

Supplementary Materials: INMU-NET: Advancing Multi-modal Intent Detection via Information Bottleneck and Multi-sensory Processing

Zhihong Zhu
Peking University
Beijing, China
zhihongzhu@stu.pku.edu.cn

Xuxin Cheng
Peking University
Beijing, China
chengxx@stu.pku.edu.cn

Zhaorun Chen
The University of Chicago
Chicago, USA
zhaorun@uchicago.edu

Yuyan Chen
Fudan University
Shanghai, China
chenyuyan21@m.fudan.edu.cn

Yunyan Zhang
Jarvis Research Center, Tencent
YouTu Lab
Shenzhen, China
yunyanzhang@tencent.com

Xian Wu*
Jarvis Research Center, Tencent
YouTu Lab
Shenzhen, China
kevinxwu@tencent.com

Yefeng Zheng
Medical Artificial Intelligence Lab,
Westlake University
Hangzhou, China
Jarvis Research Center, Tencent
YouTu Lab
Shenzhen, China
zhengyefeng@westlake.edu.cn

Bowen Xing
Beijing Key Laboratory of Knowledge
Engineering for Materials Science,
School of Computer and
Communication Engineering,
University of Science and Technology
Beijing
Beijing, China
bwxing714@gmail.com

1 Challenges in MID

We want to highlight the unique challenges associated with multi-modal intent detection (MID). The following two points correspond to the challenges outlined in our manuscript.

❶ **The issue of information redundancy in MID is significantly pronounced.** Specifically, scenarios in MID often involve distinguishing the visual information of the speaker, especially in situations where multiple individuals are present. However, manual annotation for this purpose would be highly costly. When we look at datasets in the multimodal community for tasks such as sentiment analysis and sarcasm detection, their visual content is generally simple, featuring only one person or object. This allows models to quickly locate and capture key information. While the noise and redundancy in visual and audio modalities are also present in other multimodal tasks, it is particularly pronounced in MID and urgently needs to be addressed. ❷ **The large number of intent categories in MID presents another challenge.** Specifically, multimodal sentiment analysis uses metrics such as binary and seven-category accuracy, while multimodal sarcasm detection primarily involves simple binary classification accuracy. In contrast, MID requires discrimination among 20 intent categories, and the existing benchmarks tend to exhibit a long-tail distribution.

2 Details for Cross-task Scenario

To verify the generalizability of the proposed model, we conduct preliminary experiments on multi-modal sentiment analysis (MSA).

❶ For **datasets**, we select two publicly MSA benchmarks: MOSI [6] and MOSEI [7]. MOSI contains 2,198 utterance segments, and MOSEI contains 23,453 annotated clips from YouTube. Each sample is manually annotated with a sentiment score ranging from -3 to +3 to indicate the sentiment polarity. ❷ For **evaluation metrics**, we follow previous works to adopt MAE (mean absolute error), Corr (Pearson correlation) and Acc-7 (seven-class classification accuracy) ranging from -3 to 3. ❸ For **comparison baselines**, we select five state-of-the-art models, including ICCN [3], MISA [2], Self-MM [5], MMIM [1] and DBF [4]. ❹ For **implementation details**, we have utilized grid search to determine the optimal values for the parameters α and β on the validation set of MOSI and MOSEI, respectively. The grid search is performed with a step size of 0.1 and a range spanning from 0 to 1. The results are averages of 5 random runs.

References

- [1] Wei Han, Hui Chen, and Soujanya Poria. 2021. Improving Multimodal Fusion with Hierarchical Mutual Information Maximization for Multimodal Sentiment Analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 9180–9192.
- [2] 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*. 1122–1131.
- [3] Zhongkai Sun, Prathusha Sarma, William Sethares, and Yingyu Liang. 2020. 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. 8992–8999.
- [4] Shaoxiang Wu, Damai Dai, Ziwei Qin, Tianyu Liu, Binghui Lin, Yunbo Cao, and Zhifang Sui. 2023. Denoising Bottleneck with Mutual Information Maximization for Video Multimodal Fusion. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 2231–2243.

*Corresponding author.

- [5] Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. 2021. 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. 10790–10797.
- [6] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Mosei: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259* (2016).
- [7] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2236–2246.