

Supplementary Materials for DQ-Former

Anonymous Authors

1 ADDITIONAL EXPERIMENTS

1.1 Comparison Results on MERBenchmark

Due to inconsistencies among various methods in feature extractors, evaluation manners, and experiments settings, Lian et al. [3] propose a benchmark for Multimodal Emotion Recognition (MER). For a fair comparison, we utilize their code¹ to reproduce DQ-Former base under identical settings. We employ DeBERTa-base as the textual encoder and wav2vec2.0-base as the acoustic encoder. Additionally, we exclude the incorporation of dialogue history utterances.

Table 1 presents the comparison results of our models and baselines on the MELD and IEMOCAP datasets. Our model consistently outperforms the baselines on both benchmarks, indicating a significant improvement in WF1 scores. Compared to Attention, DQ-Former exhibits a 3.15% improvement in the MELD dataset, a 3.23% improvement in IEMOCAP 4-way, and a 6.22% improvement in IEMOCAP 6-way. Notably, a significant performance decrease is observed when the dialog history is disregarded, highlighting the necessity of considering context in the textual modality.

In a nutshell, these findings underscore the efficacy of our multimodal fusion framework, establishing DQ-Former as a superior model in leveraging diverse modalities effectively.

1.2 Impact of Unimodal Feature Extractors

The transformer architecture efficiently handles various types of sequence inputs uniformly, leading to the emergence of advanced pre-trained models tailored to specific modalities [9]. Since the choice of unimodal features significantly influences multimodal results, a thorough evaluation is conducted to analyze the performance of DQ-Former under different unimodal feature extractors. The unimodal pre-trained models for the text modality include BERT and Roberta. For the acoustic modality, we consider Wav2Vec2, HuBERT, and WavLM. The comparison results are presented in Table 2.

From the results, we observe that the performance of DQ-Former varies when different feature extraction models are used. In general, the large model tends to outperform the base model in the unimodal results. For textual, RoBERTa often outperforms BERT. Moreover, the performance of different unimodal features varies across datasets. However, selecting the best-performing unimodal model does not necessarily yield optimal results in multimodal fusion. The quality of multimodal representation depends on various factors, which have not yet been fully estimated.

1.3 Visualization of Fusion Representations

In figure 1, we visualize the final fusion representation h_f in a 2D feature space using t-SNE [8] on the IEMOCAP dataset. The results demonstrate that the multimodal representation learned by our model effectively distinguishes between different types of emotions

Table 1: Performance comparison (WF1 %) with different multimodal fusion strategies on MELD and IEMOCAP. The best result is highlighted in bold; baseline results are from Lian et al. [3]. Both baselines and DQ-Former are tested under the same setting. The feature extractors are SENet-FER2013, wav2vec2.0-base and DeBERTa-large for baselines.

Methods	MELD	IEMOCAP (4-way)	IEMOCAP (6-way)
MCTN [5]	56.31	63.08	48.66
MFM [6]	54.55	70.84	54.12
GMFN [12]	56.73	71.22	55.14
MFN[11]	57.80	72.53	56.02
MuT[7]	57.63	71.13	55.07
MISA [2]	58.40	72.84	45.48
MMIM[1]	59.03	72.71	56.69
LMF [4]	58.24	72.14	56.52
TFN [10]	58.58	72.36	56.34
Attention	59.16	73.00	56.70
DQ-Former	62.31 (3.15 ↑)	76.23 (3.23 ↑)	62.92 (6.22 ↑)

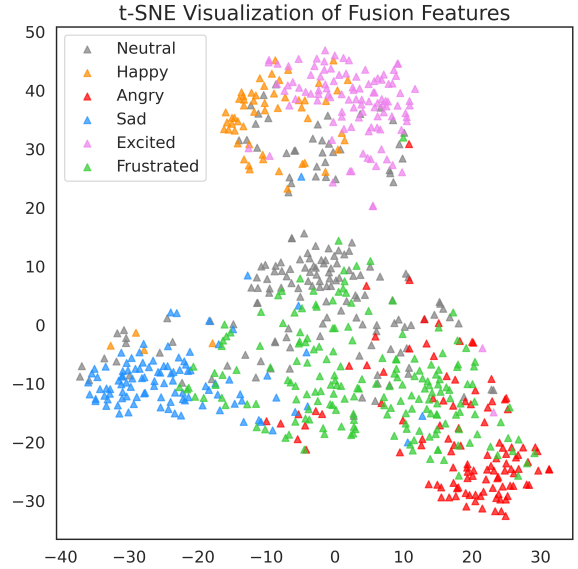


Figure 1: Visualization of fusion representations in 2D space by using t-SNE

in the space. Particularly notable is the almost complete separation achieved for emotion categories with significant differences, such as sad, angry, and excited. This suggests that the multimodal representations learned by our model contains effective emotional information.

¹The code is at <https://github.com/zeroQiaoba/MERTools>

Table 2: The performance of DQ-Former with various unimodal features. Legend: B (BERT), R (RoBERTa), W2 (Wav2Vec 2.0), W (WavLM), H (HuBERT). Best result highlighted.

Feature Extraction Model	IEMOCAP(4-way)		IEMOCAP(6-way)		MELD	
	WAA	WAF1	WAA	WAF1	WAA	WAF1
<i>Acoustic Only</i>						
WavLM-base	73.32	73.17	52.85	51.92	47.82	31.26
WavLM-large	78.21	78.08	56.78	56.48	44.21	32.62
wav2vec2-base	69.86	68.84	59.35	58.9	48.12	31.27
wav2vec2-large	74.13	73.18	56.37	56.48	48.05	34.06
HuBERT-base	75.15	75.18	52.17	51.32	44.6	36.07
HuBERT-large	70.88	70.28	55.28	54.6	53.22	48.06
<i>Textual Only</i>						
BERT-base	81.67	81.72	67.48	66.98	61.65	60.94
BERT-large	82.69	82.46	67.89	67.72	63.07	62.26
Roberta-base	83.3	83.22	64.63	64.82	63.26	60.28
Roberta-large	84.32	84.38	71.14	70.85	65.02	62.07
<i>Multimodal</i>						
B+W2	87.98	88.07	72.76	72.63	63.83	63.15
B+W	87.58	87.59	73.17	73.18	65.21	63.93
B+H	85.74	85.58	70.46	70.06	65.63	63.82
R+W2	85.54	85.36	71.82	71.8	64.14	63.83
R+W	87.17	87.17	69.92	69.43	64.25	62.8
R+H	85.34	85.18	73.31	73.19	65.4	64.45

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, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*. Association for Computational Linguistics, 9180–9192. <https://doi.org/10.18653/V1/2021.EMNLP-MAIN.723>
- [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 (Seattle, WA, USA) (MM '20)*. Association for Computing Machinery, New York, NY, USA, 1122–1131. <https://doi.org/10.1145/3394171.3413678>
- [3] Zheng Lian, Licai Sun, Yong Ren, Hao Gu, Haiyang Sun, Lan Chen, Bin Liu, and Jianhua Tao. 2024. MERBench: A Unified Evaluation Benchmark for Multimodal Emotion Recognition. *arXiv preprint abs/2401.03429* (2024). <https://doi.org/10.48550/ARXIV.2401.03429> arXiv:2401.03429
- [4] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2018. Efficient Low-rank Multimodal Fusion With Modality-Specific Factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*. Association for Computational Linguistics, 2247–2256. <https://doi.org/10.18653/V1/P18-1209>
- [5] Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. 2019. Found in Translation: Learning Robust Joint Representations by Cyclic Translations between Modalities. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*. AAAI Press, 6892–6899. <https://doi.org/10.1609/AAAI.V33I01.33016892>
- [6] Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Learning Factorized Multimodal Representations. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net. <https://openreview.net/forum?id=rygqqsA9KX>
- [7] 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 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 6558–6569. <https://doi.org/10.18653/v1/P19-1656>
- [8] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008). <https://jmlr.org/papers/v9/vandermaaten08a.html>
- [9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (Long Beach, California, USA), Vol. 30*. Curran Associates, Inc., 6000–6010. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
- [10] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor Fusion Network for Multimodal Sentiment Analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*. Association for Computational Linguistics, 1103–1114. <https://doi.org/10.18653/V1/D17-1115>
- [11] Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Memory fusion network for multi-view sequential learning. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence (New Orleans, Louisiana, USA)*. AAAI Press, Article 691, 8 pages.
- [12] 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.