Uncertainty-aware Gradient Modulation and Feature Masking for Multimodal Sentiment Analysis

Yuxian Wu, Chengji Wang (✉), Jingzhe Li,
Wenjing Zhang, and Xingpeng Jiang (✉)

Hubei Provincial Key Laboratory of Artificial Intelligence and Smart Learning,
National Language Resources Monitoring and Research Center for Network Media,
School of Computer Science, Central China Normal University, Wuhan, China
{wyx_natural,lijingz,Dongri_z}@mails.ccnu.edu.cn,
{wcj,xpjinag}@ccnu.edu.cn

Abstract. Multimodal Sentiment Analysis (MSA) aims to analyze the attitudes of speakers from video content. Previous methods focus on exploring consistent and cross-modal sentiment representations by multimodal interactions, they treat each modality equally. However, modalities are incomplete and uncertain, *e.g.*, noise, semantic ambiguity. Modality with low uncertainty contributes more to the final loss, suppressing the optimization of modalities with high uncertainties. To address this problem, we propose a new Uncertainty-aware Gradient modulation and Feature masking model (UGF) for MSA, which aims to assist optimization of modalities with high uncertainty. We propose a novel modal uncertainty estimation method, which considers both the intra- and intermodality consistency to estimate modal uncertainty. We improve the model by two aspects: First, we design a dynamic gradient modulation module (DGM) to amend the optimization process of each modality, it dynamically modulates the gradients of modality encoders according to their uncertainties. Second, we propsoe a uncertainty guided feature masking (UFM), it adaptively adds noise to the deterministic modality, making model pay more attention on uncertain modalities. We conducted extensive experiments on three popular datasets, *e.g.*, MOSI, MOSEI and CH-SIMS. Experimental results show that our propose UGF achieves competitive results, the ablation studies demonstrate the effectiveness of the proposed components.

Keywords: Multimodal sentiment analysis · Modal Uncertainty · Gradient Modulation · Feature Masking.

1 Introduction

Multimodal Sentiment Analysis (MSA) aims to interpret attitudes and opinions expressed in video content [20,2,18]. The proliferation of smart devices has led to an increase in personal opinions shared via video platforms. These videos incorporate diverse sources of data (text, visuals, and audio) that enrich sentiment analysis. MSA plays a vital role in understanding public attitudes and behaviors with significant applications across healthcare [1], marketing management [13], and human computer interaction [12].

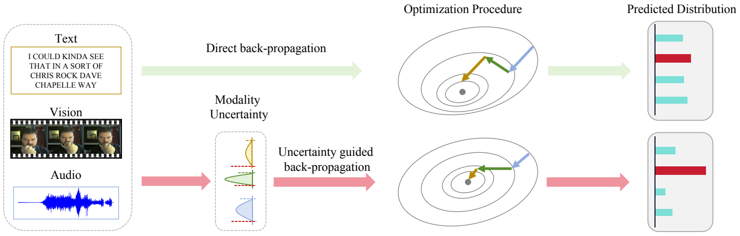

Fig. 1: Comparison between modal uncertainty guided backpropagation and direct backpropagation.

MSA is a very challenging problem. Firstly, the heterogeneous gap between modalities impedes multimodal information fusion. Secondly, data noise and semantic ambiguity make modalities to be uncertain. It is hard to fully exploit the potential of multiple modalities. Additionally, data noise and semantic ambiguity introduce uncertainty across these modalities, complicating the full exploitation of their combined potential. Most of the MSA approaches rely on intra- and inter-modal interactions for exploring consistent and cross-modal sentiment information. Self-MM [20] utilizes self-supervised learning to exploit consistent sentiment information from different modalities. MISA [2] constructs a common embedding space to obtain cross-modal features. ConFEDE [18] performs intra- and inter-modal contrastive learning to align multimodal features. These methods ignore the modality uncertainty and treat all modalities equally. However, data noise and semantic ambiguity lead to the uncertainties of unimodal encoders. As is shown in Fig. 1, multimodal information are inconsistent, various modalities converge at different rates. The contributions of modalities are unbalanced, one modality with low uncertainty contributes more to the final loss, which suppresses the optimization of high uncertain modalities [11].

To address above issue, we introduce a new Uncertainty-ware Gradient modulation and Feature masking model (UGF) for multimodal sentiment analysis. Under the guidance of modality uncertainties, the encoders of uncertain modalities will be more optimized. We propose a novel uncertainty estimation method to dynamically estimates modality uncertainties for varying quality data, it simultaneously considers the consistencies of predicted distributions in both intra-modality and inter-modality. Firstly, based on the modality uncertainties, we generate a scaling factor for each modality, a dynamical gradient modulation module is designed to control the optimization of each modality, the uncertain modalities will have bigger gradients. Secondly, the model will automatically depend on the features of modality with low uncertainty, lack of attention to other modal features. We propose the uncertainty guided feature masking (UFM) to restrain the contribution of low uncertain modality. In training, UFP adaptively adds interference to features based on according to the modal uncertainties, increasing the importance of features from uncertain modalities. Thus, the uncertain modalities will be optimized better. The main contributions of our work can be summarized as follows:

- We propose a novel modality uncertainty estimation method for regression problem, it utilizes the normalized absolute distance between predictions and labels to estimates the modality uncertainties.
- We design a dynamically gradient modulation module to magnify the intensity of back-propagation signals to help uncertain modalities optimization.
- We propose a uncertainty guided feature masking method, it applies a larger disturbance on the modality with low uncertainty, forcing model pays more attention on the features from uncertain modalities.
- Experimental results on three popular datasets show that UGF outperforms baselines in most metrics, and the extensive experiments demonstrate the effectiveness of the proposed method.

2 Related Work

2.1 Multimodal Sentiment Analysis

MSA mostly focused on multimodal fusion and representation learning[15]. A notable advancement was made by Zadeh et al. [21] , who introduced the Tensor Fusion Network (TFN). This model enhances MSA performance by leveraging complementary information from different modalities. However, this method ignores interactions and mutual influences between these modalities. To address these issues, Huang et al. [3] developed the Deep Multimodal Attentive Fusion (DMAF), which employs both unimodal and multimodal attention mechanisms to emphasize discriminative features and internal correlations.

Subsequent studies identified the invariance and specificity properties of modalities. Hazarika et al. [2] presented the Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis (MISA), a model that effectively bridges modality gaps by projecting each modality into respective invariant and specific subspaces, thus refining the fusion process and boosting performance in sentiment classification and humor detection.

The modality-invariant and specific representations mentioned in the previous research, will lead to the conflict between consistency and differentiation before the Modality. To address this challenge, Yu et al. [20] introduce an innovative self-supervised method (Self-MM). They propose a label generation module that automatically generates unimodal supervisions, enabling the learning of both consistency and differentiation among modalities. Additionally, contrastive learning has proven effective in distinguishing modal information for different sentiments. Yang et al. [18] proposed a framework that utilizes contrastive representation learning and feature decomposition (ConFEDE) to skillfully isolate and learn modality-invariant and modality-specific features across text, video, and audio modalities. Li et al. [7] further enhance this by introducing MDSE, which combines private feature learning and modality-agnostic contrastive loss for robust sentiment classification.

2.2 Modality Uncertainty

Modal uncertainty in multimodal learning presents significant challenges due to the integration of diverse and ambiguous information from various sources like text, images, and audio. This uncertainty arises from inconsistencies in the amount and type of information within and between modalities, which can adversely affect model performance. To tackle these issues, researchers have developed various strategies aimed at resolving the uncertainties associated with information inconsistency. Ji et al. [4] explore the inherent ambiguities in multimodal semantic understanding and highlight the importance of models capable of managing both intra-modal and inter-modal uncertainties. They emphasize the need for sophisticated techniques that can adapt to the varying degrees of uncertainty within each modality. Similarly, Wang et al. [16] introduce an Uncertainty-aware Multi-modal Learner that incorporates Cross-modal Random Network Prediction. This method dynamically adjusts the integration process based on the uncertainty levels of each modality, helping to enhance the robustness and accuracy of multimodal learning. Kim et al. [6] develop an uncertainty-aware framework that incorporates region of interest and predictive uncertainties to improve alignment and modality integration, enhancing detection performance on multispectral datasets in multispectral pedestrian detection.

Despite these advancements, few methods consider modal uncertainty in MSA. Many MSA methods treat different modalities equally, while ignoring the uncertainty. To address the uncertainty issue in MSA, we propose a novel modal uncertainty estimation method, combined with gradient modulation and feature perturbation strategies.

3 Methodology

We propose a novel uncertainty-aware gradient modulation and feature masking model designed for multimodal sentiment analysis. Figure 2 illustrates the proposed pipeline. Our approach involves estimating uncertainty for each modality. To control the optimization process effectively, we design a dynamic gradient modulation module. Additionally, we develop an uncertainty-aware feature masking module that compels the model to prioritize the optimization of more uncertain modalities.

3.1 Problem Formulation

Given the features of the i-th sample $M_i = \{m_i^s \in \mathbb{R}^{T_s \times d_s}\}$, where $s \in \{t, v, a\}$ represents the modalities: text (t), visual (v), and audio (a), multimodal sentiment analysis aims to predict a sentiment score that reflects affect polarity (Positive or Negative) and intensity. m^s denotes the feature set for modality s,

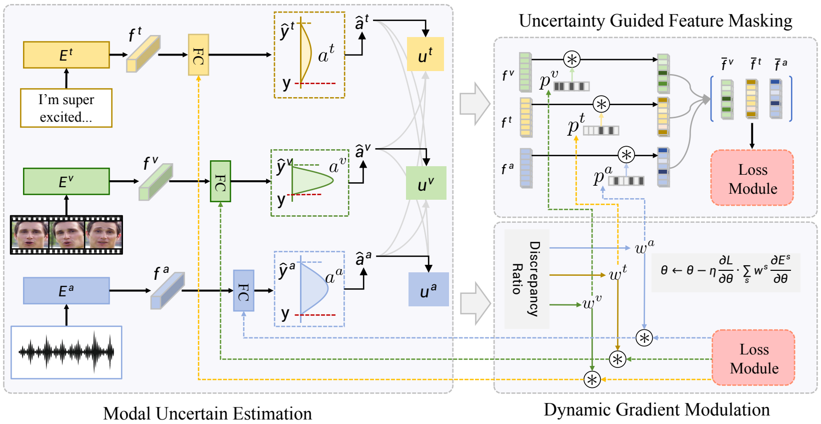

Fig. 2: Overview of our proposed UGF model. UGF aims to address the under-optimal problem in MSA. We propose a new modal uncertainty estimation method to evaluate the uncertainty of each modality. A dynamic gradient modulation module (DGM) dynamically adjusts the intensity of back-propagation signals to control the training process. The uncertainty aware feature masking (UFM) forcing model to optimize the uncertain modalities.

and y_i is the corresponding labeled sentiment score. The multimodal sentiment analysis pipeline is formulated as follows:

$$\begin{aligned} f_i^s &= E^s(m_i^s, \theta^s) \\ \hat{y}_i^s &= FC(f_i^s; \theta^s) \end{aligned} \tag{1}$$

where, E^s is the unimodal feature extractor, FC is a fully connected layer for predict unimodal sentiment score $\hat{y}_i^s$. θ^s are parameters. The multimodal sentiment score is computed by:

$$\hat{y}_i = FC([f_i^t, f_i^v, f_i^a]; \theta^m) \tag{2}$$

where $\hat{y}_i$ is the predict multimodal sentiment score, $[\dots]$ stands for feature concatenation. θ^m are the parameters. For simplicity, all model parameters are uniformly represented as θ. We assume that the unimodal and multimodal models share the same sentiment score. The parameters θ of the overall model are optimized as follows:

$$\begin{aligned} L &= \frac{1}{N}\sum_i^N (\|\hat{y}_i - y_i\|_2^2 + \sum_s \|\hat{y}_i^s - y_i\|_2^2) \\ \theta &\longleftarrow \theta - \eta \frac{\partial L}{\partial \theta} \cdot \sum_s \frac{\partial E^s}{\partial \theta} \end{aligned} \tag{3}$$

where, N is the batch size, η is the learning rate.

3.2 Modality Uncertainty Estimation

In this section, we introduce our proposed method for estimating modal uncertainty. Uncertainty is a pivotal concern in machine learning, owing to the inherent randomness in the processes involved. Multimodal sentiment analysis is particularly susceptible to the modal uncertainty problem due to the sparsity of affective information, noise in the data, and semantic ambiguities.

To quantify the confidence of our predictions, we measure the absolute distance between the predicted scores and the target sentiment scores. Initially, we compute these distances for predictions from individual modalities against the annotated sentiment scores. Subsequently, we calculate the statistical average of these distances for each modality, which provides a measure of the uncertainty associated with each.

$$a^s = \frac{1}{N}\sum_{i}^{N}\frac{1}{\alpha + tanh(|\hat{y}_i^s - y_i|)} \tag{4}$$

where α is a hyper-parameter, $tanh()$ normalize the result to between 0 and 1. a^s reflects the amount of information attributed to modality s. A bigger a^s indicates modality s has a lower uncertainty level.

However, a^s is susceptible to sample selection. In here, we adopt the moving average strategy,

$$\hat{a}_t^s = \hat{a}_{t-1}^s\frac{t-1}{t} + \frac{a_t^s}{t} \tag{5}$$

$\hat{a}_t^s$ is the statistical average value of modality s at $t-$th iteration, a_t^s represents the a^s at $t-$th iteration. It reveals the overall uncertain level of modality s in the training data. The higher the overall uncertainty level of modality s, the smaller the $\hat{a}_t^s$.

Actually, the uncertainty levels of modalities are relative. In here, we compute relative modality uncertainties by considering the inter-modality differences of predicted distributions,

$$u_t^s = exp(\frac{1}{K-1}\sum_{s'\neq s;s\in\{t,v,a\}}(\hat{a}_t^s - \hat{a}_t^{s^{'}})) \tag{6}$$

The u_t^s stands for the uncertainty of modality s. The larger the u_t^s, the lower the uncertainty of modality s.

3.3 Dynamic Gradient Modulation

Intuitively, if the uncertainty of modality s is lower than modality $s^{'}$, the benefits derived from optimizing modality s are less than those from optimizing modality $s^{'}$. However, the optimization process in multimodal discriminative models is typically dominated by the modality with lower uncertainty [11]. This dominant, less uncertain modality contributes to a lower joint loss. However, it also limits the gradient propagated to other modalities. As a result, the higher uncertainty modality may end up in an under-optimized state.

In order to boost the performance of uncertain modalities, we design a dynamic gradient modulation module. It adjust the intensity of back-propagation signals to control the training process,

$$\theta_{t+1} \longleftarrow \theta_t - \eta \frac{\partial L}{\partial \theta} \cdot \sum_s w_t^s \frac{\partial E^s}{\partial \theta} \tag{7}$$

where θ_t and θ_{t+1} denote the parameters before and after updating at iteration t. The w_t^s is the gradient modulation factor of modality s, where,

$$w_t^s = exp(u_t^s - h_t^s) \tag{8}$$

and

$$h_t^s = exp(\frac{1}{K-1} \sum_{s' \neq s; s \in \{t,v,a\}} (a_t^s - a_t^{s^{'}})) \tag{9}$$

h_t^s reflects the uncertain level of modality s in current batch. u_t^s stands for the average uncertain level of modality s in training data. We compute the discrepancy between h_t^s and u_t^s in Eq. 8. When $h_t^s < u_t^s$, the uncertainty of modality s in current batch is higher than the overall data, we should pay more attention on modality s. Inversely, when $h_t^s > u_t^s$, the uncertainty of modality s in current batch is lower than the overall data, we should suppress its update signal.

3.4 Uncertainty Guided Feature Masking

Modalities with lower uncertainty significantly reduce the overall loss and dominate the optimization process, potentially leading to the underutilization of uncertain modalities. During training, the model tends to rely predominantly on features from the modality exhibiting low uncertainty, thereby neglecting the optimization of features from modalities characterized by higher uncertainty.

To mitigate the model's reliance on a single modality, we introduce an uncertainty guided feature masking module. This module probabilistically discards features from the dominant modality, thereby compelling the model to focus more on optimizing features from uncertain modalities.

As discussed above, the dominant modality s has a small weight w_t^s, a uncertain modality s has a large weight w_t^s. Consequently, a modality with a smaller weight w_t^s is more likely to be discarded, whereas an uncertain modality is preferentially retained. The probability of discarding features from a modality is calculated as follows::

$$p^s = \frac{1 - w_t^s}{\sum_s (1 - w_t^s)} \tag{10}$$

where, p^s is the probability of discarding modality s'features.

We apply dropout to perform feature masking,

$$\overline{f}^s = Dropout(f^s, p^s) \tag{11}$$

where, f^s is the original features, $\overline{f}^s$ is the masked features, $s \in \{t, a, v\}$.

During training, as same as Eq. 1 and Eq. 2, we use the masked features to predict the unimodal and multimodal sentiment scores. This masking module is employed exclusively during the training phase.For testing, we revert to using the original multimodal features f^s and Eq. 2 to obtain multimodal sentiment score.

4 Experiements

4.1 Datasets and Experiment Setting

Datasets. We mainly evaluate UGF on three popular multimodal sentiment analysis datasets, *e.g.*, MOSI[22], MOSEI[23], and SIMS[19].

1. **MOSI.**The CMU-MOSI (Multimodal Corpus of Sentiment Intensity) dataset is a public resource for multimodal sentiment analysis, including three modalities (text, visual, audio), consisting of 2,199 video clips from YouTube that cover comments by 89 speakers on 89 movies. Each clip is annotated with sentiment score from -3 (strong negative) to 3 (strong positive).
2. **MOSEI.**The CMU-MOSEI (Multimodal Opinion Sentiment and Emotion Intensity) dataset is an extension of the CMU-MOSI dataset, containing 23,453 video clips from over 1,000 speakers commenting on more than 250 movies. As same as CMU-MOSI, MOSEI annotates the sentiment scores ranging from -3 to 3.
3. **SIMS.**The CHI-SIMS dataset is designed for Chinese multimodal sentiment analysis, including text, audio, and visual modalities, featuring 340 video clips from 170 human-computer interaction tasks. Each clip is annotated with a sentiment score from -1 (strong negative) to 1 (strong positive).

Implementation Details. In our work, we delineated our methodology by first initiating pre-training encoders for the individual modalities. The optimization of the models was conducted using the AdamW algorithm. The parameters for the pre-training phase included a learning rate of 1×10^{-4} and a weight decay of 0.3, while the multimodal fusion phase employed a learning rate of 2×10^{-4}. Experimental setups were standardized with a batch size of 32. All computations were performed on an NVIDIA GeForce RTX 4090 GPU.

Evaluation Metrics. Following previous works [2,18], we evaluate the performance of UGF by following metrics: Mean Absolute Error (MAE$\downarrow$), Pearson correlation coefficient (Corr$\uparrow$), 7-class accuracy (Acc-7$\uparrow$), binary accuracy (Acc-2$\uparrow$) and F1 score (F1$\uparrow$).

4.2 Comparison with the state-of-the-art

We conduct comprehensive experiments on the CMU-MOSI, CMU-MOSEI and SIMS datasets. The UGF is compared with the current state-of-the-art MSA methods under the same dataset settings.

Table 1, Table 2 and Table 3 show the experimental results on the three datasets, respectively. As we can see that our UGF achieves the best performance on the most metrics, which demonstrates the effectiveness of our proposed method.

Table 1: Comparison with baselines on CMU-MOSI dataset. Note that the left side of "/" is "negative/non-negative" and the right is "negative/positive". The best results are highlighted in bold.

Models	Acc-7↑	Acc-2↑	F1 ↑	MAE↓	Corr↑
LF-DNN[19]	34.52	77.52/78.63	77.46/78.63	0.955	0.658
GCD-CMR[9]	41.11	82.65/84.3	82.61/84.31	0.757	0.798
MAG-BERT[14]	41.43	82.13/83.54	81.12/83.58	0.790	0.766
HCT-DMG[17]	41.80	-/85.10	-/84.80	0.855	0.732
DMD[8]	41.90	-/83.50	-/83.50	0.727	-
ConFEDE[18]	42.27	82.22/83.84	82.15/83.83	0.742	0.784
MISA[2]	42.30	81.80/83.40	81.70/83.60	0.783	0.776
Self-MM[20]	45.79	82.54/84.77	82.68/84.91	**0.712**	**0.795**
MSEN[5]	46.09	83.20/84.23	82.70/84.22	0.740	0.784
UGF(Ours)	**47.81**	**84.4/85.82**	**84.35/85.81**	0.716	0.792

Table 2: Comparison with baselines on CMU-MOSEI dataset.

Models	Acc-7↑	Acc-2↑	F1↑	MAE↓	Corr↑
MAG-BERT[14]	50.41	79.86/86.86	80.47/83.88	0.583	0.741
LF-DNN[19]	50.83	80.60/82.74	80.85/82.52	0.580	0.709
MSEN[5]	52.55	83.39/84.89	83.05/84.74	0.544	0.759
Self-MM[20]	52.46	82.66/84.96	82.95/84.93	0.541	0.767
MISA[2]	52.20	83.60/85.50	83.80/85.30	0.555	0.756
DMD[8]	52.80	-/84.80	-/84.70	0.540	-
HCT-DMG[17]	53.20	-/84.20	-/84.00	0.535	0.752
GCD-CMR[9]	53.89	80.92/85.5	81.61/85.66	0.547	0.722
ConFEDE[18]	**54.86**	81.65/85.82	82.17/**85.83**	**0.522**	**0.780**
UGF(Ours)	54.58	**83.75/85.91**	**83.88**/85.74	0.536	0.767

From Table 1, it is evident that Unified Graph Fusion (UGF) model achieves an Acc-7 accuracy of 47.81%, Acc-2 accuracies of 84.4%/85.82%, and F1 scores of 84.35%/85.81%. Our UGF adopts a similar pipeline with Self-MM. In the three metrics Acc-7, Acc-2 and F1 scores, UGF significantly improves 2.02%, 1.86%/1.05% and 1.67%/0.9%, respectively, and shows competitive performance on other metrics. Compared with ConFEDE, which extracts both modality-shared and modality specific features and utilize cross-modal contrastive learning, UGF beat ConFEDE on all metrics. MSEN finds the text modality control the training process. We find that audio and visual modalities have higher uncertain levels than text modality. We achieves a better performance in simpler way. Especially, in Acc-7, Acc-2 and F1 scores, UGF improves previous state-of-the-art modelsby 1.72%, by 1.2%/0.72% and 1.67%/0.9%, respectively, which demonstrates the importance of modal uncertainty estimation. We automatically adjust the training process according to modal uncertainties. Thus, the uncertain modalities can win more attentions and be optimized better.

On the CMU-MOSEI dataset, our UGF achieves the best performance on Acc-2 (83.70%/85.91%) and F1 scores (83.88%/85.53%). On the SIMS dataset, our model beat all baselines in five of the six metrics. The results on three datasets prove that our UGF has strong generalization ability. This also shows that the modal uncertainty problem is universal, modal uncertainty estimation is an importance issue for multimodal sentiment analysis. Moreover, our UGF does not have any complex feature extraction or multi-modal feature interaction operations. It is proved that UGF provides a solution to the modal uncertainty problem.

Table 3: Comparison with baselines on the SIMS dataset.

Models	Acc-5↑	Acc-3↑	Acc-2↑	F1↑	MAE↓	Corr↑
TFN[21]	39.30	65.12	78.38	78.62	0.432	0.591
LF-DNN[19]	39.74	64.33	77.02	77.27	0.446	0.555
LMF[10]	40.53	64.68	77.77	77.88	0.441	0.576
MAG-BERT[14]	-	-	74.44	71.75	0.492	0.399
MISA[2]	-	-	76.54	76.59	0.447	0.563
Self-MM[20]	41.53	65.47	80.04	80.44	0.425	0.595
ConFEDE[18]	42.45	69.36	81.58	81.42	0.669	0.391
MSEN[5]	41.62	67.50	81.17	81.08	**0.413**	0.6030
Ours	**43.33**	**69.37**	**82.06**	**81.98**	0.486	**0.6033**

4.3 Ablation Study

We conduct extend ablation experiments to evaluate the contributions of proposed modules on CMU-MOSI dataset. The experimental results are shown in Table 4.

Table 4: Ablation results of dynamic gradient modulation (DGM) and uncertainty guided feature masking (UFM) on CMU-MOSI.

DGM	UFM	Acc-7↑	Acc-2↑	F1↑	MAE↓	Corr↑
×	×	37.21	81.99/84.10	81.81/84.01	0.950	0.743
✓	×	47.08	83.24/85.21	83.09/85.13	0.720	0.790
✓	✓	**47.23**	**84.4/85.82**	**84.35/85.81**	**0.716**	**0.792**

Dynamic Gradient Modulation. As can see in Table 4, DGM enhances MSA performance significantly. This indicates dynamically modulate gradients is helpful to optimize sub-optimal modalities. Suppressing the intensity of backpropagation signals from dominant modality and scaling the intensity of signals from uncertain modalities is effective. As we can observe, when using the DGM, the Acc-7 accuracy improves from 37.21% to 47.08%, MAE reduces 0.23, the other metrics also have significant improvements. These results demonstrate the uncertain modalities are sub-optimal.Weaken the dominant role of low uncertain modality is important for multimodal sentiment analysis.

Table 5: Comparison of UFM and random dropout.

dropout rate	Acc-7↑	Acc-2↑	F1↑	MAE↓	Corr↑
0.1	36.44	81.92/83.08	81.93/83.14	1.071	0.756
0.2	36.01	81.34/82.32	81.37/82.4	1.080	0.756
0.3	37.17	82.07/83.23	82.07/83.28	1.064	0.757
0.4	37.17	81.92/83.08	81.93/83.14	1.052	0.758
0.5	36.88	82.07/83.23	82.08/83.29	1.060	0.757
UFM	**47.81**	**84.4/85.82**	**84.35/85.81**	**0.716**	**0.792**

Uncertain Aware Feature Masking. We evaluate the contribution of UFM module, the results are shown in Table 4. When incorporating the UFM module, as we can observe that all metrics have improvements, such as Acc-7 improves from 47.08% to 47.23%, Acc-2 improves from 84.40%/85.82% to 84.40%/85.82%, MAE go down from 0.720 to 0.716. These results demonstrate the UFM module can force model pay more attention on uncertain modalities. We also can observe that the DGM and UFM complement each other, DGM controls the optimization process, UFM modifies features to effect the sentiment predictions. We also compare our UFM with random dropout, random dropout randomly mask features. The results are shown in Table. 5. We can see that uncertainty guided feature masking is absolutely better than random dropout, which demonstrates the uncertainty guidance is important.

Table 6: Comparison of different modalities on CMU-MOSI.

modality	Acc-7↑	Acc-2↑	F1-Score↑	MAE↓	Corr↑
T	46.21	83.38 / 85.21	83.24 / 85.13	0.719	0.789
A	**47.81**	83.09 / 85.37	82.92 / 85.28	0.719	0.788
V	47.23	82.94/84.91	82.83/84.85	0.727	0.789
T+A	46.65	83.38 / 85.52	83.22 / 85.43	**0.716**	0.789
T+V	47.23	83.38 / 85.37	83.24 / 85.29	0.722	0.790
A+V	46.79	83.53 / 85.52	83.38 / 85.44	0.725	0.789
T+A+V	**47.81**	**84.4/85.82**	**84.35/85.81**	**0.716**	**0.791**

Modality Ablation Results. We experimented with various modal combinations for analyzing their influence on performance on CMU-MOSI dataset. Results are shown in Table 6. 'T', 'A', 'V' represent the text, audio and visual as input. Overall, more modality can provide richer information. We can see that the performance of bimodal are better than unimodal, the performance of three modalities are best. We also can observe that text modality is the most predictive. This is because text modality has a low uncertainty. But the visual modality is uncertain, the potential of visual modality has yet to be fully explored.

Parameter Analysis. In Fig. 3, we compare the performance of Acc-2 and MAE with vary setting of α on CMU-MOSI. It can be observed that the performance of the model is insensitive to α. The fluctuations of Acc-2 and MAE are less than 0.01. The role of α is to control a^s within a reasonable range, ensuring model can accurately estimate uncertainty of each modality.

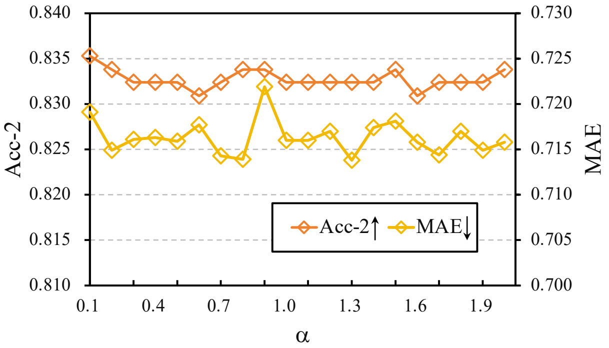

Fig. 3: Performance comparison with the change of α on CMU-MOSI dataset.

4.4 Case Study

In order to further show the effectiveness of the proposed method, we selected two multimodal examples from the CMU-MOSI dataset, as shown in Fig. 4. We can observe that 1) the predictions of multimodal are more accurate than unimodal. 2) DGM and UFM can effectively improve the accuracy of unimodal and multimodal. These results demonstrate the effectiveness of uncertainty guided optimization. With the guidance of modal uncertainty, model can balance optimize all modalities, improving the learning ability of uncertain modalities.

Examples	DGM	UFM	Text	Audio	Vision	Fusion	Sentiment Score
	✓	✓	**1.4075**	0.5789	**2.349**	**1.5697**	1.6
	✓	✗	1.3212	**0.7291**	3.4067	1.4618	
And so that's pretty entertaining	✗	✗	1.3907	0.7121	3.4070	1.5628	
	✓	✓	-2.3913	**0.0775**	**2.5249**	**-2.0556**	-2.25
	✓	✗	-2.4382	0.2632	3.6451	-1.8803	
Why you have no skills what sover	✗	✗	**-2.1333**	0.2639	3.1456	-1.8713	

Fig. 4: Visualization of Unimodal and Multimodal Sentiment Scores.

5 Conclusion

Modal uncertainty come from the data noise and semantic ambiguity. In joint optimization, uncertain modalities are sub-optimal. However, it is difficult to effectively estimate modal uncertainties in regression task. In this paper, we propose a new Uncertainty aware Gradient modulation Feature masking model, a novel method designed to address above issue in multimodal sentiment analysis. We propose a novel modal uncertainty estimation method by considering both intra-modal and inter-modal consistencies. According to the uncertainty of each modality, we improve the model by 1) controlling the back-propagation process 2) adaptively modifying feature to optimize the classifiers. We design a dynamic gradient modulation module, it suppresses the gradients of dominant modality and increases the signal intensity of uncertain modalities, making uncertain modalities won more attention. We also propose a uncertain guided feature masking module, it reduces the dependence of model on one certain modality to improve the performance of uncertain modalities. Extensive experiments on CMU-MOSI, CMU-MOSEI, and CH-SIMS verify the contributions of proposed modules.

Acknowledgement

This work was partially supported by the China Postdoctoral Science Foundation (2023M741305), the Fundamental Research Funds for the Central Universities (CCNU23XJ001), the National Natural Science Foundation of China (62372205 and 61932008), the National Language Commission Key Research Project (ZDI145-56), the Fundamental Research Funds for Central Universities (KJ02502022-0450).

References

1. Denecke, K., Deng, Y.: Sentiment analysis in medical settings: New opportunities and challenges. Artificial Intelligence in Medicine **64**(1), 17–27 (2015)
2. Hazarika, D., Zimmermann, R., Poria, S.: Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 1122–1131 (2020)
3. Huang, F., Zhang, X., Zhao, Z., Xu, J., Li, Z.: Image–text sentiment analysis via deep multimodal attentive fusion. Knowledge-Based Systems **167**, 26–37 (2019)
4. Ji, Y., Wang, J., Gong, Y., Zhang, L., Zhu, Y., Wang, H., Zhang, J., Sakai, T., Yang, Y.: Map: Multimodal uncertainty-aware vision-language pre-training model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23262–23271 (2023)
5. Jin, C., Luo, C., Yan, M., Zhao, G., Zhang, G., Zhang, S.: Weakening the dominant role of text: Cmosi dataset and multimodal semantic enhancement network. IEEE Transactions on Neural Networks and Learning Systems pp. 1–15 (2023). https://doi.org/10.1109/TNNLS.2023.3282953
6. Kim, J.U., Park, S., Ro, Y.M.: Uncertainty-guided cross-modal learning for robust multispectral pedestrian detection. IEEE Transactions on Circuits and Systems for Video Technology **32**(3), 1510–1523 (2022). https://doi.org/10.1109/TCSVT.2021.3076466
7. Li, J., Wang, C., Luo, Z., Wu, Y., Jiang, X.: Modality-dependent sentiments exploring for multi-modal sentiment classification. In: ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 7930–7934 (2024). https://doi.org/10.1109/ICASSP48485.2024.10445820
8. Li, Y., Wang, Y., Cui, Z.: Decoupled multimodal distilling for emotion recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6631–6640 (2023)

9. Liang, H., Xie, W., He, X., Song, S., Shen, L.: Circular decomposition and cross-modal recombination for multimodal sentiment analysis. In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 7910–7914. IEEE (2024)
10. Liu, Z., Shen, Y., Lakshminarasimhan, V.B., Liang, P.P., Zadeh, A., Morency, L.P.: Efficient low-rank multimodal fusion with modality-specific factors. arXiv preprint arXiv:1806.00064 (2018)
11. Peng, X., Wei, Y., Deng, A., Wang, D., Hu, D.: Balanced multimodal learning via on-the-fly gradient modulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8238–8247 (2022)
12. Poria, S., Cambria, E., Bajpai, R., Hussain, A.: A review of affective computing: From unimodal analysis to multimodal fusion. Information Fusion **37**, 98–125 (2017)
13. Poria, S., Hazarika, D., Majumder, N., Mihalcea, R.: Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research. IEEE Transactions on Affective Computing **14**(1), 108–132 (2020)
14. Rahman, W., Hasan, M.K., Lee, S., Zadeh, A., Mao, C., Morency, L.P., Hoque, E.: Integrating multimodal information in large pretrained transformers. In: Proceedings of the Conference of the Association for Computational Linguistics. vol. 2020, p. 2359 (2020)
15. Soleymani, M., Garcia, D., Jou, B., Schuller, B., Chang, S.F., Pantic, M.: A survey of multimodal sentiment analysis. Image and Vision Computing **65**, 3–14 (2017)
16. Wang, H., Zhang, J., Chen, Y., Ma, C., Avery, J., Hull, L., Carneiro, G.: Uncertainty-aware multi-modal learning via cross-modal random network prediction. In: European Conference on Computer Vision. pp. 200–217 (2022)
17. Wang, Y., Li, Y., Bell, P., Lai, C.: Cross-attention is not enough: Incongruity-aware hierarchical multimodal sentiment analysis and emotion recognition
18. Yang, J., Yu, Y., Niu, D., Guo, W., Xu, Y.: 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). pp. 7617–7630 (2023)
19. Yu, W., Xu, H., Meng, F., Zhu, Y., Ma, Y., Wu, J., Zou, J., Yang, K.: 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. pp. 3718–3727 (2020)
20. Yu, W., Xu, H., Yuan, Z., Wu, J.: Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 10790–10797 (2021)
21. Zadeh, A., Chen, M., Poria, S., Cambria, E., Morency, L.P.: Tensor fusion network for multimodal sentiment analysis. In: Palmer, M., Hwa, R., Riedel, S. (eds.) Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 1103–1114. Copenhagen, Denmark (sep 2017). https://doi.org/10.18653/v1/D17-1115
22. Zadeh, A., Zellers, R., Pincus, E., Morency, L.P.: Mosi: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. arXiv preprint arXiv:1606.06259 (2016)
23. Zadeh, A.B., Liang, P.P., Poria, S., Cambria, E., Morency, L.P.: 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). pp. 2236–2246 (2018)