

Supplementary Materials: GLoMo: Global-Local Modal Fusion for Multimodal Sentiment Analysis

Anonymous Authors

1 DATASETS

In this section, we elaborate on the experimental datasets utilized across three distinct tasks: multimodal sentiment analysis (MSA), multimodal humor detection (MHD), and multimodal emotion recognition (MER).

MSA involves predicting the intensity of emotions in spoken utterances. We evaluated GLoMo on two datasets: CMU Multimodal Opinion-level Sentiment Intensity Dataset (CMU-MOSI) [25] and CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) [26].

For the MHD task, which identifies humor in utterances, we utilize UR-FUNNY [5] and Multimodal Sarcasm Detection Dataset (MUSTARD) [1].

MER focuses on classifying the emotional content of utterances into multiple categories. We evaluated GLoMo on CHinese Emotion Recognition dataset with Modality-wise Annotations (CHERMA) [17], which includes seven emotions, with both unimodal and multimodal annotations.

The detailed introduction of the datasets are as follows:

CMU-MOSI consists of 93 English-language videos from 89 speakers, sourced from YouTube. These videos are segmented into 2,195 utterances, each rated on a scale from -3 to 3. We follow prior work [6, 13] in our dataset split: 1,281 for training, 229 for validation, and 685 for testing.

CMU-MOSEI expands on CMU-MOSI with 3,228 videos from 1,000 speakers. In line with previous studies [6, 13], we utilize 16,265 utterances for training, 1,869 utterances for validation, and 4,643 utterances for testing.

UR-FUNNY includes 1,866 videos from 1,741 speakers. Following [4], we use an updated version of the dataset, which has been cleaned of noisy and overlapping instances and more context sentences. It includes 9,588 utterances, split into 7,614 for training, 980 for validation, and 994 for testing.

MUSTARD consists of 690 videos sourced from TV shows. Following [4], we utilize 539 utterances for training, 68 utterances for validation, and 68 utterances for testing.

CHERMA features 28,717 Chinese utterances from various media, classified according to Ekman’s six basic emotions plus neutrality [2]. Each utterance is labeled with three unimodal labels and one multimodal label, distributed in a 6:2:2 ratio for training, validation, and testing.

It is important to note that due to varying utterance lengths, the feature lengths from different modalities may differ. To standardize this, we truncated the features to a maximum sequence length defined by the parameter *max seq. length*. Features shorter than this limit were zero-padded to reach the requisite length, as in [6, 13, 17].

2 BASELINE MODELS

In our study, we have selected a variety of multimodal fusion methods as baselines to conduct a comprehensive comparison for the given tasks. Due to the difference of the tasks, we choose different baselines. For CMU-MOSI and CMU-MOSEI datasets, we have considered a variety of methods that incorporate models such as TFN [24], LMF [9], MFM [20], GFN [13], and ICCN [18]. These models are designed to fuse global representations across the three modalities. In addition, we have taken into account approaches like MULT [19] and BBFN [3], as well as M3SA [27]. These methods initially fuse pairs of global representations and subsequently integrate them together. We also delve into techniques such as MISA [6], which segregate the global representations of modalities into components that are either specific to a modality or common across modalities. Moreover, we examine the significance of modality-specific tokens within each modality using algorithms like PRISA [12], and consider CubeMLP [16], which employs token-level fusion strategies. The state-of-the-art C-MIB [14] is also compared, which utilizes mutual information for denoising purposes.

For the UR-FUNNY and MUSTARD datasets, following [4], we opt for modified versions of MISA [6] and MAGBERT [15]. In these adaptations, BERT [7] is replaced with ALBERT [8] and XLNet [22] as text feature extractors. For CHERMA dataset, we select EFT [17] and LFT [17], which adapt transformer models instead of the models in [21] and [23]. Furthermore, we include models like PMR [11] and LFMIM [17] in our comparison. These models leverage unimodal labels of each modality for emotion prediction while our GLoMo not.

The details of the baseline models mentioned are as follows:

TFN [24] disentangles unimodal, bimodal and trimodal dynamics by modeling each of them explicitly using three-fold Cartesian product.

LMF [9] feeds three modality-specific representations into three unimodal networks, then performs the low-rank multimodal fusion with modality-specific factors.

MFM [20] introduces a model that separates representations into shared discriminative factors for prediction tasks and unique generative factors for each modality.

GFN [13] utilizes adversarial training to learn a unified embedding space for different modalities, bridging the gap between them and enhancing multimodal fusion.

MULT [19] repeats reinforcing one modality’s features from the another modality using pair-wise cross-modal attention to handle the problem caused by non-alignment and long-range dependencies.

MAGBERT [15] integrates the text, visual and acoustic modalities into a multimodal transformer for finetuning through generating a shift to internal representation.

M3SA [27] employs a modulation loss to fine-tune the learning process based on the confidence of each modality and a modality filter module to sift out irrelevant noise, leading to enhanced unimodal and cross-modal learning.

ICCN [18] utilizes deep canonical analysis to discover hidden correlations across text, audio and video.

CubeMLP [16] introduces a novel MLP-based framework that integrates information from various modalities using feature-mixing techniques.

MISA [6] projects each modality to their modality-specific subspace and modality-invariant subspace, thus obtaining holistic view of the multimodality.

BBFN [3] enhances representation by simultaneously fusing and separating pairwise modality representations, with a gated control mechanism in the transformers to refine the output.

PriSA [12] mitigates false correlations in text by employing preferential fusion and distance-aware contrastive learning. Initially, it calculates inter-modal correlations guided by text, followed by processing these features through distance-aware contrastive learning to determine mixed-modal correlations. Ultimately, sentiment information is identified by combining these mixed-modal correlations with discriminative intra-modal features extracted from visual and audio modalities using a self-attention module.

C-MIB [14] uses the Information Bottleneck (IB) constraint to get free-of-noisy unimodal representation.

EFT [17] and **LFT** [17] replace the deep neural networks (DNN) with transformer in Early Fusion DNN [21] and Later Fusion DNN [23], respectively.

PMR [11] introduces a message hub sending common messages to each modality and reinforces their features via cross-modal attention. Besides, the reinforced features from each modality are collected to generate a reinforced common message to progressively complement each other.

LFMIM [17] uses the modality-specific transformer encoder to learn the unimodal information and use a multimodal transformer encoder to learn the multimodal representation.

3 IMPLEMENTATION DETAILS

In this section, we provide additional details of the experiment setups. All experiments were conducted on a GTX3090 GPU with CUDA version 11.5 and PyTorch version 1.12.1. The AdamW [10] optimizer was employed for all runs, with a fixed random seed of 5576. Due to inherent differences across datasets, specific implementation procedures also varied accordingly. Specifically, for the CMU-MOSI and SMU-MOSEI datasets, we adhered strictly to the methodologies described. For the UR-FUNNY and MUSTARD datasets, which incorporate contextual information alongside the original text, we followed the precedent works [4]. This involved concatenating the contextual data with the original utterances prior to their introduction into unimodal encoder networks for representation learning. For CHERMA, which does not provide raw text but only textual modality features, the encoders for all three modalities—text, audio, and video—were identical, aligning with the processing methods used in audio encoder as described.

To determine the suitable hyperparameters, we employed a grid-search methodology across the hyperparameter space to identify

the model that yields the lowest validation loss for classification or regression tasks as in [6]. Specifically, we explored finite sets of hyperparameter values, including learning rate from $\{1e-5, 2e-5, 3e-5, 4e-5\}$, hidden dimensions from $\{48, 96, 112, 160, 192, 256\}$, max seq. length from $\{50, 60, 70, 80\}$, and transformer encoder layers in modality-specific encoder from $\{3, 4, 5, 6, 7\}$. The final hyperparameters GLoMo used throughout datasets are listed in Table 1.

4 MORE RESULTS

In this section, we present additional experimental results that illustrate the performance of GLoMo on the CHERMA dataset as the number of modality-specific experts increases. Specifically, we conducted experiments with varying numbers of experts for text, visual, and audio modalities, set at 1, 2, 3, and 4, resulting in a total of 64 different configurations, as depicted in the Fig. 2. When the number of experts is set to one, the MoEs simplifies to a two-layer MLP network. The mean F1 scores for each modality and number of experts on the CHERMA are depicted in Fig. 1. As the number of experts for each modality grows, we observe a consistent improvement in performance. This trend could be attributed to the fact that each expert focuses on different local representations, and a greater number of experts allows for the integration of more detailed information pertaining to specific types of sentiments.

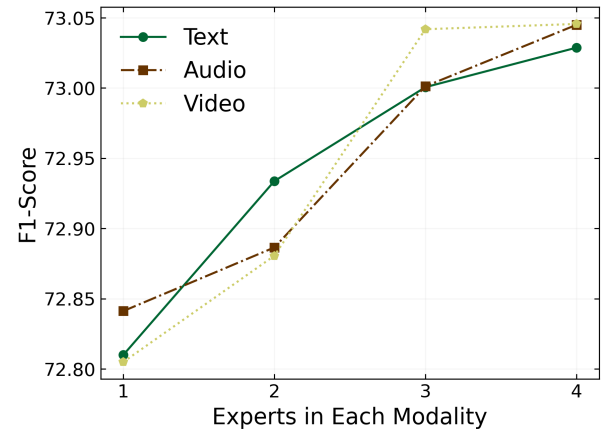


Figure 1: Ablation studies on the number of local representations and experts on CHERMA.

REFERENCES

- [1] Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. 2019. Towards multimodal sarcasm detection (an _obviously_ perfect paper). *arXiv preprint arXiv:1906.01815* (2019).
- [2] Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion* 6, 3-4 (1992), 169–200.
- [3] Wei Han, Hui Chen, Alexander Gelbukh, Amir Zadeh, Louis-philippe Morency, and Soujanya Poria. 2021. Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis. In *Proceedings of the 2021 international conference on multimodal interaction*. 6–15.
- [4] Md Kamrul Hasan, Sangwu Lee, Wasifur Rahman, Amir Zadeh, Rada Mihalcea, Louis-Philippe Morency, and Ehsan Hoque. 2021. Humor knowledge enriched transformer for understanding multimodal humor. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 12972–12980.

Table 1: Final hyperparameter values in each dataset. Here ‘transformer encoder layers’ denotes the transformer encoder layers of modality-specific encoder in unimodal encoding module.

Hyper-param	MOSI	MOSEI	UR-FUNNY	MUSTARD	CHERMA
batch size	240	64	220	64	400
transformer encoder layers	4	3	3	3	7
max seq. length	60	80	80	70	50
hidden dimensions d	48	192	112	160	256
learning rate	4e-5	1e-5	2e-5	2e-5	2e-5
num. of local representations	3	3	3	3	3
experts of MoEs	3	3	3	3	3
activated experts k	2	2	2	2	2
ω	1e-2	1e-2	1e-2	1e-2	1e-2

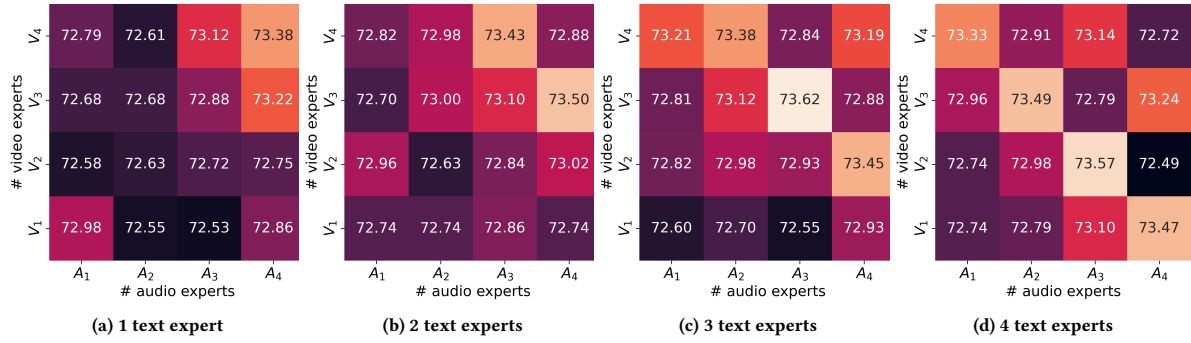


Figure 2: F1 score when increasing the number of the experts of text, audio and video.

- [5] Md Kamrul Hasan, Wasifur Rahman, Amir Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, et al. 2019. UR-FUNNY: A multimodal language dataset for understanding humor. *arXiv preprint arXiv:1904.06618* (2019).
- [6] 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.
- [7] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*. 4171–4186.
- [8] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *International Conference on Learning Representations*.
- [9] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher 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 (Volume 1: Long Papers)*. 2247–2256.
- [10] Ilya Loshchilov and Frank Hutter. 2018. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.
- [11] Fengmao Lv, Xiang Chen, Yanyong Huang, Lixin Duan, and Guosheng Lin. 2021. Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2554–2562.
- [12] Feipeng Ma, Yueyi Zhang, and Xiaoyan Sun. 2023. Multimodal Sentiment Analysis with Preferential Fusion and Distance-aware Contrastive Learning. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 1367–1372.
- [13] Sijie Mai, Haifeng Hu, and Songlong Xing. 2020. Modality to modality translation: An adversarial representation learning and graph fusion network for multimodal fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 164–172.
- [14] Sijie Mai, Ying Zeng, and Haifeng Hu. 2022. Multimodal information bottleneck: Learning minimal sufficient unimodal and multimodal representations. *IEEE Transactions on Multimedia* (2022).
- [15] Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. 2020. Integrating multimodal information in large pretrained transformers. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, Vol. 2020. NIH Public Access, 2359.
- [16] Hao Sun, Hongyi Wang, Jiaqing Liu, Yen-Wei Chen, and Lanfen Lin. 2022. CubeMLP: An MLP-based model for multimodal sentiment analysis and depression estimation. In *Proceedings of the 30th ACM international conference on multimedia*. 3722–3729.
- [17] Jun Sun, Shoukang Han, Yu-Ping Ruan, Xiaoning Zhang, Shu-Kai Zheng, Yulong Liu, Yuxin Huang, and Taihao Li. 2023. Layer-wise Fusion with Modality Independence Modeling for Multi-modal Emotion Recognition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 658–670.
- [18] 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.
- [19] 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*, Vol. 2019. NIH Public Access, 6558.
- [20] Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Learning Factorized Multimodal Representations. In *International Conference on Representation Learning*.
- [21] Jennifer Williams, Steven Kleinogesse, Ramona Comanescu, and Oana Radu. 2018. Recognizing emotions in video using multimodal dnn feature fusion. In *Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML)*. 11–19.
- [22] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems* 32 (2019).
- [23] Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. 2020. Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th annual meeting of the association for computational linguistics*. 3718–3727.
- [24] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor Fusion Network for Multimodal Sentiment Analysis. In

349	<i>Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing</i> . 1103–1114.		
350	[25] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. <i>arXiv preprint arXiv:1606.06259</i> (2016).		
351	[26] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: Cmu-mosei		
352		dataset and interpretable dynamic fusion graph. In <i>Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> . 2236–2246.	407
353		[27] Ying Zeng, Sijie Mai, and Haifeng Hu. 2021. Which is Making the Contribution: Modulating Unimodal and Cross-modal Dynamics for Multimodal Sentiment Analysis. In <i>Findings of the Association for Computational Linguistics: EMNLP 2021</i> . 1262–1274.	408
354			409
355			410
356			411
357			412
358			413
359			414
360			415
361			416
362			417
363			418
364			419
365			420
366			421
367			422
368			423
369			424
370			425
371			426
372			427
373			428
374			429
375			430
376			431
377			432
378			433
379			434
380			435
381			436
382			437
383			438
384			439
385			440
386			441
387			442
388			443
389			444
390			445
391			446
392			447
393			448
394			449
395			450
396			451
397			452
398			453
399			454
400			455
401			456
402			457
403			458
404			459
405			460
406			461