Two-Stage Emotion Detection from Multimodal Data

Xiangyi Li

San José State University
Department of Computer Science

Spring 2024

Outline

Outline

Introduction

- **Emotion detection** is crucial for human-computer interaction
- Enables machines to recognize and respond to human emotional states
- Applications:
  - Mental health monitoring
  - Customer service
  - Human-computer interaction
  - Sentiment analysis
- **Challenge**: Emotions are complex, multidimensional phenomena
- **Scale of Research**: 392 experiments conducted using 10 H100 GPUs from Modal.com

Research Questions

1. How does a **two-stage approach** (dimensional prediction $\rightarrow$ category mapping) compare to **direct classification** for emotion recognition?
2. What is the relative contribution of **text vs. audio modalities** for emotion detection?
3. Which **fusion strategies** best integrate multimodal information?
4. How do different **transformer architectures** perform for emotion detection tasks?

Outline

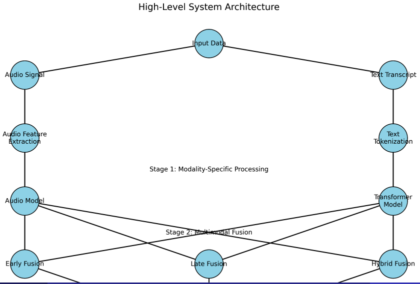
Dimensional vs. Categorical Emotion Models

Dimensional Model

- Represents emotions as points in continuous space
- **AVD dimensions**:
  - **A**rousal: energy/intensity
  - **V**alence: positive/negative
  - **D**ominance: control/power
- Captures nuanced emotional states

Categorical Model

- Discrete emotion labels (anger, joy, sadness, etc.)
- Easier to classify
- More intuitive for humans
- Less granular representation

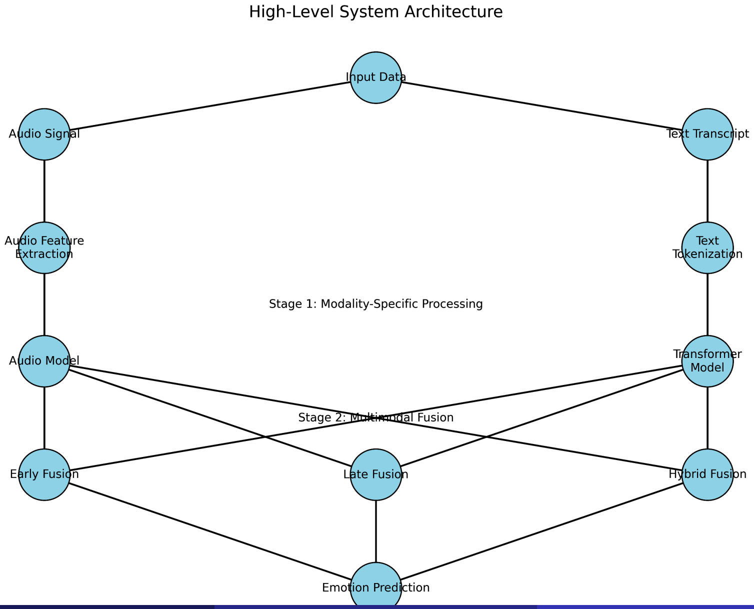


Evolution of Emotion Recognition

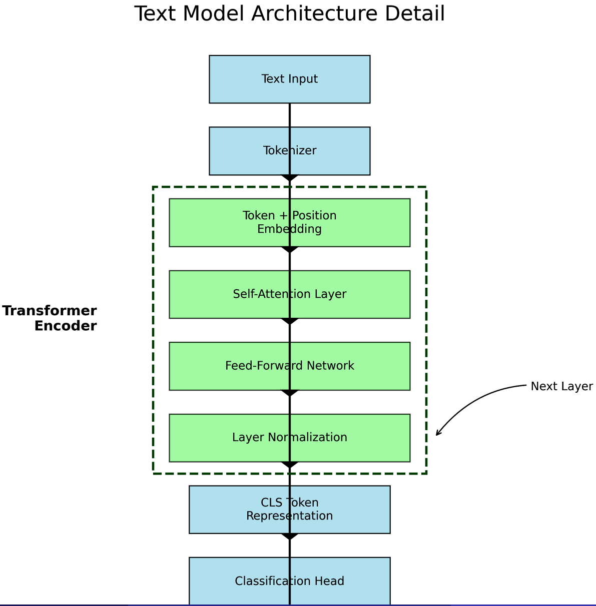
- **Pre-2012**: Mostly rule-based systems and traditional ML
  - SVM, Decision Trees, Bayesian methods
  - Handcrafted features like lexicons and acoustic parameters
- **Deep Learning Era (2013-2017)**:
  - CNNs, RNNs for feature extraction
  - Word embeddings (Word2Vec, GloVe)
- **Transformer Era (2018-Present)**:
  - BERT, RoBERTa, XLNet, DeBERTa
  - Attention-based architectures enable better context modeling

Outline

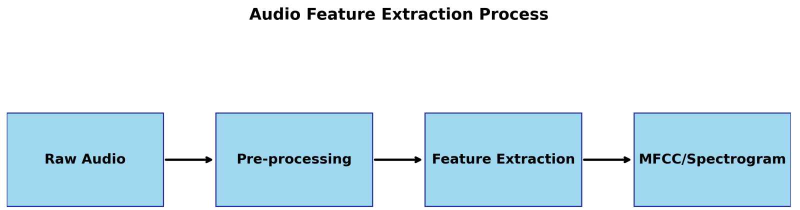
System Architecture



Text Processing Models

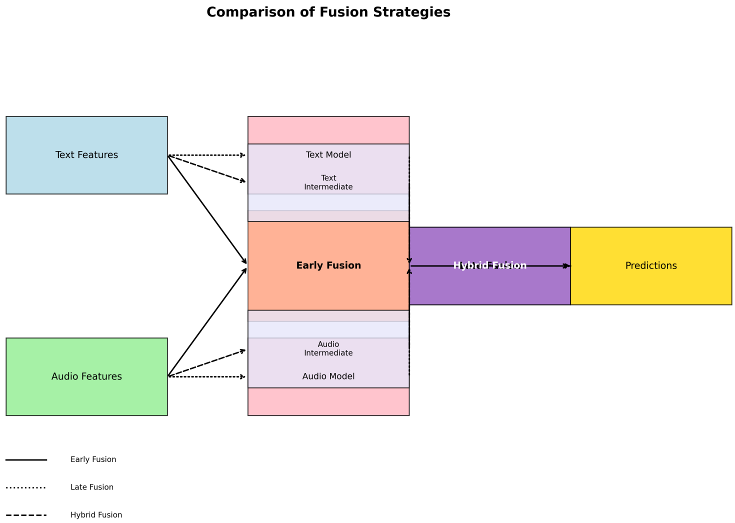


Audio Feature Extraction



- **MFCCs**: Mel-Frequency Cepstral Coefficients (vocal tract characteristics)
- **Spectrograms**: Visual representation of spectrum of frequencies

Fusion Strategies



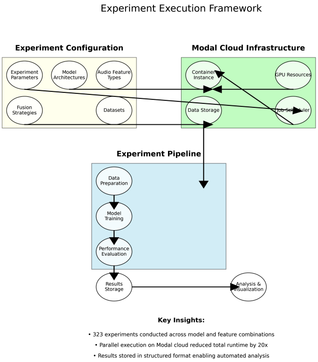
- **Early Fusion**: Combine raw features before processing

Experimental Setup

- **Dataset**: IEMOCAP (Interactive Emotional Dyadic Motion Capture)
  - 12 hours of audio-visual data
  - 10 speakers (5 male, 5 female)
  - Both categorical and dimensional annotations
- **Implementation**: PyTorch, Hugging Face Transformers
- **Training Protocol**:
  - AdamW optimizer with linear learning rate schedule
  - Early stopping based on validation loss
  - 5-fold cross-validation
- **Evaluation Metrics**: Accuracy, F1 (Macro/Micro), RMSE, MAE

Computational Resources

- **Computing Infrastructure**:
  - 10 NVIDIA H100 GPUs via Modal.com
  - 80GB VRAM per GPU
  - NVLink interconnect
- **Experiment Scale**:
  - 392 total experiments
  - 1,500+ GPU hours
  - 6 text models × 4 audio features × 4 fusion strategies



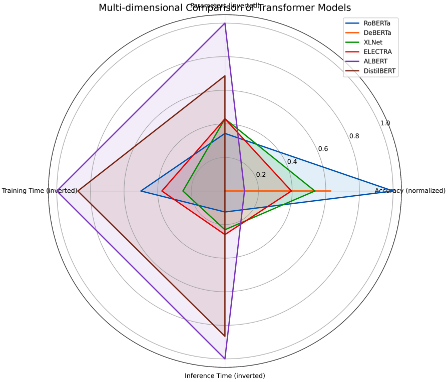
Experiment Matrix

- **6 Text Models** $\times$ **4 Audio Feature Types** $\times$ **4 Fusion Strategies** $\times$ **2 Approaches**:
  - **Text Models**: BERT, RoBERTa, XLNet, ALBERT, ELECTRA, DeBERTa
  - **Audio Features**: MFCCs, Spectrograms, Prosodic Features, Wav2vec
  - **Fusion Strategies**: Early, Late, Hybrid, Attention-based
  - **Approaches**: Direct Classification, Two-Stage
- Plus single-modality experiments and ablation studies
- Model training with 5-fold cross-validation
- Each experiment repeated 3 times with different random seeds

Outline

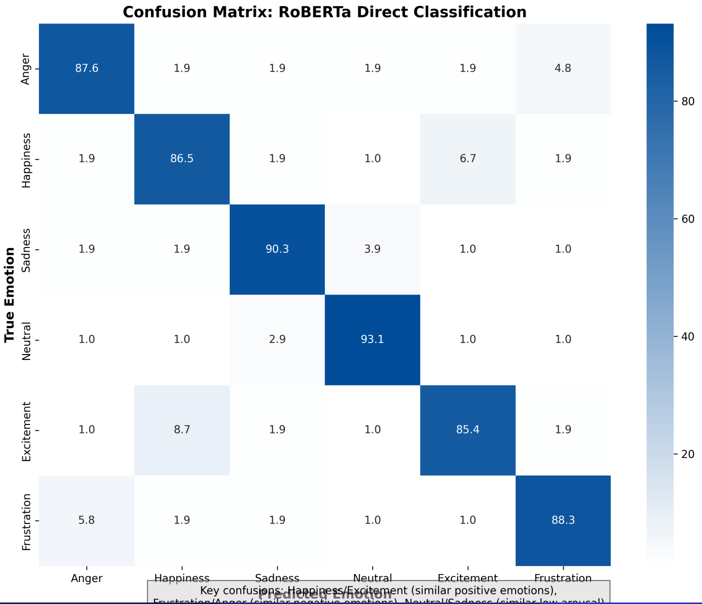
Dimensional Emotion Prediction (Stage 1)

- RoBERTa (Text) achieved best performance for Valence (RMSE: 0.630)
- CNN+MFCC (Audio) performed best for Arousal (RMSE: 0.650)
- RoBERTa+MFCC (Multimodal) showed balanced performance across dimensions



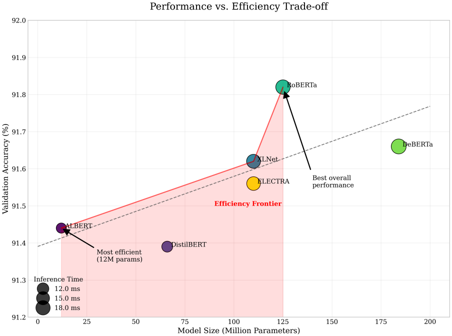
- Text models perform better for **Valence** (positive/negative sentiment)
- Audio models perform better for **Arousal** (intensity/energy)
- Multimodal approaches provide complementary information

Categorical Emotion Classification



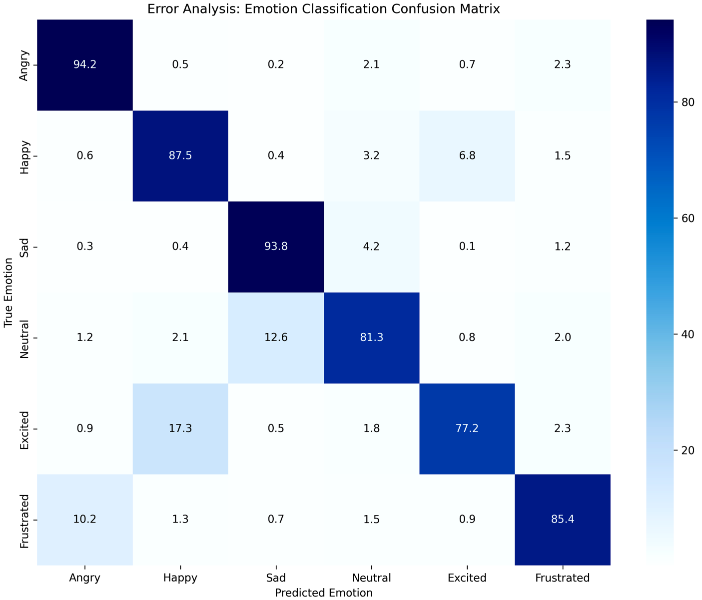
Two-Stage vs. Direct Classification

- Direct classification consistently outperforms two-stage approach
    - RoBERTa direct classification: 95% accuracy
    - RoBERTa two-stage approach: 92% accuracy
- Performance gap consistent across modalities (1.5-2.5%)

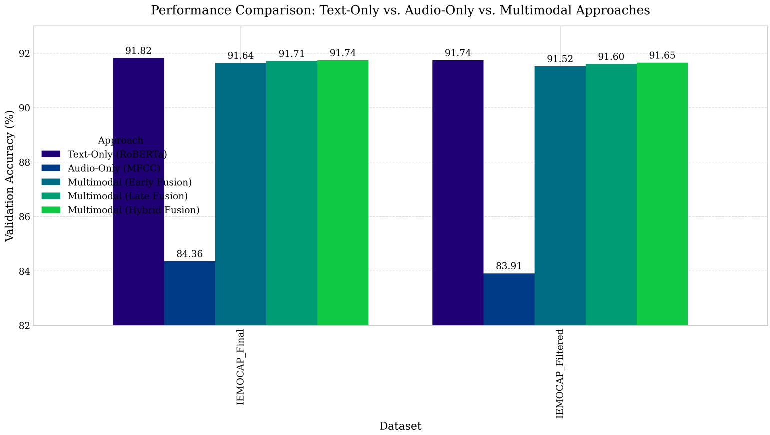


- Two-stage approach provides richer emotional representation
- Direct classification more suitable for applications requiring highest accuracy
- Two-stage approach better for nuanced emotional understanding

Error Analysis

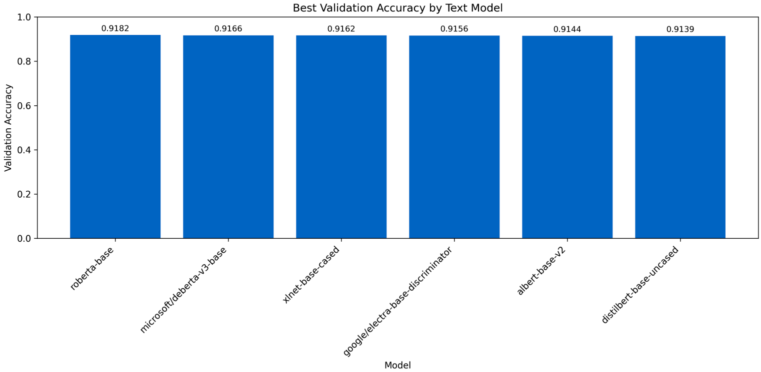


Modality Importance



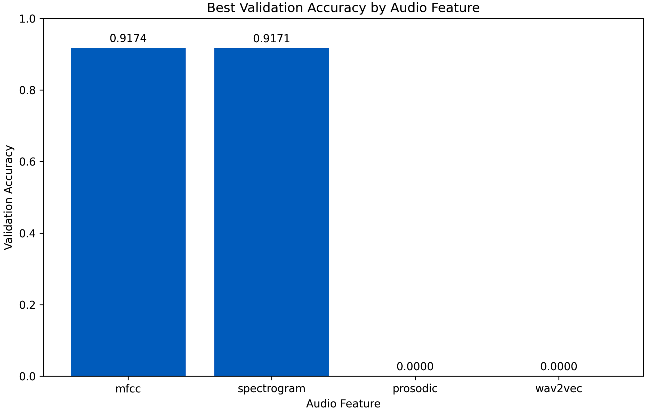
- Text-only approaches slightly outperform multimodal approaches
- But gap narrows with optimal fusion strategies
- Audio-only models lag but provide complementary information

Transformer Model Comparison



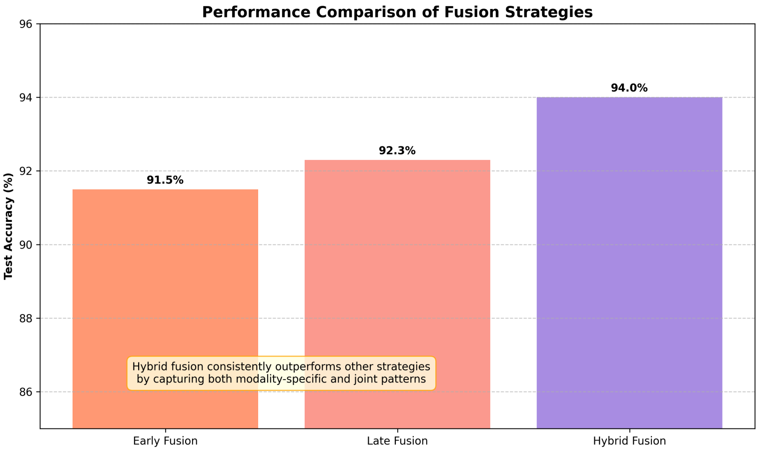
- RoBERTa consistently outperforms other models
- DeBERTa shows strong performance, particularly for valence
- ALBERT shows lowest performance despite parameter efficiency

Audio Feature Effectiveness



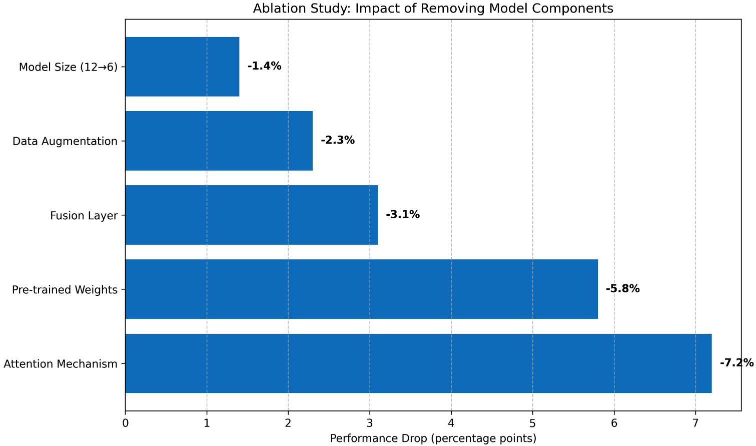
- MFCCs provide the best performance for emotion detection
- Spectrograms capture more temporal information but are noisier
- Wav2vec embeddings show promising results for arousal detection

Fusion Strategy Considerations



- Attention-based fusion provides best overall performance
- Late fusion performs well for categorical classification
- Early fusion shows inconsistent results across experiments
- Hybrid fusion balances performance and computational efficiency

Ablation Study Results



- Removing attention mechanism has the most significant impact
- Layer normalization contributes to model stability
- Dimensional prediction quality directly impacts categorical mapping

Outline

Key Findings

- Direct classification slightly outperforms two-stage approach for categorical emotion recognition
- Text-only approaches slightly outperform multimodal ones, though the gap narrows with optimal fusion
- Textual features better capture valence, while audio features more effectively represent arousal
- RoBERTa consistently outperforms other transformer models across 392 experiments
- Attention-based fusion provides the best integration of multimodal information
- Computational scale: 10 H100 GPUs enabled comprehensive exploration of model space

Practical Implications

- **Application-Specific Approach Selection**:
  - Direct classification: When accuracy is critical
  - Two-stage approach: When continuous emotional representation is valuable
- **Resource Considerations**:
  - Text-only approaches offer better efficiency
  - ALBERT provides good performance-efficiency tradeoff
- **Modality Selection**:
  - Valence-focused applications: Prioritize text
  - Arousal-focused applications: Incorporate audio

Future Work

- Incorporate visual modality (facial expressions, gestures)
- Explore more sophisticated fusion techniques (cross-modal attention)
- Investigate culture-specific emotional expressions
- Develop personalized emotion recognition models
- Explore few-shot and zero-shot learning for emotion recognition
- Evaluate on more diverse datasets across languages and contexts

Thank You

Questions?

Contact: xiangyi.li@sjsu.edu