MulT Model Technical Report

Generated on: 2025-05-08 20:37:45

MulT Model Architecture

The MulT model is a multimodal transformer-based architecture designed for analyzing audio-visual data. Here are the key components:

Model Structure:

• Input projections for both audio and video modalities

• Transformer encoder layers with multi-head attention

• Cross-modal attention mechanism

• Output layers for predicting valence and arousal

Key Parameters:

Parameter	Value
Audio dimension	40 (mel filterbanks)
Video dimension	3 * 224 * 224 (RGB image flattened)
Hidden dimension	128
Number of attention heads	4
Number of transformer layers	2
Dropout rate	0.1
Maximum sequence length	1000

Feature Extraction Methods

Audio Feature Extraction:

• Model: TorchAudio's MelSpectrogram

• Feature Type: Mel-frequency cepstral coefficients (MFCCs)

• Parameters:

- Sample rate: 16000 Hz

- Number of mel filterbanks: 40

- FFT window size: 400

- Hop length: 160

- Window type: Hann window

• Processing steps:

- Loads audio file using torchaudio.load()

- Resamples if necessary using torchaudio.transforms.Resample

- Converts to mono if stereo using mean pooling

- Extracts mel spectrogram using torchaudio.transforms.MelSpectrogram

- Converts to decibels using torchaudio.transforms.AmplitudeToDB

- Output shape: [T, n_mels]

• Libraries used:

- torchaudio for audio processing

- torch for tensor operations

Video Feature Extraction:

• Model: OpenCV (cv2) with PyTorch transforms

• Feature Type: Raw RGB frames with ImageNet normalization

• Processing steps:

- Reads video frames using cv2.VideoCapture

- Converts BGR to RGB using cv2.cvtColor

- Resizes frames to 224x224 using cv2.resize

- Applies ImageNet normalization using torchvision.transforms:

* Mean: [0.485, 0.456, 0.406]

* Std: [0.229, 0.224, 0.225]

- Flattens frames to 1D vectors

• Output shape: [T, 3*H*W]

• Libraries used:

- OpenCV (cv2) for video reading and preprocessing

- torchvision.transforms for normalization

- torch for tensor operations

Training Configuration

Parameter	Value
Batch size	4
Number of epochs	50
Learning rate	1e-4
Weight decay	1e-5
Early stopping patience	5

Model Output

The model predicts two continuous values:

1. Valence (emotional positivity/negativity)
2. Arousal (emotional intensity)

Both outputs are normalized to the range [0, 1] using sigmoid activation.

Implementation Details

Dataset Handling:

- Uses PyTorch's Dataset class
- Processes audio-video pairs
- Handles padding and truncation
- Supports 5-minute duration clips

Training Process:

- Uses Adam optimizer
- Implements learning rate scheduling
- Includes early stopping
- Uses MSE loss for both valence and arousal

Evaluation:

- Processes results in temporal chunks
- Supports batch processing
- Saves analysis results to CSV