Modal Feature Optimization Network with Prompt for Multimodal Sentiment Analysis

Xiangmin Zhang[1,2] , **Wei Wei**[1,2]* , **Shihao Zou**[1,2]

[1] Cognitive Computing and Intelligent Information Processing (CCIIP) Laboratory,
School of Computer Science and Technology, Huazhong University of Science and Technology
[2] Joint Laboratory of HUST and Pingan Property & Casualty Research (HPL)
{xmz, weiw, sh_zou}@hust.edu.cn

Abstract

Multimodal sentiment analysis(MSA) is mostly used to understand human emotional states through multimodal. However, due to the fact that the effective information carried by multimodal is not balanced, the modality containing less effective information cannot fully play the complementary role between modalities. Therefore, the goal of this paper is to fully explore the effective information in modalities and further optimize the under-optimized modal representation.To this end, we propose a novel **M**odal **F**eature **O**ptimization **N**etwork (MFON) with a **M**odal **P**rompt **A**ttention (MPA) mechanism for MSA. Specifically, we first determine which modalities are under-optimized in MSA, and then use relevant prompt information to focus the model on these features. This allows the model to focus more on the features of the modalities that need optimization, improving the utilization of each modality's feature representation and facilitating initial information aggregation across modalities. Subsequently, we design an intra-modal knowledge distillation strategy for under-optimized modalities. This approach preserves the integrity of the modal features. Furthermore, we implement inter-modal contrastive learning to better extract related features across modalities, thereby optimizing the entire network. Finally, sentiment prediction is carried out through the effective fusion of multimodal information. Extensive experimental results on public benchmark datasets demonstrate that our proposed method outperforms existing state-of-the-art models.

1 Introduction

With the development of multimedia technology, information existing forms have become more and more diverse, for example, text modality, visual modality, and acoustic modality. In many research areas people have shifted from using unimodal data to using multimodal data for research, such as recommender systems (Wei et al., 2023; Fu et al., 2024; Zhong et al., 2024), video understanding (Buch et al., 2022; Ren et al., 2023) and dialogue generation (Sun et al., 2022; Kong et al., 2024). In addition, it has been shown that humans are more inclined to understand the world from multiple modalities, from which they perceive and express emotions. Multiple modal features collected from different sensors can complement each other and help humans understand the world better.

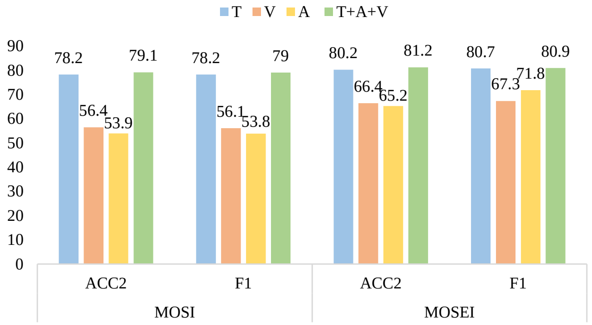

Figure 1: Mult (Tsai et al., 2019) experiments results on MOSI and MOSEI dataset for each unimodal (T, V, A) and modal combinations (T+A+V).

Multimodal sentiment analysis (MSA) leverages the relevance and complementarity of multimodal data for emotion detection and analysis. Specifically, compared to traditional text-based sentiment analysis, MSA utilizes additional information of visual and acoustic modalities in addition to text, whose core idea is more information resources can enhance model performance. Existing works focus on two perspectives: designing complex fusion methods (Zadeh et al., 2017; Mai et al., 2020) and extracting effective modal representation (Colombo et al., 2021; Yang et al., 2022; Zhang et al., 2023).

The above methods have achieved good results and it has been proved to some extent that the use of multiple modalities is indeed better than the use of only a single modality. However, we ob-

[*] Corresponding author.

Proceedings of the 31st International Conference on Computational Linguistics, pages 4611–4621
January 19–24, 2025. ©2025 Association for Computational Linguistics

serve that in the MSA task, there is a big difference between the representations of multimodal features, for which we first analyze the possible problems. First, text modality is less disturbed by noise and contains more emotional information, while visual and acoustic modalities are susceptible to noise and contain less emotional information. For example, visual images can be blurred by overillumination. Second, when the model is trained using three modalities, text modality containing more information is easy to optimize or overfit, which may create a shortcut between the model's predictions and text features. As a result, the model fails to fully integrate the emotional information contained in visual and acoustic modalities. In other words, visual and acoustic modalities are in a state of under-optimization.

The above phenomenon is also reflected in existing studies, it is specifically shown in Fig. 1. In the experiments of single modality, text performs better compared to the results of the other two modalities, from which we believe that text modality contains more emotional information than non-text modality. Moreover, the results of combining single text with three modalities are not much different, which makes one wonder if the multimodal really plays a complementary role. Obviously, the contribution of non-text modalities is very limited in the current example. For the above phenomenon, we believe that due to the under-optimization of visual and acoustic modalities, the information thus conveyed does not play a complementary role effectively, and may even affect the representation of text modality in some cases.

Based on the above problems and inferences, we believe that in order to achieve good sentiment analysis results, we need to further optimize the modalities with insufficient feature representations. Based on this, we propose the Modal Feature Optimization Network (MFON). First, to initially establish the information correlation between modalities and facilitate the subsequent optimization, we design Modal Prompt Attention (MPA) to carry out the information interaction between text modality and modalities to be optimized. Specifically, text modality is used as a guide, and modal prompts to remind the model which of the current modal needs to optimize, and to obtain a preliminary optimized and linked modal feature representation by this method. Then, in order to pay full attention to intra- and inter-modal information mining, we design an intra-modal knowledge distillation, which on the one hand ensures that the modal features to be optimized obtain more effective information, and on the other hand serves as a supervisory information to ensure that modal features to be optimized continue even when the model loss is converged. In addition, in order to further explore the feature representations between optimized modalities, we design inter-modal contrastive learning to make better use of the information and thus optimize the modal feature representations in the whole model. Experimental results demonstrate the effectiveness of our method. The code is released at https://github.com/123sprouting/MFON/. In summary, our contributions are as follows:

- We propose a modal feature optimization network, which can effectively optimize the multi-modal feature representations and can further explore the correlation between modal features.

- The Modal Prompt Attention (MPA) module is designed to carry out the information interaction between text modal and the modal to be optimized, through which the first optimization of the modal to be optimized is realized.

- We optimize modal features by intra-modal knowledge distillation to obtain more information, and perform inter-modal contrastive learning to further explore the relationship between optimized modal features to obtain a better representation of the optimized features.

- Experimental results demonstrate that our method achieves competitive performance on MSA benchmarks.

2 Related Work

Multimodal sentiment analysis has become a popular research topic that utilizes information from text, visual, and acoustic modalities to comprehend human sentiment (Caschera et al., 2016). Previous research focuses on multimodal fusion and modal representation.

For multimodal fusion, early fusion methods include feature fusion and decision fusion. With the development of deep learning, Zadeh et al. (2017) develops the tensor fusion network that uses outer product to model interactions between intermodality and intra-modality. Based on this tensor

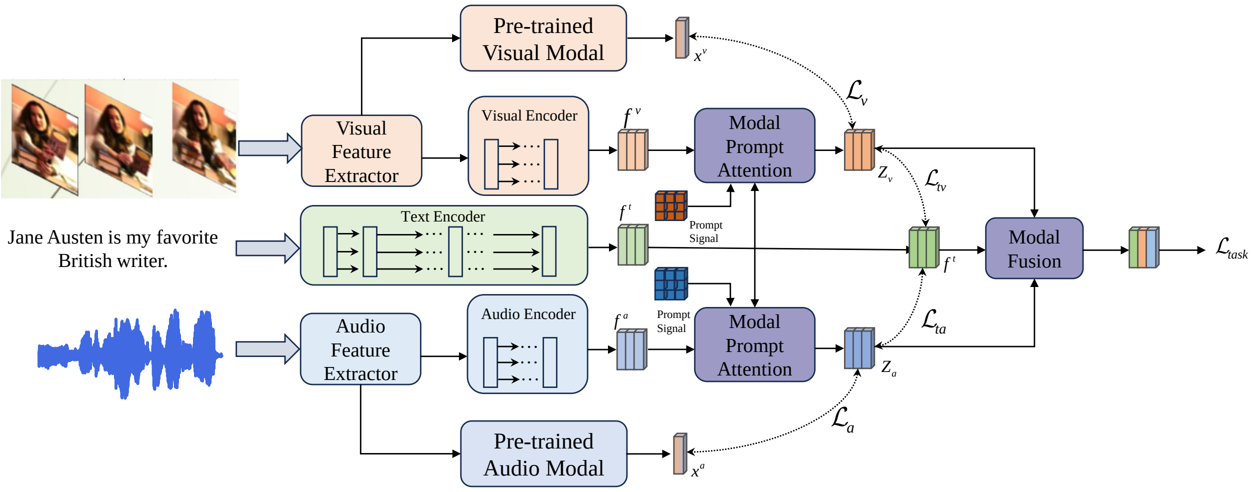

Figure 2: The structure of MFON framework. First, we still follow the previous work (Tsai et al., 2019) to obtain the modal features in the feature extraction part. Then we obtain modality-related prompt features first in order to optimize the modal features, and then use the prompt features and the original modal features as inputs for text-guided cross-modal information interaction in the MPA module. Then, in order to preserve the integrity of the modal features, intra-modal knowledge distillation is designed. Finally, inter-modal contrastive learning is employed to extract inter-modal correlation features, which enables further optimization of the entire network.

fusion network, Sun et al. (2020) uses deep canonical correlation analysis to reduce dimension of bimodal outer product. Considering that modal features can serve as vertices of graphs and modal interactions can be designed as edge relationships, fusion methods based on dynamic graphs are also explored (Zadeh et al., 2018b; Mai et al., 2020).

Modal Representation can be divided into three categories: (i) Translation-based solutions assume that the process of translating a modality into a target modality generates a representation that contains common information between modalities. Pham et al. (2018) uses hierarchical seq2seq translation. Specifically, they translate modality A into modality B, and then translate modality C using the intermediate representation generated by A and B. In practice, they find that modality A works best when it is text modality, and in fact, this translation order reuses text information. Pham et al. (2019) and Tang et al. (2021) perform cyclic translations between two modalities to learn consistency information.

(ii)Distribution-based solutions aim to learn a consistent representation of all modalities (Poklukar et al., 2022; Mai et al., 2022; Colombo et al., 2021; Li et al., 2023). However, some works suggest that learning specific representations of modalities is equally important. Wu et al. (2021) takes text as central modality and performs text-to-visual and text-to-acoustic translation. They think features with high attention weights during translation process belong to shared features, and features with high translation losses belong to modality-specific features.

(iii)Attention-based solutions model interaction processes of modalities using importance. Tsai et al. (2019) extends transformer (Vaswani et al., 2017) to multimodal domain, which utilizes bidirectional cross-modal attention to capture interactions between multimodal sequences across different time steps. Lv et al. (2021) introduces a message center to achieve cross-modal attention with each modality, solving the problem that attention can only be executed between two modalities.

3 Method

In this section, we describe in detail the modal feature optimization network to address the problem of under-optimization of modal features. Specifically, we design modal prompt attention to perform the first optimization of modal features (Sec. 3.2), and then we perform intra-modal knowledge distillation (Sec. 3.3) and inter-modal contrastive learning (Sec. 3.4) to further optimize the feature representations. Finally, we fuse the optimized features to do sentiment analysis. The detailed model structure is shown in Figure 2.

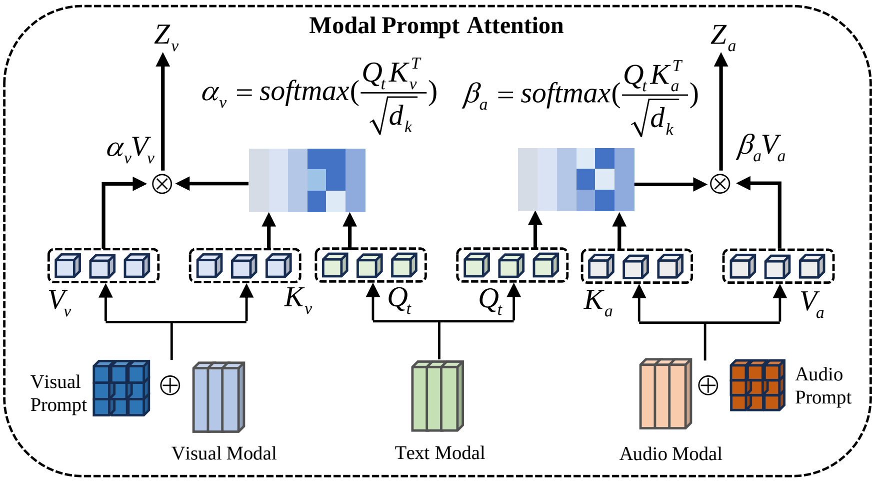

Figure 3: The structure of MPA module.

3.1 Problem State

Given a multimodal video clip that contains three modal information, denoted as $S = [u^a, u^t, u^v]$, where $u^t \in R^{T_t \times d_t}$, $u^a \in R^{T_a \times d_a}$, $u^v \in R^{T_v \times d_v}$ are text, acoustic and visual modalities, respectively. $T_m, m \in \{t, a, v\}$ denotes the length of the sequence for each modality, and d_m denotes the feature dimension of each modality. We take the three modalities as inputs to obtain the emotional polarity of the current utterance. To align with previous work methods (Yu et al., 2021; Wu et al., 2024), we also consider MSA as a regression task to ensure fair comparisons.

3.2 Modal Prompt Attention

As we all know, different modalities are collected by different sensors and the information contained in them is not necessarily the same. In order to give full play to the modal complementarity between multiple modalities, we need to first mine the amount of effective information in each modality according to the task. But the amount of information that each modality needs to contribute varies according to the task. We first need to determine the modal features that can be used as a guide among multiple modalities, i.e., the modality whose feature representation is already better among multiple modalities.

In MSA, through mutual information calculation, text modality contains the most effective information, and it can predict the other modal information to a certain extent. And we choose text modality as the guiding modality to perform the modality-guided attention for the information interaction between the text modality and the modality to be optimized.

First, we encode the original data input, for text we use BERT for encoding, for the other two modalities we follow the methodology of previous studies to encode, thus obtaining the feature representation of multiple modalities:

$$f^t = MLP(BERT(u^t)) \in R^{T_t \times d} \quad (1)$$

$$f^v = MLP(u^v) \in R^{T_v \times d} \quad (2)$$

$$f^a = MLP(u^a) \in R^{T_a \times d} \quad (3)$$

where d is the feature dimension after alignment, the feature extraction method for three modalities still follows the previous work (Tsai et al., 2019).

In the field of Natural Language Processing (NLP), prompt-based learning methods have made good progress, especially in pre-training models, where the pre-training model does not need to fine-tune the whole model after having prompt learning. Specifically, prompts are generally pre-positioned in the input, and the main purpose is to guide the model to pay more attention to the details of the features through the prompt signals, so as to get better result prediction. To this end, we design special prompt signals $p^m, m \in \{a, v\}$, for the modal features to be optimized, to guide the MPA to better guide the initial interactions between modalities, the module detail shown in Figure 3. We use visual modality as an example for modal optimization and information interaction. We fuse the prompt signal p^v first with our feature representation obtained via the encoder.

$$\hat{f}^v = f^v + p^v \quad (4)$$

After adding prompt signals, the model guides which modality should be further optimized for better information interaction. In MPA, $f^t, \hat{f}^v$ will be used as input to perform attention. We denote Query as $Q_t = f^t W_Q^t$, Key as $K_v = \hat{f}^v W_K^v$, and Value as $V_v = \hat{f}^v W_V^v$. To capture deep semantic information, we use multilayer attention to obtain preliminary results for multimodal interactions, computed as follows:

$$\begin{aligned} Z_v &= softmax(\frac{Q_t K_v^T}{\sqrt{d_k}})V_v \\ &= softmax(\frac{f^t W_Q^t (\hat{f}^v)^T {W_K^v}^T}{\sqrt{d_k}})\hat{f}^v W_V^v \end{aligned} \tag{5}$$

3.3 Intra-modal Knowledge Distillation

When using label prediction loss as a training objective, the optimization process for visual and acoustic modalities is susceptible to the influence of the guiding modality, text. To address this problem, we introduce an additional loss to supervise the non-text modalities, i.e., by performing intra-modal knowledge distillation to obtain more useful information. Specifically, distillation is used to add unimodal knowledge obtained from a pretrained model to the interaction process of multimodal features, thereby effectively enhancing visual and acoustic features.

First, we train two unimodal pretrained encoders, one for vision and the other for acoustic, by the label predict loss. The unimodal network consists of an MLP, an encoder, and an MLP predictor. For multimodal training, we use only the MLP with frozen parameters and the transform encoder. Then, during multimodal training, we input visual and acoustic features into parameter-frozen pretrained encoders to obtain features x^v and x^a, respectively. Finally, considering that f^m contains text information in addition to unimodal m, we choose the KL loss to allow the model to focus more on the information distribution than on specific numerical values. The KL divergence between the pretrained unimodal features and the text-guided modal features is calculated as follows:

$$\begin{aligned} \mathcal{L}_{va} &= \mathcal{L}_v + \mathcal{L}_a \\ &= KL(Z_v || x^v) + KL(Z_a || x^a) \end{aligned} \tag{6}$$

3.4 Inter-modal Contrastive Learning

Although relevant modal interactions have been performed by MPA, allowing under-optimized modal features to access relevant information from text modality with better feature representations, such interactions have not explored the number of effective complementary relationships. So the purpose of Inter-modal Contrastive Learning (ICL) is to learn the dynamic relationship between modalities through contrastive learning, so as to establish more discriminative boundaries in the feature space, which can further help the model to make a better prediction of emotional polarity. We therefore follow (van den Oord et al., 2018) and use the score function $Score(\cdot)$ with normalized prediction and truth vectors to measure the relationship between modalities:

$$\begin{aligned} &Score(f^t, Z_m) = exp(\overline{f}^t (\overline{Z}_m)^T) \\ &\overline{Z}_m = \frac{Z_m}{||Z_m||_2}, \overline{f}^t = \frac{f^t}{||f^t||_2} \end{aligned} \tag{7}$$

where $||\cdot||_2$ is the Euclidean norm. We treat all other representations of the modality in the same batch as negative samples, and thus calculate the loss between text modality and non-text features:

$$\mathcal{L}(Z_m, f^t) = -E_s\Big[log\frac{Score(Z_m, f_i^t)}{\sum_{f_j^t \in f^t} Score(Z_m, f_j^t)}\Big] \tag{8}$$

Finally, the loss function of ICL consists of the losses between the non-text features Z_m and text, respectively:

$$\mathcal{L}_{ICL} = \mathcal{L}_{tv} + \mathcal{L}_{ta} \tag{9}$$

3.5 Fusion & Predict

To do the final sentiment polarity prediction, we combine the optimized three modal features to obtain a multimodal fusion feature representation for each sentence, and then co-optimize the multimodal optimization network using the multiple losses mentioned above.

$$F_{fusion} = [Z_a; f^t; Z_v] \tag{10}$$

$$\hat{y} = predictor(F_{fusion}) \tag{11}$$

$$\mathcal{L}_{task} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{12}$$

Finally, minimizing $\mathcal{L}$ is our training goal:

$$\mathcal{L} = \alpha\mathcal{L}_{va} + \beta\mathcal{L}_{ICL} + \mathcal{L}_{task} \tag{13}$$

where α, β are hyperparameters that control each module's importance to the overall loss $\mathcal{L}$.

Dataset	#Train	#Valid	#Test	#All
MOSI	1284	229	686	2199
MOSEI	16326	1871	4659	22856
CH-SIMS	1368	456	457	2281

Table 1: The statistics of MOSI, MOSEI, and CH-SIMS.

4 Experiments

4.1 Datasets and Metrics

Datasets. We evaluate our model on three datasets that are popularly used to benchmark multimodal sentiment analysis[1]. The statistics are shown in Table 1. MOSI (Zadeh et al., 2016) extracts 2199 video clips from YouTube videos. Each clip includes visual frames, acoustic segments, and text transcripts, and is manually labeled with a -3 (strongly negative) to + 3 (strongly positive) emotion label. MOSEI (Zadeh et al., 2018b) is the larger version of MOSI which contains 22856 video clips, with more utterances, more topics, and more samples. These clips come from 5000 videos, with 1000 different speakers and 250 different issues. Each clip has an emotion label from -3 to 3. CH-SIMS (Liu et al., 2022) is a Chinese dataset containing 2281 video clips. Each clip not only provides a multimodal label from -1 (strongly negative) to 1 (strongly positive) but also provides a specific label for each modality, which also ranges from -1 to 1.

Metrics. Following existing works (Yu et al., 2023; Wu et al., 2024), we report results in two forms: regression and classification. For regression, we focus on Pearson correlation(Corr) and Mean Absolute Error (MAE). For classification, we calculate weighted F1 score (F1) and binary classification accuracy (Acc2). Specifically, we calculate Acc2 and F1 in negative/positive (exclude zero) and negative/non-negative (include zero) and accuracy in seven-class classification (Acc7) on MOSI and MOSEI. The larger value of all metrics except for MAE represents better results.

4.2 Baseline

We evaluate our model with some state-of-the-art models in MSA:

TFN (Zadeh et al., 2017) uses tensor outer products to capture uni-, bi-, and tri-modal interactions and perform feature fusion.

LMF (Liu et al., 2018) addresses low computational efficiency for high-dimensional tensors by low-rank tensor decomposition.

Mult (Tsai et al., 2019) performs cross-modal attention in each of the two modalities to capture alignment relationships.

MISA (Hazarika et al., 2020) decomposes modality features into shared and private representations and supervises their learning by various losses.

Self-MM (Yu et al., 2021) uses a self-supervised strategy to construct a label generation module and uses generated labels to promote extraction of modality-specific information.

MMIM (Han et al., 2021) hierarchically maximizes mutual information within unimodal features and between multimodal fusion features and unimodal features to obtain emotion-related information.

FDMER (Yang et al., 2022) introduces a modality discriminator for modality-invariant and -shared features, guiding parameter learning of a general encoder and a private encoder in an adversarial manner.

AMML (Sun et al., 2023) obtains better unimodal representation via meta-training on unimodal tasks and fusion unimodal representation via adding distribution transformation layers.

ConKI (Yu et al., 2023) uses the adapter approach to inject specific knowledge into each modality and combines it with general knowledge learned by the model to promote MSA effects.

HyDiscGAN (Wu et al., 2024) builds acoustic and visual generators based on shareable de-identified text data to generate multimodal features, and regulates the learning process through discriminators.

4.3 Implementation Details

For text encoder, we use bert-base-chinese[2] on CH-SIMS and bert-base-uncase[3] on MOSI and MOSEI. We use COVAREP (Degottex et al., 2014) and Facet[4] to extract acoustic and visual expression features respectively.

All experiments are conducted on a single NVIDIA RTX 4090 GPU. During experiments, we use the Adam optimizer and set batch size of 128. To determine the best-performing hyperparameters, we conduct 100 random grid searches and save the

[1] These datasets can be found from https://github.com/thuiar/Self-MM

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

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

[4] https://imotions.com/

Models	MOSI					MOSEI				
	Acc2↑	F1↑	Acc7↑	Corr↑	MAE↓	Acc2↑	F1↑	Acc7↑	Corr↑	MAE↓
TFN	-/80.8	-/80.7	34.9	0.698	0.901	-/82.5	-/82.1	50.2	0.700	0.593
LMF	-/82.5	-/82.4	33.2	0.695	0.917	-/82.0	-/82.1	48.0	0.677	0.623
Mult	81.5/84.1	80.6/83.9	-	0.711	0.861	-/82.5	-/82.3	-	0.703	0.580
MISA	80.79/82.10	80.77/82.03	-	0.764	0.804	82.59/84.23	82.67/83.97	-	0.724	0.568
Self-MM	82.54/84.77	82.68/84.91	45.79	0.795	0.712	82.68/84.96	82.95/84.93	53.46	0.767	0.529
MMIM	84.14/86.06	84.00/85.98	46.65	0.800	0.700	82.24/85.97	82.66/85.94	54.24	0.772	0.526
FDMER	-/84.6	-/84.7	44.1	0.788	0.724	-/86.1	-/85.8	54.1	0.773	0.536
AMML	-/84.9	-/84.8	46.3	0.792	0.723	-/85.3	-/85.2	52.4	0.776	0.614
ConKI	84.37/86.13	84.33/86.13	**48.43**	**0.816**	**0.681**	**82.73**/86.25	83.08/86.15	54.25	**0.782**	0.529
HyDiscGAN	84.1/86.7	83.7/86.3	43.2	0.782	0.749	81.9/86.3	82.1/86.2	**54.4**	0.761	0.533
MFON	**84.84/86.89**	**84.75/86.86**	44.90	0.797	0.725	82.70/**86.32**	**83.13/86.29**	53.72	0.780	**0.528**

Table 2: Results on MOSI and MOSEI. In Acc2 and F1, the left of "/" is "negative/non-negative" and the right means "negative/positive". Results of FDMER,AMML, HyDiscGAN come from Yang et al. (2022), Sun et al. (2023), Wu et al. (2024) respectively, and results of other baseline come from Yu et al. (2023).

Models	Acc2↑	F1↑	Corr↑	MAE↓
TFN	75.27	75.56	0.496	0.488
LMF	75.36	75.78	0.502	0.487
Mult	75.62	75.84	0.504	0.485
MISA	75.49	75.85	0.542	0.472
Self-MM	77.37	77.54	0.535	0.458
MMIM	69.37	58.00	-	0.607
ConKI	77.94	78.17	0.542	0.454
MFON	**78.56**	**78.51**	**0.594**	**0.420**

Table 3: Results on CH-SIMS. Baseline results come from Yu et al. (2023).

hyperparameter settings that achieve the best results. After the grid search, each model is retrained five times with the same optimal hyperparameters, and we save the average result as the final result.

4.4 Results

Result comparisons of all methods are reported in Table 2 and Table 3. We find that our model achieves better or comparable results to the state-of-the-art models, which indicates the effectiveness of our approach in multimodal sentiment analysis.

Specifically, MFON is significantly better than other models in Acc2 and F1 on MOSI. Corr and MAE indicators are in a sub-optimal position, and Acc7 result is also comparable. On MOSEI dataset, MFON outperforms the other models in Acc2, F1, and MAE metrics, and the remaining metrics rank second. On CH-SIMS, MFON outperforms other models in all metrics, with particularly significant improvements in Corr and MAE.

We notice that our model has a significant improvement in binary classification, while the seven-class classification is not so effective. We speculate that during multimodal feature fusion, using text modality to guide the extraction of shared emotional features from visual and acoustic modalities may, to some extent, weaken the performance in the seven-class classification. Compared with MOSI, MOSEI contains more training data and has a better Acc7 result. We propose that the inter-modal distillation guidance provides more information related to modality-specific emotion. The improvement in MAE and Corr also shows that our model learns more optimized visual and acoustic features.

4.5 Ablation Study

To show the benefits of our proposed module, we carry out some ablation experiments on three datasets. Results under different ablation settings are reported in Table 4 and Table 5.

4.5.1 Unimodal Study

We use unimodal data to conduct sentiment analysis and results are reported in the first three rows of Table 4 and Table 5. Compared to multimodal results, the performance drops sharply when using visual or acoustic while the performance reduction is insignificant when using text modality.

This result supports our hypothesis that text modality contains more emotional information and is easier to learn good representations. Visual and acoustic modalities contain less information and are more challenging for representation learning and they are in an under-optimized state. Therefore, we need to fully utilize the information from text modality and design reasonable methods to extract more emotional information from visual

Models	MOSI					MOSEI				
	Acc2↑	F1↑	Acc7↑	Corr↑	MAE↓	Acc2↑	F1↑	Acc7↑	Corr↑	MAE↓
T	82.36/84.45	82.27/84.43	43.21	0.790	0.734	82.27/85.83	82.75/85.83	52.21	0.763	0.534
V	59.48/59.42	59.02/59.09	18.95	0.211	1.383	65.23/65.63	65.74/64.84	40.48	0.272	0.820
A	51.89/49.47	50.76/48.51	16.03	0.143	1.445	67.31/65.49	64.69/61.5	41.15	0.205	0.827
w/o $\mathcal{L}_{va}$	82.51/84.62	82.44/84.58	43.73	0.787	0.752	82.17/86.16	82.45/86.07	49.77	0.773	0.565
w/o $\mathcal{L}_{ICL}$	82.36/84.45	82.33/84.49	41.25	0.793	0.746	82.76/86.09	83.21/86.12	52.72	0.778	0.530
w/o $\mathcal{L}_{ICL}\mathcal{L}_{va}$	81.92/83.84	81.84/83.83	42.75	0.785	0.751	82.22/85.94	**83.47**/85.81	48.94	0.773	0.571
w/o prompt	81.49/83.23	81.39/83.2	43.10	0.789	0.734	81.22/86.05	81.84/86.09	53.61	0.769	0.536
Visual as query	82.65/84.76	82.59/84.76	41.63	0.794	0.730	82.72/85.28	83.06/85.20	52.61	0.756	0.549
Acoustic as query	83.53/85.52	83.47/85.51	43.15	0.792	0.737	81.63/85.31	82.14/85.31	51.32	0.756	0.559
MFON	**84.84/86.89**	**84.75/86.86**	**44.90**	**0.797**	**0.725**	**82.70/86.32**	83.13/**86.29**	**53.72**	**0.780**	**0.528**

Table 4: Ablation experiments on MOSI and MOSEI.

Models	Acc2↑	F1↑	Corr↑	MAE↓
T	74.40	74.85	0.534	0.461
V	67.61	57.65	0.238	0.618
A	67.83	54.83	0.093	0.543
w/o $\mathcal{L}_{va}$	**78.77**	78.31	0.519	0.488
w/o $\mathcal{L}_{ICL}$	75.71	76.21	0.573	0.463
w/o $\mathcal{L}_{ICL}\mathcal{L}_{va}$	77.24	77.15	0.521	0.484
w/o prompt	77.68	77.89	0.575	0.455
Visual as query	75.93	75.93	0.568	0.429
Acoustic as query	75.93	76.19	0.559	0.466
MFON	78.56	**78.51**	**0.594**	**0.420**

Table 5: Ablation experiments on CH-SIMS.

and acoustic modalities in multimodal sentiment analysis tasks.

4.5.2 Effect of Multiple Loss Learning

Lines 4, 5, and 6 of Table 4 and Table 5 present experimental results when removing intra-modal distillation loss, inter-modal contrastive loss, and both removed. The model performance decreases when the intra-modal distillation loss is removed indicating that the model learns useful visual and acoustic features from the pretrained model. When the inter-modal contrastive loss is removed, performance drops due to the lack of implicit cross-modal interactions. These results suggest that minimizing both intra-modal and inter-modal losses helps the model learn emotion-related features and ensures continuous feature optimization.

Additionally, the performance metrics on MOSI that show a larger decline are Acc2 and F1, while on MOSEI and CH-SIMS, the larger decline is Acc7. The performance degradation of inter-modal contrastive loss is more pronounced compared to intra-modal loss on Acc7 with multiple data, which suggests that we have designed intra-modal contrastive learning to capture relevant sentiment information between modalities, and thus effectively optimize the informative representation of the fused features based on similar features learned by contrastive learning, which can lead to better results in terms of multiclassification accuracy at a finer level of granularity.

4.5.3 MPA Analysis

The MPA module contains two key components in its implementation: prompt signal and text modality guide.

We first investigate prompt signal's impact(w/o prompt). We note that removing the prompt signal leads to a significant decrease in model performance. This indicates that the added prompt information guides under-optimized visual and acoustic modalities to pay more attention to emotion-related information. Removing this component may cause the model to focus on under-optimized features when fusing text and non-text modalities, thus causing irrelevant noise to affect the final performance.

When we use non-text modality as the query (Visual as query, Acoustic as query), model performance degrades. This phenomenon demonstrates the effectiveness of using text modality to guide non-text modality learning. It also shows the emotional representation of acoustic and visual modalities is in an under-optimized state. They do not provide enough emotional information and may introduce irrelevant noise when used as the query. In contrast, text modality contains clear information which effectively guides visual and acoustic modalities to focus on emotion-related information.

4.5.4 Modal Representations Analysis

In Figure 4, we use t-SNE (Hinton and Roweis, 2002) to visualize the visual modality feature representations extracted by the model in 3D space on the MOSI test sets. In the clustering results for the

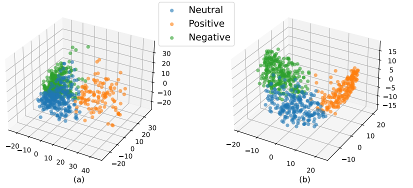

Figure 4: Visualization of the visual modality's feature representation in 3D space using t-SNE on the MOSI test sets. (a) Three-class emotion clustering results of the visual feature extracted using the model that removes modal prompt attention modules, intra-modal distillation loss modules, and inter-modal contrastive loss modules. (b) Three-class emotion clustering results of the visual feature representation extracted using MFON which includes all modules.

three emotion classes (neutral, positive, and negative), the visual modality features extracted by the model without modal prompt attention modules, intra-modal distillation loss modules, and inter-modal contrastive loss modules show significant overlap between the features of neutral and negative emotions. This indicates that the features from visual and acoustic modalities are not fully extracted. However, when the complete set of modules is added to the MFON model, the performance of visual emotion classification improves significantly, with the three emotion categories becoming clearly separable and the overlapping regions reduced. This demonstrates that MFON extracts emotion-related features from the visual modality, addressing the under-optimization issue in the acoustic and visual modalities.

5 Conclusion

In this paper, we propose MFON to address the problem of under-optimized emotional representation of visual and acoustic modalities in MSA. We leverage a modal prompt attention mechanism to guide the model to focus on under-optimized modalities and facilitate initial information aggregation across modalities. Then we design intra-modal distillation and inter-modal contrastive learning for under-optimized modalities. Experimental results prove our approach achieves comparable to state-of-the-art models.

Limitations

Our approach explores the under-optimization of visual and acoustic in MSA, but it still has some limitations. Specifically, our method considers that text is the dominant modality while visual and acoustic are in an under-optimized state, but does not strictly consider which modality is most suppressed by the dominant modality. In future work, we will investigate the problem of different degrees of under-optimization in visual and acoustic.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant No. 62276110, No. 62172039 and in part by the fund of Joint Laboratory of HUST and Pingan Property & Casualty Research (HPL). The authors would also like to thank the anonymous reviewers for their comments on improving the quality of this paper.

References

Shyamal Buch, Cristóbal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei, and Juan Carlos Niebles. 2022. Revisiting the "video" in video-language understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2907–2917.

Maria Chiara Caschera, Fernando Ferri, and Patrizia Grifoni. 2016. Sentiment analysis from textual to multimodal features in digital environments. In *Proceedings of the 8th International Conference on Management of Digital EcoSystems*, pages 137–144.

Pierre Colombo, Emile Chapuis, Matthieu Labeau, and Chloé Clavel. 2021. Improving multimodal fusion via mutual dependency maximisation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 231–245.

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

Junchen Fu, Xuri Ge, Xin Xin, Alexandros Karatzoglou, Ioannis Arapakis, Jie Wang, and Joemon M. Jose. 2024. IISAN: efficiently adapting multimodal representation for sequential recommendation with decoupled PEFT. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024*, pages 687–697. ACM.

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

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

Geoffrey E Hinton and Sam Roweis. 2002. Stochastic neighbor embedding. *Advances in neural information processing systems*, 15.

Fanheng Kong, Peidong Wang, Shi Feng, Daling Wang, and Yifei Zhang. 2024. TIGER: A unified generative model framework for multimodal dialogue response generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 16135–16141. ELRA and ICCL.

Dongyuan Li, Yusong Wang, Kotaro Funakoshi, and Manabu Okumura. 2023. Joyful: Joint modality fusion and graph contrastive learning for multimoda emotion recognition. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16051–16069.

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, 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.

Fengmao Lv, Xiang Chen, Yanyong Huang, Lixin Duan, and Guosheng Lin. 2021. Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 2554–2562.

Sijie Mai, Haifeng Hu, and Songlong Xing. 2020. Modality to modality translation: An adversarial representation learning and graph fusion network for multimodal fusion. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020*, pages 164–172.

Sijie Mai, Ya Sun, and Haifeng Hu. 2022. Curriculum learning meets weakly supervised multimodal correlation learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3191–3203.

Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. 2019. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019*, pages 6892–6899.

Hai Pham, Thomas Manzini, Paul Pu Liang, and Barnabás Poczós. 2018. Seq2Seq2Sentiment: Multimodal sequence to sequence models for sentiment analysis. In *Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML)*, pages 53–63.

Petra Poklukar, Miguel Vasco, Hang Yin, Francisco S. Melo, Ana Paiva, and Danica Kragic. 2022. Geometric multimodal contrastive representation learning. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162, pages 17782–17800.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. 2023. Timechat: A time-sensitive multimodal large language model for long video understanding. *CoRR*, abs/2312.02051.

Qingfeng Sun, Yujing Wang, Can Xu, Kai Zheng, Yaming Yang, Huang Hu, Fei Xu, Jessica Zhang, Xiubo Geng, and Daxin Jiang. 2022. Multimodal dialogue response generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2854–2866.

Ya Sun, Sijie Mai, and Haifeng Hu. 2023. Learning to learn better unimodal representations via adaptive multimodal meta-learning. *IEEE Trans. Affect. Comput.*, 14(3):2209–2223.

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

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

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

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748.

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.

Wei Wei, Chao Huang, Lianghao Xia, and Chuxu Zhang. 2023. Multi-modal self-supervised learning for recommendation. In *Proceedings of the ACM Web Conference 2023*, pages 790–800.

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.

Zhuojia Wu, Qi Zhang, Duoqian Miao, Kun Yi, Wei Fan, and Liang Hu. 2024. Hydiscgan: A hybrid distributed cgan for audio-visual privacy preservation in multimodal sentiment analysis. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 6550–6558. International Joint Conferences on Artificial Intelligence Organization. Main Track.

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*, page 1642–1651.

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. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114.

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. 2018b. 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.

Haoyu Zhang, Yu Wang, Guanghao Yin, Kejun Liu, Yuanyuan Liu, and Tianshu Yu. 2023. Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 756–767.

Shanshan Zhong, Zhongzhan Huang, Daifeng Li, Wushao Wen, Jinghui Qin, and Liang Lin. 2024. Mirror gradient: Towards robust multimodal recommender systems via exploring flat local minima. In *Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, May 13-17, 2024*, pages 3700–3711. ACM.