GLoMo: Global-Local Modal Fusion for Multimodal Sentiment Analysis (Appendix)

ACM Reference Format:
. 2024. GLoMo: Global-Local Modal Fusion for Multimodal Sentiment Analysis (Appendix). In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24), October 28-November 1, 2024, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3664647.3681527

1 DATASETS

In this section, we elaborate on the experimental datasets utilized across three multimodal 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) [28] and CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) [29].

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) [20], 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, 15] 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, 15], 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 prior study [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, 15, 20].

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 [27], LMF [11], MFM [23], GFN [15], and ICCN [21]. These models are designed to fuse global representations across the three modalities. In addition, we have taken into account approaches like MULT [22] and BBFN [3], as well as M3SA [30]. 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 [14], and consider CubeMLP [19], which employs token-level fusion strategies. The state-of-the-art C-MIB [16] 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 [17]. In these adaptations, BERT [7] is replaced with ALBERT [8] and XLNet [25] as text feature extractors. For CHERMA dataset, we select EFT [20] and LFT [20], which adapt transformer models instead of the models in [24] and [26]. Furthermore, we include models like PMR [13] and LFMIM [20] 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 [27] disentangles unimodal, bimodal and trimodal dynamics by modeling each of them explicitly using three-fold Cartesian product.

LMF [11] feeds three modality-specific representations into three unimodal networks, then performs the low-rank multimodal fusion with modality-specific factors.

MFM [23] introduces a model that separates representations into shared discriminative factors for prediction tasks and unique generative factors for each modality.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
MM '24, October 28-November 1, 2024, Melbourne, VIC, Australia
© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0686-8/24/10
https://doi.org/10.1145/3664647.3681527

GFN [15] utilizes adversarial training to learn a unified embedding space for different modalities, bridging the gap between them and enhancing multimodal fusion.

MULT [22] 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 [17] integrates the text, visual and acoustic modalities into a multimodal transformer for finetuning through generating a shift to internal representation.

M3SA [30] 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 [21] utilizes deep canonical analysis to discover hidden correlations across text, audio and video.

CubeMLP [19] 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 [14] mitigates false correlations in text by employing preferential fusion and distance-aware contrastive learning.

C-MIB[16] uses the Information Bottleneck (IB) constraint to get free-of-noisy unimodal representation.

MSG [10] utilizes a multi-stage fusion framework with a hybrid-modal attention mechanism to dynamically refine multimodal representations.

EFT [20] and **LFT** [20] replace the deep neural networks (DNN) with transformer in Early Fusion DNN [24] and Later Fusion DNN [26], respectively.

PMR[13] 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[20] 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 [12] 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.

Algorithm 1: GLoMo Framework

Input: Unimodal representations U_m, $m \in \{t, a, v\}$, hyperparameters
Output: prediction $\hat{y}$

Initialize the network ;
while *not done* **do**
  Sample a batch of utterances;
  for *utterances in the batch* **do**
    Module: Unimodal Coding ;
    $X_t^g = \text{FF}\left(CLS^{-1}\left(U_t\right)\right)$;
    $X_t^l = \text{CAT}\left(\text{AdaMaxPool}_n\left(\text{Conv1D}\left(\left[BERT^{-1}, BERT^{-2}\right]\right)\right)\right)$;
    $X_a^g = \text{MAX}\left(\text{TE}\left(\text{Conv1D}\left(U_a\right)\right)\right)$;
    $X_v^g = \text{MAX}\left(\text{TE}\left(\text{Conv1D}\left(U_v\right)\right)\right)$;
    $X_a^l = \text{CAT}\left(\text{AdaMaxPool}_n\left(\text{TE}\left(\text{Conv1D}\left(U_a\right)\right)\right)\right)$;
    $X_v^l = \text{CAT}\left(\text{AdaMaxPool}_n\left(\text{TE}\left(\text{Conv1D}\left(U_v\right)\right)\right)\right)$;
    Module: Modality-specific MoEs ;
    for *modality m, m* $\in \{t, a, v\}$ **do**
      $G(X_m^l) = \text{Softmax}(\text{KeepTopK}(X_m^l W_g, \text{k}))$;
      $\hat{X}_m^l = \sum_{i=1}^{s} G(X_m^l)_i E_i(X_m^l)$;
    end
    Module: Global-guided Fusion ;
    $M = [X_t^g, X_a^g, X_v^g, \hat{X}_t^l, \hat{X}_a^l, \hat{X}_v^l]$;
    $\left[Z_t^g, Z_a^g, Z_v^g, Z_t^l, Z_a^l, Z_v^l\right] = \text{TE}\left(M\right)$;
    for *modality m, m* $\in \{t, a, v\}$ **do**
      $Z_m = \text{MLP}_m\left(\left[Z_m^g, Z_m^l\right]\right)$;
    end
    $W_g, Z_g = \text{ATTN}_g\left(\left[Z_t^g, Z_a^g, Z_v^g\right]\right)$;
    $W_l, Z_l = \text{ATTN}_l\left(\left[Z_t^l, Z_a^l, Z_v^l\right]\right)$;
    $Z_1 = \text{MLP}_f\left(\left[Z_g, Z_l\right]\right)$;
    $Z_2 = \text{MIXUP}\left(\left(Z_t, Z_a, Z_v\right), W_g\right)$;
    $\hat{y} = \text{MLP}\left(\left[Z_1, Z_2\right]\right)$;
    Optimization object ;
    for *modality m, m* $\in \{t, a, v\}$ **do**
      $\mathcal{L}_{MoE}^m(X_m^l) = \omega \cdot (\text{CV}(\text{Importance}(X_m^l))^2 + \text{CV}(\text{Load}(X_m^l))^2)$;
    end
    if *regression task* **then**
      $\mathcal{L}_{\text{task}} = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}_i|$;
    else
      $\mathcal{L}_{\text{task}} = \frac{1}{N}\sum_{i=1}^{N} -y_i log\left(\hat{y}_i\right)$;
    end
    $\mathcal{L} = \mathcal{L}_{task} + \mathcal{L}_{MoE}^t + \mathcal{L}_{MoE}^a + \mathcal{L}_{MoE}^v$;
    Update the network according to $\mathcal{L}$;
  end
end

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

For CHERMA, the encoders for all three modalities—text, audio, and video—were identical, aligning with the processing methods used in audio encoder as described.

To illustrate our GLoMo framework clearly, we provide pseudocode in this section, as shown in Algorithm 1. Detailed introduction of each module and formula can be found in the main text, with additional explanations here for clarity. The $Conv1D(\cdot)$ uses a kernel size of 3x3, $AdaMaxPool_n(\cdot)$ dynamically selects kernel sizes based on n for max pooling along the length dimension, $CAT(\cdot)$ denotes concatenation, $MAX(\cdot)$ performs global max pooling across the length dimension. $TE(\cdot)$ signifies transformer encoder layers, $KEEPTOPK(\cdot)$ selects the top k experts, and each expert $E_i(\cdot)$ is composed of two MLP layers and a ReLU activation function. $CV(\cdot)$ denotes the coefficient of variation, $Importance(\cdot)$ refers to the weighted importance scores of various expert networks, which is the output of gating network G, and $Load(\cdot)$ calculates the number of load samples currently present in each of the expert networks, which is defined in [18], ω is hyperparameter with default value $1e-2$. $MIXUP$ is the weighted average function and $FF(\cdot)$ is a linear layer. Note that all MLP-related layers, including $MLP_m(\cdot)$, $MLP_f(\cdot)$ and $MLP(\cdot)$, share the same structure but with different parameters.

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 ADDITIONAL RESULTS

In this section, we present additional experimental results of the proposed GLoMo, including the performance on UR-FUNNY and MUStARD, and performance on the CHERMA dataset as the number of modality-specific experts increases.

4.1 Multimodal Humor Detection

Tables 2 presents the performance of the GLoMo in multimodal humor detection. GLoMo notably surpassed existing top-performing models on both benchmarks, establishing a new state-of-the-art. Specifically, GLoMo achieved a 2.52% improvement on the UR-FUNNY dataset and a significant gain of 7.39% on the MUStARD dataset. This improvement can be attributed to the extended duration of these datasets, where the target speaker's contribution is a small part of the overall context. Global representations, such as average pooling, may overlook critical details like speaker-specific nuances and contextual transitions. In contrast, local representations focus on smaller segments, preserving detailed and pertinent information necessary for accurate classification. This underscores the significance of local representations in capturing nuanced details that global methods might miss.

Table 2: The comparison with baselines on UR-FUNNY and MUStARD, in terms of ACC-2. Models in parentheses indicates the textual features used. ⋆ from [4]

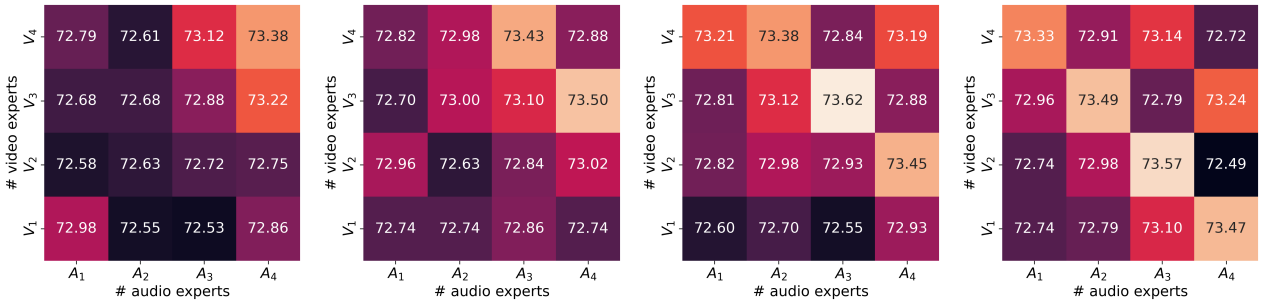
	UR-FUNNY (↑)	MUStARD (↑)
MISA [6] (BERT) ⋆	69.62	66.18
MISA [6] (ALBERT) ⋆	69.82	66.18
MAGBERT [17] (ALBERT) ⋆	67.20	69.12
MAGBERT [17] (XLNet) ⋆	<u>72.43</u>	<u>76.47</u>
AGM (BERT) [9]	65.97	-
MIL (BERT) [31]	-	76.36
GLoMo (BERT)	**74.95**	**83.86**

4.2 Varying numbers of MoEs in CHERMA

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. 1. When the number of experts is set to one, the MoEs simplifies to a two-layer MLP network. 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.

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).



(a) 1 text expert **(b) 2 text experts** **(c) 3 text experts** **(d) 4 text experts**

Figure 1: F1 score when increasing the number of the experts of text, audio and video.

[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.

[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] Hong Li, Xingyu Li, Pengbo Hu, Yinuo Lei, Chunxiao Li, and Yi Zhou. 2023. Boosting Multi-modal Model Performance with Adaptive Gradient Modulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 22214–22224.

[10] Ronghao Lin and Haifeng Hu. 2023. Dynamically shifting multimodal representations via hybrid-modal attention for multimodal sentiment analysis. *IEEE Transactions on Multimedia* (2023).

[11] 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.

[12] Ilya Loshchilov and Frank Hutter. 2018. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

[13] 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.

[14] 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.

[15] 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.

[16] Sijie Mai, Ying Zeng, and Haifeng Hu. 2022. Multimodal information bottleneck: Learning minimal sufficient unimodal and multimodal representations. *IEEE Transactions on Multimedia* (2022).

[17] 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.

[18] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2016. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. In *International Conference on Learning Representations*.

[19] Hao Sun, Hongyi Wang, Jiaqing Liu, Yen-Wei Chen, and Lanfen Lin. 2022. Cube-MLP: An MLP-based model for multimodal sentiment analysis and depression estimation. In *Proceedings of the 30th ACM international conference on multimedia*. 3722–3729.

[20] 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.

[21] 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.

[22] 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.

[23] 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*.

[24] Jennifer Williams, Steven Kleinegesse, 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.

[25] 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).

[26] 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.

[27] 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*. 1103–1114.

[28] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259* (2016).

[29] 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.

[30] Ying Zeng, Sijie Mai, and Haifeng Hu. 2021. Which is Making the Contribution: Modulating Unimodal and Cross-modal Dynamics for Multimodal Sentiment Analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. 1262–1274.

[31] Yazhou Zhang, Yang Yu, Dongming Zhao, Zuhe Li, Bo Wang, Yuexian Hou, Prayag Tiwari, and Jing Qin. 2023. Learning multi-task commonness and uniqueness for multi-modal sarcasm detection and sentiment analysis in conversation. *IEEE Transactions on Artificial Intelligence* (2023).