Multi-Metric Representation Learning Strategy Based on Clustering for Fine-Grained Multimodal Sentiment Analysis

Supplementary Material

A. Outline

This supplementary material provides additional details of MMRest from four perspectives. In Sec. B, we present the specific optimization settings and hyperparameter configurations on the CMU-MOSI and CMU-MOSEI datasets. In Sec. C, we conduct an experiment on variance to further analyze the issue of class center overlap. In Sec. D, we conduct a qualitative case study on several representative examples of the CMU-MOSI dataset, comparing MCL-MCF with different training variants of MMRest. In Sec. E, we perform a systematic hyperparameter study to analyze the sensitivity of MMRest to key hyperparameters and to justify the final choices adopted in the main paper. In Sec. F, we briefly discuss the ethical implications of multimodal sentiment analysis.

B. Settings in MMRest

Our model is optimized using both traditional Adam and improved Adam, with the improved Adam specifically designed for optimizing metric matrices with a learning rate of 1e-4. On CMU-MOSI dataset, (k, v_1, v_2, ξ, α) are set as $(7, 0.8, 1.0, 3.0, 1.0)$. On CMU-MOSEI dataset, they are set as $(10, 0.7, 0.5, 70.0, 0.7)$, respectively. On both datasets, the dimensions (d_1, d_2) of intermediate representations are set to (64, 50). (γ, η) are set as $(0.05, 0.01)$.

C. Variance analysis

To further validate the advantages observed in the visualizations, we first calculate the total variance by M_0, between-cluster variance, and within-cluster variance of the multimodal representations by ΔM_i. The results are illustrated in Tab. 6, demonstrating that our strategy effectively achieves high within-cluster compactness and between-cluster separation. Subsequently, we then conduct a fair comparison by calculating the variances of the fused representations using Euclidean distance after partitioning the representations into seven fine-grained sentiment categories based on their label value ranges. The results indicate that all our variance values are lower than those of MCL-MCF, suggesting that our method can learn clearer emotional decision boundaries more effectively.

D. Case Study

We randomly select five samples from the test set in CMU-MOSI dataset for the case study, as illustrated in Tab. 7.

Table 6. Variance interpretation of visualization on CMU-MOSI.

Metric	MCL-MCF	MMRest
MM(w/ M_0,w/ ΔM_i)	-	11.28/11.23/0.05
MM(w/ M_0,w/o ΔM_i)	-	11.29/11.26/0.03
SM	11.85 / 9.44 / 2.41	7.38 / 6.10 / 1.28

SM represents that the model uses a single-metric matrix, *i.e.*, the Euclidean distance metric matrix, MM represents that the model uses multi-metric matrices, *i.e.*, the $M_0, \Delta M_i$ distance metric matrices. The values in the table represent total variance/inter-class variance/intra-class variance.

We conduct a detailed comparison between MCL-MCF and the three training scenarios of our model, *i.e.*, MMRest(*w/o* $\mathcal{L}_{\text{MMC}}$, *w/o* $bias$), MMRest(*w/* $\mathcal{L}_{\text{MMC}}$, *w/o* $bias$) and MMRest(*w/* $\mathcal{L}_{\text{MMC}}$, *w/* $bias$). For the first three samples, although both MCL-MCF and our model demonstrate excellent sentiment prediction ability, our model still has smaller prediction errors. By comparing the three training scenarios of our model longitudinally, it can be clearly observed that our proposed MMC module and PDLF module can greatly improve the sentiment prediction ability of the model when used in combination. For the last two difficult samples, comparing our complete model (MMRest(*w/* $\mathcal{L}_{\text{MMC}}$, *w/* $bias$)) with MCL-MCF, our model can still achieve smaller errors when dealing with relatively hard samples. This case study fully demonstrates the necessity and effectiveness of enhancing the interaction between different sentiments while maintaining sufficient interaction between modalities.

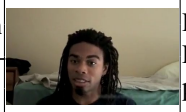
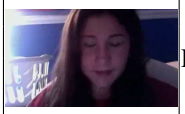
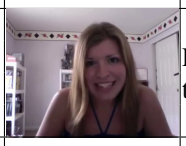
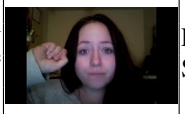
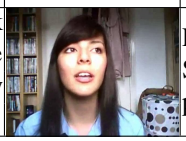
E. Hyperparameter Study

As shown in Fig. 4, we conduct a hyperparameter study on the hyperparameters in our model on the CMU-MOSI dataset. We first make a rough estimate of each hyperparameter range, and then determine a smaller range for grid search. To improve clarity, we present only the Acc-5 and Acc-7 metrics on the vertical axis. Due to the fact that the optimal values of each indicator may be distributed across different hyperparameters, such as α in Fig. 4, we comprehensively determine the final value of α based on several other evaluation metrics. Finally, k, v_1, v_2, ξ, α, γ are set as 7, 0.8, 1.0, 3.0, 1.0, 0.05, respectively.

F. Ethics discussion

Multimodal sentiment analysis seeks to infer human sentiments by jointly modeling signals from language, vision, and audio. When carefully designed and responsibly deployed,

Table 7. Case Study on CMU-MOSI dataset.

Language	Vision	Audio	Models	Prediction	Label	Error
And quite honestly I wish I've seen this over the summer		Peaceful tone, Fast speaking pace	MCL-MCF	1.9093	1.6000	0.3093
			MMRest(*w/o* $\mathcal{L}_{\text{MMC}}$, *w/o bias*)	1.9766	1.6000	0.3766
			MMRest(*w/* $\mathcal{L}_{\text{MMC}}$, *w/o bias*)	1.7891	1.6000	0.1891
			MMRest(*w/* $\mathcal{L}_{\text{MMC}}$, *w/ bias*)	1.5981	1.6000	**0.0019**
But very predictable		Peaceful tone	MCL-MCF	-1.5106	-1.5000	0.0106
			MMRest(*w/o* $\mathcal{L}_{\text{MMC}}$, *w/o bias*)	-1.2638	-1.5000	0.2362
			MMRest(*w/* $\mathcal{L}_{\text{MMC}}$, *w/o bias*)	-1.4731	-1.5000	0.0268
			MMRest(*w/* $\mathcal{L}_{\text{MMC}}$, *w/ bias*)	-1.5045	-1.5000	**0.0045**
And I hate to say that		Emphasizing tone	MCL-MCF	-1.3653	-1.0000	0.3653
			MMRest(*w/o* $\mathcal{L}_{\text{MMC}}$, *w/o bias*)	-1.0237	-1.0000	0.0237
			MMRest(*w/* $\mathcal{L}_{\text{MMC}}$, *w/o bias*)	-0.7595	-1.0000	0.2405
			MMRest(*w/* $\mathcal{L}_{\text{MMC}}$, *w/ bias*)	-0.9955	-1.0000	**0.0045**
He um had all the charm of a narcissist XXX boy the whole film		Peaceful tone, Slight pause	MCL-MCF	1.1243	-2.8000	3.9243
			MMRest(*w/o* $\mathcal{L}_{\text{MMC}}$, *w/o bias*)	1.1493	-2.8000	3.9493
			MMRest(*w/* $\mathcal{L}_{\text{MMC}}$, *w/o bias*)	1.2938	-2.8000	4.0938
			MMRest(*w/* $\mathcal{L}_{\text{MMC}}$, *w/ bias*)	0.9428	-2.8000	**3.7428**
But as an actor I really think especially an actor in the lead leading role you really should prepare for that role		Peaceful tone, Significant pause	MCL-MCF	1.5153	-2.0000	3.5153
			MMRest(*w/o* $\mathcal{L}_{\text{MMC}}$, *w/o bias*)	1.3044	-2.0000	3.3044
			MMRest(*w/* $\mathcal{L}_{\text{MMC}}$, *w/o bias*)	1.5323	-2.0000	3.5323
			MMRest(*w/* $\mathcal{L}_{\text{MMC}}$, *w/ bias*)	1.1299	-2.0000	**3.1299**

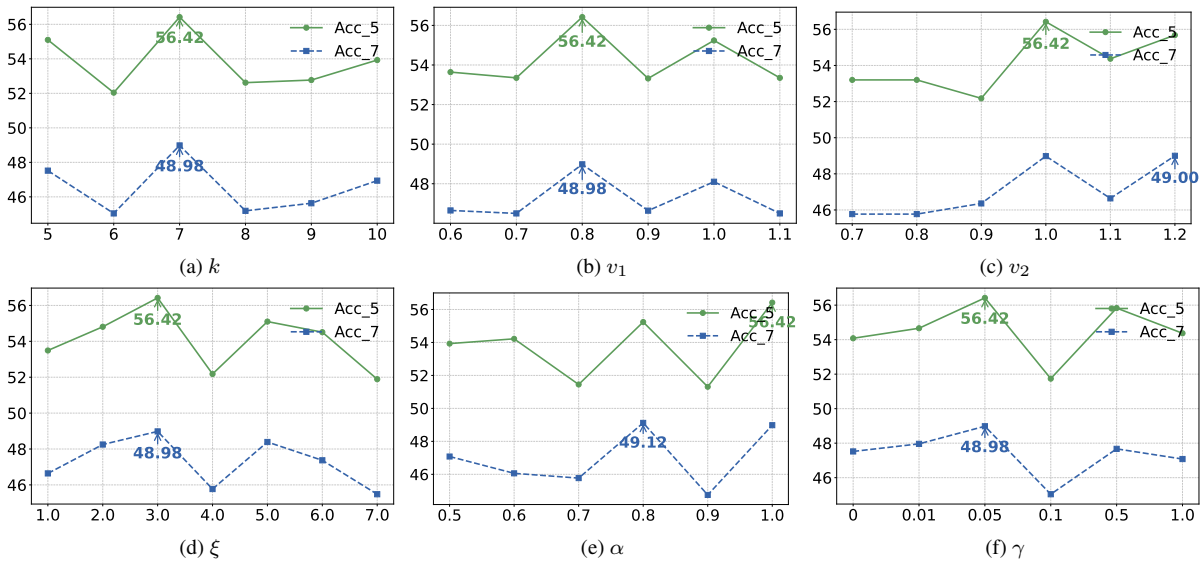


Figure 4. Visualization of the impact of hyperparameter change in CMU-MOSI dataset on Acc-5 and Acc-7.

such systems have the potential to enhance user experience in applications such as assistive technologies, education, and healthcare. However, they also raise important ethical concerns, including risks of intrusive surveillance, misuse of sensitive affective information, and the propagation of existing social biases if the underlying data or models are not appropriately audited and governed.