2023 IEEE International Conference on Multimedia and Expo (ICME) | 978-1-6654-6891-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICME55011.2023.00237

Multimodal Sentiment Analysis with Preferential Fusion and Distance-aware Contrastive Learning

Feipeng Ma, Yueyi Zhang, Xiaoyan Sun*
University of Science and Technology of China, Hefei, China
mafp@mail.ustc.edu.cn, {zhyuey, sunxiaoyan}@ustc.edu.cn

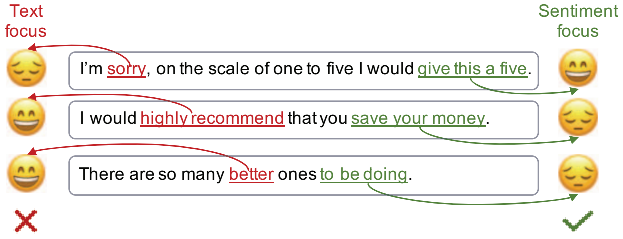
***Abstract*—Recent efforts on multimodal sentiment analysis (MSA) leverage data from multiple modalities, among which the text modality is heavily relied on. However, the text modality often contains false correlations between text tokens and sentiment labels, leading to errors in sentiment analysis. To address this issue, we propose a new framework, PriSA, which incorporates the preferential fusion and distance-aware contrastive learning. Specifically, we first propose a preferential inter-modal fusion method, which utilizes the text modality to guide the calculation of the inter-modal correlations. Then the resulting inter-modal features are further used to calculate mixed-modal correlations through our proposed distance-aware contrastive learning, which leverages the distance information of the sentiment labels. At last, we identify the sentiment information based on both the mixed-modal correlations and the discriminative intra-modal features extracted from the visual and audio modalities via a self-attention module. Experimental results show that our proposed PriSA achieves the state-of-the-art performance on four datasets, including MOSEI, MOSI, SIMS, and UR-FUNNY. The code is available at https://github.com/FeipengMa6/PriSA.**

***Index Terms*—sentiment analysis, multimodal, contrastive learning, attention, feature fusion**

I. Introduction

Over the last 20 years, Sentiment Analysis (SA) has been widely adopted in applications ranging from consumer products to national security. Instead of single modality (e.g. textual or acoustic information), recent efforts in SA pay much attention to multimodal sentiment analysis (MSA), in which data from multiple modalities are utilized to better understand human sentiments. Heterogeneity is universal among modalities. Each modality may reveal different sentiment information, with different information densities and different noise levels. Taking an MSA task with textual, visual, and audio inputs as an example, the same word with different gestures may suggest different attitudes, while the same word with different tones may suggest different emotions. Moreover, text modality is usually highly semantic and information-dense [1], while visual and audio modalities are relatively redundant in sentiment representation. Furthermore, each modality may be asynchronous in presenting the same sentiment.

Researchers have proposed a variety of fusion methods for MSA. In traditional MSA schemes, both early fusion [2] and late fusion [3], [4] approaches have been studied to integrate multimodal features. Later, deep neural networks have been adopted to extract and fuse features for MSA [5]–[7]. Inter-modal correlations have also been explored by learning features in pairs of modalities [8]. Recently, complex models with various guidances, such as attention [8], memory [9], and recurrent components [10] have been introduced to further enhance the performance of MSA.



I'm sorry, on the scale of one to five I would give this a five.

I would highly recommend that you save your money.

There are so many better ones to be doing.

Fig. 1. Three representative sentences are presented to demonstrate that excessive reliance on text modality may lead to errors in MSA. The red texts indicate the text focus which provide wrong sentiments. The green texts are the parts that should be paid attention to for correct sentiments.

Previous methods have made significant progress in exploiting the complementary and shared information of different modalities, but often rely heavily on text and may learn spurious correlations between text and sentiment [11]. As shown in Fig. 1, the red text "sorry" is often associated with negative sentiments, while "highly recommend" and "better" are usually associated with positive sentiments. However, they do not reflect the true sentiment in these sentences. And over relying on the text modality may lead to the model focusing on these spurious sentiment words and neglect the true sentiment parts that should be paid attention to, such as the green texts. Sun *et al.* [11] noticed this problem and proposed a counterfactual framework to subtract direct effect of textual modality. Different from them, we consider this problem from a fusion perspective.

In this paper, we propose a **pr**eferential fus**i**on strategy for **MSA** and present a new framework ***PriSA*** to address the issues. In our preferential fusion strategy, illustrated in Fig. 2 (b), we use the text modality as the primary modality to implicitly guide the inter-modal learning. A transformer-based attention module is devised to calculate the inter-modal correlation. The input key-value pairs and queries of the transformer-based module [12] come from the text modality and other modalities, respectively. In our framework, text modality is no longer used directly as evidence for sentiment analysis. Instead, it serves as an implicit guide for inter-modal learning with other modalities. This means that our

* The corresponding author: sunxiaoyan@ustc.edu.cn

1945-788X/23/$31.00 ©2023 IEEE
DOI 10.1109/ICME55011.2023.00237

approach removes the direct influence of text on the sentiment analysis and uses text in conjunction with other modalities. The resulting inter-modal features are further utilized to generate the mixed-modal correlations through distance-aware contrastive learning. Along with the deep intra-modal features extracted from each secondary modality, our proposed PriSA outperforms state-of-the-art methods on four datasets.

The main contributions are summarized as follows:

(1) We propose the PriSA framework and introduce the preferential fusion strategy for MSA.

(2) We propose distance-aware contrastive learning, which incorporates distance information of sentiment labels to explore mixed-modal correlations among three modalities.

(3) Experimental results show that our method achieves state-of-the-art performance on four benchmark datasets, including MOSEI, MOSI, SIMS and UR-FUNNY.

II. Related Work

Methods utilized in MSA can be broadly categorized into two groups: representation learning and multimodal fusion. With respect to representation learning, Yu *et al.* [7] designed a self-supervised label generation module to acquire independent unimodal supervision. Hazarika *et al.* [5] use both the similarity loss and difference loss to project each modality to two distinct subspaces, the modality-invariant and modality-specific subspaces. Tensor-based fusion methods fuse different modalities by tensor fusion network. For attention-based fusion, Tasi *et al.* [8] proposed a directional pairwise cross-modal attention method that attends to the interactions between multimodal sequences.

Contrastive learning is an emerging method for self-supervised learning. Its fundamental concept involves bringing the anchor and positive samples closer, while pushing away the anchor and negative samples. In the MSA task, recent works adopt supervised contrastive learning to explore the interactions between or among different modalities. To reduce modality gap in MSA, Mai *et al.* [13] proposed a hybrid contrastive learning framework. Similarly, Lin *et al.* [14] proposed a novel hierarchical graph contrastive learning framework, which performs intra-modal and inter-modal graph contrastive learning.

III. Method

In this section, we will first present the overview of our PriSA framework. Then several modules including the preferential fusion, distance-aware contrastive learning and objective function are described in detail.

A. Overview

An overview of our PriSA framework is shown in Fig. 2 (a). Given the three input modalities, we first extract deep features of each modality with the corresponding pretrained feature extractors. Then, feature-level augmentation is performed on all the secondary modalities to facilitate the subsequent contrastive learning. Next, the augmented deep features are fed into the preferential fusion module, in which the inter-modal correlations are exploited between the primary modality (text) and each secondary modality (audio or visual). The resulting inter-modal features are further delivered to the distance-aware contrastive learning to learn the mixed-modal correlations. Finally, we identify the sentiment information based on both the mixed-modal features and the intra-modal features, which are extracted from secondary modalities by their private encoders.

B. Preferential Fusion

Our preferential fusion module, shown in Fig. 2 (b), utilizes the primary modality, i.e. the text modality, to implicitly guide the inter-modal learning. We perform the inter-modal fusion with a cross-attention module and a transformer encoder [12]. The cross-attention function has three input matrices, Q as query, K as key, and V as value. The standard attention function is formulated as

$$Att(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \sigma\left(\frac{\boldsymbol{Q}\boldsymbol{W}^{(Q)}(\boldsymbol{K}\boldsymbol{W}^{(\boldsymbol{K})})^{\mathsf{T}}}{\sqrt{d}}\right)\boldsymbol{V}\boldsymbol{W}^{(V)}, \quad (1)$$

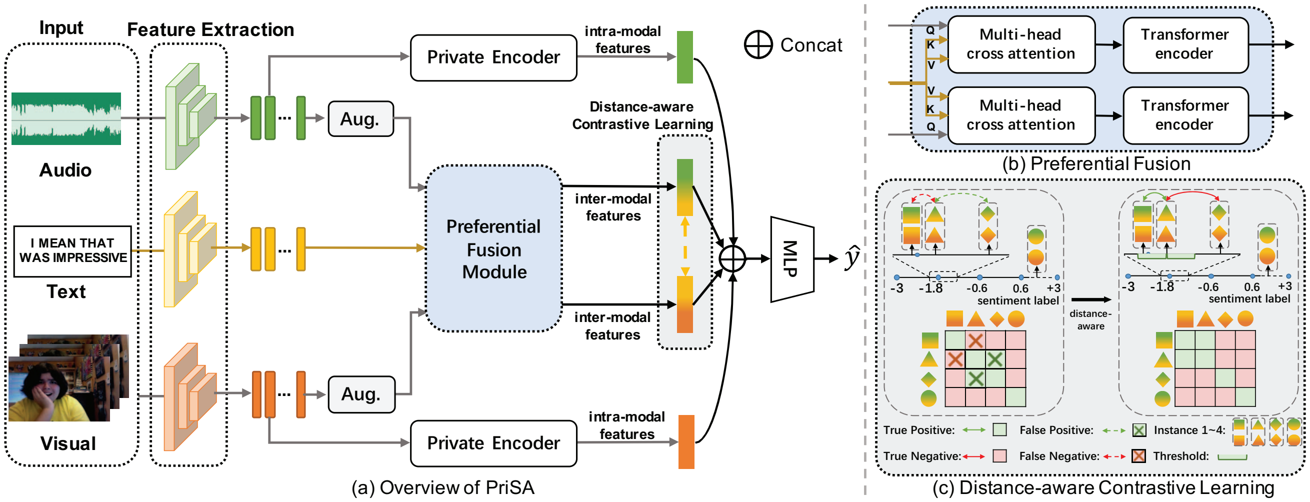
where $\boldsymbol{W}^{(Q)}$, $\boldsymbol{W}^{(K)}$ and $\boldsymbol{W}^{(V)}$, σ are the linear projection weight matrices of Q, K and V, and the softmax activation function, respectively, and d is the common dimension for three modalities.

Inspired by [8], we explore the inter-modal correlation by involving both the primary and secondary modalities in the attention operation. As shown in Fig. 2, the three matrices in our fusion module are from the primary and secondary modalities. Thus the attention function in our fusion module becomes

$$Att(\boldsymbol{Q}^s,\boldsymbol{K}^p,\boldsymbol{V}^p) = \sigma\left(\frac{\boldsymbol{Q}^s\boldsymbol{W}^{(\boldsymbol{Q}^s)}(\boldsymbol{K}^p\boldsymbol{W}^{(\boldsymbol{K}^p)})^{\mathsf{T}}}{\sqrt{d}}\right)\boldsymbol{V}^p\boldsymbol{W}^{(\boldsymbol{V}^p)}, \quad (2)$$

where the superscripts s and p indicate that the matrices are from the secondary modality and the primary modality. In other words, we use secondary modalities (audio and visual) as the query and the primary modality (text) as key and value in the attention module to enforce the modality invariance on the common space of queries and key-value pairs, as well as reduce the distribution gap between modalities. Finally, we obtain two fusion features, $\boldsymbol{h}^{at}$ and $\boldsymbol{h}^{vt}$. The former is for the fusion of audio and text, and the latter is for the fusion of visual and text.

As the primary modality is used as key and value, we reinforce the secondary modality by attending to the correlated elements of the primary modality in our preferential fusion. Thus the distribution mismatch between different modalities is well bridged. In addition, the primary modality in our fusion module does not have an independent branch and will not be directly involved in the final prediction. This means that information in the primary modality can only be propagated forward through the preferential fusion regarding the similarity of the distribution between the modalities. Thus, the shortcut from the primary modality to the final prediction will be cut off and the distribution of the secondary modality will be forced



(a) Overview of PriSA

(b) Preferential Fusion

(c) Distance-aware Contrastive Learning

Fig. 2. (a) Overview of our proposed PriSA. We extract deep features from each modality and perform augmentation on secondary modalities. We then exploit inter-modal and mixed-modal correlations through preferential fusion and distance-aware contrastive learning. The resulting features are used to identify sentiment with intra-modal features extracted by the private encoder. (b) The preferential fusion module which uses the primary modality to implicitly guide the inter-modal feature learning between the primary modality and secondary modality. (c) The distance-aware contrastive learning is proposed to avoid sampling false positive and negative samples by introducing distance information of sentiment labels into traditional contrastive learning.

closer to the primary modality to obtain more task-relevant information.

C. Distance-aware contrastive learning

We apply contrastive learning to mixed-modal features, which are fused with text, to calculate the mixed-modal correlations, as shown in Fig. 2. Previous methods using supervised contrastive learning typically treat the MSA task as a classification problem by dividing sentiment labels into discrete classes and sampling positive and negative samples based on the classes. However, these methods ignore the continuous nature of sentiment labels and the distance information between them, which can lead to the sampling of false positive and negative samples. For example, as shown in the left of Fig. 2 (c), the label of instance 2 (*ins.2*) is closer to instance 1 (*ins.1*) than instance 3 (*ins.3*). However, traditional methods would treat *ins.2* and *ins.3* as a positive pair and *ins.2* and *ins.1* as a negative pair because the labels of *ins.2* and *ins.3* are in the same segment. Zolfaghari *et al.* [15] exclude false negative samples in contrastive learning according to input embeddings. Inspired by this, we introduce distance information of labels into our distance-aware contrastive learning, which allows us to avoid both false positive and negative samples. When it comes to classification tasks, it is important to note that our loss will degenerate into a regular contrastive loss. As shown in the right of Fig. 2 (c), we select positive and negative samples based on the distance between their labels within a batch. Specifically, we calculate the label distance between the anchor sample and the other samples in the batch. Samples with label distances less than a threshold c are considered positive samples, while those with distances greater than c are considered negative samples.

Formally, the distance-aware contrastive loss $\mathcal{L}_{DACL}$ (based on InfoNCE [16]) can be derived as

$$\mathcal{L}_{DACL}(\boldsymbol{h}_i^{at}, \boldsymbol{h}_i^{vt}) = -\log \frac{\sum_{j \in P} \exp(\Phi(\boldsymbol{h}_i^{vt}, \boldsymbol{h}_j^{at})/\tau)}{\sum_{k \notin P} \exp(\Phi(\boldsymbol{h}_i^{vt}, \boldsymbol{h}_k^{at})/\tau)}, \quad (3)$$

where Φ is a cosine similarity scoring function, τ is the temperature, $\boldsymbol{h}^{vt}$ is the fused feature of visual and text, $\boldsymbol{h}^{at}$ is the fused feature of audio and text, P represents the index set of positive samples that are selected based on label distance, and i, j, k represent the indices of different samples. We consider using different fusion pairs as anchors, so the final contrastive loss $\mathcal{L}_{CL}$ is

$$\mathcal{L}_{CL} = \mathcal{L}_{DACL}(\boldsymbol{h}_i^{at}, \boldsymbol{h}_i^{vt}) + \mathcal{L}_{DACL}(\boldsymbol{h}_i^{vt}, \boldsymbol{h}_i^{at}). \quad (4)$$

D. Objectives Function

Our objective function is composed of task loss and contrastive loss. The task loss is designed for different tasks and serves as the primary training objective for our model. In our experiments, we consider two tasks with different objectives: a regression task and a classification task. For different tasks, the losses are calculated as

$$\mathcal{L}_{task} = \begin{cases} \frac{1}{N}\sum_i^N |\hat{y}_i - y_i| & \text{for regression} \\ -\frac{1}{N}\sum_i^N y_i \log \hat{y}_i & \text{for classification} \end{cases} \quad (5)$$

where N is the size of a minibatch, y_i and $\hat{y}_i$ represent the true label and the predicted label of the i^{th} sample. For each branch, we use the task loss to guide its training. After the preferential fusion module, we use the contrastive loss $\mathcal{L}_{ct}$ between the inter-modal features which is described in Eq. (4). The final objective function can be represented as

$$\mathcal{L} = \mathcal{L}_{task} + \beta \mathcal{L}_{CL}, \quad (6)$$

where β is the weight of the contrastive loss.

TABLE I

Performance on the MOSEI dataset. (B): the text features are extracted by pretrained BERT; *: the results are reproduced under the same conditions; ◇: the model uses unaligned data.

Model	MAE(↓)	Corr(↑)	Acc-7(↑)	Acc-2(↑)	F1-Score(↑)
RAVEN [10]	0.614	0.662	50.0	79.1/-	79.5/-
MCTN [19]	0.609	0.670	49.6	79.8/-	80.6/-
MulT [8]	0.580	0.703	51.8	-/82.5	-/82.3
TCSP [20]	0.576	0.715	-	-/82.8	-/82.7
TFN(B)◇ [21]	0.593	0.700	50.2	-/82.5	-/82.1
LMF(B)◇ [22]	0.623	0.677	48.0	-/82.0	-/82.1
MFM(B) [23]	0.568	0.717	51.3	-/84.5	-/84.3
ICCN(B) [24]	0.565	0.713	51.6	-/84.2	-/84.2
MISA(B) [5]	0.555	0.756	52.2	83.6/85.5	83.8/85.3
Self-MM(B)◇ [7]	0.530	0.765	-	82.81/85.17	82.53/85.30
MISA(B)* [5]	0.554	0.747	52.0	80.17/84.62	80.78/84.61
Self-MM(B)*◇ [7]	0.536	0.764	53.59	80.75/84.57	81.28/84.56
PriSA(B)◇	**0.523**	**0.772**	**54.65**	**82.84/85.93**	**83.18/85.87**

TABLE II

Performance on the MOSI dataset. (B): the text features are extracted by pretrained BERT; *: the results are reproduced under the same conditions; ◇: the model uses unaligned data.

Model	MAE(↓)	Corr(↑)	Acc-7(↑)	Acc-2(↑)	F1-Score(↑)
TFN◇ [21]	0.970	0.633	32.1	73.9/-	73.4/-
RAVEN [10]	0.915	0.691	33.2	78.0/-	76.7/-
MCTN [19]	0.909	0.676	35.6	79.3/-	79.1/-
MulT [8]	0.871	0.698	40.0	-/83.0	-/82.8
TCSP [20]	0.908	0.710	-	-/80.9	-/81.0
TFN(B)◇ [21]	0.901	0.698	34.9	-/80.8	-/80.7
LMF(B)◇ [22]	0.917	0.695	33.2	-/82.5	-/82.4
MFM(B) [23]	0.877	0.706	35.4	-/81.7	-/81.6
ICCN(B) [24]	0.860	0.710	39.0	-/83.0	-/83.0
MulT(B) [8]	0.861	0.711	-	81.5/84.1	80.6/83.9
MISA(B) [5]	0.783	0.761	42.3	81.8/83.4	81.7/83.6
Self-MM(B)◇ [7]	0.713	0.798	-	84.00/85.98	84.42/85.95
MISA(B)* [5]	0.821	0.740	43.44	81.8/83.40	81.7/83.60
Self-MM(B)*◇ [7]	0.718	**0.796**	46.27	82.77/84.36	82.73/84.38
PriSA(B)◇	**0.714**	0.792	**47.3**	**83.38/85.52**	**83.24/85.45**

TABLE III

Performance on the SIMS dataset. *: the results are reproduced under the same conditions.

Model	MAE(↓)	Corr(↑)	Acc-2(↑)	F1-Score(↑)
TFN* [21]	0.434	0.584	78.12	77.90
LMF* [22]	0.437	0.588	77.53	77.44
MulT* [8]	0.440	0.582	77.53	77.40
Self-MM(B)* [7]	0.421	0.590	77.68	77.72
PriSA(B)	**0.407**	**0.591**	**79.29**	**79.41**

IV. Experiments

In this section, we will first introduce the experimental settings, including the datasets, metrics, and implementation details. Then we report our results on four benchmark datasets followed by the ablation study.

A. Datasets and Implementatoin Details

We use four datasets to evaluate the performance of the PriSA, including CMU-MOSEI [17], CMU-MOSI [4], SIMS [6] and UR-FUNNY [18]. The CMU-MOSI dataset is one of the most prevalent benchmarks for evaluating the performance on MSA. It is collected from video blogs on YouTube and contains 2199 video segments sliced from 93 videos. The CMU-MOSEI dataset is thus far the largest dataset on MSA which contains 23453 video segments from 5000 videos. The SIMS dataset is a Chinese MSA dataset. It has fine-grained annotations for each modality. The UR-FUNNY dataset is a multimodal humor detection dataset. Video samples in the UR-FUNNY dataset are collected from TED talks. Unlike other datasets, the samples in the UR-FUNNY dataset are labeled with binary labels indicating whether they are humorous or non-humorous. In our experiments, we use *mean absolute error* (MAE), *Pearson correlation* (Corr), *seven-class accuracy* (Acc-7), *binary accuracy* (Acc-2) and *F1-Score* as metrics to evaluate the performance. The task on the UR-FUNNY dataset is a binary classification task, so we only use *binary accuracy* (Acc-2) to evaluate our method on the UR-FUNNY dataset.

We use the Adam optimizer in conjunction with the StepLR scheduler during training. To avoid overfitting, we implement an early stopping strategy with a patience of 5 epochs and utilize the MAE metric for evaluation. All the experiments are conducted on one NVIDIA RTX 3090 GPU.

B. Results

The experimental results are presented in Tables I-IV. Because all three modalities in these datasets are sequences, there is an alignment problem. To ensure a fair and detailed comparison, we follow the unaligned setting. As the results shown in MulT [8], models using aligned corpus usually achieve better results. In our experiments, we run three trials with different random seeds and calculate the average of the results to obtain the final results. As shown in Tables I and II, our PriSA achieves the state-of-the-art or comparable results on all metrics on both the MOSEI and MOSI datasets. Notably, our model shows significant improvement on the MOSEI dataset,, likely because the dataset is larger, enabling the model to learn more correlations between the primary modality and secondary modalities and align them more effectively. Methods that use BERT typically achieve better results than those that do not, and our PriSA outperforms all other methods that use BERT. On the SIMS dataset, we reproduce TFN [21], LMF [22], MulT [8] and Self-MM [7], and compare them under the same conditions. In Table III, our PriSA obtains the state-of-the-art results compared with previous models under the same conditions. On the UR-FUNNY dataset, the task is to determine whether the given sample is humorous or not. For each sample, both the punchline and context are provided. Our model uses only punchline information to determine whether

TABLE IV
PERFORMANCE ON THE UR-FUNNY DATASET.

Model	Setting	Acc-2(↑)
C-MFN [18]	context	58.45
C-MFN [18]	target	64.47
TFN [21]	punchline	64.71
LMF [22]	punchline	65.16
C-MFN [18]	punchline+context	65.23
LMF(B) [22]	punchline	67.53
TFN(B) [21]	punchline	68.57
MISA(B) [5]	punchline	70.61
PriSA(B)	punchline	**71.67**

TABLE V
ABLATION STUDY ON THE MOSEI AND MOSI DATASETS.

	Model	MOSEI		MOSI	
		MAE(↓)	F1-Score(↑)	MAE(↓)	F1-Score(↑)
1	PriSA	**0.523**	**85.87**	**0.714**	**85.45**
2	full combination	0.529	85.47	0.727	83.70
3	text as query	0.528	85.67	1.105	73.59
4	audio as primary	0.532	85.63	0.720	84.10
5	visual as primary	0.528	85.58	0.716	84.40
6	(-) v-private encoder	0.534	85.43	0.747	83.05
7	(-) a-private encoder	0.538	85.47	0.726	84.83
8	(-) $\mathcal{L}_{CL}$	0.528	85.58	0.733	83.76

it is humorous or not, and achieves a significant improvement compared to the state-of-the-art methods as shown in Table IV. Our PriSA achieves state-of-the-art results on datasets of different sizes, languages, scenarios and tasks, indicating that our method can be applied to different data scenarios.

C. Ablation Study

We conduct a detailed analysis of each component of our framework. These components include Preferential fusion (rows 2-5), Private encoder (rows 6, 7) and Distance-aware contrastive learning (row 8). In particular, we discuss two parts of preferential fusion: inter-modal learning (rows 2, 3) and text modality as implicit guidance (rows 4, 5). All these ablation experiments are performed on the MOSEI and MOSI datasets. The results of the ablation study are shown in Table V.

Inter-modal Learning. The core design of our framework is inter-modal learning, which learns correlations between the primary modality and the secondary modalities only. We use the primary modality text as the key and value in the fusion to aid in learning these correlations. In row 2, we implement a *full combination* approach that use pairwise fusion between every two modalities, similar to the method used in MulT [8]. In row 3, we use *text as query*, which means that we use text as query in our PriSA framework. We found that even though *full combination* has a larger number of parameters and involves more interaction between every two modalities, it does not perform as well as our method. This demonstrates that not all correlations between modalities need to be learned, and it is better to learn only the relationship between primary modality and secondary modalities. And *text as query* cause the primary modality to cease to be an implicit guidance role. We can observe a decrease in row 3. The reason is that the model heavily rely on the text modality and learn the false correlation and ambiguous words in it.

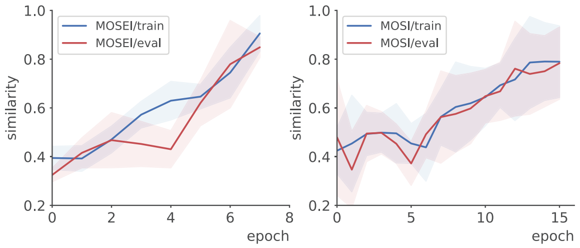


Fig. 3. The similarity of the secondary modality (audio) and primary modality (text) varies with epochs during training and evaluation on MOSEI and MOSI datasets.

Text Modality as Implicit Guidance. In our framework, we choose text modality as the primary modality because we believe it serves as an implicit guidance among three modalities. Rows 4 and 5 compare the performance of models using audio and visual modalities as the primary modality, and both show worse results than row 1, which uses text as the primary modality. These results demonstrate that text modality is the most suitable primary modality to implicitly guide the inter-modal learning between the primary modality and secondary modalities in our framework.

Intra-modal Learning. We aim to investigate the effectiveness of intra-modal features extracted by private encoders from secondary modalities. We remove the private encoder of the visual and audio modalities in rows 6 and 7 of Table V, respectively. The significant decrease in performance suggests that the intra-modal features of secondary modalities can effectively compensate for the loss of sentiment information during implicit alignment of inter-modal learning.

Distance-aware contrastive learning. With the help of contrastive learning, our framework is able to further obtain mixed-modality correlations. In row 8 of Table V, we demonstrate the impact of removing the distance-aware contrastive loss on the performance of the model. Our results show that removing this loss will decrease the model's performance in all metrics. Despite the drop, the performance of the model is still comparable to other state-of-the-art methods. This indicates that our preferential fusion strategy is effective, and contrastive learning can further build mixed-modality correlations to improve performance on this basis.

Visualization of Implicit Alignment. Our method can implicitly align the primary modality and secondary modalities. In Fig. 3, we plot the similarity between the secondary modality (audio) and the primary modality (text) during training on the training and validation sets. The similarity increases with

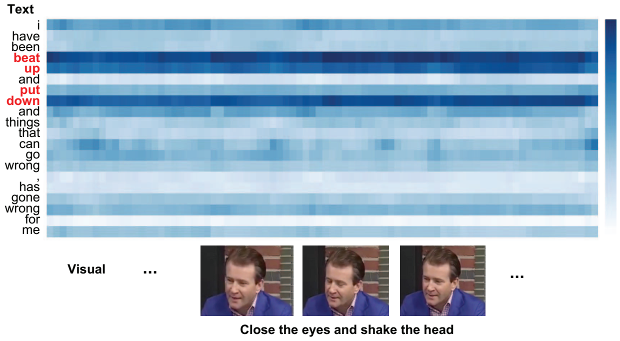


Fig. 4. Visualization of the attention map between visual and text in preferential fusion.

training and the best validation results are obtained at high similarity epochs. We notice that the similarity on the MOSEI dataset is higher than that on the MOSI dataset. This is because the MOSI dataset has a smaller dataset size and lower feature dimensions, making it difficult to learn the similarity between different modalities. This is also why our model's improvement on the MOSI dataset is not as significant as the improvement on the MOSEI dataset. These visualization results illustrate that our method can align the secondary modality and primary modality under the implicit guidance of the primary modality and a high similarity is beneficial.

Visualization of Attention Map. In Fig. 4, we draw an attention map between the visual and text modalities, where the vertical axis represents the words in the text and the horizontal axis represents key frames. We found that the frames showing the action of "closing eyes and shaking head" have a higher similarity to the phrases "beat up" and "put down" but a lower similarity to other non-emotional words. This indicates that our method can avoid learning the bias in the text modality, as well as the interference of words in the text that are not related to sentiment.

V. Conclusion

In this paper, we address the issue that an excessive reliance on the text modality may lead to the learning of false correlations between textual tokens and sentiment labels, resulting in errors in sentiment analysis. To mitigate this issue, we propose a framework with a preferential inter-modal fusion strategy, in which a primary modality is selected to implicitly guide intermodal learning. We also propose a distance-aware contrastive learning method to learn mixed-modality correlations, which leverages the distance information of the sentiment labels. Finally, we evaluate our PriSA on four datasets of MSA and our PriSA outperforms other state-of-the-art methods on all four datasets.

Acknowledgment

This work is in part supported by the National Key R&D Program of China under Grant 2020AAA0108600, and the National Natural Science Foundation of China under Grant 62032006.

References

[1] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *CVPR*, 2022, pp. 16 000–16 009.

[2] S. Poria, I. Chaturvedi, E. Cambria, and A. Hussain, "Convolutional mkl based multimodal emotion recognition and sentiment analysis," in *ICDM*. IEEE, 2016, pp. 439–448.

[3] H. Wang, A. Meghawat, L.-P. Morency, and E. P. Xing, "Select-additive learning: Improving generalization in multimodal sentiment analysis," in *ICME*. IEEE, 2017, pp. 949–954.

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

[5] D. Hazarika, R. Zimmermann, and S. Poria, "Misa: Modality-invariant and-specific representations for multimodal sentiment analysis," in *ACM MM*, 2020, pp. 1122–1131.

[6] W. Yu, H. Xu, F. Meng, Y. Zhu, Y. Ma, J. Wu, J. Zou, and K. Yang, "Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality," in *ACL*, 2020, pp. 3718–3727.

[7] W. Yu, H. Xu, Z. Yuan, and J. Wu, "Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis," in *AAAI*, vol. 35, no. 12, 2021, pp. 10 790–10 797.

[8] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *ACL*, 2019, pp. 6558–6569.

[9] A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria, and L.-P. Morency, "Memory fusion network for multi-view sequential learning," in *AAAI*, vol. 32, no. 1, 2018.

[10] Y. Wang, Y. Shen, Z. Liu, P. P. Liang, A. Zadeh, and L.-P. Morency, "Words can shift: Dynamically adjusting word representations using nonverbal behaviors," in *AAAI*, vol. 33, no. 01, 2019.

[11] T. Sun, W. Wang, L. Jing, Y. Cui, X. Song, and L. Nie, "Counterfactual reasoning for out-of-distribution multimodal sentiment analysis," in *ACM MM*, 2022, pp. 15–23.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *NeurIPS*, vol. 30, 2017.

[13] S. Mai, Y. Zeng, S. Zheng, and H. Hu, "Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis," *IEEE Trans. Affective Comput*, 2022.

[14] Z. Lin, B. Liang, Y. Long, Y. Dang, M. Yang, M. Zhang, and R. Xu, "Modeling intra-and inter-modal relations: Hierarchical graph contrastive learning for multimodal sentiment analysis," in *COLING*, 2022, pp. 7124–7135.

[15] M. Zolfaghari, Y. Zhu, P. Gehler, and T. Brox, "Crossclr: Cross-modal contrastive learning for multi-modal video representations," in *ICCV*, 2021, pp. 1450–1459.

[16] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[17] A. Zadeh and P. Pu, "Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph," in *ACL*, 2018.

[18] M. K. Hasan, W. Rahman, A. Zadeh, J. Zhong, M. I. Tanveer, L.-P. Morency *et al.*, "Ur-funny: A multimodal language dataset for understanding humor," in *ACL*, 2019, pp. 2046–2056.

[19] H. Pham, P. P. Liang, T. Manzini, L.-P. Morency, and B. Póczos, "Found in translation: Learning robust joint representations by cyclic translations between modalities," in *AAAI*, 2019.

[20] Y. Wu, Z. Lin, Y. Zhao, B. Qin, and L.-N. Zhu, "A text-centered shared-private framework via cross-modal prediction for multimodal sentiment analysis," in *ACL-IJCNLP*, 2021, pp. 4730–4738.

[21] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *EMNLP*, 2017, pp. 1103–1114.

[22] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," in *ACL*, 2018, pp. 2247–2256.

[23] Y.-H. H. Tsai, P. P. Liang, A. Zadeh, L.-P. Morency, and R. Salakhutdinov, "Learning factorized multimodal representations," in *ICLR*, 2018.

[24] Z. Sun, P. Sarma, W. Sethares, and Y. Liang, "Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis," in *AAAI*, vol. 34, no. 05, 2020, pp. 8992–8999.