Proceedings of the 5th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/115/2025.18482

Fine-Tuning Methods of Multimodal Pretraining Models for Emotion Recognition

Zekai Lin[1,a,*]

[1]*School of Computer Science and Artificial Intelligence, Zhengzhou University, Zhengzhou City, Henan Province, 450001, China*
a. zekailin.bob@gmail.com
**corresponding author*

Abstract: With the continuous progress of artificial intelligence, single-modal models are increasingly falling short of meeting the requirements of complex tasks. As a result, there has been a burgeoning interest in multimodal pretraining models. Traditional single-modal models, like those in natural language processing that concentrate solely on text or computer vision models that focus only on images, have limitations when dealing with multimodal data. They are unable to fully exploit the complementary and associative information among different modalities. The development of multimodal models has been driven by significant growth in computational power, the availability of large-scale multimodal datasets, and breakthroughs in deep learning technology. These advancements empower models to learn feature representations from diverse modalities and establish cross-modal associations and fusion. This paper centers on the current research scenario of multimodal pretraining models in the field of emotion recognition. It analyzes mainstream model architectures and methodologies, as well as their efficacy in identifying various emotional expressions. Through a comparison of existing fine-tuning techniques, this study explores strategies for attaining optimal performance in multimodal emotion recognition. Additionally, this paper discusses the principal challenges that current multimodal pretraining models face in emotion recognition, such as data scarcity and model interpretability. It also contemplates potential future research directions. This study aims to provide a comprehensive reference for researchers in the field, thereby facilitating further advancements and applications of multimodal emotion recognition.

Keywords: Multimodal Pretraining Models, Emotion Recognition, Fine-tuning, Cross-modal Learning.

1. Introduction

In recent years, deep-learning techniques based on pretraining models have exhibited tremendous potential in multimodal emotion recognition. Through training on large - scale datasets, pretrained models can effectively capture the associative information among various modalities. Moreover, with the application of fine - tuning techniques, these models are able to achieve excellent performance in specific tasks. Typical multimodal pretraining models, such as CLIP and VLMo, have already made significant progress in multiple emotion-recognition tasks. The techniques employed in multimodal pretraining models for emotion recognition have extensive applications in several fields. These

DOI: 10.54254/2755-2721/115/2025.18482

applications emphasize the vast potential of emotion-recognition technology in various aspects of daily life and industry. Nevertheless, there are still numerous challenges in the application of multimodal pretraining models in emotion recognition. Issues such as data scarcity, effective cross-modal feature fusion, and model interpretability remain unresolved. Therefore, this paper intends to review the current research situation of multimodal pretraining models in the field of emotion recognition. It will explore the strengths and weaknesses of mainstream methods and examine the specific applications and future directions of fine-tuning techniques in this area. This review not only provides a reference for researchers in related fields but also illuminates the paths for future innovative research.

2. Fundamental concepts of multimodal pretraining models

2.1. Definition of multimodality

Multimodality involves the combined learning from various data sources, including text, images, and audio, with the aim of enhancing model performance in complex tasks. Each data modality encompasses unique features and structures. For example, images offer spatial information, audio records temporal characteristics, and text imparts linguistic and semantic details. A primary challenge in multimodal learning is to effectively integrating these disparate modalities in a complementary manner. Recent research indicates that multimodal learning enriches model perception and comprehension by combining the complementary information of each modality. This is especially beneficial in tasks such as emotion analysis and object recognition[1]. The objective of multimodal learning is to seize the inter-modal dependencies and shared characteristics. This enables the model to handle not only single-modal data but also a combinations of multiple modalities, thereby achieving a more comprehensive understanding. Furthermore, factorized representation techniques assist in disentangling the shared and specific information among modalities. This facilitates the capture of complex inter-modal relationships and improves the model's performance in multimodal tasks. This approach allows models to integrate multiple modalities more efficiently, strengthening their robustness in scenarios where data is incomplete or certain modalities are absent[2].

2.2. Basic principles of pretraining models

Pretraining models have emerged as a revolutionary technology in recent years within the domains of natural language processing (NLP) and other AI-related fields. The basic principle of pretraining is to conduct training on large-scale datasets for the purpose of learning generalizable knowledge and features, thereby laying a solid foundation for subsequent task-specific learning and refinement. The concept underpinning pretrained models is transfer learning, in which the knowledge acquired during one task can be effectively applied to new tasks, greatly reducing the need for extensive training time and vast datasets.

The pretraining process generally consists of two stages. In the Pretraining Stage, the model undergoes training on a large - scale and unlabeled dataset so as to learn foundational patterns and structures. Common pretraining tasks include Masked Language Modeling (MLM) as utilized in BERT. In this process, certain words are masked for the model to predict, and thus semantic context relationships can be captured[3]. Additionally, another renowned model, GPT, employs autoregressive language modeling to generate text by sequentially predicting the next word[4]. Through these pretraining methods, the models acquire powerful transferable features during this stage, which enables them to perform excellently in downstream tasks. Subsequently, in the Fine - Tuning Stage, after the pretraining process, the model is fine - tuned using a labeled dataset for specific tasks, such as text classification or sentiment analysis. Since the model has already obtained extensive knowledge during pretraining, relatively less labeled data and training time are required for

DOI: 10.54254/2755-2721/115/2025.18482

fine - tuning[5]. For example, GPT - 3, due to its extensive pretraining, can achieve good performance in various tasks even without further fine - tuning. The success of pretrained models is largely due to their deep learning capabilities and efficient use of large datasets, which significantly reduce reliance on labeled data. In multimodal pretraining models, this principle is extended further by integrating multiple modalities, such as images, audio, and text, allowing the model to learn from diverse information sources[6]. This approach has proven highly effective in multimodal tasks, such as emotion recognition and vision-language tasks, and offers vast potential for future applications.

2.3. Characteristics and processing methods of multimodal data

Multimodal data encompasses information from varied sources, including text, images, and audio, each possessing unique features. Text carries semantic information, while images convey visual details. When handling multimodal data, effective feature extraction and fusion techniques are indispensable. Common fusion methods consist of early fusion, which integrates data from different modalities during the feature extraction phase, and late fusion, which combines the results after each modality has been processed independently. Moreover, guaranteeing the alignment of both temporal and spatial signals among modalities is crucial for maintaining data consistency[1].

2.4. Construction of multimodal pretraining models

The construction of multimodal pretraining models aims to integrate data from multiple modalities—such as text, images, and audio—to enhance learning capacity and generalization. Key steps include data collection and preprocessing, feature extraction, modality fusion, model training and fine-tuning, and evaluation. Data collection and preprocessing are foundational, ensuring consistency across multimodal data from various sources through cleaning and standardization. Feature extraction uses specialized networks for each modality—such as Transformers for text, CNNs for images, and spectral analysis for audio—capturing essential features. Modality fusion combines data to enable richer learning; early fusion merges modalities during feature extraction, while late fusion integrates them after independent processing. Following fusion, the model is pretrained on large datasets to learn general features and then fine-tuned on task-specific data to improve performance in applications like emotion recognition. Finally, rigorous evaluation and optimization fine-tune the model for accuracy and robustness, enabling strong performance and adaptability across complex tasks[7].

3. Mainstream pretrained models in multimodal emotion recognition

3.1. Overview of current mainstream multimodal pretrained models

In multimodal emotion recognition, mainstream models combine data from text, audio, and visual modalities using various feature extraction and fusion strategies, boosting emotion analysis accuracy and generalization. The Res-ViT model is a classic architecture, employing RoBERTa for text feature extraction and combining ResNet with Vision Transformer (ViT) to capture local and global image features. On the MVSA-Multiple dataset, it achieved a 71.66% accuracy and a 69.42% F1 score, demonstrating effectiveness in image-text fusion tasks[8]. The multimodal fusion model integrates audio, text, and action modalities, enhancing emotion recognition by combining audio features from deep waveform extrapolation and LSTM, emotion semantics from a Transformer model, and action cues from a three-layer bidirectional LSTM analyzing facial expressions and movements. This approach achieved an 84.5% accuracy and an 82.3% F1 score on the IEMOCAP dataset, making it suitable for complex emotional cues[9]. The multitask model based on perceptual fusion employs BERT, wav2vec 2.0, and OpenFace 2.0 for feature extraction from text, audio, and images,

DOI: 10.54254/2755-2721/115/2025.18482

respectively, and uses cross-modal attention for enhanced contextual fusion. On the CH-SIMS dataset, it improved accuracy by 1.59% and F1 score by 1.67%, underscoring its robustness in emotion recognition[10].

3.2. Performance of models in emotion recognition tasks

The performance of different models in emotion recognition tasks varies significantly, depending on their modality fusion strategies and feature extraction capabilities. The following is a summary of the performance of some mainstream models on commonly used datasets.

Table 1: Analysis of Model Performance

Model Name	Dataset	Modality Fusion	Accuracy	F1 Score
Res-ViT Feature Enhancement Model[8]	MVSA-Multiple	Text, Image	71.66%	69.42%
Audio, Text, and Action Fusion Model[9]	IEMOCAP	Audio, Text, Action	84.5%	82.3%
Perceptual Fusion Multitask Model[10]	CH-SIMS	Audio, Text, Image	78.9%	77.5%

From the comparison presented in the table, it can be observed that different multimodal emotion recognition models exhibit distinct characteristics in emotion analysis tasks. Hence, it is especially well - suited for emotion recognition tasks in multiple contexts. In conclusion, the performance of these three models in different tasks is affected by their modality integration strategies, feature extraction capabilities, and dataset characteristics, with each model having its own unique strengths.

4. Application of model fine-tuning techniques in multimodal emotion recognition

Fine-tuning techniques constitute a core step within multimodal emotion recognition models. By optimizing a pretrained model for a specific task, fine - tuning empowers the model to better adapt to the target data. In the context of multimodal emotion recognition, the model has to handle multiple modalities, including text, audio, and visual information. Fine - tuning intensifies the interaction among these modalities, thereby further enhancing the model's performance in emotion classification tasks.

4.1. Overview of fine-tuning techniques

Fine-tuning is essential for adapting pretrained models to specific tasks, allowing them to leverage prior knowledge while minimizing the need for task-specific data. It typically involves three strategies: full model fine-tuning, partial fine-tuning, and freezing. Full model fine-tuning adjusts all parameters to fit the target task, making it suitable for scenarios with ample labeled data but computationally expensive. Partial fine-tuning selectively adjusts certain layers, reducing computational burden while retaining the pretrained model’s feature learning ability. Freezing, where most parameters remain

DOI: 10.54254/2755-2721/115/2025.18482

unchanged, is ideal for tasks with limited data, such as emotion recognition, to prevent overfitting and reduce costs. The choice of strategy depends on data availability and computational resources, optimizing performance for specific tasks.

4.2. Specific applications of fine-tuning in multimodal models

In multimodal emotion recognition tasks, fine - tuning techniques are extensively utilized to make pretrained models adapt to the specific emotional data characteristics within the task. The following are typical examples of fine - tuning applications for different pretrained models, which illustrate their fine - tuning strategies and the resultant performance enhancements on diverse datasets.

The MMAF (Multimodal Affective Fusion) model excels in multimodal emotion recognition by learning emotional feature representations from text, audio, and visual data during pretraining. Fine-tuning optimizes the interactions between these modalities, enhancing the model's ability to capture subtle emotional cues. Experiments on the CMU-MOSI dataset showed a 6.17% increase in accuracy and a 6.11% improvement in F1 score, demonstrating the effectiveness of fine-tuning in improving emotion recognition performance[10]. The Res-ViT model combines ResNet and Vision Transformer for emotion recognition using image and text features. During fine-tuning, a partial fine-tuning strategy was employed, freezing most ResNet layers while fine-tuning the Transformer part. This approach preserves the image feature extraction capability of ResNet while adjusting the Transformer for emotional task data. On the MVSA-Multiple dataset, the fine-tuned model achieved 71.66% accuracy and a 69.42% F1 score, demonstrating the practicality and effectiveness of selective layer fine-tuning in resource-limited scenarios[8]. The CLIP model is pretrained using cross-modal contrastive learning and fine-tuned on emotion recognition datasets like IEMOCAP to align its visual and textual features. Fine-tuning focuses on the model's last layers to improve the semantic alignment between modalities, enhancing performance in emotion classification tasks. This process highlights the importance of cross-modal semantic alignment, providing insights for future multimodal emotion recognition tasks[4].

5. Challenges and limitations of multimodal emotion recognition

Although multimodal emotion recognition has achieved remarkable progress in recent years, it still encounters numerous challenges and limitations in both practical applications and research. These issues include data scarcity, the complexity of cross-modal feature fusion, and the interpretability of models.

5.1. Data scarcity and annotation challenges

Data scarcity constitutes a primary challenge within multimodal emotion recognition. Collecting and processing large - scale multimodal datasets usually demand vast amounts of time and resources. High - quality annotated datasets, particularly those covering multiple modalities such as text, audio, and images, are even scarcer. For example, commonly used datasets such as CMU-MOSEI and IEMOCAP are widely used in research, yet their limited scale is insufficient to meet the complex requirements of multimodal emotion analysis[2][11]. Annotation challenges also play a critical part. Emotion recognition tasks call for highly precise annotations. Since emotions are subjective, they can present distinctively across different cultures, languages, or contexts. Consequently, annotating emotional categories requires expert knowledge and significant human effort. For instance, in the audio modality, emotion annotations depend on features such as tone and speech rate, while the visual modality involves details like facial expressions and eye movements[9]. These annotation processes are time-consuming and labor-intensive, exacerbating the issue of data scarcity.

DOI: 10.54254/2755-2721/115/2025.18482

5.2. Difficulties in cross-modal feature fusion

The core of multimodal emotion recognition models lies in effectively fusing data from diverse modalities. Text, audio, and visual modalities each possess distinct feature spaces, and combining these features effectively without losing crucial information remains a formidable challenge. For example, emotional information in audio is largely dependent on sound features like pitch and rhythm. In contrast, text - based emotions rely on semantic structure and contextual information. Visual information, like facial expressions, involves both spatial and temporal sequence analysis[8]. Common fusion methods include early fusion (where fusion occurs at the input stage) and late fusion (combining after independent processing of each modality).. However, both methods have their drawbacks. Early fusion may result in information redundancy, while late fusion might cause the loss of modality interdependencies. To better capture modality complementarity, researchers have introduced cross-modal alignment techniques, which align data from different modalities to achieve better fusion. However, cross-modal alignment still faces technical bottlenecks, such as computational complexity and temporal consistency[12].

5.3. Model interpretability issues

As the complexity of multimodal emotion recognition models increases, the issue of model interpretability has become increasingly prominent. The majority of deep-learning models, especially those based on Transformer or neural networks for multimodal tasks, are commonly regarded as “black-box models”. This makes it arduous to elucidate the reasoning processes within the model. In the case of emotion recognition tasks, the emotional classification results output by models often lack transparency, making it hard for users to understand how a particular emotional judgment was made based on the input modalities[13]. This poses substantial hurdles for practical applications, especially in fields like medical diagnosis and mental health, which heavily rely on emotion recognition The interpretability of models is a prerequisite for ensuring their trustworthiness and broad application. Therefore, improving the interpretability of multimodal emotion recognition models, allowing researchers and users to trace the model’s decision-making process, is an important direction for future research[14].

6. Conclusion

This paper reviews key pre-trained models and fine-tuning techniques in multimodal emotion recognition, evaluating their performance and limitations. Models like Res-ViT, MMAF, and CLIP show high accuracy by effectively combining text, image, and audio modalities. Their performance variations are influenced by differences in feature extraction, modality fusion, and cross-modal cooperation. Fine-tuning techniques, including full model, partial, and freezing strategies, help models adapt to specific emotional data, balancing computational resources with generalization ability, thus improving accuracy. Despite progress, challenges such as data scarcity, high annotation costs, and cross-modal feature fusion complexity hinder the widespread adoption of these models. Furthermore, model interpretability and generalization across domains require improvement. Future research could focus on cross-modal learning advancements, incorporating external knowledge, and enhancing robustness through methods like ensemble learning. In conclusion, multimodal emotion recognition has significant potential in AI emotional computing and, with continued advancements, can provide robust support for applications in education, healthcare, and smart homes.

DOI: 10.54254/2755-2721/115/2025.18482

References

[1] *Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., & Ng, A. Y. (2011). "Multimodal Deep Learning." Proceedings of the 28th International Conference on Machine Learning (ICML).https://arxiv.org/abs/1006.1003*

[2] *Tsai, Y. H. H., Liang, P. P., Zadeh, A., Morency, L. P., & Salakhutdinov, R. (2019). "Learning Factorized Multimodal Representations." International Conference on Learning Representations (ICLR). https://arxiv.org/pdf/1806.06176*

[3] *Devlin, J., et al. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805*

[4] *Radford, A., et al. (2018). Improving Language Understanding by Generative Pre-Training. OpenAI blog*

[5] *Brown, T., et al. (2020). Language Models are Few-Shot Learners. arXiv preprint arXiv:2005.14165*

[6] *Li, L. H., et al. (2020). Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. arXiv preprint arXiv:2004.06165*

[7] *Baltrušaitis, T., Ahuja, C., & Morency, L. P. (2019). Multimodal Machine Learning: A Survey and Taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence.*

[8] *Ruyun, Y, JIng, M. (2023) A feature-enhanced multimodal sentiment recognition model incorporating knowledge and Res-ViT[J]. Data Analysis and Knowledge Discovery,*

[9] *Ning, J, Chunjjun, J. (2023) Multimodal emotion recognition by fusing audio, text, and expressive actions[J]. Journal of Applied Science, Vol. 41(1): 55-70.*

[10] *SiSi, W, JIng M. (2023) A multi-task multimodal sentiment analysis model based on perceptual fusion[J]. Data Analysis and Knowledge Discovery, 7(10): 74-84.*

[11] *Busso, C., et al. "IEMOCAP: Interactive Emotional Dyadic Motion Capture Database." Journal of Language Resources and Evaluation.*

[12] *Poria, S., et al. "Emotion Recognition in Conversations Using Semantic-Aware Graph Convolutional Networks." arXiv preprint arXiv:1906.02772.*

[13] *Samek, W., et al. "Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models." ITU Journal: ICT Discoveries.*

[14] *Adadi, A., & Berrada, M. "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)." IEEE Access.*