Interspeech 2024
1-5 September 2024, Kos, Greece

LoRA-MER: Low-Rank Adaptation of Pre-Trained Speech Models for Multimodal Emotion Recognition Using Mutual Information

Yunrui Cai[1], *Zhiyong Wu*[1,2,3,†], *Jia Jia*[1], *Helen Meng*[1,2]

[1] Shenzhen International Graduate School, Tsinghua University, Shenzhen, China
[2] The Chinese Unviersity of Hong Kong, Hong Kong SAR, China
[3] Peng Cheng Lab, Shenzhen, China

caiyr22@mails.tsinghua.edu.cn, jjia@tsinghua.edu.cn, {zywu, hmmeng}@se.cuhk.edu.hk

Abstract

Multimodal emotion recognition (MER) is crucial for machines to understand human intentions. Although many deep learning models have been proposed, MER still faces practical challenges. The key challenge is how to extract high-dimensional features that are more relevant to emotions. Another challenge is how to effectively model multimodal features, achieving a balance between similarity and diversity. In this paper, we propose the method of LoRA-MER[1] using mutual information. We fine-tune a pre-trained speech model with Low-Rank Adaptation (LoRA) strategy and utilize a frozen pre-trained text model to robustly extract emotional features. Additionally, we adopt a multimodal fusion approach based on Mutual Information Neural Estimation (MINE) to enhance their correlation. Experimental results demonstrate the effectiveness of each module proposed in our method, and the performance of our model surpasses that of state-of-the-art speaker-independent approaches on IEMOCAP dataset.

Index Terms: multimodal emotion recognition, parameter-efficient fine-tuning, pre-trained model, mutual information

1. Introduction

Accurate recognition of human emotions plays an important role in harmonious human-machine interactions (HCI) [1]. The task of multimodal emotion recognition (MER) requires perceiving and comprehending human emotional states using audio and text modalities, and classifies each utterance to a specific category from a predetermined set [2].

At present, a typical MER system mainly includes three modules: emotional feature extractor of each modality, multimodal feature fusion strategy and an emotion classifier [3]. However, current MER techniques still face challenges in meeting the demands of practical applications: (1) Accurate emotion classification heavily relies on good high-dimensional feature representation; (2) While different modalities contribute to the expression of emotions in their own ways, the consistency achieved through the integration of multimodal signals is of paramount importance [4].

As mentioned above, a key challenge in MER is how to let the model learn better emotional representations in each modality. In previous research of speech emotion recognition (SER), spectrograms are the most widely used emotional features and have demonstrated excellent performance [5, 6]. For the textual modality, word2vec [7] is commonly employed to calculate word embeddings. Recently, pre-trained models have become popular over the past few years. Self-supervised models trained on extensive unlabeled data can acquire more emotion representations. HuBERT [8] and WavLM [9] have demonstrated competitive performance in acquiring robust acoustic representations. Pretrained language models (PLMs) such as BERT [10] and RoBERTa [11] can acquire exceptional text representations by leveraging extensive textual data, thereby benefiting a wide range of downstream natural language processing (NLP) tasks. In the field of MER, many studies have also started utilizing these pre-trained models as encoders to extract emotional features from audio and text, achieving excellent classification performance [12, 13].

Fine-tuning specific tasks and datasets on pre-trained models is currently a popular practice for parameter-efficient model adaptation, with approaches like LoRA [14] being widely adopted. While there have been studies on fine-tuning pre-trained models to dynamically adjust the extraction of emotion features in SER domain [15], the research in this area remains largely unexplored in the context of MER. This gap may be attributed to the inherent challenges in training for multimodal classification tasks [16]. Consequently, there is an urgent need for research efforts to address this gap.

Another important problem in MER is how to let the model more effectively utilize and integrate multimodal emotional feature information. Prior studies have investigated diverse fusion techniques for the task of multimodal emotion recognition [17, 18]. Early fusion is a straightforward approach that combines multimodal information streams by integrating time-aligned features at the front-end. Nonetheless, the heterogeneity among modalities can lead to potential semantic confusion in the reconstructed feature space [18]. Conversely, late fusion, which involves employing independent encoders to extract high-level representations of diverse modalities before combining them, has proven to be effective in previous MuSe challenges [19, 20]. Furthermore, deep learning models have emerged as prominent techniques due to their ability on capturing temporal dependencies and enhancing the representation of multimodal data [20].

In addition to focusing on the fusion of feature spaces, the correlation of information between modalities should also be paid attention to. A hierarchical mutual information between input and output maximization framework was used in MER to reduce the loss of task-related information [21], which is regarded as the first attempt to combine mutual information and MER. The combination of mutual information maximization and minimization was applied on emotion recognition in conversations (ERC) to improve the effectiveness and robustness of the fusion of multimodal features [22]. Although some studies have been conducted in this area, they have not delved deep enough and more effective methods are needed to improve it.

In this paper, we propose the method of LoRA-MER us-

† Corresponding author.

[1] The code is available at https://github.com/caiyunrui/LoRA-MER.

10.21437/Interspeech.2024-1793

ing mutual information for multimodal emotion recognition. Different from the traditional extraction of emotional features completely offline, we can use a small amount of additional computing resources to obtain better emotional expression by fine-tuning the pre-trained model as an encoder during training. It is difficult to balance the weights by fine-tuning two pre-training models at the same time, and too many parameters in a multi-modal model can easily lead to training overfitting [16]. Therefore, we only perform LoRA on the feature extraction of the audio modality, which can more directly affect the expression of emotion[13]. Meanwhile, in order to utilize multi-modal features more effectively, we calculate the mutual information between audio and text features. And then we use the method of Mutual Information Neural Estimation (MINE) [23] to calculate the lower band of mutual information and maximize the similarity of feature modeling between different modalities, so that they can predict more accurate and unified emotions. The experimental results on the IEMOCAP dataset demonstrate the effectiveness of our proposed method in extracting and modeling multimodal features, and show improved performance compared to state-of-the-art speaker-independent approaches. The sensitivity of parameters is also fully verified.

2. Methods

2.1. MER Model details

The framework of the proposed model is depicted in Figure 1. Suppose the dataset consists of N speech files, denoted as $D_l = \{(\mu_{a_1}, \mu_{t_1}, e_1), ..., (\mu_{a_N}, \mu_{t_N}, e_N)\}$, where $\mu_{a_i}, \mu_{t_i}, e_i$ represent the audio signal, text and emotion label of the i-th speech. Hence, the problem can be formulated as how to build a recognition model between (μ_a, μ_t) and e using this dataset.

Firstly, we use Automatic Speech Recognition (ASR) technology to recognize audios into texts. Then the input signals μ_a and μ_t from two modalities are encoded through pre-trained models for emotional features. For audio modality, which has greater emotional impact, the pre-trained speech model using LoRA fine-tuning strategy (in Session 2.2) is applied to extract acoustic features $x_a \in \mathbf{R}^{T_a \times D_a}$. And for text modality, the frozen RoBERTa is applied to extract textual features $x_t \in \mathbf{R}^{T_t \times D_t}$. Here we denote the time frame as $T_{\{a,t\}}$ and the dimension of extracted features as $D_{\{a,t\}}$.

Afterwards, x_a and x_t are passed through the Attentive Statistics Pooling [24] layer which leads to a vector. The outputs from the pooling layer are then fed into two pointwise 1D convolutional layers, which is to project the sequences into semantic space with unified dimension:

$$f_{\{a,t\}} = Pool(Conv1D(x_{\{a,t\}})) \in \mathbf{R}^D \tag{1}$$

where we denote the processed features by $f_{\{a,t\}}$, the dimension of processed features as D.

Next is the modality fusion stage. On the one hand, in order to maximize the similarity of emotional features between different modalities, we adopt the method of multimodal mutual information maximization (in Session 2.3). On the other hand, we utilize the Audio-Text Fusion (AT-Fusion) [12] method to summarize the salient emotional information of each modality, which can be mathematically described as:

$$f_{cat} = Concat(f_a, f_t) \in \mathbf{R}^{D \times 2} \tag{2}$$

$$\alpha_{fuse} = softmax(w_f^T tanh(W_f f_{cat})) \in \mathbf{R}^{1 \times 2} \tag{3}$$

$$h = f_{cat} \alpha_{fuse}^T \in \mathbf{R}^{D \times 1} \tag{4}$$

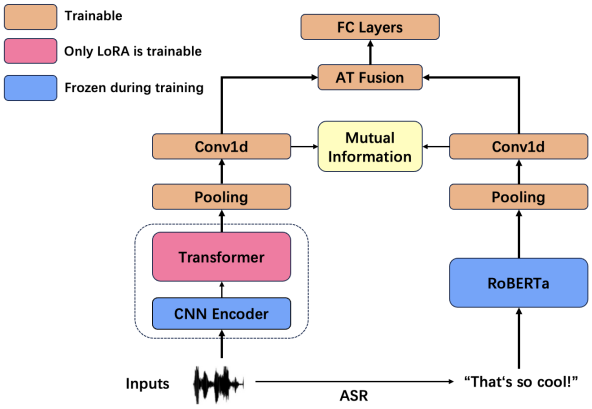

Figure 1: *Overall structure of the proposed model.*

where $w_f \in \mathbf{R}^{D \times 1}$ and $W_f \in \mathbf{R}^{D \times D}$ are trainable parameters, and h is the fusion representation output.

Finally, there is an emotion classifier, which contains several linear layers and a softmax layer, to indicate the expected posterior probabilities $\hat{e}$ for the relevant emotion categories. And the performance is optimized by minimizing the cross entropy loss L_c.

2.2. Low-Rank Adaptation

In this session, we provide details on the Low-Rank Adaptation (LoRA) [14] fine-tuning strategy applied in our model. LoRA is one of the parameter-efficient fine-tuning methods, which is to adapt to new tasks and new datasets by fine-tuning only part of the parameters of the pre-trained model. Parameter updates are approximated by a low-rank decomposition on the bypass of the original pre-trained weighted matrix. Specifically, For a pre-trained weight matrix $W_0 \in \mathbf{R}^{d \times k}$, where d and k are input and output dimensions, the original matrix is replaced by:

$$W_0 + \Delta W = W_0 + BA \tag{5}$$

where $B \in \mathbf{R}^{d \times r}$, $A \in \mathbf{R}^{r \times k}$, and the rank $r \ll min(d, k)$. The parameters in W_0 is frozen during training, while the matrix A and matrix B are trainable. We initialize matrix A with random Gaussian values, while matrix B is initialized with zeros. Assuming that the original input and output of this module are h and x, where $h = W_0 x$, then the forward propagation process is modified as:

$$h = W_0 x + \Delta W x = W_0 x + BAx \tag{6}$$

In our experiments, the pre-trained speech models we need to fine-tune are mainly Transformer-based models such as WavLM and HuBERT. Therefore, we apply the LoRA approach on the W_q and W_v matrices related to the Self-Attention layers of Transformer. As shown in Figure 2, a low-rank decomposition bypass mentioned above is added to W_q and W_v respectively for fine-tuning, while W_q, W_v themselves and other modules of Transformer are frozen.

2.3. Multimodal Mutual Information Maximization

Mutual information can be used to measure the complex correlation between two random variables. Here we propose a method based on mutual information maximization to perform

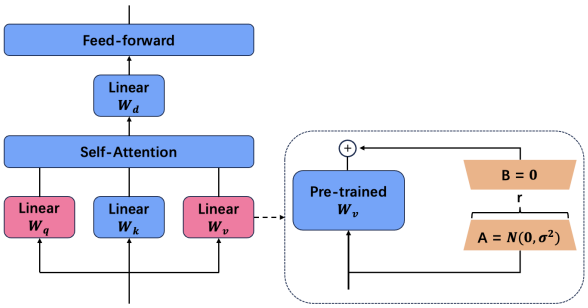

Figure 2: *System architecture of LoRA applied in Transformer.*

information fusion between multi-modalities. The mutual information of two discrete random variables X and Z can be defined as:

$$I(X,Z) = \sum_{z\in Z}\sum_{x\in X} p(x,z)\log\left(\frac{p(x,z)}{p(x)p(z)}\right) \tag{7}$$

where $p(x,z)$ is the joint distribution, $p(x)$ and $p(z)$ are the marginal distributions.

The training goal of the multimodal emotion recognition model is to make the emotional features f_a and f_t as similar as possible to improve the prediction results after feature fusion. Therefore, the mutual information between the emotional features extracted from different modalities should be as large as possible, which is the problem of maximizing mutual information. However, in the process of modal fusion, the input feature data is often high-dimensional and continuous, and we cannot directly calculate mutual information without the accurate distribution $p(x,z)$, $p(x)$ and $p(z)$ in Eq (7).

Therefore, we introduce Mutual Information Neural Estimation (MINE) [23] into multimodal emotion recognition, which is a neural network-based method for estimating the lower bound of mutual information. After having the lower bound, we can increase the mutual information between multimodal emotional features. Specifically, The original expression Eq (7) of mutual information can be seen as the Kullback-Leibler (KL-) divergence between $p(x,z)$ and $p(x)p(z)$:

$$I(X,Z) = D_{KL}\left(p(x,z)||p(x)p(z)\right) \tag{8}$$

According to Donsker-Varadhan representation, We can get a consequence from Eq (8):

$$I(X,Z) \geq \sup_{T\in\mathcal{F}} E_{p(x,z)}[T] - \log(E_{p(x)p(z)}[e^T]) \tag{9}$$

where T is any function that satisfies the condition of integrability, and F can be any class of functions $T : \Omega \to \mathbf{R}$. At this point, we have obtained a lower bound of mutual information between variables X and Z. If a neural network with parameters $\theta \in \Theta$ is used to fit the function T, Eq (9) can be rewritten as:

$$I(X,Z) \geq \sup_{\theta\in\Theta} E_{p(x,z)}[T_\theta] - \log(E_{p(x)p(z)}[e^{T_\theta}]) \tag{10}$$

In the actual training process, we can only sample from each batch of the dataset, and we denote by $\hat{p}^{(n)}_{xz}$, $\hat{p}^{(n)}_{x}$ and $\hat{p}^{(n)}_{z}$ as the empirical distribution of $p(x,z)$, $p(x)$ and $p(z)$ associated to n $i.i.d.$ samples. The maximization of mutual information with n as batch size is calculated as follows:

$$\widehat{I(X,Z)}_n = \sup_{\theta\in\Theta} E_{\hat{p}^{(n)}_{xz}}[T_\theta] - \log(E_{\hat{p}^{(n)}_{x}\hat{p}^{(n)}_{z}}[e^{T_\theta}]) \tag{11}$$

In our experiments, a neural network consisting of simple linear layers and ReLU activation function is employed to iteratively train and fit a lower bound of mutual information between the features from two modalities. And we replace X, Z with the emotional features f_a and f_t, then the loss function for multimodal mutual information maximization is:

$$\mathcal{L}_{mi} = -\widehat{I(f_a, f_t)}_n \tag{12}$$

Finally, the loss function of the entire model can be calculated as:

$$\mathcal{L} = \mathcal{L}_c + \lambda\mathcal{L}_{mi} \tag{13}$$

where λ is the hyperparameter to adjust the importance of $\mathcal{L}_{mi}$.

3. Experiments

3.1. Corpus Description

We evaluated the effectiveness of the proposed method on the Interactive Emotional Dyadic Motion Capture (IEMOCAP) [25] dataset. This dataset is widely utilized for studying English speech emotions in the SER and MER task. It consists of approximately 12 hours of audio data from ten speakers, who express different emotions in specific scenarios. While the original dataset comprises nine distinct emotions, to elicit more genuine emotional responses, it is more meaningful to group certain similar emotions together. Thus, we merged similar emotions to form four categories ($happiness$: 1636, $neutral$: 1084, $anger$: 1103, $sadness$: 1708).

3.2. Experiment Setup

We implement our model using the PyTorch deep learning library in Python. For audio modality, we use the HuBERT-base[2] (94.68M parameters) and WavLM-base[3] (94.70M parameters) model with LoRA fine-tuning strategy as the audio encoder. For text modality, the transcript is segmented into word pieces using a tokenizer and passed through the text encoder RoBERTa-base[4] (125.0M parameters) without fine-tuning. The hidden states from the last layer of the pre-trained model are adopted as final emotion representations in both modalities, and their feature dimensions $D_{\{a,t\}}$ are all 768.

The training process is conducted on a machine equipped with a GeForce RTX 4090, utilizing the Adam optimizer. During training, a learning rate of 0.0003 with a cosine annealing schedule is set. The batch size is set to 16 and the max training epoch is set to 40 with early stopping mechanism. The output channel (D) of Conv1D is 256.

To evaluate the performance of our model, we employ an original train-test-split to do 5-fold cross-validation (CV) and present the average results. Moreover, in order to address the class imbalance present in the IEMOCAP dataset, we utilize both unweighted accuracy (UA) and weighted accuracy (WA) as measures to compare the model's performance in our experiments.

3.3. Experimental Results

3.3.1. Experiment on Multimodal Features

Firstly, we conducted experiments to evaluate the effectiveness of both unimodal and multimodal feature combinations. In the

[2]https://huggingface.co/facebook/hubert-base-ls960
[3]https://huggingface.co/microsoft/wavlm-base-plus
[4]https://huggingface.co/FacebookAI/roberta-base

unimodal experiments, we applied pooling and conv1d layers to the extracted emotion features, followed by classification. For the multimodal experiments, we compared four combinations and set the value of λ to 0.25. The experimental results are shown in Table 1.

Table 1: *The UA and WA on different modality features and combinations*

Modality	Model	UA(%)	WA(%)
Audio	Frozen HuBERT	68.52	66.82
	Frozen WavLM	68.57	67.09
	LoRA HuBERT	71.78	70.80
	LoRA WavLM	72.39	71.61
Text	Frozen RoBERTa	60.94	60.01
Audio+Text	Frozen HuBERT + Frozen RoBERTa	76.46	75.86
	Frozen WavLM + Frozen RoBERTa	77.59	76.58
	LoRA HuBERT + Frozen RoBERTa	79.07	78.21
	LoRA WavLM + Frozen RoBERTa	**80.18**	**79.39**

Based on the results of the unimodal experiments, the classification performance of the model using only audio modality was significantly better than using only text modality. Moreover, the audio modality model was able to significantly improve the learning of emotional representations after applying the LoRA fine-tuning strategy, which reveals the applicability of LoRA in emotion recognition task. Additionally, the use of WavLM yielded slightly better results compared to HuBERT.

In terms of multimodal experiments, the results demonstrate significant improvements in classification performance for both multimodal combinations compared to unimodal approaches. The multimodal models fine-tuned with LoRA strategy also consistently outperform the frozen models in terms of classification performance, while only introducing an additional 74K trainable parameters to the model. This indicates the effectiveness of our proposed LoRA-MER method, which can effectively learn emotional representations from different modalities and combine them with minimal additional computational resources. Additionally, the combination of LoRA WavLM + Frozen RoBERTa outperforms the other models, which achieves the highest UA (80.18%) and WA (79.39%) scores respectively.

3.3.2. Parameter Sensitivity

In order to investigate the stability of MINE and how λ impacts model performance,, we test the LoRA WavLM + Frozen RoBERTa model using values of λ ranging from 0.0 to 2.0.

Table 2: *Comparison among different* λ

λ	UA(%)	WA(%)
2.0	70.95	69.62
1.0	76.41	75.41
0.5	79.23	78.57
0.25	**80.18**	**79.39**
0.1	79.07	78.03
0.0	78.60	76.67

The results shown in Table 2 indicate that both UA and WA are maximized when λ is set to 0.25. When λ is set to 0, which means no mutual information is introduced and only cross-entropy loss is used, the performance is better than using a single modality but significantly lower than when λ is set to 0.25. This can be considered as an ablation study that demonstrates the effectiveness of this module. When λ is set to a large value (1.0 and 2.0), the model's performance actually degrades. This may be because the model overly emphasizes the similarity between modalities and ignores their differences as well as the classification loss itself, leading to suboptimal results. When λ is set to 0.5, 0.25 and 0.1, the classification performance is improved to some extent, indicating the importance of reasonable parameter settings.

3.3.3. Comparison with State-of-the-art Approaches

To evaluate the superiority of our model, we also compared its performance with some state-of-the-art methods that have demonstrated outstanding results on the IEMOCAP dataset. As certain methods incorporate speaker information, which has a significant impact on emotion recognition [26], we categorized some state-of-the-art methods into speaker-dependent and speaker-independent approaches. The results are presented in Table 3.

Table 3: *Comparison with state-of-the-art approaches on the IEMOCAP dataset.*

	Approach	UA(%)	WA(%)
Speaker independent	Atmaja et al. (2019) [27]		75.5
	Li et al. (2019) [28]	79.2	
	Santoso et al. (2022) [29]	76.8	76.6
	Ma et al. (2023) [30]		77.64
	Proposed method	**80.18**	**79.39**
Speaker dependent	Lian et al. (2020) [12]		**82.68**
	Ghosh et al. (2022) [31]		81.2

In comparison to our approach, the approach [30] focuses on a pre-trained model for emotion feature extraction solely on the audio modality, without considering the text modality. The approaches [27, 28] employ traditional methods for feature extraction, using spectrograms and word2vec respectively. The approaches [12, 29] consider fusion strategies but do not account for the correlation between modalities. The results demonstrate that our model outperforms all the speaker-independent models in terms of classification performance. However, the approaches [12, 31], which incorporate speaker information, outperform our proposed model in terms of the WA metric. This observation inspires us to consider incorporating speaker information to further enhance the classification performance. In conclusion, our proposed method of LoRA-MER using mutual information can be proven as an extremely effective approach for multimodal emotion recognition task.

4. Conclusions

In this paper, we propose an approach to effectively extract and fuse multimodal emotional features. Experimental results demonstrate that fine-tuning a pre-trained speech model with LoRA strategy can better represent emotional features, and the mutual information maximization method based on MINE enhances the correlation between multimodal information. Therefore, our proposed model achieves a significant performance improvement in emotion recognition task. In future research, we will explore more pre-training models and fine-tuning strategies. We will also optimize the algorithm for maximizing mutual information to reduce model training time. Additionally, we consider incorporating speaker information into the model to further enhance classification performance.

5. Acknowledgements

This work is supported by National Natural Science Foundation of China (62076144), National Social Science Foundation of China (13&ZD189) and the Major Key Project of PCL (PCL2022D01, PCL2023AS7-1).

6. References

[1] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Communications of the ACM*, vol. 61, no. 5, pp. 90–99, 2018.

[2] C. Chen and P. Zhang, "Integrating cross-modal interactions via latent representation shift for multi-modal humor detection," in *Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge*, 2022, pp. 23–28.

[3] K. Ezzameli and H. Mahersia, "Emotion recognition from unimodal to multimodal analysis: A review," *Information Fusion*, p. 101847, 2023.

[4] M. Hou, Z. Zhang, C. Liu, and G. Lu, "Semantic alignment network for multi-modal emotion recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[5] N. Cummins, S. Amiriparian, G. Hagerer, A. Batliner, S. Steidl, and B. W. Schuller, "An image-based deep spectrum feature representation for the recognition of emotional speech," in *2017 ACM on Multimedia Conference (MM 2017), Mountain View, CA, USA, October 23-27, 2017*. ACM, 2017, pp. 478–484.

[6] P. Tzirakis, J. Zhang, and B. W. Schuller, "End-to-end speech emotion recognition using deep neural networks," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018), Calgary, AB, Canada, April 15-20, 2018*. IEEE, 2018, pp. 5089–5093.

[7] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[8] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing(TASLP 2021)*, vol. 29, pp. 3451–3460, 2021.

[9] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[11] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[12] Z. Lian, J. Tao, B. Liu, J. Huang, Z. Yang, and R. Li, "Context-dependent domain adversarial neural network for multimodal emotion recognition." in *INTERSPEECh 2020*, 2020, pp. 394–398.

[13] H. Wang, Y. Xi, H. Chen, J. Du, Y. Song, Q. Wang, H. Zhou, C. Wang, J. Ma, P. Hu *et al.*, "Hierarchical audio-visual information fusion with multi-label joint decoding for mer 2023," in *Proceedings of the 31st ACM International Conference on Multimedia(MM 2023)*, 2023, pp. 9531–9535.

[14] E. J. Hu, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, "Lora: Low-rank adaptation of large language models," in *International Conference on Learning Representations (ICLR 2021)*, 2021.

[15] T. Feng and S. Narayanan, "Peft-ser: On the use of parameter efficient transfer learning approaches for speech emotion recognition using pre-trained speech models," in *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII 2023)*. IEEE, 2023, pp. 1–8.

[16] W. Wang, D. Tran, and M. Feiszli, "What makes training multi-modal classification networks hard?" in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR 2020)*, 2020, pp. 12 695–12 705.

[17] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information fusion*, vol. 37, pp. 98–125, 2017.

[18] S. Poria, N. Majumder, D. Hazarika, E. Cambria, A. Gelbukh, and A. Hussain, "Multimodal sentiment analysis: Addressing key issues and setting up the baselines," *IEEE Intelligent Systems*, vol. 33, no. 6, pp. 17–25, 2018.

[19] L. Sun, M. Xu, Z. Lian, B. Liu, J. Tao, M. Wang, and Y. Cheng, "Multimodal emotion recognition and sentiment analysis via attention enhanced recurrent model," in *Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge*, 2021, pp. 15–20.

[20] C. Cai, Y. He, L. Sun, Z. Lian, B. Liu, J. Tao, M. Xu, and K. Wang, "Multimodal sentiment analysis based on recurrent neural network and multimodal attention," in *Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge*, 2021, pp. 61–67.

[21] W. Han, H. Chen, and S. Poria, "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)*, 2021, pp. 9180–9192.

[22] J. Zheng, S. Zhang, X. Wang, and Z. Zeng, "Multimodal representations learning based on mutual information maximization and minimization and identity embedding for multimodal sentiment analysis," *arXiv preprint arXiv:2201.03969*, 2022.

[23] M. I. Belghazi, A. Baratin, S. Rajeshwar, S. Ozair, Y. Bengio, A. Courville, and D. Hjelm, "Mutual information neural estimation," in *International conference on machine learning (ICML 2018)*. PMLR, 2018, pp. 531–540.

[24] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive statistics pooling for deep speaker embedding," *arXiv preprint arXiv:1803.10963*, 2018.

[25] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[26] N. Antoniou, A. Katsamanis, T. Giannakopoulos, and S. Narayanan, "Designing and evaluating speech emotion recognition systems: A reality check case study with iemocap," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2023)*. IEEE, 2023, pp. 1–5.

[27] B. T. Atmaja, K. Shirai, and M. Akagi, "Speech emotion recognition using speech feature and word embedding," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC 2019)*, 2019, pp. 519–523.

[28] R. Li, Z. Wu, J. Jia, Y. Bu, S. Zhao, and H. Meng, "Towards discriminative representation learning for speech emotion recognition." in *IJCAI 2019*, 2019, pp. 5060–5066.

[29] J. Santoso, T. Yamada, K. Ishizuka, T. Hashimoto, and S. Makino, "Speech emotion recognition based on self-attention weight correction for acoustic and text features," *IEEE Access*, vol. 10, pp. 115 732–115 743, 2022.

[30] Z. Ma, Z. Zheng, J. Ye, J. Li, Z. Gao, S. Zhang, and X. Chen, "emotion2vec: Self-supervised pre-training for speech emotion representation," *arXiv preprint arXiv:2312.15185*, 2023.

[31] S. Ghosh, U. Tyagi, S. Ramaneswaran, H. Srivastava, and D. Manocha, "Mmer: Multimodal multi-task learning for speech emotion recognition," *arXiv preprint arXiv:2203.16794*, 2022.