Supplementary Materials: Multimodal Emotion Recognition Calibration in Conversations

Anonymous Authors

1 MERC BASELINES

DialogueRNN [13] is a computational model designed to capture the emotional dynamics within conversations. It operates through three principal components: firstly, each participant is depicted by a dynamic "party state" that evolves in response to their utterances. Secondly, the conversational context is encapsulated by a "global state" that is collectively maintained by all participants, integrating prior utterances and party states to construct a comprehensive contextual representation. Lastly, the model derives an emotional representation from the current speaker's state, alongside the states of previous speakers, which is subsequently utilized for emotion classification.

DialogueGCN [5] adeptly captures both speaker-agnostic and speaker-dependent contextual information within dialogues. The model employs a sequence-level context encoder to derive comprehensive discourse-level representations. Subsequently, it incorporates a specialized speaker-level context encoder, structured as a graph network, to elucidate nuanced, speaker-specific contextual cues. This encoder accounts for both the inter-dependencies and intra-dependencies among participants, thereby proficiently capturing the dynamic nature of the dialogue. The interactions within the dialogue are further represented through a directed graph. Furthermore, DialogueGCN utilizes a locality-based convolutional feature transformation process to enhance the refinement of speaker-level context encoding features. This methodology facilitates the extraction of more detailed contextual information, thereby enriching the understanding of dialogues.

MMGCN [9] represents a sophisticated approach in the realm of multimodal fusion using graph convolutional networks. This model encapsulates the encoding of multimodal contextual information through a spectral-domain graph convolution network, enhanced by the addition of multiple stacked layers to deepen the GCN architecture. MMGCN incorporates learned speaker embeddings, aiming to capture speaker-level contextual information. Such embeddings are pivotal for simulating dependencies both among different speakers and within the same speaker. Consequently, MMGCN is adept at leveraging multimodal dependencies and utilizing speaker information to model interpersonal and intrapersonal relational dynamics effectively.

MMDFN [7] leverages modal encoders to monitor the state and context of a speaker across various modalities. It enhances graph convolutional layers with a gating mechanism and introduces an innovative graph-based dynamic fusion module to integrate multimodal contextual information. This module employs graph convolution operations to aggregate inter-modal and intra-modal contextual information within designated semantic spaces of each layer. Simultaneously, it uses a gating mechanism to discern the inherent sequential patterns of contextual information across adjacent semantic spaces. MMDFN effectively regulates the flow of information between layers, reduces redundancy, and enhances the complementarity among modalities. By embedding multimodal context features into dynamic semantic spaces, it realizes a comprehensive integration of contextual and semantic information.

M3Net [2] addresses the limitations of the existing MERC framework, which has exhibited constrained capabilities in handling the multivariate relationships and multi-frequency information inherent to dialogic contexts. M3Net introduces a novel Hypergraph representation and a tailored Multivariate Propagation module to effectively model the intricate interconnections present in conversational data. By concurrently implementing both low-pass and high-pass filtering mechanisms, M3Net is able to selectively extract salient signals from the node features, thereby achieving state-of-the-art performance on relevant benchmarks.

2 CONFIDENCE CALIBRATION BASELINES

T-Scale [6] Short for Temperature scaling, is recognized as a straightforward yet efficacious calibration technique that enhances the precision of a model's confidence assessments. The core of this method is a single scalar parameter that is fine-tuned by minimizing the negative log-likelihood (NLL) on a validation subset. The calibration process is executed in a two-step manner: first, the model's softmax logits are computed; these logits are then scaled by dividing by T, where $T > 1$. Optimization of T is conducted with the sole objective of NLL minimization on the validation data. Upon identifying the optimal T, the model is retrained, with applying the newly calibrated temperature scaling. Through this adjustment of the temperature parameter, temperature scaling harmonizes the model's confidence scores with the actual probabilities of the predicted outcomes, thereby ensuring that the confidence estimates are more representative of the model's veritable performance.

Ensemble [19] by aggregating predictions from multiple models, effectively mitigating uncertainties associated with model parameters and data. In regions of the feature space that are underrepresented by the training data, the variability in predictions from individual ensemble members inherently increases. This increased variability inversely affects the confidence associated with these predictions. The process of aggregating predictions across several models helps to diminish the impact of specific model- and data-related uncertainties, thereby enhancing the reliability and robustness of the ensemble's confidence estimates, particularly in areas of the feature space where individual models may exhibit less certainty.

CRL [14] stands for Correctness Ranking Loss. During the training process, it estimates the true class probabilities based on the number of times a sample is correctly classified, arranging them in the desirable ordinal order of confidence estimates. This helps the classifier learn ordinal relationships. CRL is an effective regularization method that can be used to train deep neural networks to mitigate the well-known issue of overconfident predictions.

FMFP [18] (Flat Minima for Failure Prediction) strategy is predicated on the theory that the confidence gap between correctly and incorrectly predicted samples is larger in flat minima compared to sharp minima. By employing Stochastic Weight Averaging (SWA)

Table 1: Detailed hyperparameter settings.

Hyperparameters	IEMOCAP	MELD
Batch size	16	16
Epochs	80	15
d_v	342	342
d_a	1582	300
d_t	1024	1024
Dropout rate	0.5	0.4
Weight decay	0.00003	0.00003
Learning rate for network learning	0.0001	0.0001
Optimizer for network learning	Adam	Adam
The weight of $\mathcal{L}_m$, $\mathcal{L}_c$, $\mathcal{L}_s$	0.05, 0.05, 0.05	0.15, 0.15, 0.02
The number of buckets	7	3
The number of graph layer	4	3
The max training step of buckets	1	1
The number of hypergraph layer	3	3
The maximum conversation length	110	33

and Sharpness-Aware Minimization (SAM) as representative methods, FMFP effectively seeks flat minima in deep neural networks (DNNs), thereby enhancing the model's calibration.

CML [12] Based on the principle that the fundamental nature of information is to reduce uncertainty, a regularization loss is implemented. This loss penalizes samples for which the estimated confidence increases following the removal of a modality. Additionally, a strategic sampling method is utilized to enhance computational efficiency. This methodology has demonstrated notable success in practical applications.

3 FEATURE EXTRACTION

Textual Features: To obtain context-independent utterance-level feature vectors, we follow [4] to fine-tune the Roberta Large model to predict emotion labels of utterances. Let an utterance $\mathbf{u}_i$ be a sequence of tokens after applying Byte Pair Encoding (BPE), denoted as $\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_N$. The emotion label associated with $\mathbf{u}_i$ is represented by $\mathbf{e}_i$, where $\mathbf{e}_i$ belongs to the set of emotion labels E. To prepare the input sequence for the RoBERTa model, we add a special token $[CLS]$ at the beginning of the original utterance. The sequence now becomes $[CLS], \mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_N$. This modified sequence is then fed into the Roberta model. The output of the last layer corresponding to the $[CLS]$ token is used as input to a small feed-forward network, which performs the classification into the appropriate emotion class. After fine-tuning the model for emotion classification, we utilize the model for generating feature vectors. We pass the BPE tokenized utterance appended with $[CLS]$ through the model and extract the outputs from the last four layers corresponding to the $[CLS]$ tokens. The four vectors are combined through averaging, resulting in a context-independent utterance-level feature vector with a dimensionality of 1024.

Acoustic Features: Regarding the extraction of acoustic features, we adopt the methodology outlined in the study by [13], utilizing the openSMILE toolkit [3] for this purpose. The chosen acoustic features undergo a subsequent normalization process, after which a fully connected layer is utilized to achieve dimensionality reduction. Notably, the dimensions of the resulting acoustic features differ between datasets: 1582 for IEMOCAP and reduced to 300 for MELD.

Table 2: Comparison of results against various LLMs. ♣, ♠, ♦ and ■ results come from [16], [15], [11] and [17], respectively.

Methods	IEMOCAP	MELD
Ours	71.98	66.85
Zero-shot		
♣ ChatGPT	40.07	54.37
♦ ChatGLM	38.60	38.80
♦ ChatGLM2	21.10	21.80
♦ Llama	0.75	9.12
♦ Llama2	2.77	16.28
Few-shot		
■ ChatGPT 1-shot	47.46	58.63
■ ChatGPT 3-shot	48.58	58.35
LORA + Backbone		
♠ Curie	57.33	65.01
♦ ChatGLM	18.94	40.54
♦ ChatGLM2	52.88	64.85
♦ Llama	55.81	66.15
♦ Llama2	55.96	65.84

Visual Features: For visual facial expression features, we employ a DenseNet architecture [10], which is pre-trained on the Facial Expression Recognition Plus (FER+) corpus [1], similar to the approach used in [9]. The dimension of the visual facial expression feature is 342 for each dataset.

4 DETAILED HYPERPARAMETER SETTINGS

All re-implementation methods have released their source codes, ensuring identical settings as the original papers. For the CMERC, hyperparameters γ_m, γ_c, γ_s, and τ are manually tuned for each dataset using hold-out validation. The specific results of the hyperparameter search are presented in Table 1. The reported results are the average score of 5 random runs on the test set. Our experiments are conducted on a single RTX 4090 GPU. The code will be open-sourced upon acceptance of the paper.

5 COMPARING RESULTS AGAINST LARGE LANGUAGE MODELS (LLMS)

Table 2 provides a clear comparison of various sentiment analysis methods on the IEMOCAP and MELD datasets. Our approach achieves notably high accuracy, reaching 71.98% and 66.85% on the respective datasets, highlighting its superiority in emotion classification tasks. In contrast, ChatGPT's performance is comparatively lower in the zero-shot scenario, with 40.07% (IEMOCAP) and 54.37% (MELD). Other zero-shot methods like ChatGLM, ChatGLM2, Llama, and Llama2 also fall short of our method. In the few-shot scenario, ChatGPT 3-shot shows a slight improvement but still lags behind our approach. In comparison with large models (LORA [8] + Backbone), our method outperforms models such as Curie, ChatGLM, ChatGLM2, Llama, and Llama2 on both IEMOCAP and MELD. Overall, our method excels in comparisons with LLMs, surpassing ChatGPT not only in zero-shot and few-shot scenarios but also demonstrating superior performance within the LORA + Backbone framework compared to other models.

REFERENCES

[1] Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. 2016. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *Proceedings of the 18th ACM international conference on multimodal interaction*. 279–283.

[2] Feiyu Chen, Jie Shao, Shuyuan Zhu, and Heng Tao Shen. 2023. Multivariate, Multi-Frequency and Multimodal: Rethinking Graph Neural Networks for Emotion Recognition in Conversation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10761–10770.

[3] Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*. 1459–1462.

[4] Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. COSMIC: COmmonSense knowledge for eMotion identification in conversations. In *Findings of EMNLP*. 2470–2481.

[5] Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. *arXiv preprint arXiv:1908.11540* (2019).

[6] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*. PMLR, 1321–1330.

[7] Dou Hu, Xiaolong Hou, Lingwei Wei, Lianxin Jiang, and Yang Mo. 2022. MM-DFN: Multimodal Dynamic Fusion Network for Emotion Recognition in Conversations. In *ICASSP*. IEEE, 7037–7041.

[8] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685* (2021).

[9] Jingwen Hu, Yuchen Liu, Jinming Zhao, and Qin Jin. 2021. MMGCN: Multimodal Fusion via Deep Graph Convolution Network for Emotion Recognition in Conversation. In *ACL*. 5666–5675.

[10] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *CVPR*. 4700–4708.

[11] Shanglin Lei, Guanting Dong, Xiaoping Wang, Keheng Wang, and Sirui Wang. 2023. Instructerc: Reforming emotion recognition in conversation with a retrieval multi-task llms framework. *arXiv preprint arXiv:2309.11911* (2023).

[12] Huan Ma, Qingyang Zhang, Changqing Zhang, Bingzhe Wu, Huazhu Fu, Joey Tianyi Zhou, and Qinghua Hu. 2023. Calibrating multimodal learning. In *International Conference on Machine Learning*. PMLR, 23429–23450.

[13] Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive rnn for emotion detection in conversations. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 6818–6825.

[14] Jooyoung Moon, Jihyo Kim, Younghak Shin, and Sangheum Hwang. 2020. Confidence-aware learning for deep neural networks. In *international conference on machine learning*. PMLR, 7034–7044.

[15] Geng Tu, Bin Liang, Xiucheng Lyu, Lin Gui, and Ruifeng Xu. 2023. Do topic and causal consistency affect emotion cognition? a graph interactive network for conversational emotion detection. In *In The 26th European Conference on Artificial Intelligence (ECAI'23)*. 2362–2369.

[16] Geng Tu, Bin Liang, Bing Qin, Kam-Fai Wong, and Ruifeng Xu. 2023. An Empirical Study on Multiple Knowledge from ChatGPT for Emotion Recognition in Conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. 12160–12173.

[17] Weixiang Zhao, Yanyan Zhao, Xin Lu, Shilong Wang, Yanpeng Tong, and Bing Qin. 2023. Is ChatGPT Equipped with Emotional Dialogue Capabilities? *arXiv preprint arXiv:2304.09582* (2023).

[18] Fei Zhu, Zhen Cheng, Xu-Yao Zhang, and Cheng-Lin Liu. 2022. Rethinking confidence calibration for failure prediction. In *European Conference on Computer Vision*. Springer, 518–536.

[19] Fei Zhu, Zhen Cheng, Xu-Yao Zhang, and Cheng-Lin Liu. 2022. Rethinking Confidence Calibration fornbsp;Failure Prediction. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXV* (Tel Aviv, Israel). Springer-Verlag, Berlin, Heidelberg, 518–536. https://doi.org/10.1007/978-3-031-19806-9_30