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

CENG 562 - Machine Learning Term Project
Melike Demirci
274420

https://github.com/melike1818/MultimodalEmotionRecognition

# Outline

1. Introduction & Motivation
2. Survey Overview
3. Implemented Method: MER-HAN
4. Implementation Details & Dataset
5. Experimental Results & Analysis
6. Challenges & Open Problems
7. Conclusion & Future Directions

# Context & Motivation

- What is Multimodal Emotion Recognition (MER)
- Why is MER Important?
  - Theoretically: Human emotions are **multimodal**.
  - Practically: Healthcare, education, customer service, conversational AI, media industries.
- Core Challenge:
  Effectively fusing diverse data streams to capture nuanced inter-modal interactions.

# Survey Overview

- Survey Scope
  - Focused on **audio** & **text** modalities for MER.
  - Comparative study of fusion methods.
  - **Early**, **late** and **model-level** fusion strategies.

- MER-HAN: Multimodal emotion recognition based on audio and text by using hybrid attention networks (Zhang et al., 2023) [3]
  - Hybrid attention network
  - 73.66% F1-score on IEMOCAP
  - Advanced attention-based fusion.

# Hybrid Attention

Combines multiple types of attention(self-attention, cross-attention etc.) to capture richer and more diverse contextual information. Enhances a model's ability to focus on relevant features from various perspectives.

# Survey Results: Fusion Approaches

| Fusion Strategy | Description | Pros | Cons |
|---|---|---|---|
| Early Fusion (Feature-level) | Concatenate or sum features from different modalities before feeding into a model. | Simple to implement | Struggles with modality heterogeneity; may miss fine-grained interactions |
| Late Fusion (Decision-level) | Combine outputs of separate unimodal models | Modular, easy to plug-and-play | Ignores deep inter-modal relationships |
| Model-Level Fusion (Hybrid/Intermediate) | Integrate modalities within model architecture | Captures complex cross-modal dependencies | Computationally intensive; requires careful design |

# Shift towards model-level fusion with attention mechanisms and pre-trained encoders (BERT, Wav2Vec).

# Overview of MER-HAN

**Audio and Text Encoder (ATE) Block:**

Learn refined unimodal representations.

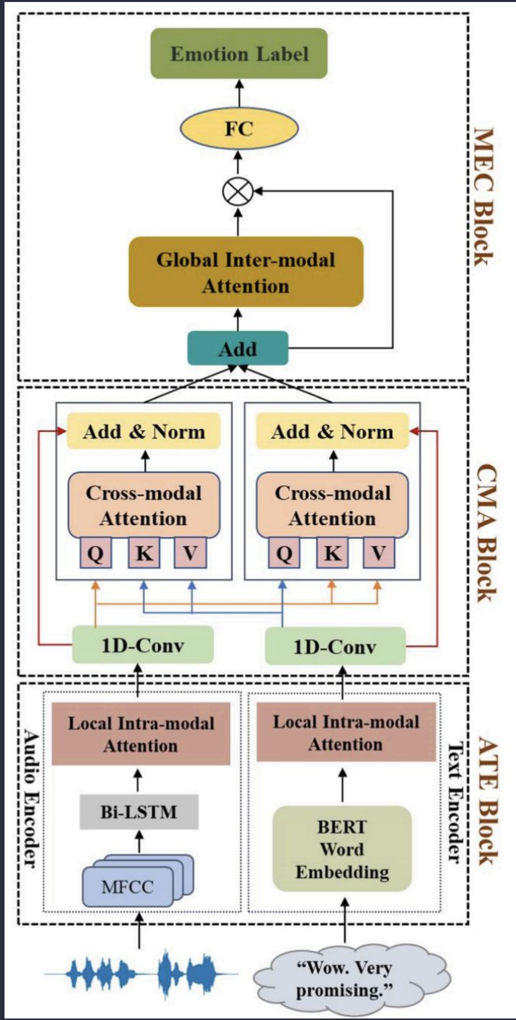
**Cross-Modal Attention (CMA) Block:**

Dynamically align and weigh information across modalities.

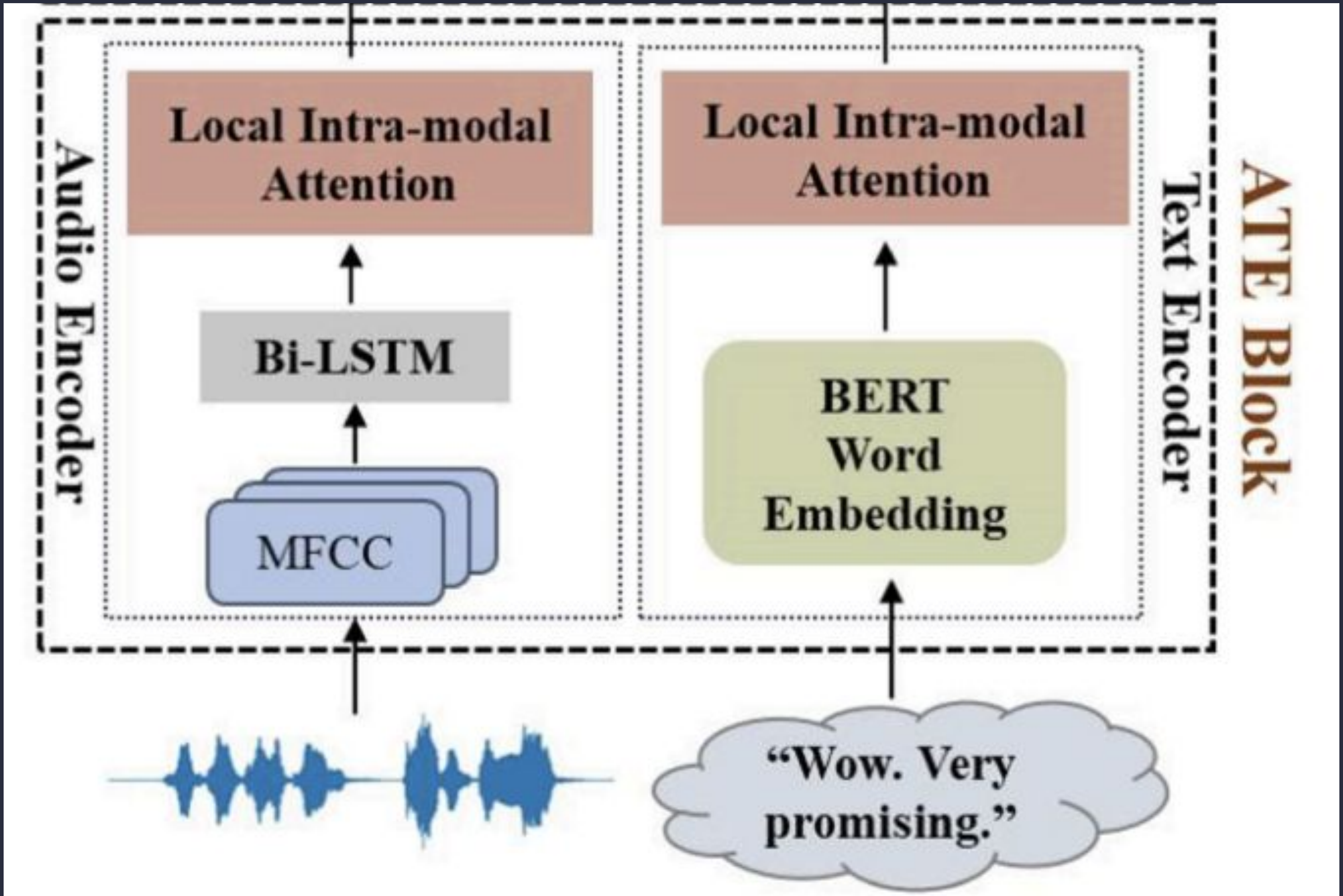
**Multimodal Emotion Classification (MEC) Block:**

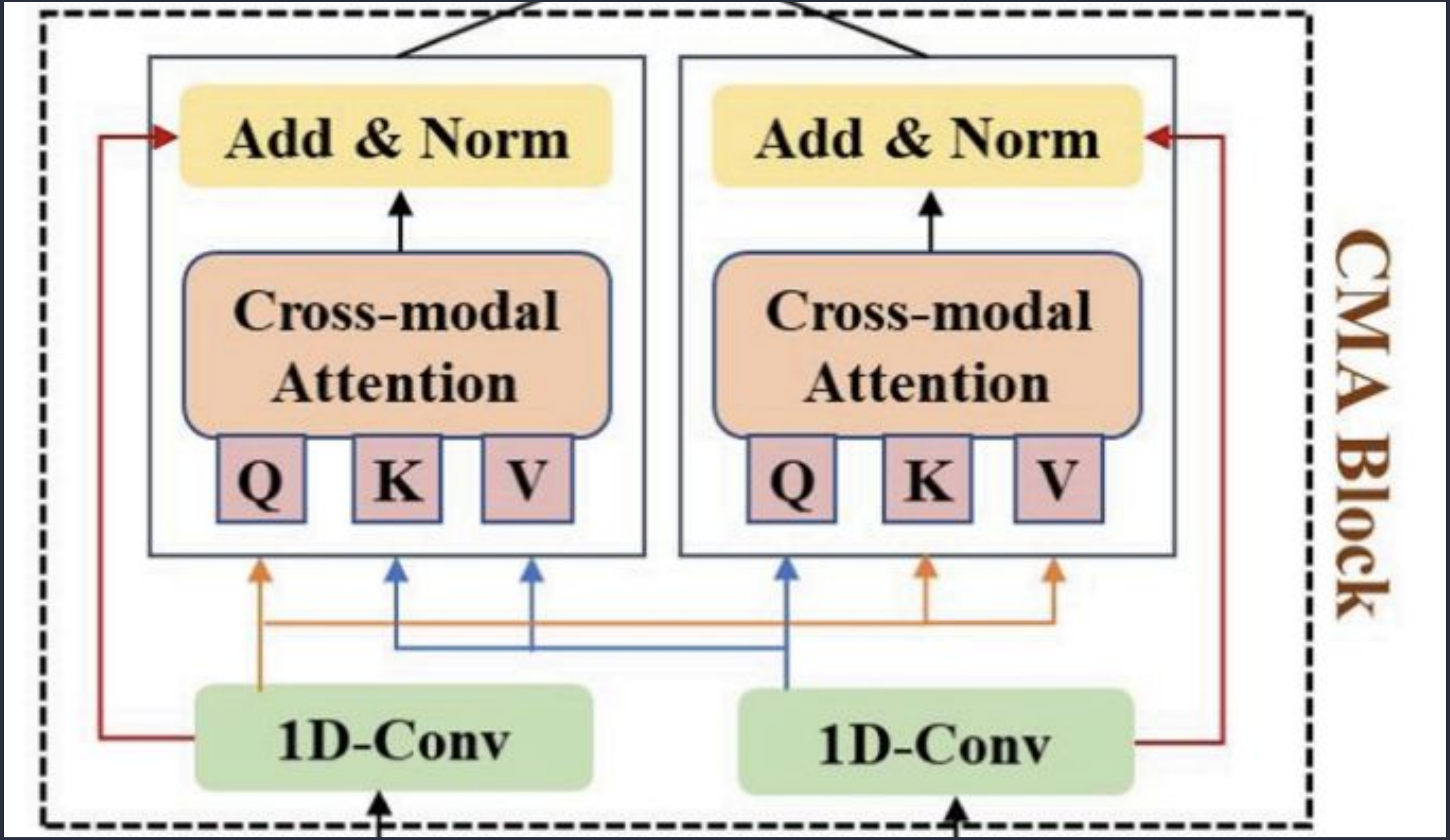
Fuse aligned features and predict emotion.

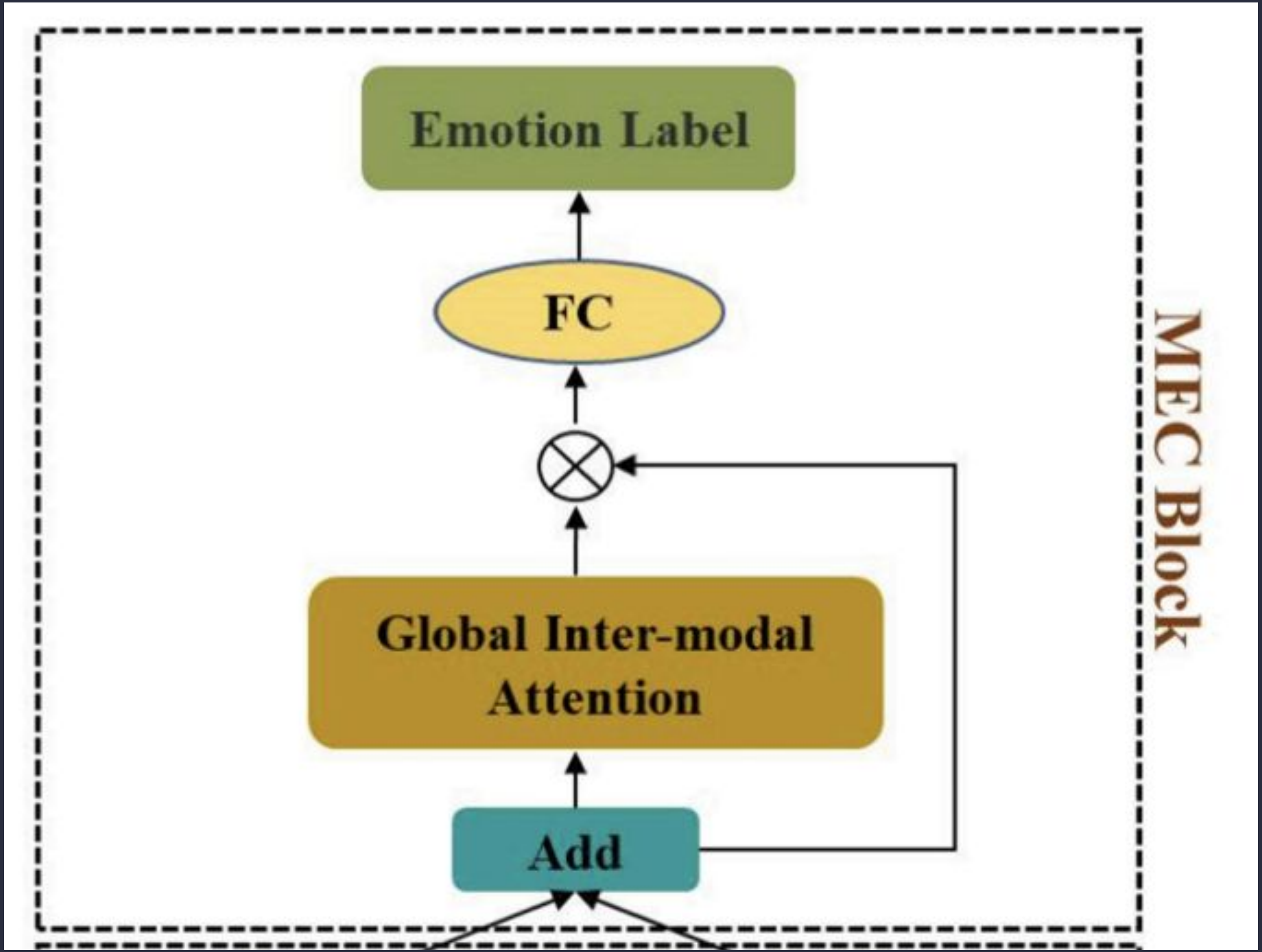
Three-tiered attention (intra-modal, cross-modal, global inter-modal) to capture comprehensive dependencies.



Emotion Label
FC
Global Inter-modal Attention
Add
MEC Block
Add & Norm
Cross-modal Attention
Q
K
V
Add & Norm
Cross-modal Attention
Q
K
V
1D-Conv
1D-Conv
CMA Block
Local Intra-modal Attention
Bi-LSTM
MFCC
Audio Encoder
Local Intra-modal Attention
BERT Word Embedding
Text Encoder
ATE Block
"Wow. Very promising."


ATE Block
Audio Encoder
Text Encoder
Local Intra-modal Attention
Local Intra-modal Attention
Bi-LSTM
MFCC
BERT Word Embedding
"Wow. Very promising."


Add & Norm
Add & Norm
Cross-modal Attention
Cross-modal Attention
Q
K
V
Q
K
V
1D-Conv
1D-Conv
CMA Block


Emotion Label
FC
Global Inter-modal Attention
Add
MEC Block

# Implementation Details & Dataset

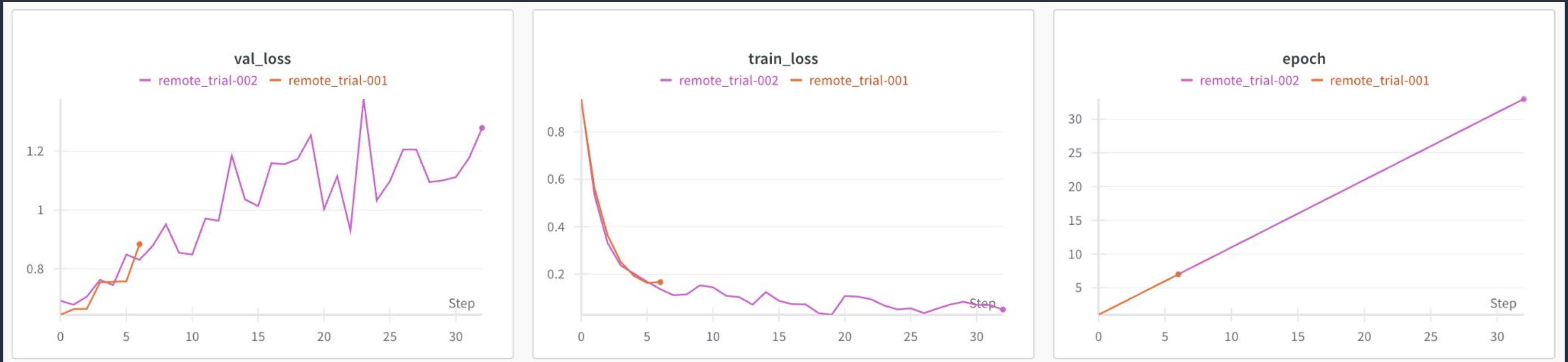
- **Dataset:** IEMOCAP (Interactive Emotional Dyadic Motion Capture) [4]
  - Audio recordings & transcriptions
  - Widely used benchmark for MER
- **Task:** 4-class emotion classification (Angry, Happy, Sad, Neutral)
- **Preprocessing:**
  - **Audio:** MFCC extraction (40-dim), 25ms frames, 10ms stride, padding/truncation.
  - **Text:** BERT tokenizer, padding/truncation.

# Experimental Setup

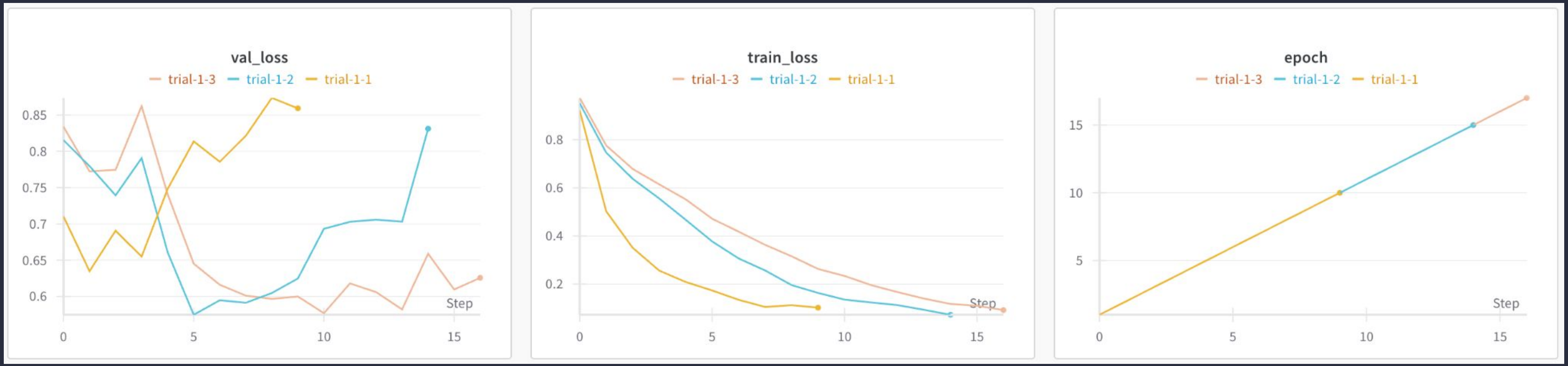
- **Framework:** Python, PyTorch [9], HuggingFace Transformers [8].
- **Model Architecture:** As described in the MER-HAN paper.
- **Training:**
  - Optimizer: Adam
  - Loss: Categorical Cross-Entropy.
  - Batch Size: 32, Epochs: 64 (with early stopping)
- **Evaluation Metrics:** Weighted Average Recall (WAR), Unweighted Average Recall (UAR), F1-score.
- **Experiment Tracking:** Weights & Biases [10]
- **GPU:** NVIDIA RTX 4090 24 GB

# Experimental Results

The initial findings showed room for improvement..



# Experimental Results



# Experimental Results

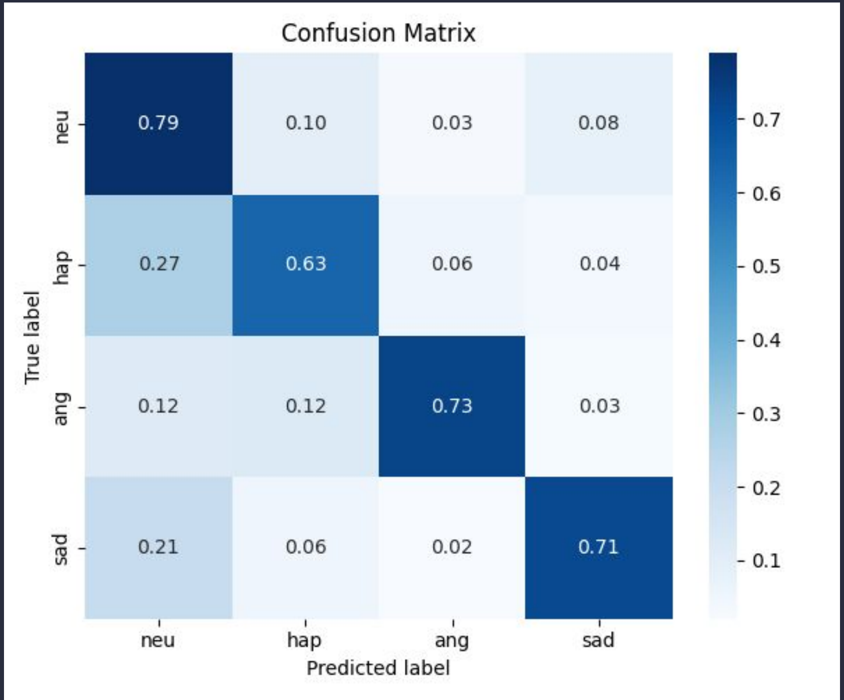
Final Evaluation Results

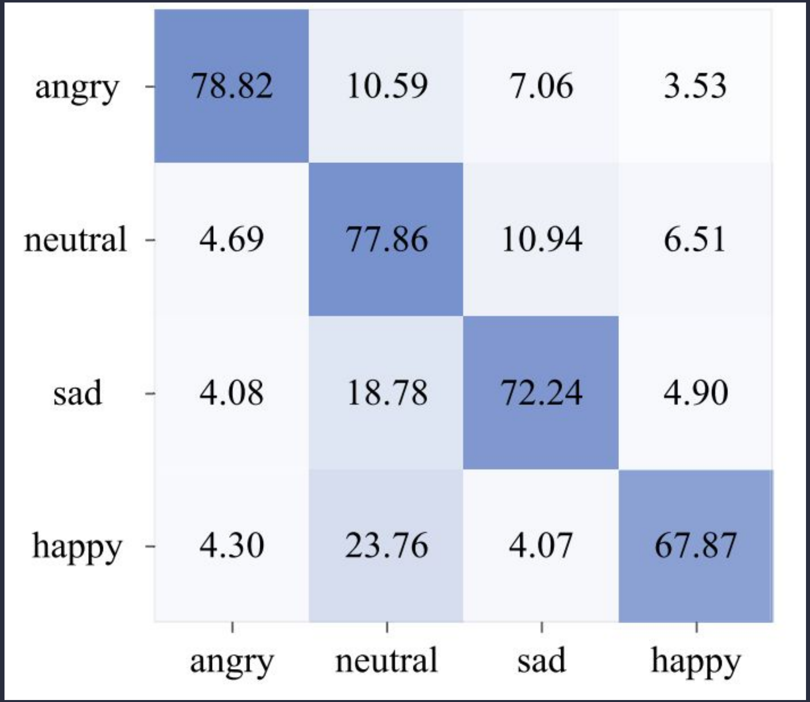
- WAR: 0.7099
- UAR: 0.7160
- Macro F1 score: 0.7172
- Weighted F1 score: 0.7108

**Table 4**
Recognition performance ( %) in an ablation study on IEMOCAP dataset.

| Methods | WAR | UAR | F1-score |
|---|---|---|---|
| Audio without local intra-modality attention | 50.60 | 52.18 | 51.05 |
| Audio without local intra-modality attention | 50.60 | 52.18 | 51.05 |
| Audio with local intra-modality attention | 55.52 | 56.98 | 56.06 |
| Text without local intra-modality attention | 67.43 | 67.72 | 66.61 |
| Text with local intra-modality attention | 68.57 | 69.54 | 68.96 |
| MER-HAN without CMA block | 70.34 | 71.87 | 70.56 |
| MER-HAN without global inter-modality attention | 71.21 | 71.43 | 71.31 |
| **MER-HAN** | **73.33** | **74.20** | **73.66** |

# Experimental Results



# Challenges in the Literature

- **Temporal Dynamics:**
  Modeling long-range dependencies
- **Cross-Modal Complementarity vs. Redundancy:**
  Balancing modality-specific strengths, suppressing noise.
- **Complexity & Interpretability:**
  Advanced attention reduces transparency.
- **Dataset Limitations:**
  Generalizability across languages (non-english)

# Challenges in My Implementation

- **Computational Resources:**

  Significant GPU memory and time required (NVIDIA RTX 4090).

- **Initial Problem of Overfitting:**

  As mentioned before, there was a overfitting problem due to unfreezed BERT layers.

# Open Problems and Future Directions

- Development of **robust fusion mechanisms** (efficiency & performance).
- Creation of **larger**, **more diverse**, **generalizable** datasets (multilingual, real life).
- Improved **temporal** and **contextual modeling**.
- **Interpretability** tools for attention mechanisms.

# Conclusion

**Summary:**

- Development of **robust fusion mechanisms** (efficiency & performance).
- Creation of **larger**, **more diverse**, **generalizable** datasets (multilingual, real life).
- Improved **temporal** and **contextual modeling**.
- **Interpretability** tools for attention mechanisms.

**Key Learnings:**

- MER-HAN's architecture is powerful for capturing inter-modal dynamics.
- However it requires high computational resource.

# 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 UsingMulti-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).

# Thank You!

Any questions? Ask away!