Multimodal Emotion Classification from Audio-Text Modalities: A Comparative Study of Transformer-Based Fusion Methods

MER-HAN Implementation: https://github.com/melike1818/MultimodalEmotionRecognition

Abstract

Multimodal Emotion Recognition (MER) addresses the critical challenge of integrating heterogeneous data from modalities like audio and text to capture the complexity of human emotions, essential for applications in healthcare, education and customer service. Traditional fusion techniques such as early and late fusion often fail to model nuanced cross-modal interactions, prompting a shift toward model-level fusion leveraging attention mechanisms. This survey explores advanced methods such as the Hybrid Attention Network (MER-HAN), which integrates intra-modal, cross-modal and global inter-modal attention to dynamically align acoustic and textual cues. MER-HAN's architecture processes audio via Bi-LSTM with self-attention and text via BERT, followed by cross-modal alignment and global feature weighting, achieving 73.66% F1-score on the IEMOCAP. Implementation of MER-HAN using IEMOCAP revealed challenges including the initial problem of overfitting and computational complexity. Freezing internal BERT layers mitigated the overfitting issue, yielding a F1-score of 71.08%. The survey also identifies persistent gaps in temporal dynamics modeling, dataset diversity and interpretability, advocating for robust fusion frameworks, multilingual datasets and efficient attention mechanisms. These insights emphasize the importance of balancing performance, computational cost and transparency to advance adaptable and generalizable MER systems.

1. Introduction

Context and Motivation: Speech Emotion Recognition (SER) has become a critical research area due to its potential to enhance conversational agents across various fields, such as healthcare, education, mental health monitoring, customer service and media industries [1].

A key challenge in Multimodal Emotion Recognition (MER) is effectively integrating multimodal data. Traditional methods such as feature-level (early) fusion and decision-level (late) fusion offer straightforward integration but often fall short in capturing intricate modality interactions. Recent advancements have introduced sophisticated fusion techniques, leveraging deep learning and attention mechanisms to overcome these limitations [2].

This survey explores diverse techniques in multimodal emotion classification from audio-text modalities. As a practical implementation, the Hybrid Attention Network (MER-HAN) method described in the paper "Multimodal emotion recognition based on audio and text by using hybrid attention networks" was selected with IEMOCAP dataset [3][4].

2. Background and Problem Definition

2.1 Concepts and Theories: Human emotions are inherently multimodal, expressed simultaneously through speech, text, facial expressions and physiological signals. While single-modal Speech Emotion Recognition (SER) has advanced significantly, it often fails to capture the full complexity of emotional states due to the complementary nature of cues across modalities. Multimodal fusion strategies can be broadly categorized into three as follows:

1. **Early Fusion (Feature-level):** Modality-specific features such as audio spectrograms, text embeddings are concatenated or summed at the input stage, creating a unified representation for classification. For example, Dehghani Tafti and BabaAli [2] concatenate audio and text embeddings early in their pipeline, while Avro et al. [7] fuse low-level acoustic and textual features via concatenation.
2. **Late Fusion (Decision-level):** Modalities are processed independently and their outputs are combined post-classification. Faria et al. [5] employ a Dynamic Bayesian Mixture Model to merge speech emotion and text sentiment predictions.
3. **Model-level Fusion:** Fusion occurs within the model architecture, leveraging shared layers or attention mechanisms to capture inter-modal dynamics. For instance:
    - Co-attention: Dutta and Ganapathy [6] use hierarchical cross-attention to align audio and text embeddings.
    - Hybrid Attention: Zhang et al. [3] propose a Hybrid Attention Network (MER-HAN) that integrates intra-modal (self-attention) and cross-modal attention, enabling joint learning of acoustic and linguistic cues.
    - Common Space Projection: Cheng et al. [1] map audio, text and visual features into a shared space fused via dot-product attention.

2.2 Problem Statement: The central challenge in MER remains effectively fusing heterogeneous multimodal data. Traditional fusion techniques like early and late fusion can miss nuanced interactions between modalities. Advanced methods must address two critical issues:

- **Temporal Context Modeling:** Acoustic and textual features evolve dynamically over time, requiring architectures such as transformers, recurrent networks that preserve sequential dependencies.
- **Cross-Modal Complementarity:** Effective fusion must balance modality-specific strengths while suppressing redundant or noisy signals.

2.3 Survey Overview: This survey provides a comprehensive review of multimodal fusion techniques. The discussed fusion strategies, particularly model-level attention mechanisms, address the above challenges by enabling adaptive interaction learning. For example, MER-HAN's hybrid attention [3] exemplifies how intra-modal and inter-modal attention can jointly refine emotion representations, achieving state-of-the-art performance on IEMOCAP. By systematically analyzing these approaches, this survey bridges theoretical frameworks with practical implementations.

3. Literature Review

3.1 Overview of Methods: The literature on multimodal emotion recognition (MER) showcases a progression from simpler fusion strategies to more complex architectures. These can be broadly categorized as follows:

- **Early Fusion (Feature-Level):**
  - **Avro et al. [7]** exemplify this with their EmoTech model, which fuses low-level acoustic features (MFCCs processed by BiLSTM and 2D CNN) and textual features before classification. They achieved 84% accuracy on IEMOCAP, demonstrating the viability of early fusion for integrating complementary modalities, especially when coupled with hybrid recurrent-convolutional architectures for feature extraction. However, early fusion can sometimes struggle with the heterogeneity of feature spaces and the optimal point of fusion.
  - **Dehghani Tafti and BabaAli [2]** also explored early fusion using pre-trained Wav2vec 2.0 for audio and BERT for text on IEMOCAP. They found that their early fusion strategy with layer-wise feature aggregation outperformed late fusion and cross-modal attention, achieving state-of-the-art results (UAR: 78.42%, WAR: 77.75%) for their setup. This highlights the potential of powerful pre-trained embeddings in an early fusion context.
- **Late Fusion (Decision-Level):** Faria et al. [5] utilized a two-layered Dynamic Bayesian Mixture Model (2L-DBMM) for late fusion of speech emotion (arousal) and text sentiment (valence) predictions. Their ensemble classifiers (1D-CNN, MLP, SVM) processed handcrafted features (MFCCs, Chroma, TF-IDF, BERT, GPT-2) and the DBMM dynamically adjusted weights, achieving high accuracy (96-98%) on EmoUERJ and ESD datasets. This approach offers robustness and modularity but may miss out on nuanced inter-modal interactions that occur at lower feature levels. Dehghani Tafti and BabaAli [2] also evaluated late fusion, finding it less effective than their early fusion approach.
- **Model-Level (Hybrid/Intermediate) Fusion:**
  - **Dutta and Ganapathy [6]** proposed the Hierarchical Cross Attention Model (HCAM), a three-stage framework. It processes audio (fine-tuned wav2vec 2.0) and text (RoBERTa, Bi-GRU) for utterance embeddings, models inter-utterance context (Bi-GRU with self-attention) and finally fuses modalities using cross-attention and self-attention. Achieving 85.9% weighted F1 on IEMOCAP, HCAM demonstrates the power of co-attention for dynamic cross-modal interaction and temporal context modeling.

  - **Cheng et al. [1]** in their MER-MCE framework, also employed model-level fusion. For MER, they used modality-specific encoders (InstructERC, HuBERT, expMAE) and attention-based fusion to map features into a shared space, weighting cross-modal interactions via dot-product attention (0.6807 weighted F1 on IEMOCAP). This aligns with the trend of leveraging attention for dynamic interaction.
  - The **Hybrid Attention Network (MER-HAN) by Zhang et al. [3]**, which is the method implemented, employed model-level fusion. MER-HAN specifically addresses multimodal fusion challenges by integrating local intra-modal attention (Bi-LSTM with multi-head self-attention for MFCCs and BERT with MHSA for text), cross-modal attention (CMA) to dynamically align audio-text embeddings and global inter-modal attention to weight complementary features. Evaluated on IEMOCAP and MELD, it achieved a 73.66% F1 score on IEMOCAP, underscoring the efficacy of its comprehensive hybrid attention approach.

3.2 Critical Review: The reviewed literature highlights a clear shift toward model-level fusion, particularly leveraging attention mechanisms, as these methods excel at capturing complex inter-modal dependencies often overlooked by early or late fusion strategies. Notably, recent approaches such as HCAM [6] and MER-HAN [3] demonstrate how attention mechanisms effectively model both intra-modal context and cross-modal interactions, while pre-trained encoders like BERT and Wav2vec 2.0 provide robust foundational features, streamlining downstream fusion tasks. Despite these advances, the previously mentioned challenges persist.

The MER-HAN method [3] was selected for implementation due to its explicit and comprehensive approach to attention-based fusion.

- **Strengths of MER-HAN:**
  - **Comprehensive Attention:** Its three-tiered attention (intra-modal, cross-modal, global inter-modal) is designed to capture a wide range of dependencies. Intra-modal attention refines modality-specific representations, cross-modal attention aligns them and global inter-modal attention selectively focuses on the most salient fused features for classification. This is a significant advantage over simpler fusion techniques.
  - **Addresses Key Challenges:** By design, MER-HAN directly tackles the challenge of effectively fusing heterogeneous multimodal data and capturing nuanced interactions, which is a limitation of traditional early/late fusion.

    - **Good Performance:** It achieved competitive results (73.66% F1 on IEMOCAP), demonstrating the practical effectiveness of its hybrid attention strategy for audio-text emotion recognition.
- **Areas for Further Investigation for MER-HAN:**
    - **Feature Extractor Dependency:** MER-HAN utilizes MFCCs for audio processed by a Bi-LSTM and BERT for text. While effective, the performance might be further improved or altered by using more advanced audio encoders like Wav2vec 2.0 directly as input to the fusion network (as seen in [2, 6]) rather than relying on handcrafted features like MFCCs as an intermediate step.
    - **Computational Cost:** The multiple layers of attention (self-attention within modalities, cross-attention between and another global attention layer) can lead to higher computational complexity compared to simpler fusion methods or even some other attention-based models.
    - **Comparison with SOTA:** While MER-HAN reported strong results, more recent models like HCAM [6] have reported even higher F1-scores (85.9% on IEMOCAP) using different architectural choices for cross-attention and context modeling. This suggests that while MER-HAN's hybrid approach is sound, there is still room for advancement in optimizing attention mechanisms and sequence modeling components.
    - **Interpretability:** While attention weights can offer some insight, fully understanding which features and interactions drive the emotion predictions in such a multi-layered attention system can still be challenging.

4. Methodology and Implementation

4.1 Overview of Hybrid Attention Network (MER-HAN):

The Hybrid Attention Network (MER-HAN) [3] was selected for implementation due to its comprehensive approach to fusing audio and text modalities for emotion recognition. The key components and workflow of MER-HAN are as follows:

1. **Audio and Text Encoder (ATE) Block:**
    - **Audio Encoder:** Raw audio signals are first processed to extract Mel-Frequency Cepstral Coefficients (MFCCs). These MFCCs are then fed into a two-layer Bidirectional Long Short-Term Memory (Bi-LSTM) network to capture temporal dependencies. A local intra-modal attention mechanism (Multi-Head Self-Attention, MHSA) is applied to the Bi-LSTM outputs to refine audio feature representations by focusing on salient parts of the audio sequence.
    - **Text Encoder:** For the text modality, a pre-trained Bidirectional Encoder Representations from Transformers (BERT) model is used to generate contextual

word embeddings. Similar to the audio branch, a local intra-modal attention mechanism (MHSA) is applied to these embeddings to derive refined textual feature representations.

2. **Cross-Modal Attention (CMA) Block:**
    - The refined audio and text features from the ATE block are first projected into a shared embedding space using 1D Convolutional Neural Networks (1D-CNNs).
    - A cross-modal attention mechanism is then applied to learn the interactions between the two modalities. This involves computing attention from audio to text (A→T) and from text to audio (T→A), allowing the model to dynamically align and weigh information from one modality based on the context of the other. This helps in capturing complementary information.
3. **Multimodal Emotion Classification (MEC) Block:**
    - The outputs from the cross-modal attention (representing audio-informed text features and text-informed audio features) are concatenated.
    - A global inter-modal attention mechanism is applied to this concatenated representation to selectively focus on the most salient fused features for the final emotion classification task.
    - Finally, the attended multimodal features are passed through a Fully Connected (FC) layer followed by a Softmax activation function to predict the emotion class.

4.2 Implementation Details

4.2.1 Dataset & Preprocessing

The **IEMOCAP (Interactive Emotional Dyadic Motion Capture) dataset** [4] was selected for this implementation. IEMOCAP provides rich, dyadic interactions with audio recordings and corresponding manually transcribed textual utterances, annotated with categorical emotions. For this implementation, four emotion categories: **angry, happy, sad and neutral** were used. Following practice in MER-HAN, a speaker-independent split was adopted, where data from four sessions were used for training and the fifth session for testing/validation [3]. For the preprocessing following applied for each different modalities:

- **Audio Modality:**
    - Audio files were processed using torchaudio.
    - 40-dimensional MFCCs were extracted from the raw audio waveforms. Utterances were divided into frames of 25ms width with a Hamming window and a frame stride of 10ms.
    - Sequences were padded or truncated to a fixed maximum length to enable batch processing.
- **Text Modality:**
    - Textual transcripts were tokenized using the pre-trained BERT tokenizer from the HuggingFace Transformers library [8].
    - Input sequences were padded or truncated to a fixed maximum length.
- **Labels:** Emotion labels were converted into one-hot encoded vectors for training with categorical cross-entropy loss.

4.2.2. Programming Language & Framework: Python with PyTorch as the deep learning framework. torchaudio was used for audio processing and HuggingFace Transformers provided the pre-trained BERT model and tokenizer [8][9].

4.2.3 Model Architecture:

1. **Audio Encoder:** A Bi-LSTM (2 layers, hidden size 256) followed by a Multi-Head Self-Attention (MHSA) layer.
2. **Text Encoder:** bert-base-uncased from HuggingFace, followed by an MHSA layer.
3. **CMA Block:** 1D-CNNs for projection, followed by scaled dot-product attention for cross-modal interactions.
4. **MEC Block:** Concatenation, global inter-modal attention and a final FC layer for classification into the 4 emotion classes.

4.2.4 Training & Evaluation Metrics:

- **Optimizer:** Adam optimizer, with an initial learning rate 5e-5
- **Loss Function:** Categorical Cross-Entropy Loss.
- **Experiment Tracking:** Weights and Biases was used to log metrics, model parameters and visualize training progress [10].
- **Evaluation Metrics:** Weighted Average Recall (WAR), Unweighted Average Recall (UAR), F1-score.

4.2.5 Challenges Faced:

- **Overfitting:** A significant challenge encountered during initial training attempts was overfitting. Training loss decreased steadily, but validation loss started to increase after a certain number of epochs. This indicated that the model was learning the training data too well, including its noise and was not generalizing effectively to unseen validation data.
- **Computational Resources:** Training deep learning models with attention mechanisms is computationally intensive and requires significant GPU memory and time. To fasten the initial training attempts GPU server with NVIDIA RTX 4090 24 GB was rented.
- **Hyperparameter Sensitivity:** The performance of such complex models can be sensitive to hyperparameter choices (learning rate, dropout rates, attention heads, etc.), making the tuning process iterative.

5. Results and Discussion

MER-HAN was implemented using the IEMOCAP dataset for a four-class emotion classification task, including angry, happy, sad and neutral categories. During training, an early stopping mechanism was employed to limit unnecessary computation and avoid overfitting. Additionally, the internal layers of the BERT model were frozen, which significantly reduced overfitting observed in initial training attempts. Following these modifications, the final evaluation metrics achieved shown in Table 1.

Metric	**Score**
Weighted Accuracy Rate (WAR)	0.7099
Unweighted Accuracy Rate (UAR)	0.7160
Macro F1 Score	0.7172
Weighted F1 Score	0.7108

Table 1: Evaluation Results

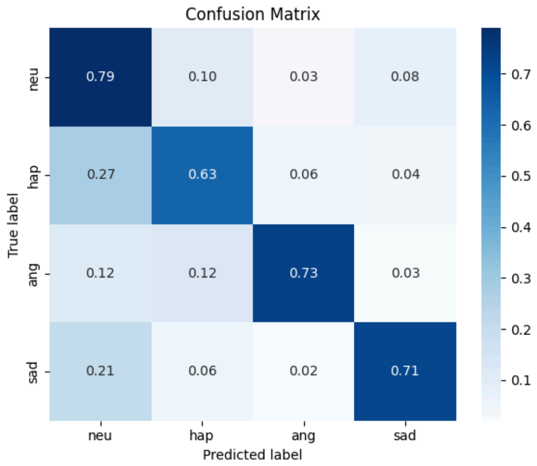

Figure 1: Confusion Matrix

Compared to the original results reported by the MER-HAN (73.66% F1 Score), the implemented model performs slightly below (71.08% F1 Score). However, this gap can be attributed to the use of early stopping. Early stopping has been implemented due to increasing validation loss during the later epochs. As illustrated in Figure 2 and 3, validation loss exhibited fluctuations during early training phases, while the training loss decreased steadily across all trials. Early stopping is beneficial for reducing computational cost and overfitting but may have prevented the model from reaching its full learning potential. Despite the limitations, the results suggest that the MER-HAN architecture is effective at fusing multimodal features. The performance levels achieved on the test set confirm the model's capability to capture complex inter-modal dynamics between audio and text. Overall, the experiment validates the architectural strengths of hybrid attention mechanisms in emotion recognition tasks.

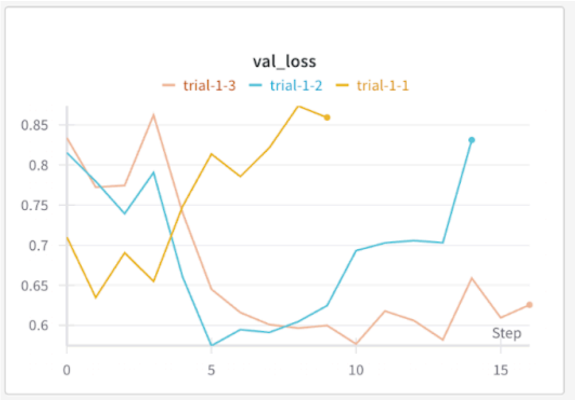

Figure 2

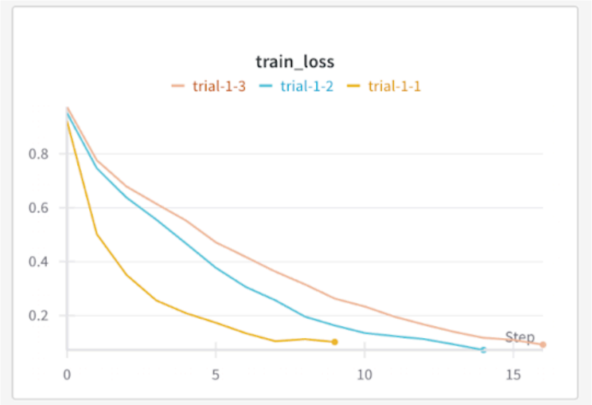

Figure 3

6. Challenges and Open Problems

6.1 Challenges in the Literature

The literature on Multimodal Emotion Recognition (MER) highlights several persistent challenges:

1. **Temporal Dynamics:** Accurately modeling long-range temporal dependencies in acoustic and textual sequences remains difficult.
2. **Cross-Modal Complementarity vs. Redundancy:** Balancing modality-specific strengths while suppressing redundant or noisy signals is a critical yet unresolved issue.
3. **Complexity and Interpretability:** Advanced attention mechanisms improve performance but increase computational costs and reduce model transparency.
4. **Dataset Limitations:** Models trained on standard datasets like IEMOCAP struggle with generalizability across languages and emotional expressions.

6.2 Challenges in Implementation of MER-HAN

During the implementation of MER-HAN, the following practical challenges arose:

1. **Overfitting:** Initially, the model exhibited significant overfitting, with training loss decreasing while validation loss increased after a few epochs. This was originally attributed to the model's complexity, limited dataset diversity or suboptimal hyperparameters. However, further investigation revealed that the primary cause was that the BERT layers were unintentionally left unfrozen during training. Freezing the BERT layers stabilized training and mitigated the overfitting problem.
2. **Computational Resources:** Training MER-HAN required substantial GPU memory and time, necessitating the use of high-end hardware. NVIDIA RTX 4090 is used as it is the most cost-effective option.

6.3 Open Problems and Future Directions

Based on the survey and implementation experience, future research should prioritize the development of **robust fusion mechanisms** that balance computational efficiency with performance. Expanding the availability of **generalizable datasets** with multilingual and diverse emotional expressions is critical to enhance cross-domain robustness and address biases in existing benchmarks like IEMOCAP.

Improving **temporal and contextual modeling** through hierarchical transformers or memory-augmented networks could better capture long-range dependencies in multimodal sequences. Additionally, **hyperparameter optimization** remains an open challenge, necessitating systematic studies to identify optimal configurations such as batch sizes, learning rates for multimodal architectures. Finally, designing **interpretability tools** to visualize attention weights and decision pathways would enhance model transparency and trustworthiness. Addressing these gaps would advance the field toward more adaptable, interpretable and generalizable multimodal emotion recognition systems.

7. Conclusion

Traditional multimodal fusion struggles with cross-modal dynamics, favoring model-level attention mechanisms. MER-HAN, combining intra-modal & cross-modal & global attention, achieved 73.66% F1-score on IEMOCAP. During implementation some challenges were encountered including computational demands and overfitting. An initial overfitting issue was traced back to unintentionally unfrozen BERT layers rather than model complexity. Freezing these layers significantly improved generalization by achieving 71.08% F1 Score. Future steps include regularization and multilingual datasets to enhance generalizability. Addressing temporal modeling and interpretability through attention visualization tools remains vital. These efforts aim to balance performance, efficiency and transparency in MER systems, bridging theoretical advancements with real-world applicability.

References

[1] Z. Cheng et al., "MIPS at SemEval-2024 Task 3: Multimodal Emotion-Cause Pair Extraction in Conversations with Multimodal Language Models," SemEval, 2024.

[2] Z. Dehghani Tafti and B. BabaAli, "Audio-Textual Emotion Recognition using Pre-trained Models: Investigating Various Representations and Fusion Techniques," Univ. Tehran, 2024.

[3] S. Zhang et al., "Multimodal emotion recognition based on audio and text by using hybrid attention networks," Biomed. Signal Process. Control, vol. 85, 2023.

[4] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower Provost, S. Kim, J. Chang, S. Lee, and S. Narayanan, “IEMOCAP: Interactive emotional dyadic motion capture database,” *Language Resources and Evaluation*, vol. 42, pp. 335–359, Dec. 2008, doi: 10.1007/s10579-008-9076-6.

[5] D. R. Faria, A. I. Weinberg, and P. P. Ayrosa, "Multimodal Affective Communication Analysis: Fusing Speech Emotion and Text Sentiment Using Machine Learning," Appl. Sci., vol. 14, 2024.

[6] Dutta and S. Ganapathy, "Hierarchical Cross Attention Model for Multi-modal Emotion Recognition," IEEE, 2024.

[7] S. B. H. Avro et al., "EmoTech: A Multi-modal Speech Emotion Recognition Using Multi-source Low-level Information with Hybrid Recurrent Network," IEEE, 2024.

[8] Hugging Face, “Transformers,” (Version 4.51.3) [Software]. [Online]. Available: https://huggingface.co/docs/transformers. (Accessed: Apr. 22, 2025).

[9] PyTorch Core Team, “PyTorch,” (Version 2.7.0) [Software]. [Online]. Available: https://pytorch.org/. (Accessed: Apr. 2, 2025).

[10] Weights & Biases, “Weights & Biases,” [Software]. San Francisco, CA, USA. [Online]. Available: https://wandb.ai. (Accessed: May. 1, 2025).