Multimodal Emotion Recognition (MER2025)

Project Overview

This project implements a Multimodal Emotion Recognition system using the MER2025 Track 1 dataset. The system integrates video, audio, and text modalities, with features extracted using ViT (video), HuBERT (audio), and RoBERTa (text). A LLaMA 2 model performs fusion and reasoning, combining the modalities into a unified representation for emotion classification.

For more details about the data set you can see in the link : https://github.com/zeroQiaoba/MERTools/tree/master/MER2025/MER2025_Track1

Dataset:

- Name: MER2025 Track 1
- Type: Self-Supervised dataset – feature extraction models were pretrained in a self-supervised manner before fine-tuning

Modalities:

- Video (frames from clips)
- Audio (speech and environmental sounds)
- Text (transcripts of speech)

Preprocessing & Feature Extraction:

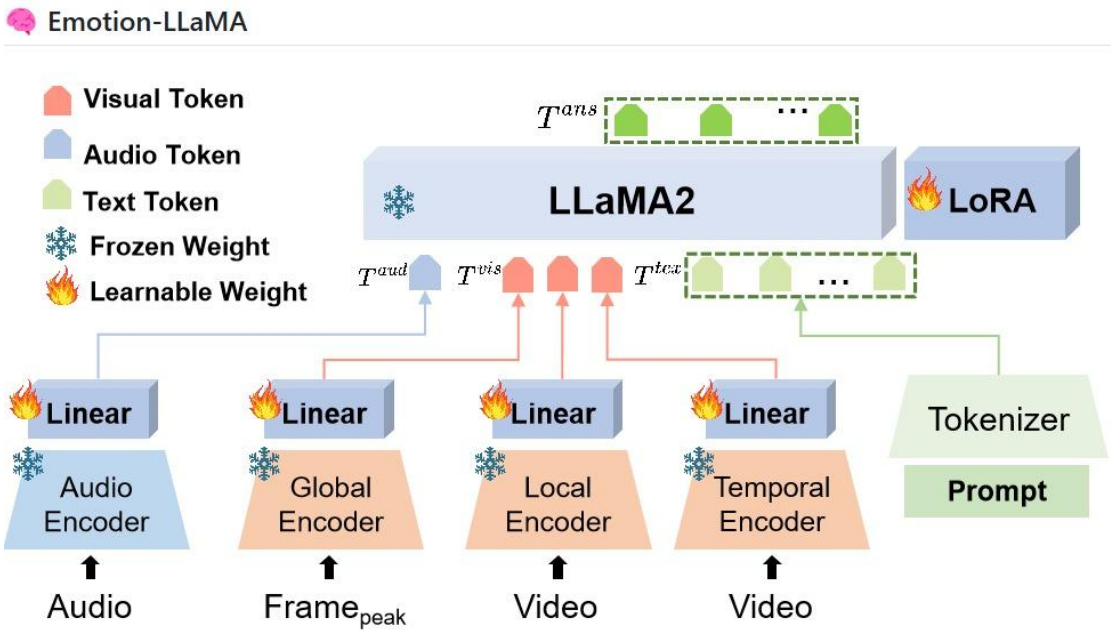
- Video: Vision Transformer (ViT) – pretrained with self-supervised learning for spatial-temporal visual features
- Audio: HuBERT – self-supervised speech/audio representation learning
- Text: RoBERTa – pretrained language model for contextual embeddings

Unimodal Baseline Testing:

- Before multimodal fusion, each modality was tested individually (Unimodal setting)
- An Attention-based model was applied to each modality separately to evaluate its standalone performance and identify the most informative modalities

Multimodal Runing:

Architecture:



Fusion :

Top-N Fusion in LLaMA 2:

- LLaMA 2 receives embeddings from all three modalities and applies cross-modal attention
- Each modality embedding is assigned an attention weight based on its relevance for the current sample
- The **Top N** highest-weighted embeddings are selected
- LLaMA 2 performs reasoning over these selected embeddings to produce a fused representation

Classifier:

- Fully connected layers for final emotion prediction

Model Output:

- Classes: Six discrete emotion categories (as defined by the dataset, without explicit explanations)

Loss Functions:

Two loss functions were used:

1. Cross-Entropy Loss – primary classification loss
2. Mean Squared Error (MSE) – secondary loss to align predicted emotion score distributions with ground truth

Implementation Steps:

1. Feature Extraction – Obtained embeddings from ViT, HuBERT, and RoBERTa (all pretrained with self-supervised learning)
2. Fusion in LLaMA 2 – Combined modality embeddings for reasoning
3. Training – Used AdamW optimizer, early stopping on validation accuracy
4. Evaluation – Measured Accuracy, F1-score, confusion matrix

Results:

Fusion	Top	Audio	Text	Video	10Epocs hours Running time	F1	Acc	2Epocs hours Running time	F1	Acc	100Epocs hours Running time	F1	Acc
AVT	1	A1	T1	V1	01:20	74.73%	59.66%				12:20	78.68%	64.86%
AVT	2	A1,A2	T1,T2	V1,V2	01:28	79.67%	66.22%	00:30	78.66%	64.33%	13:04	79.30%	65.70%
AV	1	A1	V1	V1	01:13	76.52%	61.97%				12:05	77.27%	62.96%
AV	2	A1,A2	V1,V2	V1,V2	01:21	73.72%	58.38%				12:55	76.33%	61.72%
AT	1	A1	T1	T1	01:14	73.92%	58.64%				11:55	74.25%	59.05%
AT	2	A1,A2	T1,T2	T1,T2	01:26	75.49%	60.63%				15:00	78.68%	64.86%
VT	1	T1	T1	V1	01:15	65.35%	48.53%						
VT	2	T1,T2	T1,T2	V1,V2	01:23	65.81%	49.04%	00:47	49.91%	33.25%			

Observations:

- Audio & video features were strong indicators for expressive emotions

Requirements:

Usage: