2020) or employing interactive operations (Zadeh et al. 2017; Liu et al. 2018) on the representations of individual modalities within the feature space. For multimodal alignment, researchers are committed to designing a cross-modal attention mechanism (Tsai et al. 2019) or an inter-modal temporal position prediction task (Yu et al. 2023) to capture cross-modal alignment information. The latter addresses data annotation's time-consuming and labor-intensive nature through semi-supervised methods. Liang (Liang, Li, and Jin 2020) designed a cross-modality distribution matching task to enhance the consistency of emotional representation, while Lian (Lian, Liu, and Tao 2022) proposed a Semi-supervised Multi-modal Interaction Network (SMIN) to learn multimodal interactive and contextual information. Current supervised and semi-supervised learning methods have made significant advancements, but they struggle to incorporate sentiment-related features from unlabeled data or properly separate irrelevant knowledge specific to different modalities. Figure 1 illustrates how crucial it is to maintain a balanced proportion of information passing between intra- and inter-modal interactions. On one side,

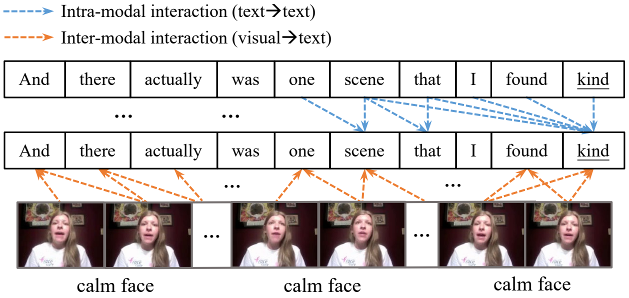


Figure 1: The importance of dynamically controlling the intra- and inter-modal interactive information. The arrows denote attention weights, whereas the blue and orange arrows indicate attention weight distributions ”between words in language modeling” and ”from visual to textual modalities”, respectively. For instance, on the semantic similarity level, the words ”scene” and ”that” refer to the same concept, resulting in higher attention scores between them (arrow: ”scene” to ”that”). On the task-oriented level, the word ”kind” is a key sentiment word and thus has a higher self-attention score (arrow: ”kind” to ”kind”).

[*] Corresponding author

the speaker's emotion can be inferred from key sentiment words such as "kind". Nevertheless, the transfer of information between visual and textual sequences in an inter-modal interactive manner can lead to confusion, as conventional attention may not distinguish between different words (due to the absence of clear positive emotional cues in image sequences) and overlook important emotional expressions. Conversely, when a speaker's comments, gestures, and intonation all convey a uniform emotional tone, inter-modal interaction becomes crucial. It can be used as additional information for interaction within the same mode. Therefore, we suggest a new framework that adjusts the proportion of intra- and inter-modal interaction based on the assumption that independent learning and dynamic selection of information are essential. This framework also utilizes self-training to gain knowledge from unlabeled data. Our contributions can be summarized as follows:

- We present a new network called Semi-IIN that combines two unique masked attention mechanisms to capture meaningful interactions among image sequences, audio frames, and text tokens.
- We use a self-training method that creates dependable pseudo-labels by a top-k confidence filtering strategy, allowing for model improvements through retraining and the extraction of emotion-related features from data without labels. Under the semi-supervised learning setting, Semi-IIN achieves improved performance.
- Experimental results on two public datasets show that Semi-IIN performs better than other current methods. To better understand how effective our approach is, we carry out thorough ablation experiments.

Related Work

Semi-supervised Sentiment Analysis

Supervised learning methods are commonly used for sentiment analysis, but their effectiveness is limited by the lack of labeled data for training. Researchers are trying to address this challenge by incorporating semi-supervised learning techniques to decrease the need for labeled data and enhance overall performance. To generate reliable pseudo-samples, researchers are committed to incorporating consistency-based pseudo-label strategy to identify misleading instances (Yuan et al. 2024), or establishing a specific threshold for prediction confidence in categories with clear and dependable characteristics (Cheng et al. 2023). Another direction is utilizing autoencoders (Lian, Liu, and Tao 2022; Zhang et al. 2020) to extract emotion-salient representations from additional unlabeled data. To narrow the heterogeneous gap between different modalities, Hu (Hu et al. 2020) designed a Semi-supervised Multimodal Learning Network, which correlates different modalities by capturing the multimodal data's intrinsic structure and discriminative correlation. Liang (Liang, Li, and Jin 2020) proposed a semi-supervised learning method based on cross-modal distribution matching. Parthasarathy (Parthasarathy and Busso 2020) employed semi-supervised ladder networks that incorporated skip connections between the encoder and decoder to extract emotion-relevant features.

Multimodal Interaction Learning

Previous research has focused on creating fusion strategies to capture interactive connections. Existing methods can be categorized into utterance-level and token-level interaction learning. To facilitate learning at the level of utterance interactions, single-mode representations are initially encoded individually and then combined by applying constraints (Hazarika, Zimmermann, and Poria 2020; Yu et al. 2021), separating (Tsai et al. 2018), analyzing correlations (Sun et al. 2020), or capturing relationships (Zadeh et al. 2017; Liu et al. 2018) to enable single-mode, dual-mode, or triple-mode interactions. Recently, Han (Han, Chen, and Poria 2021) introduced information theory to maximize the mutual information between unimodal and multimodal fusion results, while Yang (Yang et al. 2023) performed contrastive representation learning and contrastive feature decomposition to enhance the representation of multimodal information. Nevertheless, these approaches fail to account for the fact that emotions can vary throughout different points in the video as unimodal representations are averaged along the time axis to capture intricate and evolving emotional signals. Additionally, they face either high computational complexity or the introduction of extra hyperparameters. For token-level interaction learning, MAG-BERT proposed by Rahman (Rahman et al. 2020), incorporates non-verbal token-level information by generating a shift based on visual and acoustic modalities to enhance interaction learning. Nevertheless, this method necessitates coordination among modalities. To address this challenge, Tsai (Tsai et al. 2019) proposed a cross-modal attention mechanism to reinforce the target modality with emotional signals from the source modality. Recently, Chen (Chen et al. 2023a) proposed an inter-intra modal representation augmentation approach to enhance modal-representation learning ability.

Our goal is to utilize sentiment information from text, audio, and visual cues at the token level to improve the model's generalization ability with the help of semi-supervised learning. In contrast to previous studies, the processing of interactive information across different senses is carried out independently, filtering out distracting stimuli and regulated by a gate mechanism to maintain the consistency of emotional cues. Moreover, we use a self-training approach to enhance model training.

Method

Feature Encoding

This section provides a detailed description of our proposed framework, Semi-IIN. As shown in Figure 2, we first use the pre-trained 24-layer RoBERTa (Liu et al. 2019) to capture lexical features to obtain a single-modal representation, following the approach of previous studies (Lian, Liu, and Tao 2022; Chen et al. 2023a). To capture visual emotions, we use the pre-trained Fabnet (Wiles, Koepke, and Zisserman 2018; Chen et al. 2023a) to depict fundamental emotional characteristics. HuBERT (Hsu et al. 2021) is utilized for extracting the initial vector representations in the context of acoustic modality. They are formulated as:

$$h_{ti} = RoBERTa(X_{ti}; \theta_t^{RoBERTa}) \in \mathbb{R}^{l_t \times d_t} \tag{1}$$

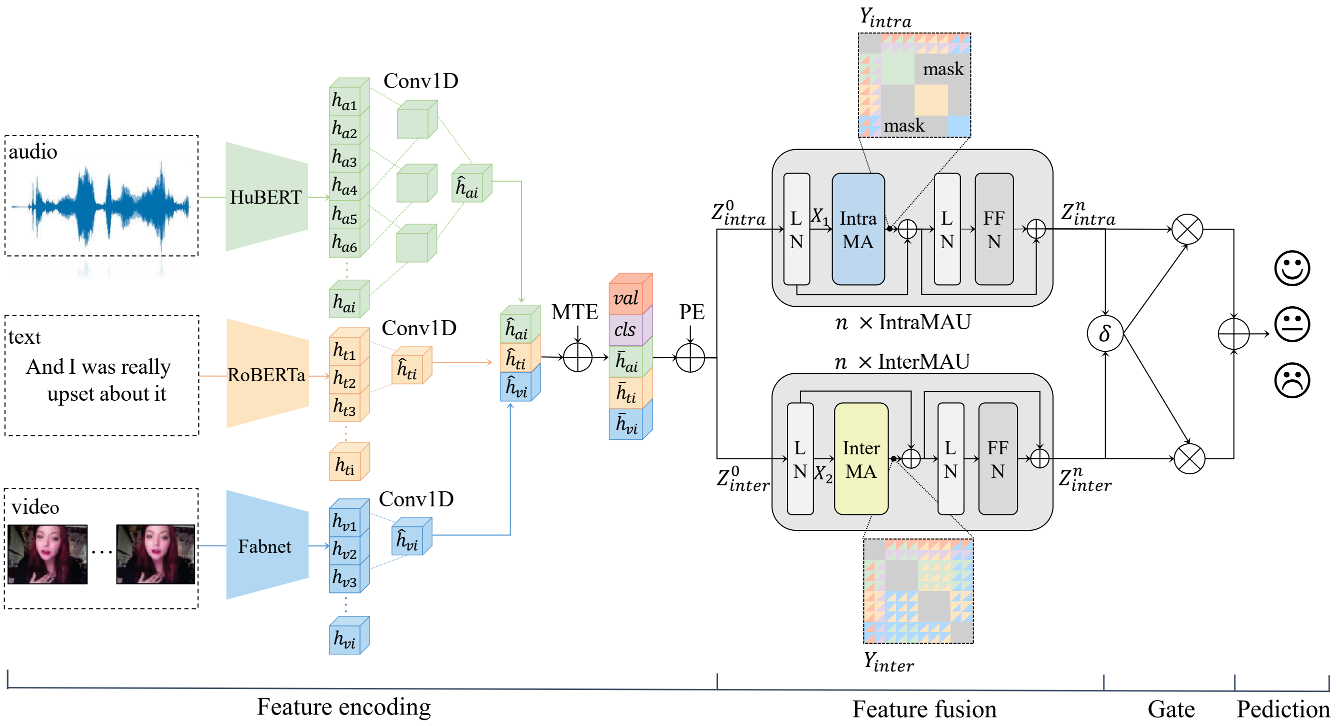


Figure 2: The overall architecture of Semi-IIN. Notably, Z^0_{inter} and Z^0_{intra} are the same as Z in equation (10).

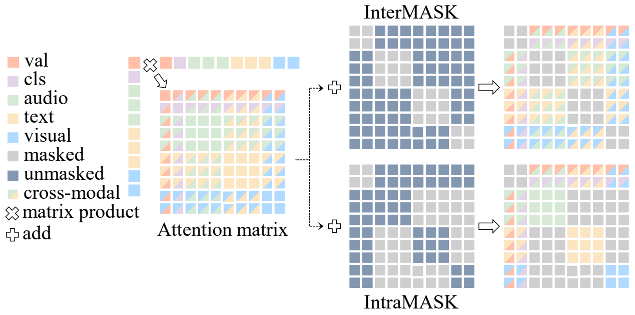


Figure 3: Implementation of InterMA(top) and IntraMA(bottom).

$$h_{vi} = Fabnet(X_{vi}; \theta_v^{Fabnet}) \in \mathbb{R}^{l_v \times d_v} \tag{2}$$

$$h_{ai} = HuBERT(X_{ai}; \theta_a^{HuBERT}) \in \mathbb{R}^{l_a \times d_a} \tag{3}$$

where h_{ti}, h_{vi}, and h_{ai} represent features corresponding to the lexical, visual, and acoustic modalities for the i-th sample, respectively. Next, a module for local feature extraction is used, which includes an 1D convolutional neural network (Conv1D) with various receptive fields to uncover the emotion-relevant features of each type of data. They are:

$$\hat{h}_{ti} = Conv1D(h_{ti}, k_t) \in \mathbb{R}^{l_t^{\prime} \times d_h} \tag{4}$$

$$\hat{h}_{vi} = Conv1D(h_{vi}, k_v) \in \mathbb{R}^{l_v^{\prime} \times d_h} \tag{5}$$

$$\hat{h}_{ai} = Conv1D(Conv1D(h_{ai}, k_{a1}), k_{a2}) \in \mathbb{R}^{l_a^{\prime} \times d_h} \tag{6}$$

where k_t, k_v, k_{a1}, and k_{a2} are the convolutional kernel sizes. d_h is the common hidden dimension. Afterward, $\hat{h}_{ti}$, $\hat{h}_{vi}$, and $\hat{h}_{ai}$ are added with their corresponding modal-type embeddings (MTE).

$$\bar{h}_{ti} = \hat{h}_{ti} + t^{type} \tag{7}$$

$$\bar{h}_{vi} = \hat{h}_{vi} + v^{type} \tag{8}$$

$$\bar{h}_{ai} = \hat{h}_{ai} + a^{type} \tag{9}$$

Next, we combine the hidden representations at the token level from three different modalities using two weight vectors (cls and val), along with introducing positional encoding (PE), to create the multimodal input sequence Z. That is:

$$Z = [val; cls; \bar{h}^1_{ai}; ...; \bar{h}^L_{ai}; ...; \bar{h}^1_{ti}; ...; \bar{h}^N_{ti}; \bar{h}^1_{vi}; ...; \bar{h}^M_{vi}] + PE \tag{10}$$

where $Z \in \mathbb{R}^{T \times d_h}$. L, N, and M represent the length of input feature sequences of corresponding modalities and the total sequence $T = L + N + M + 2$.

Feature Fusion

As shown in Figure 3, two distinct Masked Attention(MA), Intra-modal Masked Attention(IntraMA) and Inter-modal Masked Attention(InterMA) are designed to mask unrelated relations between modalities. The former focuses on exploiting interactive information within each modality, while the latter enables the cross-modal exchange of emotional cues. Specifically, we design two different attention masks: Intra-modal MASK(IntraMASK) and Inter-modal MASK(InterMASK):

$$\text{IntraMASK}_{ij} = \begin{cases} 0, & \text{if } i,j \in \text{Intra}_{pos} \\ -\infty, & \text{if } i,j \notin \text{Intra}_{pos} \end{cases} \tag{11}$$

$$\text{InterMASK}_{ij} = \begin{cases} 0, & \text{if } i,j \in \text{Inter}_{pos} \\ -\infty, & \text{if } i,j \notin \text{Inter}_{pos} \end{cases} \tag{12}$$

where IntraMASK $\in \mathbb{R}^{T \times T}$ and InterMASK $\in \mathbb{R}^{T \times T}$. Intra$_{pos}$ and Inter$_{pos}$ are two pre-defined matrices designed to separate the tokens of the interaction position from the masked ones. After that, IntraMA is achieved by adding the IntraMASK with the conventional global attention (Vaswani et al. 2017), which facilitates the extraction of the key emotional clues in each modality. It is mathematically expressed as:

$$\begin{aligned} Y_{intra} &= \text{IntraMA}(X_1) \\ &= softmax(\frac{QK^T}{\sqrt{d_k}} + \text{IntraMASK})V \\ &= softmax(\frac{X_1 W_Q W_K^T X_1^T}{\sqrt{d_k}} + \text{IntraMASK})X_1 W_V \end{aligned} \tag{13}$$

where input $X_1 \in \mathbb{R}^{T \times d_h}, W_Q \in \mathbb{R}^{d_h \times d_k}, W_K \in \mathbb{R}^{d_h \times d_k}$, and $W_V \in \mathbb{R}^{d_h \times d_v}$. Similarly to the IntraMA, InterMA is achieved as:

$$\begin{aligned} Y_{inter} &= \text{InterMA}(X_2) \\ &= softmax(\frac{QK^T}{\sqrt{d_k}} + \text{InterMASK})V \\ &= softmax(\frac{X_2 W_Q W_K^T {X_2}^T}{\sqrt{d_k}} + \text{InterMASK})X_2 W_V \end{aligned} \tag{14}$$

Subsequently, as shown in Figure 2, by replacing the global attention with the proposed IntraMA and InterMA, the Intra-modal Masked Attention Unit(IntraMAU) and the Inter-modal Masked Attention Unit(InterMAU) are constructed. Specifically, for the l-th layer input $Z_{intra}^{l} \in \mathbb{R}^{T \times d_h}$, the output of l-th layer of the IntraMAU can be calculated as:

$$\begin{aligned} \hat{Z}_{intra}^{l} &= \text{IntraMA}(\text{LN}(Z_{intra}^{l-1})) + \text{LN}(Z_{intra}^{l-1}) \\ Z_{intra}^{l} &= \text{FFN}(\text{LN}(\hat{Z}_{intra})) + \text{LN}(\hat{Z}_{intra}^{l}) \end{aligned} \tag{15}$$

where $l \in [1, n]$, FFN and LN represent the feed-forward network with ReLU as the activation function and the layer normalization, respectively. The InterMAU consists of the same modules as the IntraMAU except for the IntraMA being replaced with the InterMA.

Gate Mechanism

Two special tokens, val, and cls, serve as fusion features, aggregating information from all tokens except for the special ones. Thus, the first element of Z_{intra}^{n} and Z_{inter}^{n}, along with the second position of Z_{intra}^{n} and Z_{inter}^{n} are treated as the final fusion feature. They are processed via the following dynamic gate mechanism (Lv et al. 2021):

$$\begin{aligned} G_v &= sigmoid(Z_{intra}^{n}[0] \cdot W_1 + Z_{inter}^{n}[0] \cdot W_2 + b_1) \\ Z_v &= G_v \odot Z_{inter}^{n}[0] + (1 - G_v) \odot Z_{intra}^{n}[0] \\ G_e &= sigmoid(Z_{intra}^{n}[1] \cdot W_1^{'} + Z_{inter}^{n}[1] \cdot W_2^{'} + b_1^{'}) \\ Z_e &= G_e \odot Z_{inter}^{n}[1] + (1 - G_e) \odot Z_{intra}^{n}[1] \end{aligned} \tag{16}$$

where W_1, W_2, $W_1^{'}$ and $W_2^{'}$ all $\in \mathbb{R}^{d_h \times d_h}, b_1$ and $b_1^{'} \in \mathbb{R}^{d_h}$. The passed proportions between modality-specific and modality-complimentary knowledge are further dynamically determined through the gating approach, facilitating meaningful modality interaction learning. The predictions of sentiment intensity v and emotional category e are derived from the filtered feature Z_v and Z_e. They are

$$v = \text{FC}_v(\text{MLP}(Z_v)) \in \mathbb{R}^{d_1} \tag{17}$$

$$e = \text{FC}_e(\text{MLP}(Z_e)) \in \mathbb{R}^{d_2} \tag{18}$$

Algorithm 1: Self-training

Input: Labeled dataset(L_d) and unlabeled dataset(U_d). Model ϕ

Output: Final model ϕ'

```
while current epoch < total epoch do
    for $sample_i$, $v_i$, $e_i$ in $L_d$ do
        $\hat{v}_i$, $\hat{e}_i$ = $\phi(sample_i)$
        Using equation (19), (20) and (22) to caculate loss
        and update model $\phi$'s parameters.
    end for
end while
Initialize model $\phi'$ using $\phi$. Generate predictions based
on $U_d$ and $\phi'$. Constructing dataset $U_d'$ by selecting the
top-k highest confidence ones in each category
while current epoch < total epoch do
    for $sample_i$, $v_i$, $e_i$ in $L_d \cup U_d'$ do
        $\hat{v}_i$, $\hat{e}_i$ = $\phi'(sample_i)$
        Using equation (19), (20), (21) and (23) to caculate
        loss and update model $\phi'$'s parameters.
    end for
end while
return model $\phi'$
```

Self-training

We design a self-training strategy to distill emotional knowledge from unlabeled data. The process is shown in algorithm 1. Initially, we train the MSA model shown in Figure 2 with labeled data and then employ the trained model, referred to as ϕ, to make predictions using unlabeled data. Following recent progress in semi-supervised learning (Chen et al. 2023b), the top-k confidence method is employed to eliminate unreliable samples. Due to the significantly larger amount of data in the MOSEI dataset in comparison to the MOSI dataset, the model trained on the MOSEI dataset generates predictions with greater confidence and accuracy. As a result, we assigned a value of 40 to k for the MOSI dataset, whereas, for the MOSEI dataset, k is set to the total

		MOSI					MOSEI				
Methods	Embedding	MAE↓	Corr↑	Acc-2↑	F1↑	Acc-7↑	MAE↓	Corr↑	Acc-2↑	F1↑	Acc-7↑
LMF†	Glove	0.917	0.695	-/82.50	-/82.40	33.20	0.623	0.700	-/82.00	-/82.10	48.00
TFN†	Glove	0.901	0.698	-/80.80	-/80.70	34.90	0.593	0.677	-/82.50	-/82.10	50.20
MFM‡	Glove	0.877	0.706	-/81.7	-/81.6	35.40	0.568	0.703	-/84.40	-84.30	51.30
Mult‡	Glove	0.861	0.711	81.50/84.10	80.60/83.90	-	0.580	0.713	82.50/84.23	82.67/83.97	-
ICCN‡	Bert-base	0.862	0.714	-/83.00	-/83.00	39.00	0.565	0.704	-/84.20	-/84.20	51.60
MISA†	Bert-base	0.804	0.764	80.79/82.10	80.77/82.03	-	0.568	0.717	82.59/84.23	82.67/83.97	-
Self-MM†	Bert-base	0.713	0.798	84.00/85.98	84.42/85.95	-	0.530	0.765	82.81/85.17	82.53/85.30	-
MAG-BERT‡	Bert-base	0.712	0.796	84.20/86.10	84.10/86.00	-	-	-	84.70/-	84.50/-	-
MMIM†	Bert-base	0.700	0.800	84.14/86.06	84.00/85.98	**46.65**	0.526	0.772	82.24/85.97	82.66/85.94	54.24
ConFEDE†	Bert-base	0.742	0.784	84.17/85.52	84.13/85.52	42.27	0.522	0.780	81.65/85.82	82.17/85.83	54.86
SMIN†	Roberta-large	-	-	-/81.55	-/81.45	-	-	-	-/86.82	-/86.81	-
TCDN†	Roberta-large	0.697	0.805	-/**87.10**	-/**87.20**	-	0.521	0.782	-/87.50	-/87.20	-
Ours†	Roberta-large	**0.679**	**0.822**	**85.28**/87.04	**85.19**/87.00	46.50	**0.497**	**0.804**	**84.98/87.70**	**85.27/87.65**	**55.89**

Table 1: Results on CMU-MOSI and CMU-MOSEI dataset. The best performance is highlighted in bold, while the second-best is denoted with an underline. †: unaligned setting. ‡: aligned setting

number of unlabeled instances. Finally, we combine the labeled and unlabeled data(the labeled portion) to retrain the model(weight initialization from ϕ). Note that only the emotion classification task loss is sent back through the network for pseudo-labeled samples.

Loss Function

Emotional states can be represented either through discrete categories (such as “sad” and “happy”) or dimensional annotations (points in a continuous space). In the MSA, Lian (Lian et al. 2023) and Wang (Wang et al. 2022) highlighted a high correlation between discrete and dimensional annotations. As a result, we classify the data into seven specific emotional categories by determining how close the dimensional labels are to predefined discrete categories. After obtaining discrete labels, we adopt the Mean Squared Error (MSE) L_v and Cross-entropy Loss L_e as our optimization objectives. We have:

$$L_v = \frac{1}{N_l}\sum_{i}^{N_l}(\hat{v}^i - v^i)^2 \tag{19}$$

$$L_e = -\frac{1}{N_l}\sum_{i}^{N_l} e_i log(\hat{e}^i) \tag{20}$$

$$L_e^u = -\frac{1}{N_u}\sum_{i}^{N_u} e_i log(\hat{e}^i) \tag{21}$$

The loss function is defined in the supervised learning process as follows:

$$L_{total}^p = \lambda_1 L_v + (1-\lambda_1)L_e \tag{22}$$

while for the semi-supervised process, the loss function is as follows:

$$L_{total}^r = \lambda_1 L_v + (1-\lambda_1)L_e + \lambda_2 L_e^u \tag{23}$$

where N_l and N_u are the number of labeled and unlabeled training samples, respectively. λ_1 and λ_2 are two weighting factors.

Experiment

Dataset

CMU-MOSI (Zadeh et al. 2016), a dataset for human MSA, includes 2,199 video segments taken from 93 videos. Each segment is marked with a sentiment score between -3 and +3 to show the level of sentiment expressed in that portion.
CMU-MOSEI (Zadeh et al. 2018), an enhanced version of MOSI, consists of 22,856 video clips. Each segment is annotated with sentiment and emotion.
AMI (Carletta et al. 2005) dataset includes 100 hours of meeting recordings. It provides video recordings of each speaker, voice track, and transcripts of their speeches. We use it as the unlabeled dataset because it does not include any sentiment annotation.

Evaluation Metrics

We report Mean Absolute Error (MAE), Pearson correlation (Corr), binary classification accuracy (Acc-2), and F1-score on MOSI and MOSEI datasets. The Acc-2 and F1-score are calculated in negative/non-negative (non-exclude zero) and negative/positive (exclude zero).

Implementation Details

Semi-IIN is trained with the Adam optimizer, configured with a learning rate of 1e-4. The batch size is 32. The loss

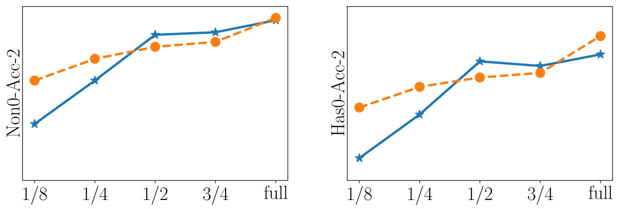


Figure 4: Results under different proportions of labeled samples on MOSI dataset. Blue: Semi-IIN without Semi. Orange: Semi-IIN

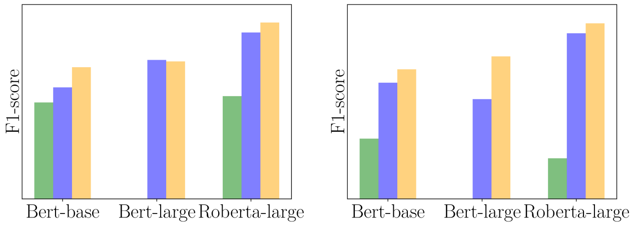


Figure 5: Comparison between different embedding on two datasets. Left: MOSEI dataset. Right: MOSI dataset. Green: SMIN-fully, the previous semi-supervised SOTA method(under fully-supervised training). Purple: ours-w/o MA, Semi-IIN without MA is trained under full supervision. Orange: ours-fully, Semi-IIN with MA is trained under full supervision.

weight factors λ_1 and λ_2 are set at 0.6 and 0.2, and the embedding size of transformer encoders is 128. We implement the proposed Semi-IIN on a single NVIDIA A100.

Comparison to State-of-the-art Methods

Table 1 illustrates the results for two datasets. For the unaligned setting (LMF, TFN, MISA, Self-MM, MMIM, ConFEDE, SMIN, and TCDN) and the aligned setting (MFM, ICCN, MulT, and MAG-BERT), our method achieves competitive performance on the MOSI and MOSEI datasets. Specifically, compared to the current SOTA method TCDN which employs the same word embeddings (Roberta-large), Semi-IIN surpasses it by 0.016 MAE and 0.017 Corr on the MOSI dataset, respectively. On the MOSEI dataset, Semi-IIN surpasses TCDN by 0.024 MAE and 0.45% accuracy. These results demonstrate the superiority of Semi-IIN in MSA.

Ablation Study

Firstly, to verify the effectiveness of the self-training strategy, we conducted experiments with varying ratios of labeled samples in a semi-supervised training scenario. Figure 4 illustrates that Semi-IIN still shows progress even with a small number of labeled samples. However, performance decreases when the ratio is adjusted to 50% or 75%. The decrease in numbers could be due to the lack of balance in the created fake samples. Additionally, experiments are carried out to confirm that the increase in performance is not a result of improving word embeddings. Since Lian (Lian, Liu, and Tao 2022) only shares findings from fully supervised learning with different embedding setups, we conducted training for Semi-IIN in a fully supervised manner to maintain a fair comparison. As illustrated in Figure 5, Semi-IIN-fully consistently outperforms Semi-IIN-fully (without MA) across different embedding configurations on both the MOSI and MOSEI datasets. This result confirms the efficacy of the masked attention strategy. Furthermore, compared to SMIN-fully, Semi-IIN-fully demonstrates superior performance, validating its suitability and scalability.

Method	MA	Semi	Params	MAE↓	Corr↑	Acc-2↑	F1↑
Baseline	-	-	1.3M	0.509	0.793	83.97/86.85	84.30/86.8
Semi-IIN	✓		1.6M	0.499	0.800	**85.04**/87.26	85.22/87.14
		✓	1.3M	0.507	0.792	84.54/86.74	84.8/86.64
	✓	✓	1.6M	**0.497**	**0.804**	84.98/**87.70**	**85.27/87.65**

Table 2: Comparison of the overall result of Semi-IIN with different settings. Baseline employs conventional global attention. MA: IntraMA and InterMA. Semi: Semi-supervised learning

Fusion mode	MAE↓	Corr↑	Acc-2↑	F1↑
dot	0.506	0.794	84.07/87.01	84.37/86.93
add	0.512	0.794	83.45/86.85	83.80/86.78
concat	0.506	0.794	83.64/86.65	84.03/86.64
gate	**0.499**	**0.800**	**85.04/87.26**	**85.22/87.14**

Table 3: The impact of different fusion modes(Based on Semi-IIN(only MA))

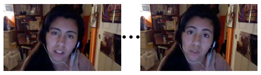
Furthermore, various ablation experiments are carried out under different conditions to showcase the effectiveness of the suggested MA, along with the semi-supervised learning approach. The results are presented in Table 2. Compared with the baseline model, despite introducing a few parameters, Semi-IIN(only MA) achieves nearly 1% accuracy improvement and lower MAE. Semi-IIN(with Semi) also results in a 0.5% increase in accuracy. The best outcome is achieved by combining Semi and MA. The findings above show that separating interactions into two branches from both intra- and inter-modal perspectives helps to better utilize consistent emotional signals across modalities. Additionally, using additional unlabeled data slightly improves the training of the model.

We also utilize various fusion methods to confirm the importance of dynamically determining the proportions of intra- and inter-modal information, as shown in Table 3. The findings show that the gating fusion method is more effective than other fusion techniques, highlighting the importance of choosing between modality-specific and modality-common information.

Qualitative Analysis

Case Study

In this section, we select two examples to verify the importance of dynamically selecting effective interactions. As

modality	content	**Ground Truth**	**Semi-IIN**	**Only Inter**	**Only Intra**
text	And I **like** how it shows.	1.2	0.97	0.55	0.94
audio	Unemtional tone				
visual	 calm face				

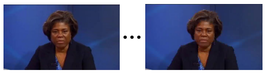
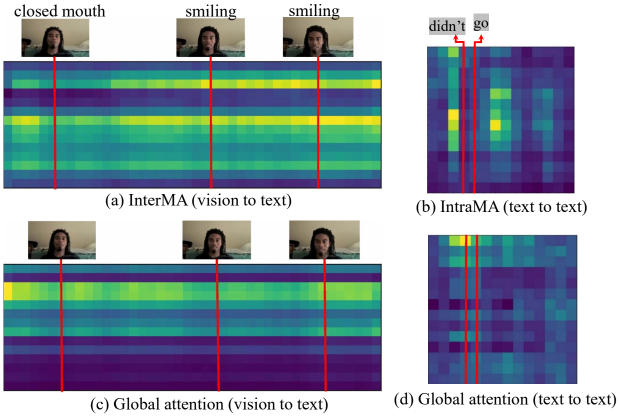
modality	content	**Ground Truth**	**Semi-IIN**	**Only Inter**	**Only Intra**
text	And I use this opportunity to call upon the people of CAR to end the violence, to find a way forward to peace	-0.33	-0.32	-0.66	0.78
audio	Unemtional tone				
visual	 straight face				

Figure 6: Case study for the Semi-IIN. The "Only Intra" and the "Only Inter" refer to the stacked IntraMAU and InterMAU prediction, respectively.



(a) InterMA (vision to text)

(b) IntraMA (text to text)

(c) Global attention (vision to text)

(d) Global attention (text to text)

Figure 7: Visualization of IntraMA and InterMA mechanisms

shown in Figure 6, in the first case, since visual and acoustic modality both contain irrelevant emotional signals such as calm face and unemotional tone, the inter-modal interactive branch is inclined to perceive the speaker's emotions as closer to neutral. In contrast, the intra-modal interactive branch offers a more precise prediction by disregarding the unuseful cross-modal information flow. In the second case, the intra-modal interactive branch is affected by the predominant lexical mode, leading to an inaccurate sentiment evaluation. Conversely, the inter-modal interactive branch fully utilizes visual modality that consists of abundant sentiment cues to reinforce lexical and acoustic modality, leading to accurate emotional polarity. In these two cases, Semi-IIN remains unaffected by irrelevant interactive noise and achieves accurate results overall by effectively exploring interactions.

Visualization of IntraMA and InterMA

Figure 7 illustrates the visualization results of InterMA and IntraMA. It is noteworthy that the speaker's emotion in this video is positive. Figure 7(a) and Figure 7(c) show that the InterMA pays more attention to important image frames with a lot of emotional content, instead of focusing on unnecessary frames, e.g., neutral facial expressions, unlike traditional global attention. Moreover, Figure 7(b) and Figure 7(d) illustrate how the IntraMA mechanism mitigates the impact of emotion-unrelated words like "didn't go", which may lead to incorrect affective polarity, by assigning less attention compared to conventional global attention. We think that IntraMA and InterMA are effective because they can use specific and complementary knowledge from different modalities to filter out unnecessary information.

Conclusion and Future Work

This paper introduces a new MSA framework, called Semi-IIN, aimed at reducing both intra- and inter-modal noise at a detailed level. Semi-IIN, along with the IntraMA and InterMA mechanisms, successfully captures important interactive information within and between modalities, making it easier to extract consistent emotional cues from multimodal data. In addition, our model decreases the need for extensive human annotations by including semi-supervised learning. The effectiveness of Semi-IIN is demonstrated through experimental results on two benchmark datasets, CMU-MOSI and CMU-MOSEI. Our proposal outperforms previous approaches, setting the new SOTA result for MSA. Our future directions mainly lie in designing semi-supervised intra-inter modal interaction learning networks for multilingual multimodal sentiment analysis, e.g., Spanish, French, and German, and enhancing interpretability.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 62202174, in part by the Basic and Applied Basic Research Foundation of Guangzhou under Grant 2023A04J1674, in part by The Taihu Lake Innocation Fund for the School of Future Technology of South China University of Technology under Grant 2024B105611004, and in part by Guangdong Science and Technology Department Grant 2024A1313010012.

References

Carletta, J.; Ashby, S.; Bourban, S.; Flynn, M.; Guillemot, M.; Hain, T.; Kadlec, J.; Karaiskos, V.; Kraaij, W.; Kronenthal, M.; et al. 2005. The AMI meeting corpus: A pre-announcement. In *International workshop on machine learning for multimodal interaction*, 28–39. Springer.

Chen, C.; Hong, H.; Guo, J.; and Song, B. 2023a. Inter-intra modal representation augmentation with trimodal collaborative disentanglement network for multimodal sentiment analysis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31: 1476–1488.

Chen, H.; Guo, C.; Li, Y.; Zhang, P.; and Jiang, D. 2023b. Semi-Supervised Multimodal Emotion Recognition with Class-Balanced Pseudo-labeling. In *Proceedings of the 31st ACM International Conference on Multimedia*, 9556–9560.

Cheng, Z.; Lin, Y.; Chen, Z.; Li, X.; Mao, S.; Zhang, F.; Ding, D.; Zhang, B.; and Peng, X. 2023. Semi-supervised multimodal emotion recognition with expression mae. In *Proceedings of the 31st ACM International Conference on Multimedia*, 9436–9440.

Han, W.; Chen, H.; and Poria, S. 2021. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. *arXiv preprint arXiv:2109.00412*.

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

Hsu, W.-N.; Bolte, B.; Tsai, Y.-H. H.; Lakhotia, K.; Salakhutdinov, R.; and Mohamed, A. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29: 3451–3460.

Hu, G.; Lu, G.; and Zhao, Y. 2021. Bidirectional hierarchical attention networks based on document-level context for emotion cause extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 558–568.

Hu, P.; Zhu, H.; Peng, X.; and Lin, J. 2020. Semi-supervised multi-modal learning with balanced spectral decomposition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 99–106.

Lian, Z.; Liu, B.; and Tao, J. 2022. Smin: Semi-supervised multi-modal interaction network for conversational emotion recognition. *IEEE Transactions on Affective Computing*.

Lian, Z.; Sun, H.; Sun, L.; Chen, K.; Xu, M.; Wang, K.; Xu, K.; He, Y.; Li, Y.; Zhao, J.; et al. 2023. Mer 2023: Multi-label learning, modality robustness, and semi-supervised learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, 9610–9614.

Liang, J.; Li, R.; and Jin, Q. 2020. Semi-supervised multimodal emotion recognition with cross-modal distribution matching. In *Proceedings of the 28th ACM international conference on multimedia*, 2852–2861.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Liu, Z.; Shen, Y.; Lakshminarasimhan, V. B.; Liang, P. P.; Zadeh, A.; and Morency, L.-P. 2018. Efficient low-rank multimodal fusion with modality-specific factors. *arXiv preprint arXiv:1806.00064*.

Lv, F.; Chen, X.; Huang, Y.; Duan, L.; and Lin, G. 2021. Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2554–2562.

Parthasarathy, S.; and Busso, C. 2020. Semi-supervised speech emotion recognition with ladder networks. *IEEE/ACM transactions on audio, speech, and language processing*, 28: 2697–2709.

Poria, S.; Hazarika, D.; Majumder, N.; and Mihalcea, R. 2020. Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research. *IEEE Transactions on Affective Computing*.

Rahman, W.; Hasan, M. K.; Lee, S.; Zadeh, A.; Mao, C.; Morency, L.-P.; and Hoque, E. 2020. Integrating multimodal information in large pretrained transformers. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2020, 2359. NIH Public Access.

Sun, Z.; Sarma, P.; Sethares, W.; and Liang, Y. 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, 8992–8999.

Tsai, Y.-H. H.; Bai, S.; Liang, P. P.; Kolter, J. Z.; Morency, L.-P.; and Salakhutdinov, R. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2019, 6558. NIH Public Access.

Tsai, Y.-H. H.; Liang, P. P.; Zadeh, A.; Morency, L.-P.; and Salakhutdinov, R. 2018. Learning factorized multimodal representations. *arXiv preprint arXiv:1806.06176*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, K.; Lian, Z.; Sun, L.; Liu, B.; Tao, J.; and Fan, Y. 2022. Emotional reaction analysis based on multi-label graph convolutional networks and dynamic facial expression recognition transformer. In *Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge*, 75–80.

Wiles, O.; Koepke, A.; and Zisserman, A. 2018. Self-supervised learning of a facial attribute embedding from video. *arXiv preprint arXiv:1808.06882*.

Yang, J.; Yu, Y.; Niu, D.; Guo, W.; and Xu, Y. 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)*, 7617–7630.

Yu, T.; Gao, H.; Lin, T.-E.; Yang, M.; Wu, Y.; Ma, W.; Wang, C.; Huang, F.; and Li, Y. 2023. Speech-Text Pre-training for Spoken Dialog Understanding with Explicit Cross-Modal Alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 7900–7913.

Yu, W.; Xu, H.; Yuan, Z.; and Wu, J. 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, 10790–10797.

Yuan, Z.; Fang, J.; Xu, H.; and Gao, K. 2024. Multimodal Consistency-Based Teacher for Semi-Supervised Multimodal Sentiment Analysis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; and Morency, L.-P. 2017. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*.

Zadeh, A.; Zellers, R.; Pincus, E.; and Morency, L.-P. 2016. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*.

Zadeh, A. B.; Liang, P. P.; Poria, S.; Cambria, E.; and Morency, L.-P. 2018. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2236–2246.

Zhang, D.; Li, S.; Zhu, Q.; and Zhou, G. 2020. Multimodal sentiment classification with independent and interactive knowledge via semi-supervised learning. *IEEE Access*, 8: 22945–22954.

Zhang, H.; Xu, H.; and Lin, T.-E. 2021. Deep open intent classification with adaptive decision boundary. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 14374–14382.