Tackling Modality Heterogeneity with Multi-View Calibration Network for Multimodal Sentiment Detection

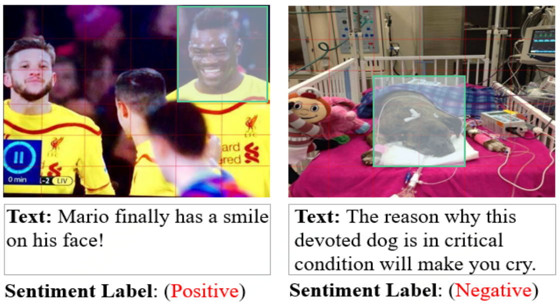
Yiwei Wei[1,4]*, **Shaozu Yuan**[2]*†, **Ruosong Yang**[5], **Lei Shen**[2]
Longbiao Wang[1,3]†, **Zhangmeizhi Li**[4], **Meng Chen**[2]
[1]Tianjin Key Laboratory of Cognitive Computing and Application,
College of Intelligence and Computing, Tianjin University, Tianjin, China
[2]JD AI Research, Beijing, China
[3]Huiyan Technology (Tianjin) Co., Ltd, Tianjin, China
[4]China University of Petroleum(Beijing) at Karamay, Karamay, China
[5]The Hong Kong Polytechnic University, Hongkong, China

Abstract

With the popularity of social media, detecting sentiment from multimodal posts (e.g. image-text pairs) has attracted substantial attention recently. Existing works mainly focus on fusing different features but ignore the challenge of **modality heterogeneity**. Specifically, different modalities with inherent disparities may bring three problems: 1) introducing redundant visual features during feature fusion; 2) causing feature shift in the representation space; 3) leading to inconsistent annotations for different modal data. All these issues will increase the difficulty in understanding the sentiment of the multimodal content. In this paper, we propose a novel Multi-View Calibration Network (MVCN) to alleviate the above issues systematically. We first propose a text-guided fusion module with novel Sparse-Attention to reduce the negative impacts of redundant visual elements. We then devise a sentiment-based congruity constraint task to calibrate the feature shift in the representation space. Finally, we introduce an adaptive loss calibration strategy to tackle inconsistent annotated labels. Extensive experiments demonstrate the competitiveness of MVCN against previous approaches and achieve state-of-the-art results on two public benchmark datasets.

1 Introduction

Multimodal sentiment detection(Xu, 2017; Xu and Mao, 2017) aims to explore the sentiment embedded in the multimodal contents, such as text, images, and videos. With the growth of social media, it shows broad applications for the understanding of one's position, attitude, or opinion towards an entity, person, or topic, which has attracted substantial attention from both academic and industrial communities. In this paper, we focus on detecting the sentiment of multimodal posts in social media. As shown in Figure 1, the model is required to infer the human sentiments given the image and text pairs.


Text: Mario finally has a smile on his face!
Sentiment Label: (Positive)

Text: The reason why this devoted dog is in critical condition will make you cry.
Sentiment Label: (Negative)

Figure 1: Examples of multimodal sentiment detection.

Previous works mainly focus on how to integrate modalities and have achieved astonishing progress. Earlier works apply concatenation to fuse different features, such as Xu (2017); Xu and Mao (2017). To promote the fusion, Yang et al. (2020); Yu and Jiang (2019); Kumar and Vepa (2020); Xu et al. (2018) construct different modules to realize deeper multimodal interaction. However, these methods of viewing different modalities in the same light neglect the **modality heterogeneity**, hindering the performance of the model. Although CLMLF (Li et al., 2022) tries to apply contrastive learning to alleviate this problem, it is in coarse granularity and only considers feature level, which is insufficient to address the issue of feature shift.

Modality heterogeneity is mainly caused by the modality gap, which has been discussed in some related multimodal research (Hazarika et al., 2020; Lin and Hu, 2022; Liang et al., 2022; Radford et al., 2021). In multimodal sentiment detection, it is because image modality often has less information and more redundant features compared to text modality. As shown in Figure 1, only a few visual

*Both authors contribute equally to this work.

†Corresponding authors. Email:{yuanshaozu@jd.com, longbiao_wang@tju.edu.cn.}

elements reveal sentiment information(e.g., people smile, ill dog), causing redundant visual elements and low information density for the sentiment. In contrast, text modality is more indicative of sentiment owing to higher information density (Tsai et al., 2019; Sun et al., 2021). Consequently, this heterogeneity across different modalities will bring difficulty in understanding the sentiment of multimodal content.

Concretely, the ignorance of modality heterogeneity will directly cause three problems: a) It takes sparse and noisy visual features into fusion and causes confusion in understanding sentimental information. b) Integrating modalities of different properties will further lead to the multimodal feature shift and makes the model capture spurious correlations between multimodal features and sentiments. c) As mentioned by Niu et al. (2016); Yang et al. (2021), it will affect the data annotators to vote inconsistent labels for unimodal, which leads to uncertain annotated labels and weaken the label confidence. Taking MVSA-Single (Niu et al., 2016) for example, approximately 47% of the samples suffer from inconsistent annotated labels.

To tackle the above problems systematically, we propose a Multi-View Calibration Network (MVCN) from three different views:

(1) To avoid sparse and redundant visual features of direct integrating modalities (Li et al., 2022), we propose a Text-Guided Fusion (TGF) module to leverage text data to dominate the fusion process. Specially, we propose Sparse-Attention mechanism using sparsemax (Martins and Astudillo, 2016) to automatically eliminate redundant visual features and capture the essential parts of the image with respect to the sentiment.

(2) To further calibrate the feature shift, we propose a Sentiment-based Congruity Constraint (SCC) task to restrain representation space. In the SCC task, we propose relative distance to gather multimodal features around the corresponding sentimental centroids estimated with samples' labels. In addition, to overcome the limitation of the mini-batch, we also introduce Accumulating Calibration (AC) strategy to accumulate sampling information, thus computing the sentiment centroids from a global perspective. Compared to contrastive learning (Li et al., 2022), SCC has more strength to calibrate the feature shift for it brings sentimental semantic labels from a global perspective.

(3) To alleviate uncertain annotated labels misleading the model during the training stage, we also introduce an adaptive loss calibration (ALC) strategy to calibrate the training loss in the sentiment detection task, where the detection model is forced to be less confident for uncertain annotated labels. The experiments conducted on different benchmark datasets (Niu et al., 2016; Cai et al., 2019) show that MVCN has significantly improved the performance on all metrics compared with previous state-of-the-art models. In summary, the key contributions of this paper are as follows:

- We introduce a novel Multi-View Calibration Network (MVCN) including Text-Guided Fusion module, Sentiment-based Congruity Constraint, and Adaptive Loss Calibration strategy for multimodal sentiment detection to systematically solve the problems of modality heterogeneity from different views.
- The thorough experiments show that MVCN improves the performance over all metrics and achieves state-of-the-art on two benchmark datasets (Niu et al., 2016; Cai et al., 2019).

2 Methodology

The architecture of the proposed multi-view calibration network (MVCN) is shown in Figure 2. Generally, MVCN consists of text-guided fusion module and two paralleled sub-tasks. The two paralleled sub-tasks are sentiment classification and sentiment-based congruity constraint respectively.

2.1 Text-Guided Fusion Module

As shown in Figure 2, Text-Guided Fusion Module contains three components: Unimodal Encoders, Text-Guided Unit, and Reduction Unit.

Unimodal Encoders. To obtain visual and textual features for multimodal fusion, we apply two different unimodal encoders to extract their representations. For text modality, we use the pretrained BERT (Jacob Devlin, 2019) model as the text encoder to obtain the text representation. Given a sequence of text $T = \{t_1, t_2, ..., t_{n_t}\}$, where n_t is the number of text length, the output of the BERT model can be defined as:

$$X_t = \{E_C, E_1, E_2, ..., E_{n_t}\} = BERT(T; \theta_t^{bert}) \quad (1)$$

where $E_C \in \mathbb{R}^{d_t}$ is the embedding of the CLS token and θ_t^{bert} denotes the parameters of the BERT model. For image modality, we use the pretrained

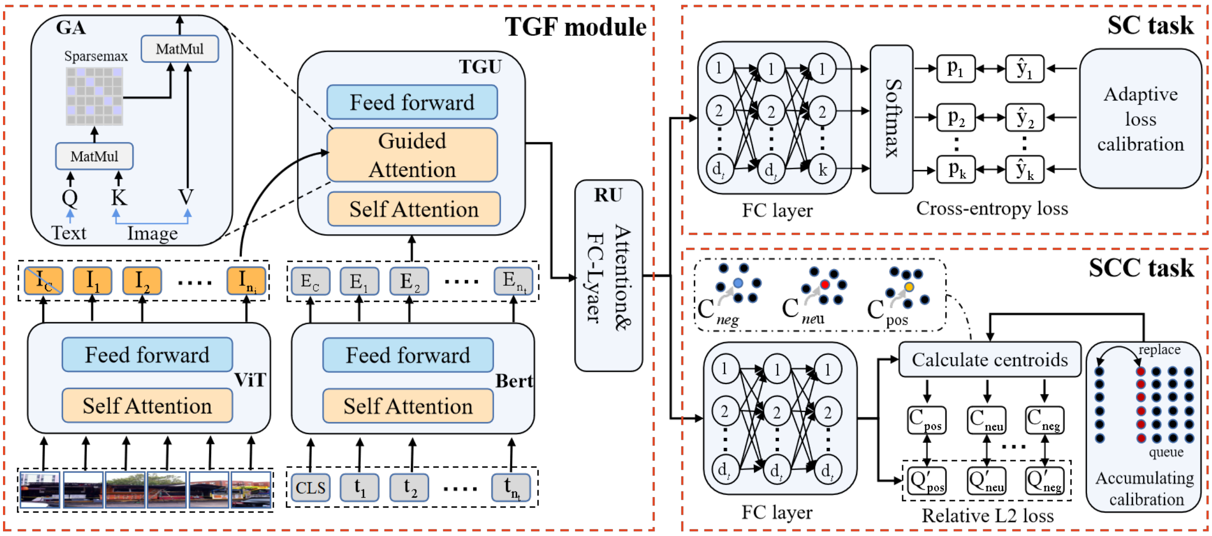


Figure 2: The overall architecture of the proposed model MVCN.

ViT model (Dosovitskiy et al., 2020) as the image encoder to obtain the image representation. Given an image I, the output of the ViT model can be defined as:

$$X_i = \{I_C, I_1, I_2, ..., I_{n_i}\} = ViT(I; \theta_i^{vit}) \quad (2)$$

where n_i is the length of the image representation and θ_t^{ViT} denotes the parameters of the ViT model.

Text-Guided Unit (TGU). Taking the extracted image features X_i and the text features X_t as inputs, we perform multimodal fusion by passing the inputs through N layers stacking Text-Guided Unit (TGU). For each layer of TGU, X_t is first fed into a self-attention (α) as query, key, and value to generate text-aware feature X_f. We then input X_i and X_f to a Sparse-Attention (γ) layer to obtain text-guided visual sparse features X_g, and finally apply a feed-forward network (FFN) for X_g to produce the TGU output. Here, we design Sparse-Attention in two places: (1) We use the textual features to attend to visual features and get text-guided features. (2) We normalize attention weight with sparsemax (Martins and Astudillo, 2016), thus obtaining sparse posterior attention weights, where the weight for redundant visual features are set to 0. Moreover, inspired by previous works (Rahman et al., 2019; Yuan et al., 2022) that leverage pretrained language model to enhance the ability of capturing multimodal context, we initialize TGU with the weight of the pretrained BERT. In the formula, the output of N^{th} layer TGU can be defined as: $TGU(X_t, X_i)^N = [X_f = \alpha(X_t, X_t, X_t), X_g = \gamma(X_f, X_i, X_i), FFN(X_g)]$.

Reduction Unit (RU). To get the multimodal representation $Q \in \mathbb{R}^{d_t}$ for sentiment classification, we follow Li et al. (2022) and use stacked attention layer and fully-connected layer with GELU activate function (Hendrycks and Gimpel, 2016) as Reduction Unit to perform dimensionality reduction for the features.

2.2 Sentiment Congruity Constraint (SCC)

Intuitively, samples' features Q with the same sentimental labels should be closer in the representation space. However, feature shift caused by modality heterogeneity hinders this trend and makes it difficult to capture correlations between multimodal features and sentiments. To address this problem, we take advantage of label information to estimate sentimental centroids, then restrict the distance between the samples and their centroids, thus calibrating the feature shift.

Here, we first employ a fully-connected layer to map multimodal representation Q to normalized representations $\acute{Q}$ for the SCC task. Then, we compute the positive centroid (C_0), the neutral centroid (C_1) and the negative centroid (C_2) of the representation $\acute{Q}$ during training stage. Let 0,1,2 denote the positive label, neutral label, and negative label, the formula for computing centroids is defined as:

$$C_{(0|1|2)} = \frac{\sum_{j=1}^{B} I(Y(j) = (0|1|2)) \cdot \acute{Q}_j}{\sum_{j=1}^{B} I(Y(j) = (0|1|2))} \quad (3)$$

where B is the number of all the candidate computing samples, $I(.)$ is an indicator function, and $\acute{Q}_j$ and $Y(j)$ are the representation and Ground-Truth label respectively of the j-th sample.

To restrict the distance between the samples and their corresponding centroids, an intuitive idea is to minimize absolute distance, such as L2 loss. However, we find it causes all the samples to get too close to their centroids and completely eliminates data distribution, which makes the training hard to converge. To tackle this issue, we propose a relative L2 loss to measure the distance. Here, we first calculate the absolute L2 distance $D = \{d_1, ..., d_B\}$ between the centroid C and the multimodal representation $\acute{Q}$ for a batch:

$$D = \frac{\{\|\acute{Q}_i - C_{Y(i)}\|_2^2\}_{i=1}^{B}}{\sqrt{\iota}} \tag{4}$$

where $Y(i)$ represents the ground-truth label of the i^{th} sample, ι is the feature dimension, and $\sqrt{\iota}$ serve as a scale factor. We then normalize the absolute distance to a relative distance and optimize the SCC task with the following formula:

$$L_{scc} = -\sum_{i=1}^{B} log(\frac{exp(-d_i)}{\sum_{i=1}^{B} exp(-d_i)}) \tag{5}$$

Accumulating calibration strategy. Nevertheless, there still exist two problems to optimize the SCC task because of the limitation of the mini-batch. On the one hand, computing centroid for iterating mini-batch results in a frequently updated centroid. On the other hand, the samples in a mini-batch are insufficient to estimate an accurate centroid. To solve this problem, we propose an accumulating calibration strategy to enlarge the computing space and narrow the change for the samples.

We first employ an auxiliary accumulating TGF module (denoted as TGF') to produce sufficient representations Q_m as candidate computing samples for centroids in advance. To accumulate computing information, we then build a queue to restore all the representations. The queue is dynamically updated by replacing the premier mini-batch with the current mini-batch during training for each iteration. Thus, we can estimate stable and slow-updated centroids with Equation 3 from a more global perspective. Here, to guarantee training stability, we leverage the momentum optimization (He et al., 2020; Li et al., 2021) to slowly update TGF', which can be defined as:

$$\theta_m \leftarrow \beta\theta_m + (1-\beta)\theta_t \tag{6}$$

where θ_m denotes the parameters of the accumulating TGF', θ_t denotes the parameters of the TGF module and $\beta \in [0, 1)$ is a balance parameter.

2.3 Sentiment Classification (SC)

For the sentiment classification task, we feed multimodal representation $Q \in \mathbb{R}^{d_t}$ into the fully connected layer with softmax function to predict the logits. However, directly optimizing this task with Cross Entropy (CE) as previous work still suffers from the issue of uncertain annotated labels during the training stage.

Adaptive loss calibration (ALC) strategy. To overcome the above issue, we design a simple but effective strategy, adaptive loss calibration. It forces the detection model to decrease confidence for the training examples of inconsistent annotated labels, thus calibrating the loss. To better elaborate the strategy, we first define cross entropy loss utilized in most previous works. Suppose that $p_i \in \{p_1, p_2, ..., p_K\}$ denotes the predicted logit and $y_i \in \{y_1, y_2, ..., y_K\}$ represents the ground-truth, CE loss can be defined as $L = -\sum_i^K y_i log(p_i)$, where $y_i \in \{0\ ,1\}$, $p_i \in [0\ ,1]$ and K denotes the number of categories. Different from label smoothing (Szegedy et al., 2016), we first leverage unimodal labels provided by the datasets to adaptively adjust the confidence factor $\alpha \in \{0, 0.1\}$, then normalize the ground-truth probability. Intuitively, it makes the model becomes confident about its ground-truth by setting α as 0. In the formula, we normalize the ground-truth probability $\hat{y}_i$ for each label as:

$$\hat{y}_i = (1-\alpha) \cdot I(y_i = 1) + (\frac{\alpha}{K-1}) \cdot I(y_i = 0) \tag{7}$$

where $I(.)$ is a indicator function. Moreover, the loss function L_{sc} can be reformatted as:

$$\begin{aligned} L_{sc} = \sum((1-\alpha) * L_i \cdot I(y_i = 1) \\ + \alpha \cdot L_i \cdot I(y_i = 0)) \end{aligned} \tag{8}$$

2.4 Training Loss

We optimize the above two tasks with total loss L:

$$L = \lambda_{sc}L_{sc} + \lambda_{scc}L_{scc} \tag{9}$$

where λ_{sc} and λ_{scc} are hyper-parameters to balance the different training losses.

3 Experiments

In this section, we first introduce the experimental setup and report the experimental results, then conduct the ablation study and visualization analysis.

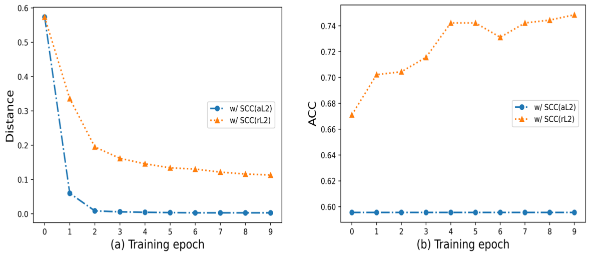


Figure 3: The training curves for the average semantic distance between the samples and corresponding centroids in (a), and the accuracy in (b) with different constraint methods. Here, aL2 and rL2 represent optimizing SCC with absolute L2 and relative L2.

3.1 Datasets

All our experiments were conducted on three public datasets: MVSA-Single (Niu et al., 2016), MVSA-Multiple (Niu et al., 2016) and HFM (Cai et al., 2019). Here, we give a brief introduction to these datasets and dataset statistics are shown in Table 2.

MVSA-Single, **MVSA-Multiple.** MVSA-Single and MVSA-Multiple are popular text-image sentiment datasets crawled from Twitter, where MVSA-Multiple is an upgraded version of MVSA-Single and contains more text-image pairs. Both of them have three categories: positive, neutral, and negative. For fair comparison, the two MVSA datasets are processed in the same way as Xu and Mao (2017).

HFM. HFM has two sentimental categories: positive and negative. We follow Cai et al. (2019) and adopt the same preprocessing method for experiments, which has been widely used in previous works.

3.2 Implementation Details

For feature extraction, we use 12-layer visual transformer ViT-B/16 (Dosovitskiy et al., 2020) as visual encoder and BERT-base (Jacob Devlin, 2019) as textual encoder. The text-guided fusion encoder is composed of 6 stacked text-guided units. In ALC, the confidence factor is set as 0 or 0.1 according to unimodal labels. For Sentiment Congruity Constraint, the queue sizes of accumulating calibration are set as 3611, 13624, and 19816 for MVSA-Single, MVSA-Multiple, and HFM respectively.

During the training stage, the learning rate is set to 2e-5. We train the model for 40 epochs with batch size of 32, 64, and 64 for MVSA-Single, MVSA-Multiple, and HFM. In addition, we adopt AdamW optimizer with ϵ of $1e-8$ and β of $(0.9, 0.999)$. For loss function, both hyper-parameters λ_{sc} and λ_{cc} are set to 1.0. Following previous settings, we adopt ACC and Weighted F1 as the evaluation metrics for MVSA datasets and Macro-F1 and ACC for HFM to evaluate the performance of the model.

3.3 Baselines

To fully validate the performance of MVCN, we select both unimodal and multimodal baselines.

Unimodal Baselines. For text modality, we choose **CNN** (Kim, 2014), **Bi-LSTM** (Zhou et al., 2016) and **BERT** (Jacob Devlin, 2019) as baselines since they are popular models for text classification. For image modality, **ResNet** (He et al., 2016) and **ViT** (Dosovitskiy et al., 2020) are selected for their superior capability for image classification.

Multimodal Baselines. For MVSA-$*$ datasets, the compared baselines include: **MultiSentiNet** (Xu and Mao, 2017) that designs an attention-based semantic network for multimodal sentiment analysis; **HSAN** (Xu, 2017) applying a hierarchical semantic attentional network for multimodal sentiment analysis; **Co-MN-Hop6** (Xu et al., 2018), employing a co-memory network to iteratively model the interactions between multiple modalities; **MGNNS** (Yang et al., 2021) utilizing a multi-channel graph neural networks with sentiment-awareness for image-text sentiment detection. **CLMLF** (Li et al., 2022) is the previous SOTA model that aligns and fuses the text and image modalities through contrastive learning. For HFM datasets, we compare two variants of **Concat** (Schifanella et al., 2016): Concat(2) concatenates text and image, while Concat(3) brings extra image attribute features; **MMSD** (Cai et al., 2019) fusing text, image, and image attributes with a multimodal hierarchical framework; and ***D&R* Net** (Xu et al., 2020) that fuses text, image, and visual attributes by the decomposition and relation network.

3.4 Main Results

The comparison between MVCN and the baselines is demonstrated in Table 1. Obviously, the text-only model is more competitive with image-only methods. It indicates that the image modality is less helpful and contains more redundant features compared to text modality, which supports our intuition of eliminating redundant visual features. In

[1] Considering CLMLF brings extra data for data augmentation, to evaluate capacity of the model, we also report the results without extra data for fair comparison.

Modality	Model	MVSA-Single		MVSA-Multiple		Model	HFM	
		Acc	F1	Acc	F1		Acc	F1
Text	CNN	0.6819	0.5590	0.6564	0.5766	CNN	0.8003	0.7532
	BiLSTM	0.7012	0.6506	0.6790	0.6790	BiLSTM	0.8190	0.7753
	BERT	0.7111	0.6970	0.6759	0.6624	BERT	0.8389	0.8326
Image	ResNet-50	0.6467	0.6155	0.6188	0.6098	ResNet-50	0.7277	0.7138
	ViT	0.6378	0.6226	0.6194	0.6119	ViT	0.7309	0.7152
Multimodal	MultiSentiNet	0.6984	0.6984	0.6886	0.6811	Concat(2)	0.8103	0.7799
	HSAN	0.6988	0.6690	0.6796	0.6776	Concat(3)	0.8174	0.7874
	Co-MN-Hop6	0.7051	0.7001	0.6892	0.6883	MMSD	0.8344	0.8018
	MGNNS	0.7377	0.7270	**0.7249**	0.6934	D&R Net	0.8402	0.8060
	CLMLF	0.7533	0.7346	0.7200	0.6983	CLMLF	0.8543	0.8487
	CLMLF[1]	0.7378	0.7291	0.7112	0.6863	CLMLF[1]	0.8489	0.8446
	MVCN	**0.7606**	**0.7455**	0.7207	**0.7001**	MVCN	**0.8568**	**0.8523**

Table 1: Experimental results of different models on MVSA-Single, MVSA-Multiple and HFM datasets.

Dataset	Label	Train	Val	Test
MSVA-S	Positive	2147	268	268
	Neutral	376	47	47
	Negative	1088	135	135
MSVA-M	Positive	9056	1131	1131
	Neutral	3528	440	440
	Negative	1040	129	129
HFM	Positive	8642	959	959
	Negative	11174	1451	1450

Table 2: Dataset statistics for MVSA-Single, MVSA-Multiple and HFM.

addition, the multimodal models surpass the unimodal models because of fusing more information. Overall, MVCN achieves state-of-the-art with a considerable performance gain over other methods, which indicates the necessity of tackling modality heterogeneity from different views. Specially, we find that MVCN achieves better results on MSVA-Single compared to the other two datasets. We conjecture that small dataset suffers more of the modality heterogeneity problem due to lack of data diversity.

3.5 Ablation Study

To investigate the effectiveness of each module, we conduct an ablation study in Table 3. Firstly, compared to the MFS model that equally fuses the image and text features, it is straightforward that TGF module can aid sentiment detection since it eliminates redundant visual features. In addition, the model equipped with Sentiment-based Congruity Constraint (SCC) brings a significant improvement, implying the importance of calibrating feature shift with congruity constraint. And accumulating calibration (AC) strategy by additionally augmenting SCC with more accurate and stable centroids, consistently improves performance. Furthermore, it can be observed that ALC strategy can further boost performance, demonstrating ALC is an effective way to reduce the impact of the uncertain annotated labels. Finally, MVCN equipped with all novel modules achieves the best performance, illustrating the effectiveness of all the above modules.

4 Analysis

4.1 Discussion of Components

Variants of Text-Guided Unit. To verify the design for Text-Guided Unit in TGF, we evaluate the performance of TGU variants in Table 4. Here, “w/ softmax” denotes Self-Attention, while “w/ spm” represents we normalize attention weight with sparsemax. “w/ pretrain” and “w/o pretrain” indicate whether the guided unit initialized with the pretrained BERT weight or not. From the table, we observe “w/ spm” shows superiority compared to “w/ softmax”, indicating Sparsemax-Attention boosts the performance of the model by eliminating noisy visual features. Additionally, “w/ pretrain” can further promote the performance, which is consistent with previous works that pretrained language models can enhance the capacity for capturing multimodal context.

Effectiveness of SCC. In Figure 3, we plot the curves during training to explore why we apply relative L2 distance to optimize SCC. From Figure 3 (a), “w/ SCC(aL2)” that applies absolute L2 dis-

Model	MVSA-Single		MVSA-Multiple		HFM	
	Acc	F1	Acc	F1	Acc	F1
BERT	0.7111	0.6970	0.6759	0.6624	0.8389	0.8326
ViT	0.6378	0.6226	0.6194	0.6119	0.7309	0.7152
MFS	0.7217	0.7205	0.7063	0.6851	0.8434	0.8375
TGF	0.7396	0.7355	0.7095	0.6887	0.8493	0.8451
TGF, SCC	0.7515	0.7403	0.7158	0.6929	0.8522	0.8499
TGF, SCC+AC	0.7563	0.7422	0.7188	0.6971	**0.8568**	**0.8523**
TGF, SCC+AC, ALC	**0.7606**	**0.7455**	**0.7207**	**0.7001**	-	-

Table 3: Ablation results of our MVCN. Here, MFS denotes equally fusing the image and text features as Li et al. (2022). And TGF, SCC, AC, ALC are respectively our four designs: text-guided fusion module, sentiment-based congruity constraint task, accumulating calibration strategy, and adaptive loss calibration strategy. Note that the experiment for the ALC model on the HFM dataset is missing due to the absence of the unimodal labels.

TGF	MVSA-Single		HFM	
	Acc	F1	Acc	F1
w/ softmax	0.7365	0.7339	0.8476	0.8427
w/ spm	0.7396	0.7355	0.8493	0.8451
w/o pretrain	0.7232	0.7148	0.8306	0.8217
w/ pretrain	0.7396	0.7355	0.8493	0.8451

Table 4: Ablation results of TGF and we remove ALC and SCC modules here.

Dataset	Strategy	ACC	F1
MSVD-Single	-	0.7563	0.7422
	LS	0.7579	0.7436
	ALC	0.7606	0.7455

Table 5: Results with adaptive loss calibration (ALC) and label smoothing (LS). Note − represents the model without any strategy.

tance dramatically narrows the distance between the samples and the centroids to an extremely low level. However, it hinders the training and eventually leads to the model being hard to optimize in Figure 3 (b). Because it will completely eliminate data distribution and make the data samples lose correlation with each other by directly minimizing L2 distance. Conversely, "w/ SCC(aL2)" that utilizes relative L2 distance can reduce the semantic distance to a reasonable extent, and more importantly, maintain the data distribution. Therefore, it ensures the SCC task can be optimized and the accuracy of "w/ SCC(rL2)" in Figure 3 (b) gradually increases during training.

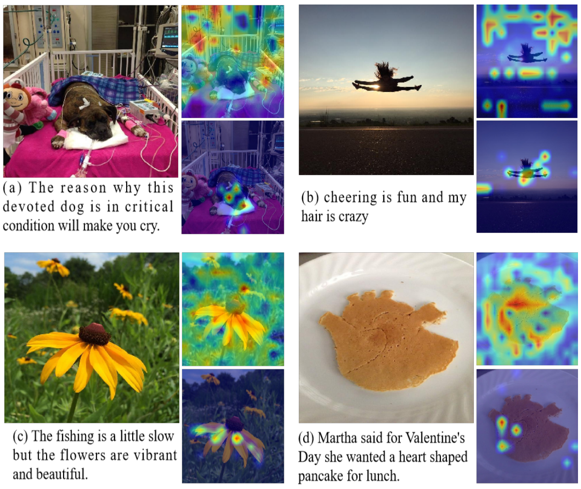
Different Loss Calibration Strategies. To explore the effectiveness of ALC, we compare different loss calibration strategies in Table 5. It shows label smoothing (LS) only brings slight improvement since it avoids overfitting and strengthens the robustness of the model. However, LS can not handle the label problem caused by modality heterogeneity. Compared to the LS, our proposed ALC exhibits superiority and outperforms LS in performance, verifying it is necessary to mitigate the adverse effects of inconsistent labels.

4.2 Visualization

Sparse-Attention Visualization. To verify the advantage of Sparse-Attention in the TGF module, we visualize the attention weight in Figure 4. Compared to Self-Attention, the sampling cases show Sparse-Attention captures the essential parts of the image with respect to the sentiment and meanwhile attenuates the negative effect of redundant visual features. As the example in Figure 4(a), the model paid more attention on the "ill dog" in the image for it reflects negative sentiment, certifying that the model can focus the sentimental regions in the image and avoid the interference of irrelevant objects. This also confirms it is necessary to eliminate redundant visual features, reinforcing the importance of Sparse-Attention.

Feature Distribution Visualization. In order to visually demonstrate the superiority of SCC task with AC strategy, we visualize the feature distribution on the Multiple-Single dataset with contrastive learning (Li et al., 2022) and SCC. Here, we apply T-SNE[2] algorithm to perform dimensionality reduction for the feature, obtaining a 2-dimensional feature vector distribution visualized in Figure 5. From Figure 5(b), we observe that the SCC task forces samples belonging to the same category to

[2] https://github.com/mxl1990/tsne-pytorch



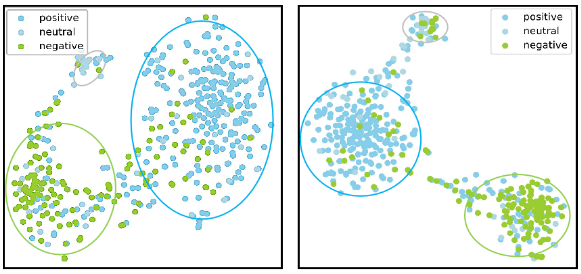
(a) The reason why this devoted dog is in critical condition will make you cry.

(b) cheering is fun and my hair is crazy

(c) The fishing is a little slow but the flowers are vibrant and beautiful.

(d) Martha said for Valentine's Day she wanted a heart shaped pancake for lunch.

Figure 4: Attention visualization for sampling cases with Self-Attention (the upper one) and Sparse-Attention in TGF (the lower one).



(a) Without SCC (b) With SCC

Figure 5: Visualization of representation. Different colored dots represent samples with different categories.

gather around their corresponding centroids. Conversely, Figure 5(a) shows that when we remove SCC task, the degree of data aggregation is less obvious. The reason for this phenomenon is that, compared to contrastive learning, the SCC taking sentimental labels into consideration to constrain the distribution from a more global perspective, is more capable of calibrating the feature shift. Hence, this strength facilitates the MVCN to learn the distinguished features in representation space and improve the performance of the model.

4.3 Case Study

In Figure 6, we conduct a case study for MVCN and the previous SOTA model CLMLF. We find that for these complicated and confusing cases, it is difficult for CLMLF to capture the user's sentiment of samples because of its limitations in handling modality heterogeneity. For example, in the first case, although the phrases (e.g., vibrant colors) in texts represent specific sentiments, CLMLF was still misled by sparse visual sentimental information. For the following two cases in Figure 6, we find the text may express negative sentiments, and images contain elements (e.g., serious man, smiling man) that reflect sentiments. However, CLMLF fails to predict the real sentiment due to modality heterogeneity. In contrast, our model can distinguish these cases for the corresponding sentiment, which also proves the advantage of tackling modality heterogeneity from different views.

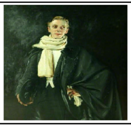
Image	Text	MVCN	CLMLF
	Why buy ordinary art when you can have your art on sailcloth. vibrant colors!!	Positive	Negative
	My job is not to simply defend a position when having a heated discussion.	Negative	Neutral
	William Bruce Ellis Ranken's ebullient portrait of the young Ernest Thesiger. Born this day in 1879.	Positive	Neutral

Figure 6: Case Study of MVCN and CLMLF.

5 Related Work

In this section, we introduce the background of multimodal sentiment detection and modality heterogeneity.

5.1 Multimodal Sentiment Detection

Multimodal sentiment detection has become a significant research topic and previous works mainly focus on fusing multimodal features with different strategies. Some earlier works fuse the multimodal features by concatenation, such as HSAN (Xu, 2017) and MultiSentiNet (Xu and Mao, 2017). Later relevant works mainly focus on how to integrate modalities and promote better inter-modal interaction. For example, CoMN (Xu et al., 2018) designs a co-memory network to iteratively model the interactions between image and text features, TomBERT (Yu and Jiang, 2019) fuses the multimodal representations with a bilinear interaction layer, Kumar and Vepa (2020) proposes to use gating mechanism and attention to perform deep multimodal interaction, and MVAN (Yang et al., 2020) adapts memory network to fuse the multimodal features via perceptron and stacking-pooling module. Recent work (Li et al., 2022) applies contrastive learning and data augmentation to handle this problem. Different from these works, we aim to improve this task by tackling modality heterogeneity from multiple views.

5.2 Modality Heterogeneity

Modality heterogeneity is reflected in that multiple modalities show different properties. Typically, heterogeneity usually exists in human-generated signals (language) with high information density and other natural signals modalities (e.g., image, video, audio) with heavy redundancy. In the past few years, some works have been done in multimodal representation learning to alleviate the impact of modality heterogeneity. Hazarika et al. (2020) project different modalities to modality-invariant representation space and learn their commonalities. Furthermore, Wu et al. (2021); Zhao et al. (2021) employs the modality translation method to convert the source modality to the target one in order to learn the commonalities between the different modalities. However, due to the huge modality gap, it is insufficient to handle multimodal heterogeneity by projecting different modal features into the same representation space. To overcome the modality gap and encourage learning useful features, Li et al. (2022); Lin and Hu (2022) design various contrastive learning tasks to predict cross-modal representation in an implicit way. However, they are feature-level alignment and in coarse granularity, which is not sufficient to address the problem of feature shift.

6 Conclusion

In this paper, we present a Multi-View Calibration Network (MVCN) to address the modality heterogeneity problem for text-image sentiment detection from different views. Specially, we respectively introduce the text-guided fusion module to calibrate multimodal fusion and reduce the negative impacts of redundant visual elements, a sentiment-based congruity constraint task to further calibrate the feature shift in representation space and an adaptive loss calibration strategy to calibrate the training loss in terms of uncertain annotated labels. The thorough experiments show the MVCN achieves state-of-the-art performance on two benchmark datasets.

Limitations

Considering Modality Heterogeneity can promote many related multimodal applications, it is worth continually exploring. In this paper, we propose Text-Guided Fusion (TGF) module equipped with Sparse-Attention to integrate different modalities in representation aspects, which is an implicit way to build the relations of fine-grained features, such as visual objects, and textual words. Previous work (Khademi, 2020; Wang et al., 2020) has proven that Graph Convolutional Network (GCN) (Scarselli et al., 2008) shows advantages in modeling the relations among visual and textual elements. Inspired by these works, we argue that explicitly introducing the relationship of fine-grained features via GCN can better guide the model to eliminate redundant features. Thus it can further narrow the modalities gap and facilitate fusion for multimodal content understanding. In the future, we will bring GCN to learn multimodal relationships and boost the performance of the model.

References

Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019. Multimodal sarcasm detection in twitter with hierarchical fusion model. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 2506–2515.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

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

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

Toutanova Kristina Toutanova Jacob Devlin, Ming-Wei Chang. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Mahmoud Khademi. 2020. Multimodal neural graph memory networks for visual question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7177–7188. Association for Computational Linguistics.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751.

Ayush Kumar and Jithendra Vepa. 2020. Gated mechanism for attention based multi modal sentiment analysis. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4477–4481. IEEE.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705.

Zhen Li, Bing Xu, Conghui Zhu, and Tiejun Zhao. 2022. CLMLF:a contrastive learning and multi-layer fusion method for multimodal sentiment detection. In *Findings of the Association for Computational Linguistics: NAACL 2022*. Association for Computational Linguistics.

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. 2022. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625.

Ronghao Lin and Haifeng Hu. 2022. Multimodal contrastive learning via uni-modal coding and cross-modal prediction for multimodal sentiment analysis. *arXiv preprint arXiv:2210.14556*.

Andre Martins and Ramon Astudillo. 2016. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *International conference on machine learning*, pages 1614–1623. PMLR.

Teng Niu, Shiai Zhu, Lei Pang, and Abdulmotaleb El Saddik. 2016. Sentiment analysis on multi-view social data. In *International Conference on Multimedia Modeling*, pages 15–27. Springer.

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

Wasifur Rahman, Md. Kamrul Hasan, Amir Zadeh, Louis-Philippe Morency, and Mohammed Ehsan Hoque. 2019. M-BERT: injecting multimodal information in the BERT structure. *CoRR*, abs/1908.05787.

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. 2008. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80.

Rossano Schifanella, Paloma De Juan, Joel Tetreault, and Liangliang Cao. 2016. Detecting sarcasm in multimodal social platforms. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 1136–1145.

Yang Sun, Nan Yu, and Guohong Fu. 2021. A discourse-aware graph neural network for emotion recognition in multi-party conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2949–2958.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

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

Youze Wang, Shengsheng Qian, Jun Hu, Quan Fang, and Changsheng Xu. 2020. Fake news detection via knowledge-driven multimodal graph convolutional networks. In *Proceedings of the 2020 on International Conference on Multimedia Retrieval, ICMR 2020, Dublin, Ireland, June 8-11, 2020*, pages 540–547. ACM.

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, Online. Association for Computational Linguistics.

Nan Xu. 2017. Analyzing multimodal public sentiment based on hierarchical semantic attentional network. In *2017 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 152–154. IEEE.

Nan Xu and Wenji Mao. 2017. Multisentinet: A deep semantic network for multimodal sentiment analysis. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 2399–2402.

Nan Xu, Wenji Mao, and Guandan Chen. 2018. A co-memory network for multimodal sentiment analysis. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 929–932.

Nan Xu, Zhixiong Zeng, and Wenji Mao. 2020. Reasoning with multimodal sarcastic tweets via modeling cross-modality contrast and semantic association. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 3777–3786.

Xiaocui Yang, Shi Feng, Daling Wang, and Yifei Zhang. 2020. Image-text multimodal emotion classification via multi-view attentional network. *IEEE Transactions on Multimedia*, 23:4014–4026.

Xiaocui Yang, Shi Feng, Yifei Zhang, and Daling Wang. 2021. Multimodal sentiment detection based on multi-channel graph neural networks. 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 328–339.

Jianfei Yu and Jing Jiang. 2019. Adapting bert for target-oriented multimodal sentiment classification. IJCAI.

Shaozu Yuan, Xin Shen, Yuming Zhao, Hang Liu, Zhiling Yan, Ruixue Liu, and Meng Chen. 2022. MCIC: multimodal conversational intent classification for e-commerce customer service. In *Natural Language Processing and Chinese Computing - 11th CCF International Conference, NLPCC 2022, Guilin, China, September 24-25, 2022, Proceedings, Part I*, volume 13551 of *Lecture Notes in Computer Science*, pages 749–761. Springer.

Jinming Zhao, Ruichen Li, and Qin Jin. 2021. Missing modality imagination network for emotion recognition with uncertain missing modalities. 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 2608–2618.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers)*, pages 207–212.