

LMCM: Large Pre-trained Model Driven Multimodal Emotion Recognition via Cross-Modal Attention

Yidong Gao *

Department of Computer Science, City University of Hong Kong, Kowloon Tong, Hong Kong, China

* Corresponding Author Email: Yidonggao2-c@my.cityu.edu.hk

Abstract. With the continuous progress of deep learning technologies, emotion recognition has become an essential research topic in fields including HCI, intelligent assistants, and mental health applications. Compared to unimodal methods, emotion recognition across multiple modalities can offer a deeper insight into human affective states. However, existing methods still suffer from insufficient cross-modal feature alignment, limited generalization ability, and underutilization of large pre-trained models. To tackle these problems, this study proposes a new framework termed LMCM. For the visual modality, CLIP is employed as the feature extractor to leverage its strong image-text alignment capability; for the textual modality, the DeBERTa-V3-base model is adopted to obtain high-quality semantic representations. In the fusion stage, a residual cross-attention and a dual-branch parallel cross-attention mechanism are designed to maximize the use of complementary cues between visual and textual modalities. Experiments conducted on the IEMOCAP demonstrate that the method of this study surpasses earlier baselines, with higher Weighted Accuracy (WA) and Weighted F1-score (WF1). This study not only validates the potential of large-scale pre-trained models for multimodal emotion recognition but also provides a reproducible paradigm for future research on cross-modal feature fusion.

Keywords: Emotion Recognition, Multimodal Learning, CLIP, Natural Language Processing, Cross-Attention.

1. Introduction

Driven by the swift advances in AI and deep learning, emotion recognition has become an essential research topic in fields including HCI, intelligent assistants, and mental health services. Human emotional expression is inherently multimodal, involving not only speech and text but also facial expressions and body gestures. Compared to unimodal methods, multimodal emotion recognition can enable a broader grasp of human affective states, thus attracting significant attention from both academia and industry [1].

In recent years, deep learning-driven multimodal emotion recognition has made remarkable progress. CNNs, RNNs, and Transformer models have shown strong performance in visual and textual modalities [2-4]. Moreover, the introduction of cross-modal attention mechanisms has further improved semantic alignment across modalities. However, several challenges remain: (1) Insufficient feature alignment: There exists a semantic gap between modalities, and simple feature concatenation is inadequate for modeling cross-modal dependencies. (2) Limited generalization ability: Existing models often lack robustness when applied to noisy and diverse real-world scenarios. (3) Underutilization of large pre-trained models: Although large-scale pre-trained models such as CLIP [5] and DeBERTa [6] have shown great potential in vision-language representation learning, their applications in emotion recognition remain limited.

To overcome these limitations, this paper investigates combining large pretrained vision-language models and advanced text encoders into multimodal emotion recognition tasks, aiming to achieve more accurate and robust cross-modal alignment and classification. The motivation can be outlined as follows: (1) Leverage the image-text alignment capability of CLIP to enhance visual representation; (2) Utilize the strong NLP capability of DeBERTa to obtain high-quality textual features; (3) Employ a dual-branch parallel cross-attention mechanism to make full use of complementary features between visual and textual modalities, thereby alleviating misalignment issues.

The key contributions of this study are summarized as: (1) Introduce a new framework (LMCM) that unifies CLIP and DeBERTa for multimodal emotion recognition; (2) Introduce residual and dual-branch cross-attention mechanisms to enhance semantic alignment between modalities; (3) Obtain competitive state-of-the-art (SOTA) results on IEMOCAP, outperforming existing multimodal baselines significantly; (4) Provide a reproducible framework that serves as guidance for future studies.

2. Related Work

2.1. Multimodal Emotion Recognition

Recent advances in emotion recognition using multiple modalities have explored diverse fusion strategies across audio, visual, and textual modalities. Fan et al. proposed a trimodal symmetric architecture that integrates facial expressions, speech, and body posture with the E-Branchformer network, achieving superior accuracy and efficiency on CREMA-D compared to baseline models [7]. Li et al. proposed JOYFUL, an approach which jointly optimizes integrating modality fusion with graph contrastive learning, improving dialogue-level emotion recognition [8]. Lee et al. developed HFU-BERT, which unifies heterogeneous features by combining BERT-based textual embeddings with audio-visual modalities through multi-attention mechanisms, achieved top-level results on CMU-MOSEI and IEMOCAP [9]. In the domain of intelligent transportation, Tan et al. designed a speech-text multimodal recognition model using AlexNet and BERT, enhanced by 5G-enabled SAGIN, to improve both recognition accuracy and real-time traffic efficiency [10]. Complementarily, Middya et al. proposed deep learning-based model-level audio-video fusion, reporting accuracies of 99% on SAVEE and 86% on RAVDESS [11]. Finally, Tsai et al. developed MulT, the Multimodal Transformer with cross-modal attention to handle unaligned multimodal sequences, significantly outperforming earlier approaches on CMU-MOSEI and IEMOCAP [12].

2.2. Emotion Recognition on Large Scaled Model

Boitel et al. conducted an empirical study of large language models applied to emotion recognition, showing that fine-tuned GPT-3 model obtained the best results on Italian tweets, while prompting-based approaches also exhibited competitive transferability [13]. Kadiyala et al. extended this line of work to cross-lingual settings, where ensembles of GPT-4, Claude-Opus, and LLaMA-3 significantly outperformed baselines in WASSA 2024 [14]. Lalk et al. applied LLMs to psychotherapy transcripts, demonstrating their ability not only to classify emotions but also to predict symptom severity and therapeutic alliance [15]. Yu et al. proposed Merlin, a multimodal LLM framework with foresight reasoning, integrating trajectory-based modeling to improve video understanding and cross-modal interaction [16]. Li et al. introduced SKIER, which combines symbolic knowledge graphs with dialogue relation modeling, thereby enhancing conversational emotion recognition through structured contextual dependencies [17]. Together, these studies highlight the growing role of LLMs and multimodal extensions in emotion recognition, addressing key challenges in cross-lingual transfer, domain adaptation, and cross-modal alignment [18, 19].

3. Methodology

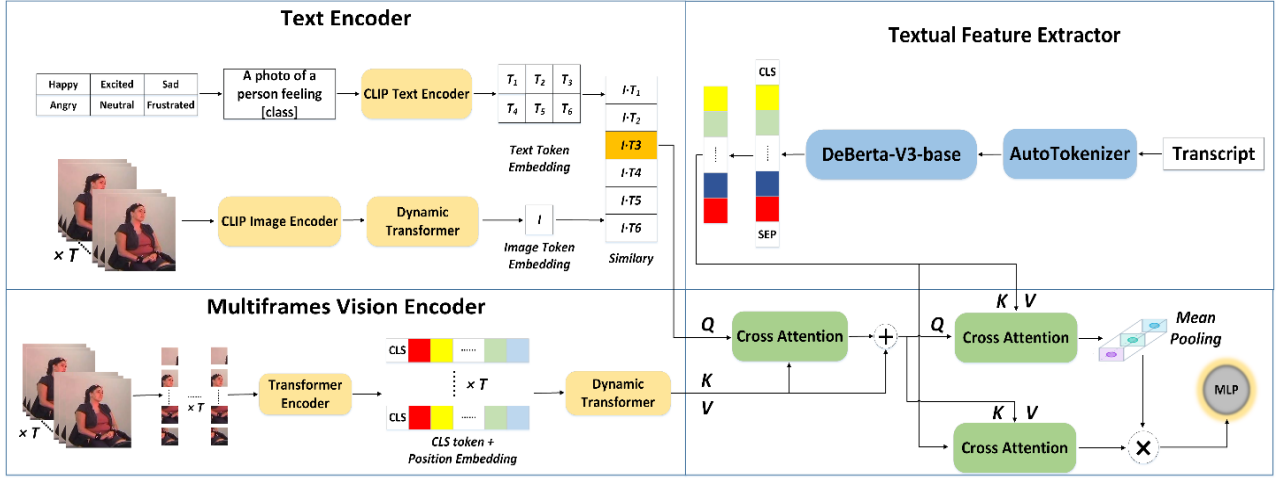


Fig. 1 Architecture of the LMCM. This study uses T frames in the video and the transcripts as input and take CLIP and DeBERTa-V3-base as backbone to extract multimodal feature. Besides, this study utilizes a residual cross-attention and dual-branch parallel cross-attention to fuse features. The text encoder in CLIP is frozen during training.

Figure 1 illustrates the overall structure of LMCM, which is composed of three main parts: (A) video feature extractor, (B) text feature extractor, and (C) fusion-classification block.

3.1. Notations and Task Definition

In conversational emotion recognition, a dataset $D = \{(U_i, Y_i)\}_{i=1}^N$ is provided, where U_i denotes the i -th dialogue consisting of a sequence of continuous utterances $U_i = \{u_1, \dots, u_n\}$, each paired with its video clip and transcript $U_i = \{(v_i, t_i)\}$, and $Y_i \in Y^m$ corresponds to one of k emotion categories $Y = \{y_1, \dots, y_k\}$. The vision and text features are X^v and X^t , respectively. The task is defined as learning a mapping $F(X^v, X^t) \rightarrow Y$. This study utilizes a widely used multimodal conversational datasets, IEMOCAP, to assess the performance of the model proposed in this study (see Section 4.1 for dataset statistics).

3.2. Visual Feature Extractor

CLIP, a large-scale pre-trained vision-language model with 400 million image-text pairs, exhibits excellent capabilities in image feature extraction and image-text alignment. Therefore, this study employs CLIP as the backbone of the video extraction module.

3.2.1. Multiframe Vision Encoder

This study adopts the pre-trained ViT-B/32 variant of CLIP, as the backbone of the visual stream. The input dimensions are denoted as $X^v \in \mathbb{R}^{B \times T \times C \times H \times W}$, which represent batch size, number of independent video frames, the number of channels, the height, and the width of images, respectively. Each video frame image is divided into 32×32 patches. All patches are flattened and fed into a linear layer. After adding positional embedding, the data is passed through transformer encoders to obtain the 512-dimensional embedded vector I_{frame} for each video frame image. After processing T video frames with image encoder of CLIP, $I' \in \mathbb{R}^{B \times T \times 512}$ is obtained. A CLS token and positional embedding are added to I' , which is then fed into a six-layer dynamic Transformer encoder [18] to model temporal dependencies. Finally, a vector $I \in \mathbb{R}^{B \times (T+1) \times 512}$ is obtained as the global visual representation, which can be expressed as:

$$I = DyT(\text{MultiframeEncoder}(X^v)) \quad (1)$$

3.2.2. CLIP Text Encoder

The CLIP text encoder component of the CLIP model can provide powerful semantic information. This study predefines the text prompt "A photo of a person feeling [prompt type]" as the input to text encoder of CLIP, where [prompt type] corresponds to six primary emotion categories: happy, excited, frustrated, sad, and neutral, based on emotion categories in the IEMOCAP. After processing through the text encoder, six 512-dimensional vectors are obtained. Cosine similarity between these six embeddings and the initial embedding of $I(\text{CLS token})$ from Section 3.2.1 are calculated to select the text embedding I_p with the highest similarity. This step can be described as:

$$I_p = \text{TextEncoder}(\text{tokenize}(\text{prompt})) \quad (2)$$

3.3. Textual Feature Extractor

Due to the DeBERTa model demonstrating exceptional performance across numerous natural language processing tasks, this study employs DeBERTa-V3-base model as the feature extractor for the textual modality. First, the raw text transcripts X^t are tokenized using the HuggingFace Transformer AutoTokenizer for DeBERTa-V3-base, which is a SentencePiece-based byte-pair encoding (BPE). The formatted token sequence $[\text{CLS}][\text{tokenized transcript}][\text{SEP}]$ is converted into input IDs, which are then transformed into the final input representation by summing the token, segment, and positional embedding. These input representations are then fed into the 12-layers transformer encoder. Finally, the remaining tokens excluding the $[\text{CLS}]$ and $[\text{SEP}]$ tokens, denoted as $T \in \mathbb{R}^{B \times N \times 768}$, which represent batch size, number of token and output dimensions, Then, it is transformed into $T' \in \mathbb{R}^{B \times N \times 512}$ through a linear layer. This step can be described as:

$$T' = \text{Linear}(\text{DeBERTaEncoder}(X^t)) \quad (3)$$

3.4. Feature Fusion Module

This subsection describes how features from different modalities are fused, specifically through a residual cross-attention and a dual-branch parallel cross-attention. All embeddings are passed through linear layers before cross-attention computation.

3.4.1. Residual Cross-attention

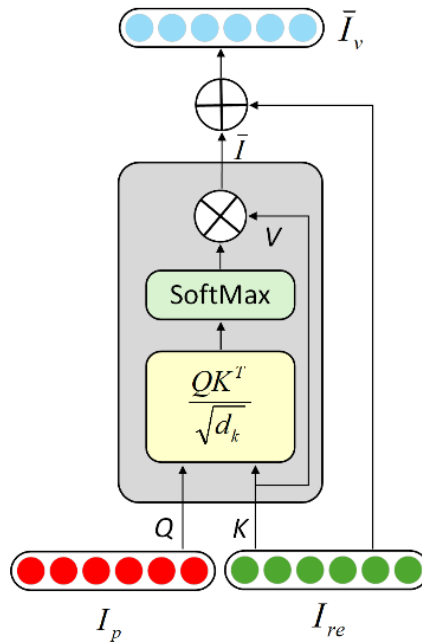


Fig. 2 Residual cross-attention. Embedding I_{re} and I_p are calculated, and a residual connection is applied to this module.

As shown in Figure 2, to enable the semantic information obtained from the CLIP text encoder to provide semantic guidance for the image information obtained from the CLIP image encoder with dynamic Transformer encoder, this study adopts a residual cross-attention. First, the remaining embedding I_{re} excluding CLS token and the embedding I_p are calculated:

$$\bar{I} = \text{Matmul} \left(\text{softmax} \left(\frac{I_p I_{re}^T}{\sqrt{D_k}} \right), I_{re} \right) \quad (4)$$

To prevent semantic information from excessively dominating the image information, a residual connection is applied after the cross-attention:

$$\bar{I}_v = \bar{I} + I_p \quad (5)$$

$\bar{I}_v$ serves as the holistic representation of the image modality.

3.4.2. Dual-Branch Parallel Cross-Attention

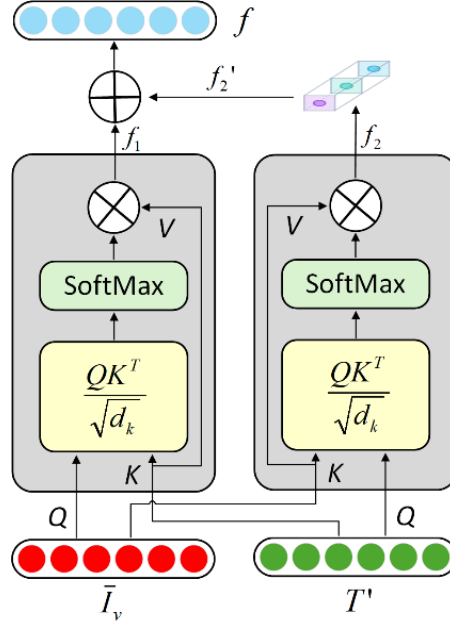


Fig. 3 Dual-branch parallel cross-attention. The embedding $\bar{I}_v$ is used as the query embedding, while the text embeddings T serve as the key and value embedding. Conversely, the roles of the two embeddings are swapped.

As shown in Figure 3, to fully leverage both image and textual information, this study employs a dual-branch parallel cross-attention. First, the embedding $\bar{I}_v$ is used as the query embedding, while the text embeddings T' serve as the key and value embedding, resulting in f_1 :

$$f_1 = \text{softmax} \left(\frac{\bar{I}_v T'^T}{\sqrt{D_k}} \right) \quad (6)$$

Then, the embeddings T' serve as the query embedding, while the embedding $\bar{I}_v$ is applied as the key and value embeddings, yielding f_2 . Global mean pooling is applied to f_2 to obtain f_2' :

$$f_2' = \text{pooliong} \left(\text{softmax} \left(\frac{T' \bar{I}_v^T}{\sqrt{D_k}} \right) \right) \quad (7)$$

Finally, f_1 and f_2' are concatenated to form the embedding f :

$$f = \text{cat}(f_1, f_2') \quad (8)$$

3.4.3. Classification Head

A classification head is implemented, consisting of three linear layers, two GELU activation functions, two dropout layers, and one normalization layer:

$$f_{final} = MLP(f) \quad (9)$$

3.5. Optimization

3.5.1. Loss Function

This study employs a weighted Cross-Entropy loss with label smoothing for training the emotion recognition model. It is expressed mathematically as:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^N w_{y_i} \cdot \left[\lambda \cdot \log(p_{i,y_i}) + (1 - \lambda) \cdot \sum_{c \neq y_i} \log(1 - p_{i,c}) \right] \quad (10)$$

Where w_{y_i} is the class weight for the ground-true label y_i , $p_{i,c}$ represents the predicted probability for class c , and λ is the label smoothing factor. This formulation enhances generalization by preventing overconfidence in the model predictions.

3.5.2. Class Balance

To address the inherent class imbalance in datasets, this study computes class weights inversely proportional to the frequency of each class:

$$weight_i = \frac{1}{(freq_i)^k} \quad (11)$$

Where k is a tunable power parameter (default $k = 1$), and $freq_i$ is the frequency of class i . The weights are normalized to sum to the number of classes to maintain loss stability:

$$weight_i^{norm} = weight_i \times \frac{C}{\sum_{j=1}^C weight_j} \quad (12)$$

Where C is the number of classes. This weighting strategy ensures that minority classes contribute more significantly to the loss, encouraging the model to learn more balanced representations across all emotions.

4. Experiments

4.1. Datasets and Preprocessing

To validate the performance of the model, this study conducted experiments on a well-known emotion recognition dataset, IEMOCAP, which is a 12-hour multimodal dyadic dialogue dataset of aligned audio-video-text utterances annotated with six categories of emotion: happy, angry, sad, neutral, excited and frustrated, for speech emotion recognition and multimodal affective computing research.

This study first discards all emotion categories other than the six primary ones (happy, sad, angry, neutral, excited, and frustrated) as well as any utterances whose emotional label the dataset authors could not determine. For the videos of IEMOCAP, this study extract frames separately for speakers in the left and right, respectively. Each frame is fed into YOLO V8 for person detection; detected individuals are cropped to 224×224 pixels. To suppress frame-to-frame jitter that can yield inconsistent visual inputs, this study applies a jitter-smoothing strategy: if the detected bounding-box center of adjacent frames moves more than 10 pixels horizontally or vertically, the midpoint between the two centers is used as the new center.

Because the videos vary in length, the resulting frame counts differ. To provide a uniform input while balancing model performance and GPU memory, this study fixes the sequence length at 60 frames. Videos with fewer than 60 frames are up-sampled by uniform duplication; those with more are down-sampled by uniform selection.

The accompanying transcripts have already been segmented by the dataset providers and are used without further preprocessing.

4.2. Evaluation Metric

For multi-class emotion recognition, this study adopts WA (Weighted Accuracy) and WF1 (Weighted F1-score) as the primary assessment measures. These metrics consider the uneven distribution of classes in the dataset, providing a more reliable reflection of model performance compared to simple accuracy.

Given a dataset with C classes, let n_i denote the sample count of class c , $\sum_{i=1}^C n_i$ the total number of instances, and Acc_i the classification accuracy for class i . The WA can be formulated as:

$$WA = \frac{1}{N} \sum_{i=1}^C n_i \cdot Acc_i \quad (13)$$

Similarly, let $F1_i$ be the F1-score for class i . The WF1 can be formulated as:

$$WF1 = \frac{1}{N} \sum_{i=1}^C n_i \cdot F1_i \quad (14)$$

A higher WA or WF1 indicates better model performance.

4.3. Implementation Details

The implementation was carried out with PyTorch, running on an NVIDIA RTX PRO 6000 GPU. This study employed 10-fold stratified cross-validation, and results are presented as averages across all folds. Besides, the text encoder of the CLIP in visual feature extractor was frozen in training.

For optimization, this study used AdamW with differential learning rates and weight decays across modules. Specifically, the visual feature extractor and the residual cross-attention module were optimized with a $5e-7$ learning rate and $5e-3$ weight decay. The textual feature extractor and dual-branch parallel cross-attention module were optimized with a $5e-6$ learning rate and $1e-5$ weight decay. The AdamW optimizer used momentum parameters $\beta_1 = 0.9, \beta_2 = 0.95$. A StepLR decay scheme was used for the learning rate with an initial warm-up phase to stabilize convergence. The training objective was the class-weighted cross-entropy loss with label smoothing. For regularization, a dropout rate of 0.4 in the visual feature extractor and 0.3 in the textual feature extractor were applied.

5. Results and Analysis

5.1. Comparison with the Baselines

To assess the performance gap between our proposed model and existing SOTA baselines. This study selected MulT, FE2E [20], DiaRNN, COSMIC [21], Af-CAN [22], AGHMN [23], RGAT [24], COGMEN [25] as SOTA baseline models. The performance was assessed using the IEMOCAP dataset with a 6-ways(F1) evaluation metric, which includes the categories of Happy (Hap), Sad (Sad), Neutral (Neu), Excited (Exc), Angry (Ang), and Frustrated (Fru). The results, as shown in Table 1, demonstrate that the LMCM model reaches SOTA results with significant improvement in both accuracy (Acc) and weighted F1 score (WF1), outperforming all the other baseline methods.

5.2. Ablation Study

To further verify the effectiveness of the framework’s modules. This study conducted three sets of ablation experiments.

Table 1. Comparison with SOTA methods. LMCM achieves SOTA results with significant improvement.

Method	IEMOCAP 6-ways(F1)						Average	
	Hap	Sad	Neu	Ang	Exc	Fru	Acc	WF1
MuT	48.23	76.54	52.38	60.04	54.71	57.51	58.04	58.10
FE2E	44.82	64.98	56.09	62.12	61.02	57.14	58.30	57.69
DiaRNN	32.88	78.08	59.11	63.38	73.66	59.41	63.34	62.85
COSMIC	53.23	78.43	62.08	65.87	69.60	61.39	64.88	65.38
Af-CAN	37.01	72.13	60.72	67.34	66.51	66.13	64.62	63.74
AGHMN	52.10	73.30	58.40	61.91	69.72	62.31	63.58	63.54
RGAT	51.62	77.32	65.42	63.01	67.95	61.23	65.55	65.22
COGMEN	51.91	81.72	68.61	66.02	75.31	58.23	68.26	67.63
LMCM	67.24	75.59	71.66	73.17	75.02	63.85	70.58	70.64

Table 2. Ablation Study for dual-branch parallel Cross-attention

Method	IEMOCAP 6-ways(F1)						Average	
	Hap	Sad	Neu	Ang	Exc	Fru	Acc	WF1
w/o Text-Q Branch	46.06	75.58	69.93	64.37	66.58	64.61	66.26	66.22
w/o Vision-Q Branch	51.91	70.43	62.47	69.93	70.02	62.94	65.39	65.08
LMCM	67.21	75.59	71.66	73.17	75.02	63.85	70.58	70.64

Table 3. Ablation Study for multimodal

Method	IEMOCAP 6-ways(F1)						Average	
	Hap	Sad	Neu	Ang	Exc	Fru	Acc	WF1
w/o Text modality	48.23	71.36	64.80	71.12	68.55	64.63	65.71	65.84
w/o Vision modality	42.86	61.86	53.05	62.26	59.60	55.19	56.34	56.36
LMCM	67.21	75.59	71.66	73.17	75.02	63.85	70.58	70.64

5.2.1. Ablation Study for Cross-attention

This section first analyzes the effectiveness of dual-branch parallel cross-attention. This section retains the backbone of the model. First, this section removes the cross-attention module that uses features extracted by the textual feature extractor as queries, while keeping the attention module that employs features from the visual encoder (i.e., the output of residual cross-attention) as queries. Subsequently, only remove the cross-attention module that adopts features extracted by the visual features extractor as queries.

Table 2 provides the corresponding results, removing either branch leads to a drop in final performance; information from each modality can learn useful complementary cues from the corresponding locations of the features extracted by the other modality.

5.2.2. Ablation Study for Multimodal

This section analysis the necessity of the visual feature extractor and the textual feature extractor. Firstly, this section only preserves visual feature extractor. And then only preserves textual feature extractor. The results present in the table

As shown in the Table 3, any modality is necessary for the model. Textual feature lacks fine-grained visual detail, while images struggle to convey abstract semantics. By allowing each modality to act alternately as Query, Key, and Value for the others, the model can disambiguate, complete, and reinforce features through complementary alignment and cross-checking, thereby enhancing robustness to noise.

Ablation experiments show that removing any modality and any cross-attention causes a significant drop in performance, confirming that multimodal synergy is not redundant but essential for achieving comprehensive, reliable, and human-like perception.

6. Conclusion

This study proposes a network based on the CLIP and DeBERTa to get accurate emotion recognition, named as LMCM. Leveraging powerful semantic–visual alignment of CLIP, this study injects extra linguistic cues to enrich the image representations; this study adopts DeBERTa—proven highly effective on NLP tasks—for robust textual feature extraction; finally, a dual-branch parallel cross-attention mechanism lets visual and textual features fully exploit their complementary information, yielding significant performance gains on the six-class IEMOCAP SOTA baselines.

However, this study has certain limitations: it has not yet been evaluated on more recent, large-scale datasets to fully unleash the potential of the large models, and the audio modality has not been incorporated. These points provide clear directions for future work.

References

- [1] Yuan Y, Li Z, Zhao B. A survey of multimodal learning: Methods, applications, and future. *ACM Computing Surveys*, 2025, 57 (7): Article 167. doi:10.1145/3713070.
- [2] Jaiswal A, Raju A K, Deb S. Facial emotion detection using deep learning. In: 2020 International Conference for Emerging Technology (INCET). 2020: 1-5. doi:10.1109/INCET49848.2020.9154121.
- [3] Majumder N, Poria S, Hazarika D, Mihalcea R, Gelbukh A, Cambria E. DialogueRNN: An attentive RNN for emotion detection in conversations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, 33 (1): 6818-6825.
- [4] Delbrouck J B, Tits N, Brousmiche M, Dupont S. A transformer-based joint-encoding for emotion recognition and sentiment analysis. In: Second Grand-Challenge and Workshop on Multimodal Language (Challenge-HML). Association for Computational Linguistics, Seattle, USA, July 2020: 1-7.
- [5] Radford A, et al. Learning transferable visual models from natural language supervision. In: *International Conference on Machine Learning (ICML)*. 2021.
- [6] He P, Liu X, Gao J, Chen W. DeBERTa: Decoding-enhanced BERT with disentangled attention. *arXiv preprint arXiv:2006.03654*, 2020.
- [7] Fan S, Jing J, Wang C. Audio-visual learning for multimodal emotion recognition. *Symmetry*, 2025, 17 (3): 418.
- [8] Li D, Wang Y, Funakoshi K, Okumura M. Joyful: Joint modality fusion and graph contrastive learning for multimodal emotion recognition. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Singapore, December 2023: 16051-16069.
- [9] Lee S, Han D K, Ko H. Multimodal emotion recognition fusion analysis adapting BERT with heterogeneous feature unification. *IEEE Access*, 2021, 9: 94557-94572.

- [10] Tan L, et al. Speech emotion recognition enhanced traffic efficiency solution for autonomous vehicles in a 5G-enabled space–air–ground integrated intelligent transportation system. *IEEE Transactions on Intelligent Transportation Systems*, 2022, 23 (3): 2830-2842.
- [11] Middy A I, Nag B, Roy S. Deep learning based multimodal emotion recognition using model-level fusion of audio–visual modalities. *Knowledge-Based Systems*, 2022, 244: 108580.
- [12] Tsai Y H H, Bai S, Liang P P, Kolter J Z, Morency L P, Salakhutdinov R. Multimodal transformer for unaligned multimodal language sequences. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics, Florence, Italy, July 2019: 6558-6569.
- [13] Diamantini C, Mircoli A, Potena D, Vagnoni S. An experimental comparison of large language models for emotion recognition in Italian tweets, itaDATA. 2023.
- [14] Kadiyala R M R. Cross-lingual emotion detection through large language models. In: *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis (WASSA)*. Association for Computational Linguistics, Bangkok, Thailand, August 2024: 464-469.
- [15] Lalk C, et al. Employing large language models for emotion detection in psychotherapy transcripts. *Frontiers in Psychiatry*, 2025, 16.
- [16] Yu E, et al. Merlin: Empowering multimodal LLMs with foresight minds. In: *Computer Vision – ECCV 2024*. Springer Nature Switzerland, Cham, 2025: 425-443.
- [17] Li W, Zhu L, Mao R, Cambria E. SKIER: A symbolic knowledge integrated model for conversational emotion recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023, 37 (11): 13121-13129.
- [18] Zhu J, Chen X, He K, LeCun Y, Liu Z. Transformers without normalization. In: *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10-17 June 2025: 14901-14911.
- [19] Busso C, et al. IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 2008, 42 (4): 335-359.
- [20] Dai W, Cahyawijaya S, Liu Z, Fung P. Multimodal end-to-end sparse model for emotion recognition. *arXiv preprint arXiv:2103.09666*, 2021.
- [21] Ghosal D, Majumder N, Gelbukh A, Mihalcea R, Poria S. COSMIC: COMmonSense knowledge for eMotion identification in conversations. In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, November 2020: 2470-2481.
- [22] Wang T, Hou Y, Zhou D, Zhang Q. A contextual attention network for multimodal emotion recognition in conversation. In: *2021 International Joint Conference on Neural Networks (IJCNN)*. 18-22 July 2021: 1-7.
- [23] Jiao W, Lyu M, King I. Real-time emotion recognition via attention gated hierarchical memory network. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, 34 (5): 8002-8009.
- [24] Ishiwatari T, Yasuda Y, Miyazaki T, Goto J. Relation-aware graph attention networks with relational position encodings for emotion recognition in conversations. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2020: 7360-7370.
- [25] Joshi A, Bhat A, Jain A, Singh A V, Modi A. COGMEN: Contextualized GNN based multimodal emotion recognition. *arXiv preprint arXiv:2205.02455*, 2022.