Feature Decoupling with Modality Modulation for Multimodal Sentiment Analysis

Yongbo Wang[1], Jiaxiang Wang[2(✉)], Aihua Zheng[1], Wenjuan Cheng[3(✉)], and Xiaofei Sheng[4]

[1] Information Materials and Intelligent Sensing Laboratory of Anhui Province, Anhui Provincial Key Laboratory of Security Artificial Intelligence, School of Artificial Intelligence, Anhui University, Hefei, China

[2] School of Artificial Intelligence, Anhui University of Science and Technology, Hefei, China
Netizenwjx@foxmail.com

[3] School of Computer and Information, Hefei University of Technology, Hefei, China
cheng@ah.edu.cn

[4] Wuhu Simba Network Technology Co., Wuhu, China

Abstract. Multimodal sentiment analysis aims to extract and integrate meaningful information from diverse modalities to infer a speaker's emotional state. Due to the inherent heterogeneity among modalities, most existing approaches decouple modalities into specific and invariant features, which partially capture cross-modal representations. However, in multimodal tasks, certain modalities often dominate the optimization process, leading to the under-optimization of weaker modalities. To address this imbalance, we propose the Modal Modulation Adaptive Fusion Network (MMAFNet), which enhances the learning of valuable information from each modality. Specifically, for modality-specific features, we introduce a gradient modulation strategy that dynamically adjusts learning rates to prioritize weaker modalities. For modality-invariant features, we employ a parameter reset strategy based on inter-modal distances to mitigate overfitting and strengthen feature extraction in underperforming modalities. Additionally, an adaptive fusion module combines modality-specific and invariant features according to their learned weights. Our comprehensive analysis of feature characteristics and tailored modulation strategies effectively alleviates modality imbalance. Extensive experiments on two benchmark datasets demonstrate the superiority of our approach.

Keywords: Multimodal sentiment analysis · modality modulation · modality imbalance

1 Introduction

In recent years, with the rapid development of social networks, people tend to share their personal opinions and feelings online. This content exists in various forms, such as text, audio, and visual, which together reflect the various

Z.Lin et al. (Eds.): ICIG 2025, LNCS 16162, pp. 92–104, 2026.
https://doi.org/10.1007/978-981-95-3393-0_8

sentiment states of users [4]. As an individual's psychological response to a specific situation, sentiment analysis is essential for facilitating effective communication and supporting decision-making processes. For instance, companies can adjust their product strategies by capturing consumers' emotional feedback on new products. Meanwhile, researchers are working to equip machines with the ability to analyze human emotions, making related applications more human-centric and enhancing the human-computer interaction experience. Therefore, sentiment analysis plays a vital role in driving progress in the field of artificial intelligence [19].

Multimodal sentiment analysis combines text, audio, and visual information to recognize sentiment, which provides richer data sources compared to traditional single modality analysis, thus improving recognition accuracy [13,16]. Although the different data forms and information transfer ways of text, audio, and visual modalities lead to heterogeneity among modalities, they can all express the same emotional tendency from different perspectives. The heterogeneity between modalities can be mitigated by learning modality-specific and invariant features to facilitate more effective joint representation learning [5]. The core of multimodal sentiment analysis is the fusion of cross-modal information, which can be roughly categorized into three types based on the timing of fusion, early fusion, middle fusion, and late fusion. For instance, Wang *et al.* [18] propose a Recurring Attended Variation Encoding Network that utilizes local fusion modules to learn the nature of nonverbal dynamics. Shamane *et al.* [11] design a fusion method based on Transformer to achieve text and image alignment, and the overall model is self-supervised. And Xu *et al.* [21] propose a framework for a mixed-modal knowledge expert with two-stage training to dynamically fuse three unimodal information to obtain a joint representation.

However, according to recent research [10], in multimodal tasks, due to the existence of a uniform learning objective, the strong modality tends to dominate the optimization process, leading to under-optimization of the weak modality, which in turn generates the modality imbalance problem. At the same time, previous studies have confirmed that the learning rate of different modalities varies [17], and a unified optimization objective may lead to inconsistent convergence speeds of each modality.

To address the aforementioned issues, we propose a Modality Modulation Adaptive Fusion Network (MMAFNet), which modulates modality-specific and modality-invariant features separately, considering modality heterogeneity, to fully extract the relevant information from each modality. First, for each modality-specific feature, we introduce a Specific-feature Gradient Modulation (SGM) strategy. Since stronger modalities dominate the overall optimization direction, this method adaptively adjusts the encoder gradient based on the classifier results from each modality, slowing the learning speed of stronger modalities to promote the optimization of weaker modalities. Second, we propose an Invariant-feature Parameter Reset (IPR) strategy for the invariant features shared by all three modalities. This method directly focuses on the representation space of the modalities and evaluates its learning ability based on the distance between the modalities. The closer the distance between the modalities, the better the learning, and the weighting parameters are calculated accordingly.

These weight parameters are then used to reinitialize the learning process for each modality, enabling more effective extraction of invariant features and reducing the impact of noise. Finally, we design an Adaptive Fusion Module (AFM) for weighted fusion to learn invariant and specific features more effectively. The main contributions of this paper are as follows:

- We propose a multimodal sentiment analysis framework to address the problem of modal imbalance. This framework accelerates the learning of strong modalities by refining modality-specific features while enhancing weak modalities through gradient modulation.
- For modality-invariant features, we optimize the representation space by resetting parameters according to inter-modal distances, thereby improving feature extraction across all three modalities.
- We design an adaptive fusion module that dynamically combines modality-specific and invariant features based on learned weights, ensuring optimal integration of each modality's informative components.

Extensive experiments on the multimodal sentiment analysis datasets CMU-MOSI [25] and CMU-MOSEI [26] validate the effectiveness of our proposed method.

2 Methodology

This paper introduces a framework termed the Modality Modulation Adaptive Fusion Network (MMAFNet), designed to address the imbalance problem in multimodal sentiment analysis, as shown in Fig. 1. The framework decouples information from text, audio, and visual modalities into modality-specific features and modality-invariant features, which are then modulated and fused systematically. The process involves three core steps: (1) For modality-specific features, a Specific-feature Gradient Modulation (SGM) approach is employed to adjust the gradient of each modality based on its contribution, thereby mitigating the network's bias toward strong modalities. (2) For modality-invariant features, an Invariant-feature Parameter Reset (IPR) strategy is implemented, which utilizes a distance function to measure inter-modality distances and reinitialize parameters for effective relearning. (3) Finally, an Adaptive Fusion Module (AFM) is applied to weight the fusion according to feature importance, generating the final sentiment prediction results.

2.1 Multimodal Decoupling

The multimodal sentiment analysis task aims to predict the sentiment information conveyed by input multimodal data, including text (t), audio (a), and visual (v) modalities. A video sequence is defined as $S = \{(X_i, Y_i)\}_{i=1}^{N}$, where Y represents the label, N denotes the number of samples, and $X = \{X_m | m \in (t, a, v)\}$ represents the input multimodal data. Additionally, the input data dimension is defined as $X_m \in \mathbb{R}^{T_m \times d_m}$, where T_m denotes the sequence length, and d_m is the feature dimension.

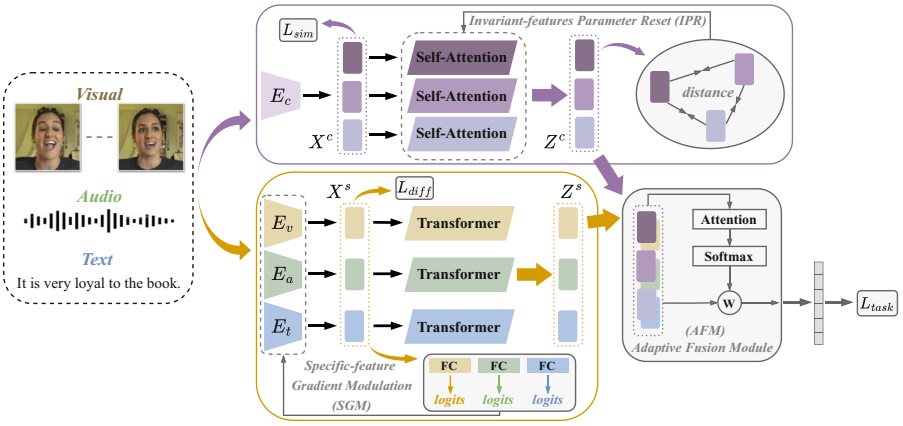

Fig. 1. The overall framework of MMAFNet.

Due to the heterogeneity of multimodal data, directly fusing information from different modalities is suboptimal. Decoupling multimodal information into modality-specific and modality-invariant features is an effective strategy to address this issue. Modality-specific features capture information unique to each modality, while modality-invariant features extract commonalities among modalities, reducing mutual interference. To achieve multimodal feature decoupling, three independent 1D convolutional layers are used to align the dimensions of the three modalities and extract low-dimensional features. For the specific feature branch, an encoder $E_m, m \in \{t, a, v\}$ is defined for each modality. The input data is passed through the encoder to obtain the specific features X_t^s, X_a^s, X_v^s, and the invariant features X_t^c, X_a^c, X_v^c are derived using a shared encoder E_c. The process is described as:

$$X_m^s = E_m(X_m), X_m^c = E_c(X_m). \tag{1}$$

To better distinguish between invariant and specific features, a corresponding loss function is introduced. Learning invariant features aims to extract commonalities among different modalities. Center Moment Discrepancy (CMD) [27] introduces a distance constraint to minimize the discrepancy between different modalities. By matching the order-wise moment differences of feature distributions, CMD ensures the effective learning of modality-invariant features. This is computed as:

$$\begin{gathered} L_{sim} = \frac{1}{3} \sum_{(m_1,m_2)} CMD(X_{m_1}, X_{m_2}), \\ (m_1, m_2) \in \{(a, v), (a, t), (v, t)\}. \end{gathered} \tag{2}$$

On the other hand, specific features are unique to each modality, and features from different modalities may contain redundant information or interfere with one another. To address this, orthogonality constraints are applied to reduce mutual interference. Both inter-modal and intra-modal orthogonality

are enforced, ensuring that invariant and specific features learn different aspects of the data. This is formulated as:

$$L_{diff} = \sum_{m \in \{t,a,v\}} \langle X_m^c, X_m^s \rangle + \sum_{\substack{(m_1,m_2) \in \\ \{(a,v),(a,t),\\(v,t)\}}} \langle X_{m_1}^s, X_{m_2}^s \rangle, \tag{3}$$

where $\langle \cdot \rangle$ denotes the orthogonal operation implemented using dot product. These loss constraints enable the network to decouple multimodal features better and mitigate modality heterogeneity.

2.2 Specific-Feature Gradient Modulation

According to previous research [10], low-confidence modalities tend to receive limited optimization during backpropagation, while better-performing modalities dominate the optimization process. This imbalance results in the under-optimization of weaker modalities when the model converges. Modality-specific features aim to learn information unique to each modality individually, and to prevent the model from over-preferring strong modalities, we propose a modulation strategy SGM for specific features, as illustrated in Fig. 1.

We define three modality-specific encoders $E_m(\sigma_m, X_m)$, where $m \in \{t, a, v\}$, used to extract features for each modality, with σ_m denoting the encoder parameters. Therefore, the gradient update for the encoders can be expressed as:

$$\sigma_m^{k+1} = \sigma_m^k - \eta \nabla_{\sigma_m^k} \mathcal{L}, \tag{4}$$

where η represents the learning rate, and $\nabla_{\sigma_m^k} \mathcal{L}$ is the gradient of the loss function for the encoder at the k^{th} iteration. To balance multimodal learning, we perform adaptive gradient adjustment for specific feature encoders. The accuracy ρ_m for each modality is measured by introducing an independent classification layer, defined as:

$$f(X_m) = W_m \cdot E_m(\sigma_m, X_m) + b, \tag{5}$$

where $f(X_m)$ represents the prediction output for modality m, W_m and b are the parameters of the classification layer. The accuracy ρ_m is computed as:

$$\rho_m = \frac{1}{N} \sum_{j=1}^{N} \mathbb{I}(f(X_m^j) = Y^j), \tag{6}$$

where $\mathbb{I}$ is an indicator function:

$$\mathbb{I}(\text{condition}) = \begin{cases} 1, & \text{if condition is true,} \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

After obtaining the accuracy ρ_m for all modalities, the modulation weights ω_m^k are calculated to reflect the contribution of each modality:

$$\omega_m^k = \frac{\sum_{j \in B_k} \rho_m}{\sum_{j \in B_k} (\rho_t + \rho_a + \rho_v)}, \tag{8}$$

where B_k denotes the randomly sampled min-batch at the k^{th} iteration. Higher accuracies indicate stronger modalities, resulting in higher modulation weights. The modulation function for the weight φ_m^k is defined as:

$$\varphi_m^k = \begin{cases} 1 - tanh(\alpha \cdot \omega_m^k), & \text{if } \omega_m^k = max(\omega_t^k, \omega_a^k, \omega_v^k), \\ 1, & \text{otherwise,} \end{cases} \tag{9}$$

where α is a hyperparameter.

This modulation reduces the gradient updates for the dominant modality, thereby slowing its optimization rate. Throughout the training process, the contributions of the three modalities are dynamically monitored, and the weights are adaptively adjusted. The updated gradient for the encoder is then computed as:

$$\sigma_m^{k+1} = \sigma_m^k - \eta \cdot \varphi_m^k \nabla_{\sigma_m^k} \mathcal{L}. \tag{10}$$

By applying φ_m^k only to the dominant modality, weaker modalities are unaffected, achieving a balanced multimodal learning state.

2.3 Invariant-Feature Parameter Reset

In contrast to modality-specific features, invariant features focus on eliminating modality differences by mapping all modalities to the common feature space and capturing their shared characteristics. Since all modalities share the same encoder for feature extraction and fusion during prediction, the specific feature modulation strategy is not applicable. When learning invariant features, it is crucial to consider the magnitude of information contributed by each modality. Over-suppressing strong modalities can lead the model to focus excessively on weaker modalities, thereby learning excessive noise [20]. To address this issue, we propose a modulation strategy IPR for invariant feature learning, as illustrated in Fig. 1. This strategy determines modulation weights by evaluating the distances between different modalities and resets parameters during invariant feature extraction to avoid overfitting strong modalities and minimize noise in weak modalities.

To implement this strategy, we utilize a self-attention mechanism to extract high-dimensional invariant features Z_m^c from the low-dimensional features produced by the shared invariant feature encoder E_c. The computation is as follows:

$$Z_m^c = Self\text{-}Attention_m(\theta_m, E_c(X_m)), \tag{11}$$

where θ_m denotes the parameters of the self-attention module. Based on the properties of invariant features, we propose a new modality contribution measure. We determine the distance D_m^c between different modality invariant features Z_m^c to calculate the corresponding weights. We use the cosine distance D_{cos} to compute this, as follows:

$$D_t^c = \frac{(D_{cos}(Z_t^c, Z_a^c) + D_{cos}(Z_t^c, Z_v^c))}{2}, \tag{12}$$

$$D_{cos}(Z_1, Z_2) = 1 - \frac{\sum Z_1 \cdot Z_2}{\sqrt{\sum Z_1^2} \cdot \sqrt{\sum Z_2^2}}, \tag{13}$$

where Z_1 and Z_2 denote the invariant features of different modalities. Similarly, we can also get the distance D_a^c for audio and D_v^c for visual based on the calculation of text modality. The objective of invariant feature learning is to reduce the distance between modalities, where a smaller distance indicates better learning outcomes. In this way, the contribution of the invariant features of different modalities to the model can be better evaluated without requiring an additional classification layer. Using the computed distances D_m^c, we calculate the weights ϕ_m^c for each modality as follows:

$$\phi_m^c = max(tanh(D_m^c), \delta), m \in \{t, a, v\}, \tag{14}$$

where the *tanh* function ensures that the weights range between 0 and 1, and δ is a hyperparameter to prevent the weights from becoming too small.

Next, we reset and relearn the parameters θ_m for every modality at each epoch based on the computed weights ϕ_m^c. At the beginning of model training, we save the corresponding initialization parameter θ_m^{init}, which is later used for resetting. The reset process is defined as:

$$\theta_m^{new} = \phi_m^c \theta_m^{init} + (1 - \phi_m^c)\theta_m^{cur}, \tag{15}$$

where θ_m^{cur} denotes the current parameters and θ_m^{new} is the updated parameter.

This approach resets the parameters of stronger modalities to a greater extent, reducing the model's reliance on them and improving the learning of weaker modalities. Additionally, parameter resetting helps mitigate overfitting in stronger modalities and prevents excessive noise from weaker ones. By considering the properties of invariant features, the approach alleviates modality imbalance and enhances the model's ability to capture commonalities across modalities, which supports subsequent predictions and improves overall performance.

2.4 Adaptive Fusion Module

By employing the proposed modality modulation strategies, the model is able to sufficiently learn the information from each modality while effectively addressing the issue of inter-modality heterogeneity. To further enhance the integration of modulated specific features (Z_m^s) and invariant features (Z^c), we introduce an adaptive fusion module. This module dynamically adjusts the weights of each feature using an attention mechanism, thereby improving the model's robustness and avoiding data redundancy or loss caused by fixed weights. The specific calculation is as follows:

$$g = Attention([Z_t^s; Z_a^s; Z_v^s; Z^c]), \tag{16}$$

$$Z = [g_t \cdot Z_t^s; g_a \cdot Z_a^s; g_v \cdot Z_v^s; g_c \cdot Z^c], \tag{17}$$

where g represents the weights obtained through the attention mechanism, $[;]$ denotes the concatenation operation, and Z is the resulting weighted fused feature. These weighted fused features are then used for prediction, significantly improving the accuracy of sentiment analysis.

Therefore, we integrate all the losses mentioned above to define the overall optimization objective:

$$L_{total} = L_{task} + \lambda_1 L_{sim} + \lambda_2 L_{diff}, \tag{18}$$

where L_{task} denotes the loss of the multimodal sentiment analysis task, and λ_1 and λ_2 denote the hyperparameters.

3 Experiments

3.1 Experimental Details

We extract original features from both datasets following the method described in [7]. For text features, we utilize the pre-trained BERT-base-uncased [6] model to obtain 768-dimensional word vectors. For audio data, 74-dimensional features are extracted by the audio analysis framework COVAREP [3]. For visual features, we focus on faces using the Facet [1] tool to extract 35-dimensional facial features. The network is trained for 50 epochs using the SGD optimizer and modality modulation is performed in the previous 25 epochs. During training, the optimal values of hyperparameters are set as follows: $\lambda_1 = 0.1$, $\lambda_2 = 0.01$, $\alpha = 0.3$, $\delta = 0.1$, and the learning rate $\eta = 0.0001$. We measure the performance of multimodal sentiment analysis based on the following metrics: binary accuracy (Acc2), F1-Score (F1), 7-class accuracy (Acc7), and mean absolute error (MAE). All experiments are implemented using the PyTorch framework and performed on a workstation equipped with an NVIDIA GeForce RTX 3090 GPU.

3.2 Comparison with State-of-the-Art Methods

We compare MMAFNet with current multimodal sentiment analysis state-of-the-art methods. These methods include pure learning-based models such as TFN [24], LMF [9], and multimodal fusion-based models MFM [15], ICCN [12], MulT [14], MISA [5], Self-MM [23], FDMER [22], and some recent competitive models PS-Mixer [8], MInD [2], DCD [7]. All methods are compared under the word-aligned setting.

Results on CMU-MOSI [25]: As shown in Table 1, our proposed model MMAFNet outperforms the compared methods across all evaluation metrics. On the MOSI dataset, MMAFNet shows consistent improvements over the state-of-the-art methods in all four multimodal sentiment analysis metrics. Notably, MMAFNet achieves a 1.3% improvement in the Acc7 metric compared to the sub-optimal results, highlighting the effectiveness of our modality modulation strategy. This improvement indicates that the proposed approach facilitates better learning of each modality's information, particularly for fine-grained sentiment analysis. These results confirm the effectiveness of our method.

Table 1. Experimental results on two datasets, where * indicates the results reproduced.

Methods	CMU-MOSI [25]				CMU-MOSEI [26]			
	*Acc*2 ↑	*F*1 ↑	*Acc*7 ↑	*MAE* ↓	*Acc*2 ↑	*F*1 ↑	*Acc*7 ↑	*MAE* ↓
TFN [24]	80.8	80.7	34.9	0.901	82.5	82.1	50.2	0.593
LMF [9]	82.5	82.4	33.2	0.917	82.0	82.1	48.0	0.623
MFM [15]	81.7	81.6	35.4	0.877	84.4	84.3	51.3	0.568
ICCN [12]	83.0	83.0	39.0	0.862	84.2	84.2	51.6	0.565
MulT [14]	83.0	82.8	40.0	0.871	82.5	82.3	51.8	0.580
MISA [5]	83.4	83.6	42.3	0.783	85.5	85.3	52.2	0.555
Self-MM [23]	85.9	85.9	-	0.713	85.1	85.3	-	0.530
FDMER [22]	84.6	84.7	44.1	0.724	86.1	85.8	54.1	0.536
PS-Mixer [8]	82.1	82.1	44.3	0.794	86.1	86.1	53.0	0.537
MInD [2]	86.0	86.0	45.8	0.705	**86.6**	**86.7**	53.9	0.529
DCD* [7]	86.2	86.2	45.1	0.724	85.6	85.5	53.4	0.543
MMAFNet (Ours)	**86.6**	**86.6**	**47.1**	**0.698**	86.3	86.2	**54.4**	**0.527**

Table 2. Ablation study of modules validity in MMAFNet on the two datasets.

Dataset	SGM	IPR	AFM	*Acc*2 ↑	*F*1 ↑	*Acc*7 ↑	*MAE* ↓
CMU-MOSI [25]	✗	✗	✗	84.6	84.5	44.3	0.735
	✓	✗	✗	85.4	85.3	45.5	0.721
	✓	✓	✗	86.2	86.1	46.3	0.711
	✓	✓	✓	**86.6**	**86.6**	**47.1**	**0.698**
CMU-MOSEI [26]	✗	✗	✗	84.8	84.9	52.6	0.547
	✓	✗	✗	85.4	85.4	53.5	0.539
	✓	✓	✗	86.0	86.0	54.0	0.530
	✓	✓	✓	**86.3**	**86.2**	**54.4**	**0.527**

Results on CMU-MOSEI [26]: As shown in Table 1, although our method works slightly overshadowed by MInD [2] with 0.3% and 0.5% decreasing on Acc2 and F1, it exhibits competitive results with the best Acc7 and MAE. Note that our modality modulation approach enhances multimodal information utilization, particularly for fine-grained sentiment analysis, and mitigates the multimodal learning imbalance problem. The larger size of the MOSEI dataset adds complexity and exacerbates the problem of sentiment category imbalance. Whereas Acc2 is a simpler binary task, this tends to bias the model toward more frequent categories, which reduces the accuracy of binary categorization. Additionally, inter-modality heterogeneity may not be fully resolved in binary tasks, resulting in slight declines in Acc2 and F1 performance.

3.3 Ablation Study

To verify the effectiveness of the proposed modules, we conduct ablation experiments on both the CMU-MOSI [25] and CMU-MOSEI [26] datasets, as shown in Table 2. Our method consists of three key components: specific feature gradient modulation (SGM), invariant feature parameter reset (IPR), and adaptive fusion

Table 3. MMAFNet modulation frequency analysis experiments on the CMU-MOSI [25] dataset.

Metrics \ Modulation	SGM				SGM+IPR			
	15	20	25	30	15	20	25	30
*Acc*2 ↑	85.1	85.5	**85.9**	85.6	85.8	86.3	**86.6**	86.2
*F*1 ↑	85.2	85.5	**85.9**	85.7	85.8	86.2	**86.6**	86.1
*Acc*7 ↑	45.1	45.7	**46.0**	45.8	45.2	46.5	**47.1**	47.0
MAE ↓	0.729	0.720	**0.715**	0.721	0.716	0.710	**0.698**	0.705

module (AFM). When all modules are removed, we observe a significant performance degradation, highlighting each component's essential role. Specifically, we find that the performance improves when either SGM is applied to modulate the specific feature branches, or IPR is used to modulate the invariant feature branches. This result demonstrates that both strategies contribute to addressing the modality imbalance problem. Finally, adding the AFM, which discriminates the importance of different features and applies weighted fusion to multimodal information, further boosts the model's performance. This reinforces the importance of feature-level modulation for effective multimodal sentiment analysis. In conclusion, the ablation experiments confirm the significance of each proposed module in improving multimodal sentiment analysis performance.

3.4 Modulation Frequency Analysis

We conduct modulation frequency analysis to achieve accurate modulation, as shown in Table 3. By keeping other parameters constant, we test frequencies of 15, 20, 25, and 30 epochs. The experimental results indicate that the SGM strategy performs best with 25 epochs, with performance declining at 30 epochs. Thus, we select 25 epochs as the optimal frequency for SGM. Building on this, we integrate the IPR strategy, which achieves the best overall performance at 25 epochs. Modulating both SGM and IPR at 25 epochs effectively enhances the model's ability to learn multimodal information and prevents overfitting during later training.

4 Conclusion

This paper proposes a novel approach, the Modality Modulation Adaptive Fusion Network (MMAFNet), designed to address the challenges of imbalance and heterogeneity in multimodal sentiment analysis tasks. In such tasks, strong modalities often dominate the network optimization process, leading to the under-optimization of weak modalities. To tackle this issue, our approach employs a modality decoupling framework, which effectively mitigates modality heterogeneity. For modality-specific features, we implement a feature gradient modulation strategy that adaptively adjusts gradients based on the relative importance

of each modality, ensuring balanced optimization. Additionally, we introduce a parameter resetting strategy for invariant features to prevent overfitting to strong modalities, reduce noise interference, and enhance the learning of shared invariant features. Finally, an adaptive fusion module is utilized to achieve weighted fusion, enabling the model to focus on the most relevant information. Extensive experiments on the CMU-MOSI and CMU-MOSEI datasets demonstrate the effectiveness of our approach. Future work will investigate the imbalance problem in scenarios with missing modalities to further improve model robustness.

Acknowledgments. This work is supported in part by the National Natural Science Foundation of China under Grants 62372003, the Natural Science Foundation of Anhui Province under Grants 2308085Y40 and 2208085J18, and the University Synergy Innovation Program of Anhui Province under Grant GXXT-2022-036.

References

1. Baltrušaitis, T., Robinson, P., Morency, L.P.: OpenFace: an open source facial behavior analysis toolkit. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision, pp. 1–10 (2016)
2. Dai, W., et al.: MinD: improving multimodal sentiment analysis via multimodal information disentanglement. arXiv preprint arXiv:2401.11818 (2024)
3. Degottex, G., Kane, J., Drugman, T., Raitio, T., Scherer, S.: COVAREP-a collaborative voice analysis repository for speech technologies. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 960–964 (2014)
4. Gandhi, A., Adhvaryu, K., Poria, S., Cambria, E., Hussain, A.: Multimodal sentiment analysis: a systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. Info. Fusion **91**, 424–444 (2023)
5. Hazarika, D., Zimmermann, R., Poria, S.: MISA: modality-invariant and-specific representations for multimodal sentiment analysis. In: Proceedings of the ACM International Conference on Multimedia, pp. 1122–1131 (2020)
6. Kenton, J.D.M.W.C., Toutanova, L.K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 4171–4186 (2019)
7. 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)
8. Lin, H., Zhang, P., Ling, J., Yang, Z., Lee, L.K., Liu, W.: PS-Mixer: a polar-vector and strength-vector mixer model for multimodal sentiment analysis. Info. Process. Manag. **60**(2), 103229 (2023)
9. Liu, Z., Shen, Y., Lakshminarasimhan, V.B., Liang, P.P., Zadeh, A., Morency, L.P.: Efficient low-rank multimodal fusion with modality-specific factors. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics, pp. 2247–2256 (2018)

10. 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)
11. Siriwardhana, S., Reis, A., Weerasekera, R., Nanayakkara, S.: Jointly fine-tuning "BERT-like" self supervised models to improve multimodal speech emotion recognition. In: Proceedings of the Annual Conference of the International Speech Communication Association, pp. 3755–3759 (2020)
12. Sun, Z., Sarma, P., Sethares, W., Liang, Y.: Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 8992–8999 (2020)
13. Tan, Q., Shen, X., Bai, Z., Sun, Y.: Cross-modality fused graph convolutional network for image-text sentiment analysis. In: International Conference on Image and Graphics, pp. 397–411 (2023)
14. Tsai, Y.H.H., Bai, S., Liang, P.P., Kolter, J.Z., Morency, L.P., Salakhutdinov, R.: Multimodal transformer for unaligned multimodal language sequences. In: Proceedings of the Conference. Association for Computational linguistics. Meeting. vol. 2019, p. 6558 (2019)
15. Tsai, Y.H.H., Liang, P.P., Zadeh, A., Morency, L.P., Salakhutdinov, R.: Learning factorized multimodal representations. In: Proceedings of the International Conference on Learning Representations (2019)
16. Vinodhini, G., Chandrasekaran, R.: Sentiment analysis and opinion mining: a survey. Int. J. **2**(6), 282–292 (2012)
17. Wang, W., Tran, D., Feiszli, M.: What makes training multi-modal classification networks hard? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12695–12705 (2020)
18. Wang, Y., Shen, Y., Liu, Z., Liang, P.P., Zadeh, A., Morency, L.P.: Words can shift: dynamically adjusting word representations using nonverbal behaviors. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 7216–7223 (2019)
19. Wankhade, M., Rao, A.C.S., Kulkarni, C.: A survey on sentiment analysis methods, applications, and challenges. Artif. Intell. Rev. **55**(7), 5731–5780 (2022)
20. Wei, Y., Li, S., Feng, R., Hu, D.: Diagnosing and re-learning for balanced multimodal learning. In: Proceedings of the European Conference on Computer Vision, pp. 71–86 (2025)
21. Xu, W., Jiang, H., Liang, X.: Leveraging knowledge of modality experts for incomplete multimodal learning. In: Proceedings of the ACM International Conference on Multimedia, pp. 438–446 (2024)
22. Yang, D., Huang, S., Kuang, H., Du, Y., Zhang, L.: Disentangled representation learning for multimodal emotion recognition. In: Proceedings of the ACM International Conference on Multimedia, pp. 1642–1651 (2022)
23. 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)
24. Zadeh, A., Chen, M., Poria, S., Cambria, E., Morency, L.P.: Tensor fusion network for multimodal sentiment analysis. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1103–1114 (2017)
25. Zadeh, A., Zellers, R., Pincus, E., Morency, L.P.: Multimodal sentiment intensity analysis in videos: facial gestures and verbal messages. IEEE Intell. Syst. **31**(6), 82–88 (2016)

26. 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 Annual Meeting of the Association for Computational Linguistics, pp. 2236–2246 (2018)
27. Zellinger, W., Grubinger, T., Lughofer, E., Natschläger, T., Saminger-Platz, S.: Central moment discrepancy (CMD) for domain-invariant representation learning. arXiv preprint arXiv:1702.08811 (2017)