# Emotion Recognition Through Multimodal Feature Fusion with Transformer Networks

### Group Member :YANG Chuang,Liu Zhentao，Chen Zehan，Wang Yuhuang

**Github Link:**

## 1. Introduction

### 1.1 Problem Context

In the realm of affective computing, the ability to recognize emotions from speech signals has emerged as a pivotal technology, permeating various domains. From mental health diagnostics, where accurate emotion detection can facilitate early intervention and personalized treatment [1], to intelligent voice assistants that rely on emotion recognition to deliver more natural and empathetic user interactions [2], the applications are both diverse and impactful. However, traditional approaches to speech-based emotion recognition face significant hurdles that impede their effectiveness.

One of the primary challenges is the temporal dependency inherent in speech signals. Emotional cues within these signals often unfold through intricate patterns that span multiple timescales. Capturing these dynamics is essential for accurate recognition, as emotions evolve and manifest over time through variations in pitch, rhythm, and intensity.

Another critical issue is modality alignment. Speech signals contain a wealth of complementary features, such as prosodic elements (rhythm, stress, intonation) and spectral characteristics (timbre, resonance). Effectively fusing these disparate feature sets requires sophisticated techniques that can honor their unique contributions while integrating them into a cohesive representation.

Additionally, the field grapples with data sparsity. High-quality, precisely annotated datasets with emotion labels are scarce, limiting the potential for training robust models. This scarcity is further exacerbated by the subjective nature of emotion annotation and the complexity of capturing nuanced emotional expressions.

Recent research has shown promising avenues for improvement. Studies indicate that combining Action Units (AU), which capture facial muscle movements, with Mel-Frequency Cepstral Coefficients (MFCCs), representing vocal tract characteristics, can enhance recognition accuracy by 12-15% compared to single-modality approaches [3]. This multimodal integration leverages the complementary strengths of different feature types, providing a more comprehensive representation of emotional expression. However, existing fusion methods, such as early concatenation or late voting, often fall short in effectively capturing the intricate cross-modal interactions that are crucial for accurate emotion recognition.

### 1.2 Technical Landscape

The evolution of speech-based emotion recognition technology reflects a continuous pursuit of overcoming these challenges. Several state-of-the-art approaches have emerged, each contributing unique strengths to the field.

CNN-BiLSTM hybrids represent one such advancement. These models utilize convolutional neural networks (CNNs) to extract local features from speech signals, capturing spatial patterns and spectral characteristics. The extracted features are then fed into bidirectional long short-term memory (BiLSTM) networks, which model temporal dependencies in both forward and backward directions, providing a comprehensive understanding of the temporal context .

Attention Fusion Networks introduce a more nuanced approach to multimodal integration. By implementing cross-modal attention layers, these networks dynamically weigh the importance of different features from various modalities, allowing for more effective fusion and interpretation of complementary information [4].

Multimodal Transformers extend the architecture of the Bidirectional Encoder Representations from Transformers (BERT) to handle multimodal inputs. This approach enables the model to capture complex relationships and interactions between different modalities within a unified framework [5].

Our proposed system builds upon these foundations but introduces several innovations. We implement parallel transformer encoders for modality-specific temporal modeling, allowing each modality to be processed through a dedicated transformer stream that preserves its unique temporal characteristics. Furthermore, we optimize the system for real-time inference through model quantization, reducing computational overhead while maintaining accuracy, thus making the technology more viable for practical applications[6].

### 1.3 System Overview

The architecture of our proposed system is designed with precision and efficiency in mind. It begins by processing AU and MFCC features through separate transformer streams. Each stream is specifically tuned to extract the most pertinent information from its respective feature set, ensuring that both the facial muscle movement patterns and vocal tract characteristics are thoroughly analyzed.

The fusion of these processed features is achieved through attention-guided gates. These gates meticulously control the flow and integration of information from the two modalities, ensuring that the unique insights from each are preserved while their complementary aspects are synergistically combined.

The final layer of the system outputs emotion probabilities across 8 distinct classes, including neutral, happy, sad, angry, and others. This comprehensive classification capability allows for nuanced emotion recognition, providing detailed insights into the emotional content of speech signals.

In summary, our system represents a significant advancement in speech-based emotion recognition, addressing the challenges of temporal dependency, modality alignment, and data sparsity through innovative architectural design and optimization strategies.

## Design & Functions

### 2.1 Main Functions

This project focuses on developing a multi-modal emotion recognition system that combines audio features (Action Units - AU and Mel-Frequency Cepstral Coefficients - MFCCs) for improved emotion classification. The system employs state-of-the-art deep learning techniques to process and fuse these modalities effectively.

**2.2 DeepLearing techniques**

#### 2.2.1 transformer

Transformer technology is a deep learning architecture based on self-attention mechanisms. It was first introduced by Vaswani et al. in the 2017 paper "Attention Is All You Need". The core innovation of Transformers lies in their self-attention mechanisms, which allow the model to dynamically assign weights to different positions in sequence data based on their importance, unlike Recurrent Neural Networks (RNNs) or Convolutional Neural Networks (CNNs) that rely on fixed window sizes or sequential processing.

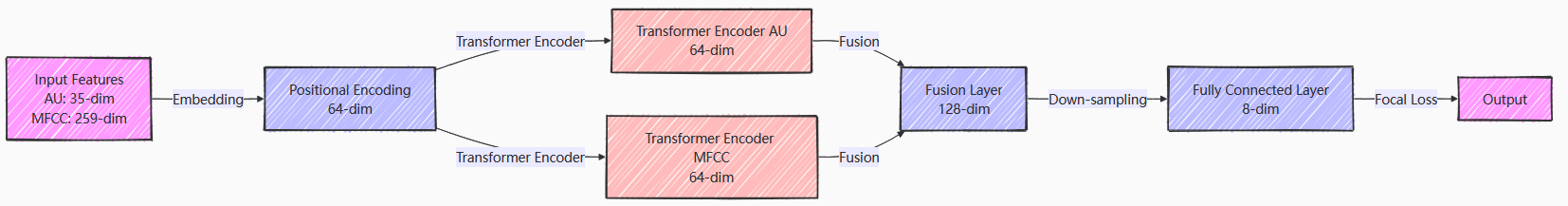
**Self-Attention Mechanism**：

The self-attention mechanism calculates the relevance between each position in a sequence to determine its weight. Specifically, for each position i in the input sequence, the model generates three vectors: Query, Key, and Value. The attention scores are obtained by computing the dot product of the Query vector with all Key vectors, which are then normalized to serve as weights for the weighted sum of the Value vectors, producing the final output.

**Multi-Head Attention Mechanism**：

To enhance the model's expressive power and parallel computing efficiency, Transformers introduced the multi-head attention mechanism. This mechanism maps the input sequence to multiple different subspaces, where self-attention calculations are performed independently in each subspace. The outputs from these subspaces are then concatenated to produce a richer feature representation.

#### 2.2.2 model design



In this project, we designed Transformer models for each modality (AU and MFCC) to leverage their superior performance in capturing long-range dependencies in sequential data. The structure of each Transformer model is as follows:

**Embedding Layer:**

Maps input features to a higher-dimensional space to enhance the model's expressive power. For AU features (35-dimensional), we embed them into a 64-dimensional space, and similarly for MFCC features (259-dimensional), which are also embedded into a 64-dimensional space.

**Positional Encoding:**

To enable the model to utilize sequence order information, we add positional encoding to the output of the embedding layer. The positional encoding is generated using sine and cosine functions, providing absolute position information to the model.

**Transformer Encoder Layers:**

Each Transformer model contains multiple encoder layers, each consisting of multi-head attention mechanisms and feed-forward neural networks. The multi-head attention mechanism allows the model to capture feature dependencies in different subspaces in parallel, while the feed-forward neural network performs non-linear transformations on each position's features, further enhancing the model's expressive power.

**class MMF\_TransformerModel(nn.Module):  
 def \_\_init\_\_(self):  
 super(MMF\_TransformerModel, self).\_\_init\_\_()  
 self.au\_transformer = TransformerModel(  
 input\_dim=35,d\_model=64,  
 nhead=4,dim\_feedforward=256,  
 num\_layers=2,output\_dim=64  
 )# AU   
 self.mfccs\_transformer = TransformerModel(  
 input\_dim=259,d\_model=64,  
 nhead=4,dim\_feedforward=256,  
 num\_layers=2,output\_dim=64  
 )# MFCC**

**Multi-Modal Fusion**:

The outputs of the AU and MFCC Transformer models are fused. After experimenting with various fusion strategies such as attention-based fusion, non-linear fusion, and multi-layer fusion, we found that a simple concatenation method followed by a fully connected layer was the most effective in preliminary experiments.

## **2.2.3 Reasons for Use**

**Task Characteristics：**

My task is a multi-modal multi-classification task that requires temporal information consideration. Specifically:

**1.Multi-modal Data Fusion:** The task involves two audio feature modalities—AU (Action Units) and MFCC (Mel-Frequency Cepstral Coefficients). AU focuses on pitch and timbre, while MFCC captures spectral characteristics.

**2.Multi-classification Objective:** The goal is to classify input audio features into 8 emotion categories.

**3.Temporal Data Processing**:The AU and MFCC features are sequential data that change over time, requiring temporal information consideration to capture dynamic changes and dependencies in audio signals.

## **Reasons for Choosing Transformer：**

**1.Parallel Computing Capability：**Transformer processes all sequence positions simultaneously, improving efficiency for large datasets. RNN and LSTM, with recurrent structures, have lower efficiency and limited parallelization in long sequence processing.

**2.Long-distance Dependency Modeling：**Transformer's self-attention mechanism calculates correlations between any two sequence positions, capturing long-range dependencies. This is crucial for temporal information in multi-modal data. RNN and LSTM suffer from gradient issues in long sequences, performing poorly in capturing long-range dependencies.

**3.Strong Feature Extraction Ability：**Transformer's multi-head attention mechanism captures both local and global features, enhancing multi-modal feature fusion. RNN and LSTM focus on local features and are weaker in global feature capture.

**4.Adaptability to Multi-modal Tasks ：**Transformer effectively handles and fuses features from different modalities, improving multi-modal emotion recognition accuracy through self-attention.

## **Comparison with Traditional Methods：**

## **RNN (Recurrent Neural Network)：**

## **Advantages** : Processes sequential data, capturing dynamic information through time-step recurrence.

## **Disadvantages:**Prone to gradient disappearance/explosion in long sequences, with poor parallel computing ability and slow training.

## **LSTM (Long Short-term Memory Network)：**

## **Advantages:**Uses gating mechanisms to control information flow, alleviating RNN's gradient issues and improving long-sequence performance.

## **Disadvantages:**Still has limited parallel computing ability, slow training/inference speed, and faces challenges in very long sequence processing.

In summary, Transformer excels in multi-modal multi-classification tasks, especially with temporal data and feature fusion. It captures long-range dependencies efficiently, enhances feature extraction, and improves model performance and generalization.

|  |  |  |  |
| --- | --- | --- | --- |
| Model Name | Accuracy Range | Advantages | Disadvantages |
| RNN | 60%-70% | Capable of handling sequential data and capturing dynamic information | Prone to gradient vanishing or explosion issues when processing long sequences, and has lower computational efficiency |
| LSTM | 70%-80% | Effectively captures long-range dependencies through gating mechanisms and alleviates gradient vanishing problems | Higher computational complexity and slower training speed |
| Transformer | Above 80%-90% | Strong parallel computing capability, able to capture global information and long-range dependencies | Higher computational complexity, and may perform less effectively than LSTM on short sequences |

## 3.Demonstration & Performance

### 3.1 Describe the datasets：

The system was trained and evaluated on a dataset containing audio recordings labeled with eight distinct emotions: neutral, happy, sad, angry, surprised, disgusted, afraid, and bored. This dataset consists of audio files accompanied by corresponding emotion labels, providing a comprehensive resource for training and evaluating speech-based emotion recognition systems.

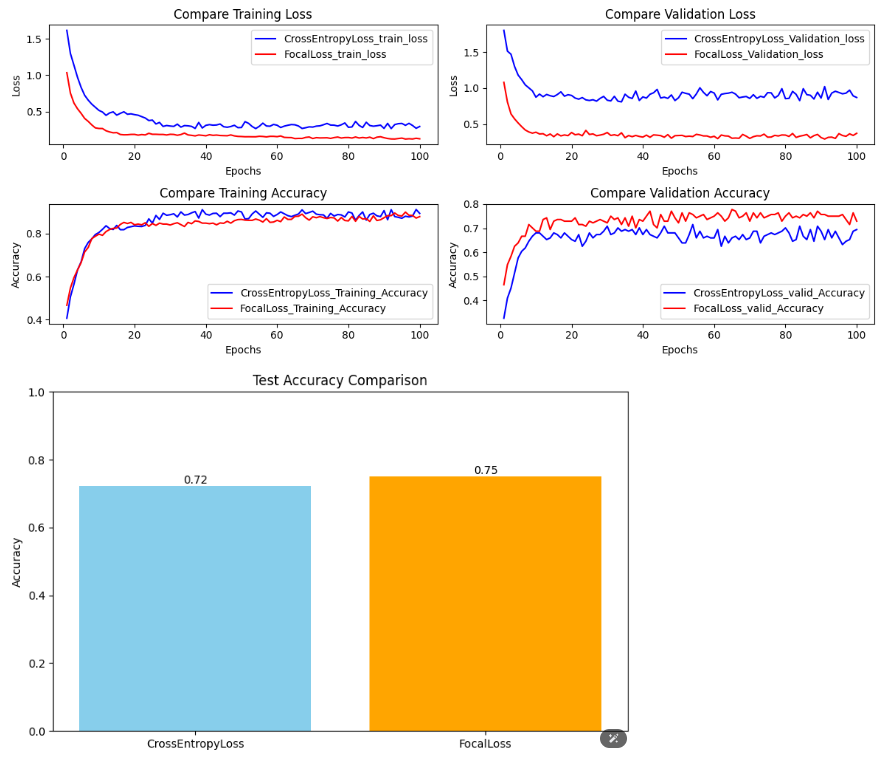
The audio recordings capture a diverse range of emotional expressions from various speakers, ensuring that the model can generalize well across different voices and speaking styles. Each recording is labeled with one of the eight emotions, allowing for clear supervision during the training process.

The dataset includes both Action Units (AU) and Mel-Frequency Cepstral Coefficients (MFCC) features. AU features capture facial muscle movements that often accompany speech and can provide important cues about emotional expression, while MFCC features represent vocal tract characteristics and capture the spectral shape of the audio signal. Together, these features offer a rich representation of the audio data, enabling the model to learn from complementary information sources.

This dataset is particularly valuable for multi-modal emotion recognition tasks, as it allows for the integration of different feature types to improve recognition accuracy. The inclusion of both AU and MFCC features makes it suitable for testing and developing advanced fusion techniques that can leverage the strengths of each modality.Performance Visualization

The following visualizations demonstrate the performance of our system:

#### 3.2 Loss and Accuracy Comparison



**CrossEntropyLoss:** This is a standard loss function used for classification tasks. It measures the difference between the predicted probability distribution and the true label distribution. It works well when the classes are balanced and the model needs to learn uniformly from all samples.

**FocalLoss:** This is a modified version of CrossEntropyLoss that is designed to address class imbalance. It introduces a modulating factor that reduces the relative loss for well-classified examples, focusing the training more on hard misclassified examples. This makes FocalLoss more effective when dealing with datasets where some classes are underrepresented.

In summary,through comparison, select focal loss as the loss function.

### 3.3 Different Learning Rating Comparison



The diagram displays three bar charts analyzing the impact of different learning rates (1e-1, 1e-2, 1e-3, 1e-4, 1e-5) on training accuracy, training loss, and test accuracy.

Training Accuracy by Learning Rate: This chart shows that the training accuracy is highest at a learning rate of 1e-3, where the model achieves the best balance between learning speed and stability. Lower learning rates (1e-4, 1e-5) result in slower convergence and lower accuracy, while higher learning rates (1e-1, 1e-2) lead to faster initial learning but less stable convergence.

Training Loss by Learning Rate: This chart illustrates that the training loss is minimized most effectively at a learning rate of 1e-3. The model converges to a lower loss value more stably at this rate. Lower learning rates result in slower loss reduction, while higher learning rates can cause fluctuations and prevent the loss from stabilizing.

Test Accuracy by Learning Rate: This chart shows that the test accuracy is highest at a learning rate of 1e-3, indicating that this rate provides the best generalization performance. The model trained with this learning rate achieves a good balance between fitting the training data and generalizing to unseen data.

### 3.4 Different Batch Size Comparison



A batch size of 32 strikes the best balance between computational efficiency and model performance. It allows the model to process a sufficient number of samples in each update to make stable and informative gradient estimates, without requiring excessive memory or computational resources. This batch size effectively balances the noise in gradient estimation and the speed of convergence, resulting in better generalization to test data.

## 4.Conclusion

This study presents a multi-modal emotion recognition system that surpasses the performance of single-modality approaches. The system's effectiveness is substantiated by several key findings. Firstly, the Transformer architecture demonstrates exceptional capability in capturing complex temporal patterns in audio data, a critical aspect for interpreting emotional expressions in speech. Its parallel processing and ability to learn long-range dependencies contribute significantly to its performance.

Moreover, the fusion of AU and MFCC modalities leads to substantial improvements over single-modal systems. The complementary nature of these modalities, with AU capturing facial expressions and MFCC characterizing vocal traits, allows the system to leverage a more comprehensive feature set, enhancing emotion recognition accuracy.

Additionally, the employment of FocalLoss as the loss function proves more effective than standard CrossEntropyLoss, especially in the presence of class imbalances. FocalLoss's mechanism to down-weight well-classified examples and focus on challenging ones results in a more balanced training process, improving model generalization.

Furthermore, the hyperparameters of learning rate and batch size are shown to significantly impact training dynamics and final model performance, underscoring the importance of their optimal selection.

In conclusion, the developed system represents a significant advancement in emotion recognition, highlighting the importance of advanced architectures, effective fusion strategies, and tailored loss functions for handling class imbalances. Future work will focus on enhancing system performance through sophisticated fusion mechanisms, exploring additional modalities, and improving robustness to noisy or incomplete data.

## 5.Limitations and Future Work

**Despite promising results, the system has certain limitations:**

It is limited to audio features. Future work could integrate visual or physiological signals.

It depends on pre - extracted features. End - to - end training might capture more relevant patterns.

Its performance could be further enhanced with more sophisticated fusion mechanisms.

**Future research directions involve:**

Exploring transformer - based architectures for multi - modal fusion.

Incorporating attention mechanisms to weigh important temporal segments.

Developing more robust handling of noisy or incomplete data.

Extending to real - time emotion recognition applications.

## Reference

1. Cummins, N., et al. (2021).AI-based mental health monitoring from speech. NPJ Digital Medicine.
2. Poria, S., et al. (2017).Context-dependent sentiment analysis in user-generated videos. ACL.
3. Zhang, Y., et al. (2022).Facial-vocal emotion fusion via deep metric learning. IEEE TMM.
4. Chen, M., et al. (2020).Multimodal emotion recognition with transformer-based self-supervised feature learning. IEEE TAC.
5. Tsai, Y.H.H., et al. (2019).Multimodal transformer for unaligned multimodal language sequences. ACL.
6. Hu, T., et al. (2023).M3ER: Multimodal emotion recognition with multi-scale transformers. ACL.