university of groningen

campus fryslân

Enhanced Multimodal Emotion Recognition using GRU and Self-Attention Mechanisms: Techniques and Applications

Jingwen Shi

university of groningen

campus fryslân

University of Groningen - Campus Fryslân

Enhanced Multimodal Emotion Recognition using GRU and Self-Attention Mechanisms: Techniques and Applications

Master's Thesis

To fulfill the requirements for the degree of
Master of Science in Voice Technology
at University of Groningen under the supervision of
Assistant Professor Dr. Shekhar Nayak (Voice Technology, University of Groningen)

Jingwen Shi (S5718902)

June 11, 2024

Acknowledgements

As my journey as an international student draws to a close, I reflect on this experience as if it were an incredibly vivid dream. Here, I have met many friends who have kept me company and provided the support that has allowed me to reach this point smoothly.

First and foremost, I would like to express my sincere gratitude to Dr. Shekhar and Dr. Matt for their invaluable assistance with my thesis. I am also deeply thankful for the guidance and support provided by Dr. Joshua, Dr. Phat, and Dr. Vass throughout my studies.

The process of discussing, sharing, and learning with Huang, Lai, Li, Liao, Su, and Wang was incredibly enjoyable. Their willingness to share and mutual companionship made the journey of writing this thesis much less lonely. I am especially grateful to Sherry for generously lending me her computer. Without it, and with my own faulty one, I would have struggled to complete my thesis on time.

A heartfelt thanks goes to Sun, who flew from the UK to accompany me. Her presence encouraged me to explore parts of the Netherlands I had never visited before. To Zhang, who is soon to arrive, thank you for giving me something to look forward to and the motivation to keep going.

I am deeply appreciative of my parents, who provided me with full financial support and encouraged me to travel and relax my mind. I am also thankful for my friend Shi back in China. Despite the time difference, we regularly talked on the phone, shared our feelings, and comforted each other, easing the frustrations of life. Her voice always made me feel like I had a solid support system behind me.

Finally, my deepest gratitude goes to Lei. She was the first person I met here and the one who has given me the most help, support, and companionship throughout this year. She helped me navigate daily life in this country. Her tolerance and optimism have profoundly influenced me, making any obstacles seem insignificant.

Thank you to everyone. I wish us all a future filled with peace, health, and success!

Abstract

This thesis makes substantial contributions to the field of multimodal emotion recognition by developing and evaluating models that integrate audio, visual, and textual data. We utilized state-of-the-art feature extraction techniques, including BERT for text, LibROSA for audio, and OpenFace for visual cues, achieving a comprehensive representation of multimodal data. A novel temporal alignment technique was introduced to synchronize features across modalities, ensuring coherent integration and enhancing the model's ability to capture intricate relationships between different modalities.

The proposed model architecture combines Gated Recurrent Units (GRUs) and self-attention mechanisms, effectively capturing both local and global dependencies, significantly improving feature extraction and emotion recognition accuracy. A stacking fusion module was implemented to amalgamate information from text, audio, and visual modalities, leading to superior performance metrics across multiple datasets, including CMU-MOSI, CMU-MOSEI, and CH-SIMS. Extensive evaluation demonstrated substantial improvements over baseline models, validating the effectiveness of the proposed methods in achieving higher accuracy and robustness in emotion recognition.

Our research has significant practical implications, setting a new benchmark for emotion recognition. The developed system enhances human-computer interactions, provides multilingual support in virtual assistants, and assists language learners, thereby contributing to the preservation of linguistic diversity and cultural heritage. Additionally, this work contributes to the development of socially intelligent and empathetic artificial systems, paving the way for more advanced applications in affective computing.

In conclusion, this thesis advances the field of multimodal emotion recognition through innovative methods and comprehensive evaluation. The findings underscore the importance of integrating multiple data modalities and provide a solid foundation for future research and practical applications, offering pathways for continued innovation in recognizing and understanding human emotions.

Contents

1 Introduction **9**

1.1 Background 9

1.2 Importance 9

1.3 Objective and Significance 10

1.4 Structure of the Thesis 10

2 Literature Review **13**

2.1 Research on Feature Extraction 13

2.1.1 Bert 13

2.1.2 MFCC 15

2.1.3 Facial landmark detection and tracking 17

2.2 Research on Model Architecture 19

2.2.1 RNNs 19

2.2.2 GRUs 20

2.3 Research on Attention Mechanisms 21

2.3.1 Attention 21

2.3.2 Self-attention 22

2.4 Research on Fusion Strategies 23

2.4.1 MLP 24

2.4.2 Rectified Linear Unit (RELU) 25

2.4.3 LeakyReLU 25

3 Methodology **27**

3.1 Datasets 27

3.1.1 Dataset Introduction 27

3.2 Models 27

3.2.1 Input Layer 28
3.2.2 Feature Alignment 28
3.2.3 GRU and Self-Attention Modules 29
3.2.4 Fusion Module 29
3.2.5 Prediction Layer 29
3.3 Innovations 29
3.4 Evaluation - Word Error Rate 30
3.4.1 Model System 30
3.4.2 Output Metric 30
3.5 Ethical considerations 31

4 Experimental Setup **34**
4.1 Datasets 34
4.1.1 PKL File Format 34
4.1.2 Process setup 35
4.2 Models 36
4.2.1 GRUWithLinear Model 36
4.2.2 MLP (Two-layered Perceptron) Model 37
4.2.3 SelfAttention Layer 37
4.2.4 EMIFusion Model 37
4.3 Supervised Learning 38
4.3.1 Models Defined 38
4.3.2 Components and Functionality 38
4.3.3 Relationships and Effects 39
4.4 Evaluation 39
4.4.1 Complexity 39
4.4.2 Performance 40

4.4.3 Robustness 40

5 Results **43**

5.1 MOSI, MOSEI, and SIMS Datasets 43

5.2 Training and Validation Losses 44

5.3 Confusion Matrices 46

5.3.1 Seven-Class Confusion Matrices 46

5.3.2 Two-Class Confusion Matrices 48

6 Discussion **52**

6.1 Validation of the First Hypothesis: Temporal Alignment of Multimodal Features . . . 52

6.2 Validation of the Second Hypothesis: Combination of Self-Attention and GRU . . . 54

6.3 Validation of the Third Hypothesis: Stacking in the Fusion Module 55

6.4 Summary and Implications 55

7 Conclusion **57**

7.1 Summary of the Main Contributions 57

7.2 Future Work 57

7.3 Impact & Relevance 58

References **60**

1 Introduction

1.1 Background

Human emotions play a crucial role in interpersonal communication and social interaction, forming an integral part of human expression. Effective recognition and understanding of emotions are essential for facilitating meaningful communication between individuals. With the growing demand in areas such as human-computer interaction and affective computing, the need for computers to comprehend and respond to user emotions is becoming increasingly significant. The emergence of deep learning and multimodal data has led to a surge of interest in multimodal speech emotion recognition, a field that combines various modalities such as speech, images, and text to accurately capture and understand human emotions, thereby enhancing the performance and user experience of affective computing systems.

1.2 Importance

Multimodal speech emotion recognition holds significant practical application value and theoretical research significance, contributing to the improvement of human-computer interaction experiences, the advancement of affective computing, and the assistance in medical diagnosis.

- **Enhancing User Experience**: In applications like human-computer interaction and intelligent customer service, recognizing user emotional states can help systems better understand user needs and intentions, thereby providing more personalized and effective services.
- **Fostering Affective Computing Development**: Affective computing is a vital direction in artificial intelligence aimed at enabling computers to understand and express emotions. Multimodal speech emotion recognition provides crucial technical support for the development of affective computing systems.
- **Assisting Medical Diagnosis**: Emotion recognition technology can be applied in the medical field to assist doctors in better understanding patients' emotional states, aiding in diagnosis and treatment. For instance, in the realm of mental health, analyzing patients' speech and images can help doctors promptly detect and diagnose emotional disorders.
- **Promoting Social Intelligence Development**: Social intelligence refers to the ability of computer systems to understand and simulate human social behavior. Multimodal speech emotion recognition can assist computer systems in better comprehending human emotional communication, thereby enhancing their social intelligence.

1.3 Objective and Significance

The objective of this paper is to address the current shortcomings in multimodal speech emotion recognition research and propose strategies to overcome them. By conducting an extensive literature review and analyzing existing research findings, this study aims to provide a theoretical and practical foundation for advancing the field of multimodal speech emotion recognition.

The significance of this study lies in its potential to:

- **Advance Research in Multimodal Speech Emotion Recognition**: By identifying and addressing existing research gaps, this study aims to contribute to the development of more robust and accurate multimodal speech emotion recognition systems.
- **Inform Future Research Directions**: Through a comprehensive analysis of the current state of the art and the challenges faced by researchers, this paper aims to guide future research efforts towards addressing key issues and improving the performance of multimodal speech emotion recognition systems.
- **Enhance Practical Applications**: By improving the accuracy and reliability of multimodal speech emotion recognition systems, this research has the potential to enhance various practical applications, including human-computer interaction, affective computing, and medical diagnosis.

This paper endeavors to shed light on the current landscape of multimodal speech emotion recognition, identify areas for improvement, and propose strategies to advance the field, ultimately contributing to the development of more effective and efficient systems for recognizing and understanding human emotions.

1.4 Structure of the Thesis

This thesis is structured to provide a comprehensive examination of the research question and hypothesis concerning multimodal emotion recognition. The organization is designed to guide the reader through the theoretical foundations, methodological approaches, experimental procedures, results, and discussions, leading to insightful conclusions and recommendations for future work.

- **Introduction** 1 The introduction provides a comprehensive overview of the thesis, beginning with the central inquiry into whether recognizing correlations between language content, vocal characteristics, and facial expressions can enhance multimodal emotion recognition accuracy. It highlights the practical applications and theoretical significance of this research, aiming to address existing gaps, inform future directions, and improve practical applications. The structure of the thesis, delineated in detail, guides the reader through key sections including literature review, methodology, experimental setup, results, discussion, conclusion, and future work, setting the stage for a systematic exploration of multimodal emotion recognition research.

- **Literature Review** 2 The literature review offers an extensive exploration of the existing body of work in the field of multimodal emotion recognition. It covers a wide range of topics, from feature extraction to fusion strategies. Advanced techniques such as BERT for text processing, LibROSA for audio feature extraction, and OpenFace for facial feature analysis are discussed in detail. The review also delves into the use of recurrent neural networks, specifically GRUs, for modeling sequential dependencies, and the application of self-attention mechanisms and multilayer perceptrons (MLPs) for integrating information across modalities. This section not only frames the research question and hypothesis but also highlights the state-of-the-art approaches and identifies gaps that the current research aims to address.

- **Methodology** 3 The methodology section outlines the research design, including the datasets used, data preprocessing techniques, model architectures, training procedures, and evaluation metrics. It describes the innovative structural elements of the approach, including a modal alignment module to ensure synchronization of modality data, the incorporation of attention mechanisms in GRUs, and an efficient multimodal feature fusion module that improves prediction performance and reduces overfitting risks.

- **Experimental Setup** 4 This section describes the experimental setup developed to validate the proposed methodologies. It includes a detailed explanation of the code structure and the functionality of various components, such as data storage and evaluation scripts. The main components include data preprocessing modules, model training and testing scripts, and the fusion module for integrating outputs from different modalities. The experimental setup ensures that the models are trained and evaluated consistently, and the results are compared against baseline methods to highlight the effectiveness of the proposed approach.

- **Results** 5 The results section presents the outcomes of the experiments conducted on the CMU-MOSI, CMU-MOSEI, and CH-SIMS datasets. It includes a comparison of the proposed approach against baseline methods, highlighting significant improvements in performance. Detailed analyses are provided for each dataset, and the accuracy of recognizing each type of emotion is evaluated. Performance metrics demonstrate the superiority of the proposed methods in achieving higher accuracy and robustness.

- **Discussion** 6 This section discusses the results in depth, drawing insights from the performance metrics and qualitative analyses. It examines the strengths and limitations of the proposed approach and identifies potential areas for improvement. The use of confusion matrices helps to analyze the results for each emotion category across the three datasets, providing a clear understanding of where the model performs well and where it may need refinement.

- **Conclusion and Future Work** 7 The thesis concludes with a summary of the key findings and contributions. It reflects on the implications of the research, emphasizing the advancements made in multimodal emotion recognition. Recommendations for future work are provided, suggesting avenues for further exploration to enhance the robustness and accuracy of these systems. The conclusion aims to guide future research endeavors and inspire continued innovation in the field of multimodal emotion recognition.

  Through this structured approach, the thesis aims to make a significant contribution to the development of more intelligent, empathetic, and context-aware artificial systems.

2 Literature Review

Effective multimodal emotion recognition hinges upon robust feature extraction and sophisticated model architectures. This review explores prominent techniques in both realms, including feature extraction methodologies such as BERT for text, MFCC for audio, and facial landmark detection and tracking for facial images. Additionally, it delves into model architectures, with a focus on the utilization of GRU-based recurrent neural networks, attention mechanisms, and fusion strategies employing MLPs with ReLU, and LeakyReLU activations. By synthesizing insights from these diverse approaches, this review sets the stage for understanding the landscape of multimodal emotion recognition and highlights avenues for further research and development.

2.1 Research on Feature Extraction

In the realm of multimodal emotion recognition, effective feature extraction plays a pivotal role in capturing the nuanced characteristics of diverse data modalities. To this end, state-of-the-art approaches often leverage advanced techniques such as BERT for text, MFCC for audio, and facial landmark detection and tracking for facial images. BERT, a pre-trained transformer model, has demonstrated remarkable proficiency in capturing contextual information from textual data, thereby enabling more nuanced understanding of linguistic cues and semantics. We use LibROSA as a tool for audio analysis, extracting insightful features from audio signals, and facilitating the representation of acoustic patterns and emotional nuances. Additionally, for robust facial landmark detection and feature extraction capabilities, we use OpenFace, it offers a rich source of facial features essential for discerning emotional expressions from images.

2.1.1 Bert

BERT, proposed by [1], represents a new model paradigm. Its full name, Bidirectional Encoder Representations from Transformers, indicates that it is a bidirectional encoder representation derived from the Transformer model. Unlike previous models such as ELMo ([2]) and the Generative Pre-trained Transformer (OpenAI GPT) ([3]), BERT introduces a novel pre-training mechanism that allows for the simultaneous consideration of left and right textual context at all layers. This design leads to significant performance improvements, enabling the BERT model to achieve state-of-the-art results across various natural language processing tasks, including the GLUE benchmark, the MultiNLI task, and the SQuAD v1.1 and SQuAD v2.0 tests.

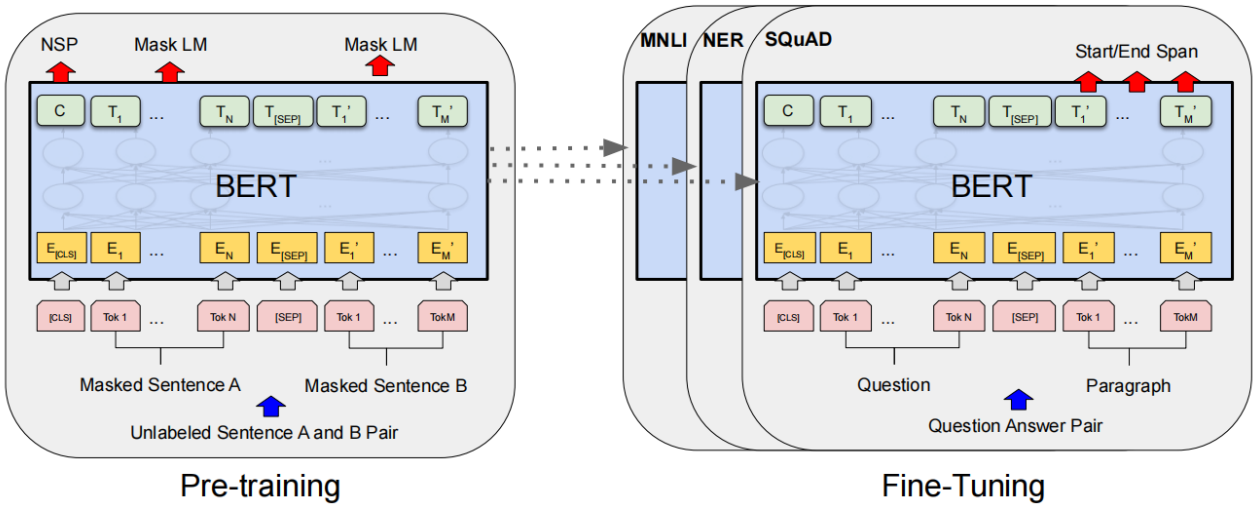


Figure 1: Overall pre-training and fine-tuning procedures for BERT. Apart from output layers, the same architectures are used in both pre-training and fine-tuning. The same pre-trained model parameters are used to initialize models for different down-stream tasks. During fine-tuning, all parameters are fine-tuned. [CLS] is a special symbol added in front of every input example, and [SEP] is a special separator token (e.g. separating questions/answers).[1]

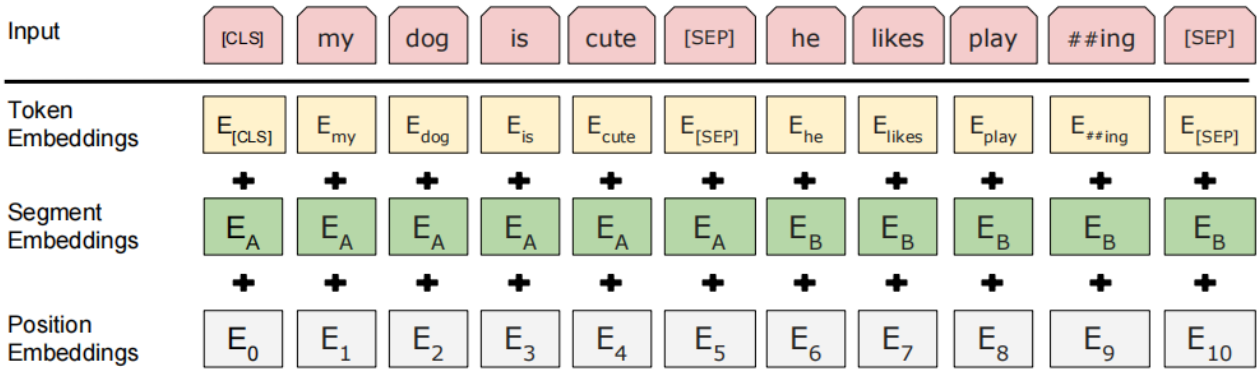


Figure 2: BERT input representation. The input embeddings are the sum of the token embeddings, the segmentation embeddings and the position embeddings.[1]

BERT's task-specific design can represent a single sentence or a pair of sentences as a contiguous token array [1]. For a given token, its input representation is constructed by summing the corresponding token, segment, and position embeddings. In classification tasks, the first word of the sequence is marked with a unique [CLS] token, and a fully connected layer is attached to the [CLS] position in the final encoder layer. The classification of the sentence or sentence pair is then completed via a softmax layer [4]. BERT has two parameter-intensive configurations: BERTbase and BERTlarge. BERTbase consists of 12 Transformer blocks, a hidden layer size of 768, 12 self-attention heads, and a total of 110 million parameters for the pre-trained model. In contrast, BERTlarge comprises 24 Transformer blocks, a hidden layer size of 1024, 16 self-attention heads, and a total of 340 million parameters for the pre-trained model. Due to the higher memory requirements of the BERTlarge

model, its maximum batch size is very small on standard GPUs with 12GB of RAM, which can affect the model's accuracy.

By visualizing the loss landscape and optimization trajectories of fine-tuning BERT on specific datasets, [5] revealed the effectiveness and impact of language model pre-training. The results indicate that pre-training facilitates the fine-tuning process by making it easier to locate broad optima, enhancing the model's generalization capabilities, and demonstrating strong robustness against overfitting. Moreover, the lower layers of the BERT model exhibit superior transfer learning abilities. These findings provide valuable insights for further optimizing the fine-tuning of pre-trained models, with the potential to advance the field of natural language processing.

[3] introduced BERT Base Uncased as well, a version within the BERT (Bidirectional Encoder Representations from Transformers) series that disregards case sensitivity. During the pre-training process, all text is converted to lowercase to enhance the model's generalization capabilities.

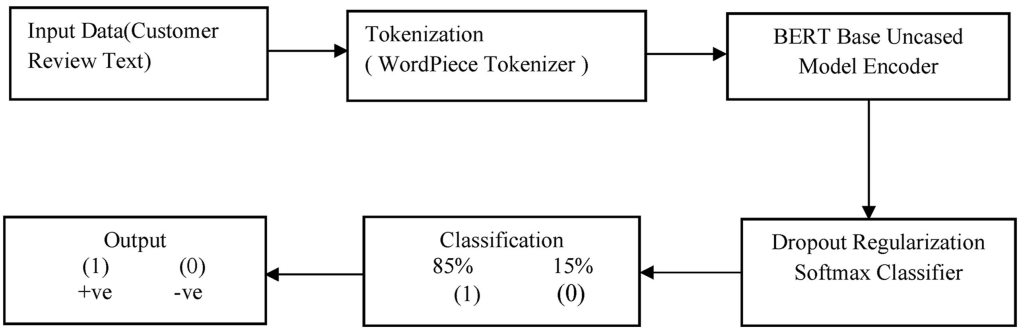


Figure 3: Bert Base uncased model architecture.[3]

2.1.2 MFCC

Mel Frequency Cepstral Coefficients (MFCC) are a crucial feature extraction technique in speech processing, leveraging the principles of human auditory perception to provide an efficient representation of speech signals. In speech and speaker recognition systems, MFCCs are extracted from speech signals during both training and testing phases. During training, MFCCs are used to learn and store the characteristics of each speaker's voice. In the testing phase, MFCCs extracted from new speech samples are compared with the stored data using similarity measures like Euclidean distance to identify the speaker.

Steps to Calculate MFCC:

- **Pre-emphasis**: A pre-emphasis filter is applied to boost the high frequencies of the signal.
- **Framing**: The signal is divided into small frames of 20-40 milliseconds, as the properties of speech are quasi-stationary over short durations.
- **Windowing**: Each frame is windowed to minimize discontinuities at the beginning and end of each frame.

- **Fast Fourier Transform (FFT)**: The FFT is applied to convert each frame from the time domain to the frequency domain.
- **Mel Filter Bank**: The frequency domain signal is then passed through a series of filters that mimic the human ear's response, known as mel filters.
- **Logarithm**: The logarithm of the filter bank outputs is taken.
- **Discrete Cosine Transform (DCT)**: The log mel spectrum is converted back to the time domain, resulting in the MFCCs.

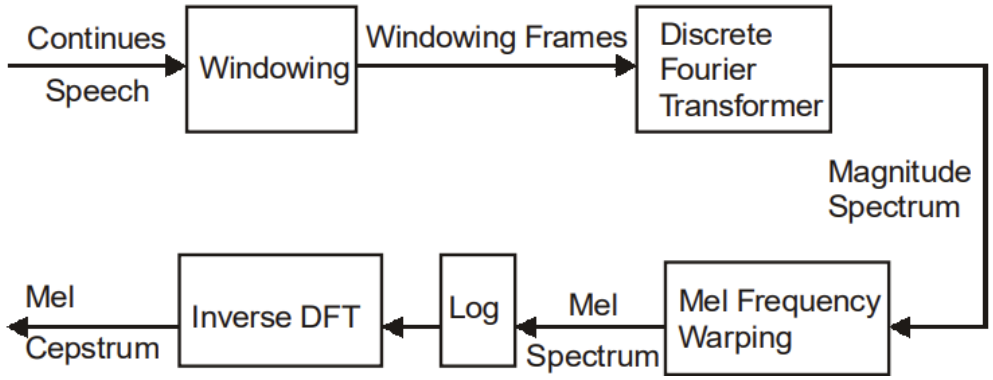


Figure 4: Complete pipeline for MFCC[6]

MFCCs are highly effective due to their simulation of the non-linear perception of the human ear, providing a compact and robust representation of the speech signal's spectral properties. This makes them a proven and efficient tool in various speech recognition applications. [6]

Developed by Brian McFee and other contributors libROSA[7] is a Python package designed for audio and music signal processing. The process of extracting audio features using MFCC (Mel-Frequency Cepstral Coefficients) with libROSA includes loading the audio, preprocessing, framing, windowing, Fourier transform, power spectrum density computation, applying a Mel filter bank, logarithmic transformation, discrete cosine transform, and feature extraction. These steps combine signal processing and spectral analysis techniques to convert audio into MFCC coefficients, which are then used for tasks such as audio classification and speech recognition. Its introduction has made audio and music signal analysis using Python more convenient and efficient. [8] Utilizing Python along with audio processing libraries such as librosa and soundfile, and employing the scikit-learn library for audio analysis, emotions in the RAVDESS dataset—anger, sadness, happiness, neutrality, calmness, fear, disgust, and surprise—were recognized.

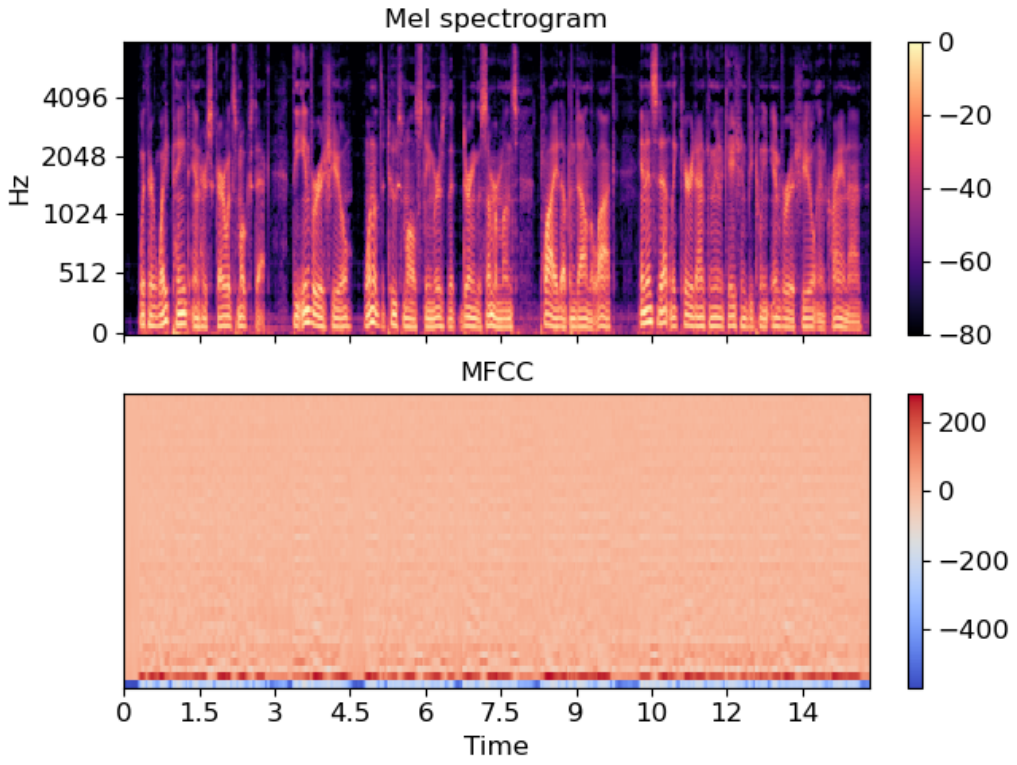


Figure 5: Librosa-Feature-MFCC
`https://librosa.org/doc/main/generated/librosa.feature.mfcc.html`

2.1.3 Facial landmark detection and tracking

Facial landmark detection is crucial for capturing the rigid and non-rigid deformations of facial components due to head movements and facial expressions, making it essential for various facial analysis tasks. Over the years, numerous algorithms have been developed to detect these key points automatically. This review [9] extensively reviews these algorithms, classifying them into three major categories: holistic methods, Constrained Local Model (CLM) methods, and regression-based methods. Holistic methods build models representing global facial appearance and shape information, CLMs leverage global shape models while building local appearance models, and regression-based methods implicitly capture facial shape and appearance information. The underlying theories and differences of algorithms within each category are discussed, along with their performance on both controlled and ”in-the-wild” benchmark datasets under varying facial expressions, head poses, and occlusions. Additionally, this review [9] includes a section on the latest deep learning-based algorithms, benchmark databases, and existing software, highlighting their respective strengths and weaknesses. Future research directions are identified, including the potential of combining different methodological categories to enhance landmark detection in diverse, real-world scenarios.

Developed by Brandon Amos, Bartosz Ludwiczuk, and Mahadev Satyanarayanan from Carnegie Mellon University, OpenFace [10] is the first open-source tool capable of performing facial landmark detection, head pose estimation, facial action unit recognition, and eye gaze estimation. At its

core, it utilizes computer vision algorithms that have demonstrated state-of-the-art results in various tasks, such as TCDCN CNN [11] and FaceTracker CLM [12]. Additionally, the tool achieves real-time performance and can operate with standard webcams without the need for any specialized hardware. Lastly, OpenFace allows for easy integration with other applications and devices through a lightweight messaging system.[10]

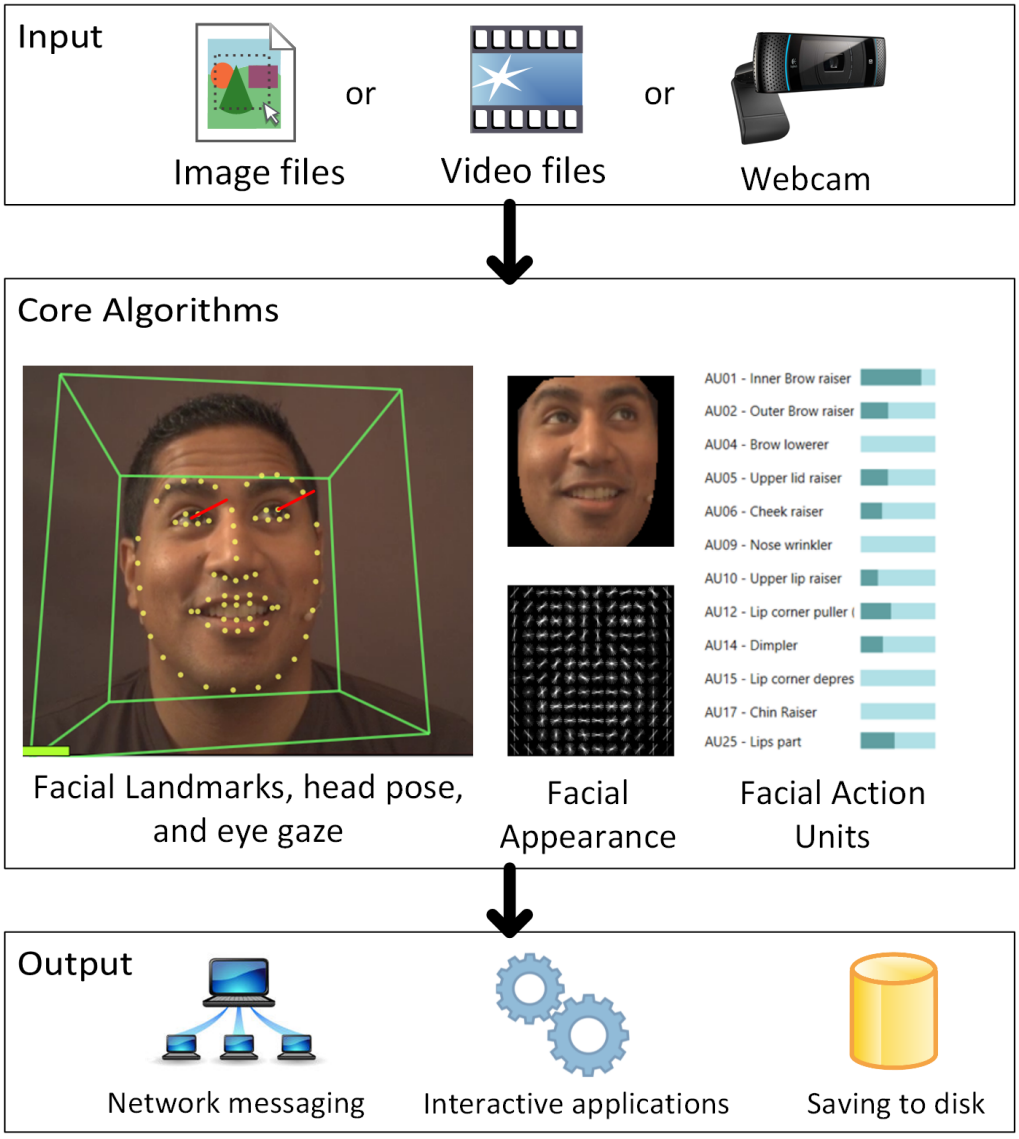


Figure 6: OpenFace is an open source framework that implements state-of-the-art facial behavior analysis algorithms including: facial landmark detection, head pose tracking, eye gaze and facial Action Unit estimation.[10]

2.2 Research on Model Architecture

In the domain of model architecture, the utilization of recurrent neural networks (RNNs) has been instrumental in capturing sequential dependencies within multimodal data streams. Specifically, the Gated Recurrent Unit (GRU), an extension of traditional RNNs, has gained prominence owing to its ability to mitigate the vanishing gradient problem and capture long-range dependencies more effectively. By dynamically updating and forgetting information over time, GRUs excel in modeling temporal sequences across multiple modalities.

2.2.1 RNNs

Recurrent Neural Networks (RNNs) were first introduced by Paul Werbos in his 1988 doctoral thesis. However, the practical application and development of RNNs are largely attributed to subsequent works. In 1990, Elman proposed the Elman network [13], a form of RNN that gained widespread use in tasks such as language modeling. Additionally, the Long Short-Term Memory network (LSTM), proposed by Hochreiter and Schmidhuber in 1997 [14], addressed the difficulty RNNs had with handling long-term dependencies, thus advancing the application of RNNs in sequence modeling.

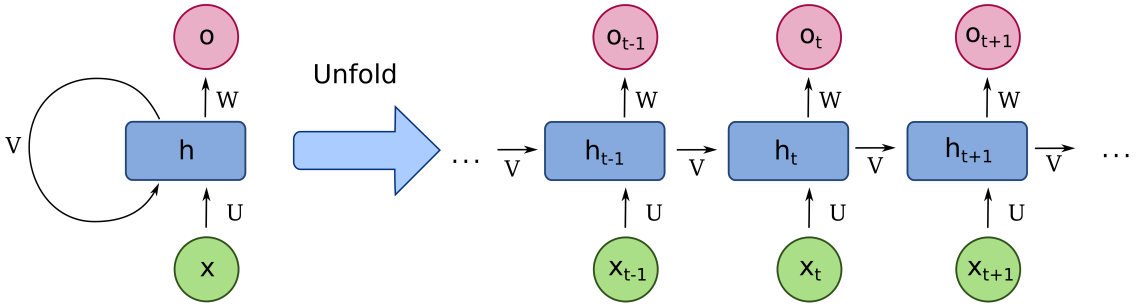


Figure 7: Recurrent neural network unfold
https://commons.wikimedia.org/wiki/File:Recurrent_neural_network_unfold.svg

A significant development in the field was the RNN Encoder-Decoder model introduced by [15]. This model comprises two RNNs: one RNN encodes a sequence of symbols into a fixed-length vector representation, and the other RNN decodes this representation into another sequence of symbols. The encoder and decoder are jointly trained to maximize the conditional probability of the target sequence given the source sequence.

Furthermore, [16] introduced the multimodal Recurrent Neural Network (m-RNN) model for image captioning. This model directly models the probability distribution of generating words, taking into account the previous words and an image to generate the caption. The m-RNN consists of two subnetworks: a deep recurrent neural network for sentences and a deep convolutional neural network

for images. These two subnetworks interact within a multimodal layer, forming the complete m-RNN model.

2.2.2 GRUs

Later, in [17], Cho et al. further developed this new RNN unit by introducing a model called the Gated Recurrent Convolutional Neural Network (grConv), designed to handle variable-length sequences. This model combines features of both Recurrent Neural Networks and Convolutional Neural Networks, using a recursive structure to process input sequences incrementally and employing a gating mechanism to learn the structure of source sentences. Although GRUs are not directly mentioned, the gating mechanisms described in the paper are similar to the reset gate and update gate used in GRUs.

- **Reset Gate** The reset gate is computed based on the previous hidden state and the current input:
$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t] + b_r)$$
- **Update Gate** The update gate is computed similarly but with different weights:
$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t] + b_z)$$
- **Candidate Hidden State** The candidate hidden state (new memory content) is computed as follows:
$$\tilde{h}_t = \tanh(W_h \cdot [r_t \odot h_{t-1}, x_t] + b_h)$$
- **Final Hidden State** The final hidden state at the current time step combines the current candidate hidden state and the previous hidden state:
$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t$$
- **Output** The output is simply the hidden state:
$$y_t = h_t$$

[18] evaluated three variants of Gated Recurrent Units (GRUs) in Recurrent Neural Networks (RNNs), aiming to reduce computational costs by simplifying the parameters of the update and reset gates. These three variants are named GRU1 (each gate uses only the previous hidden state and bias for computation), GRU2 (each gate uses only the previous hidden state for computation), and GRU3 (each gate uses only the bias for computation).

2.3 Research on Attention Mechanisms

Moreover, attention mechanisms have emerged as a crucial component in multimodal fusion frameworks, enabling the model to selectively attend to relevant modalities while disregarding irrelevant information. Traditional attention mechanisms, such as additive and multiplicative attention, facilitate the weighting of input features based on their importance. Furthermore, self-attention mechanisms, epitomized by the Transformer architecture, offer a powerful mechanism for capturing global dependencies and modeling interdependencies across different modalities without relying on sequential processing.

2.3.1 Attention

In 2014, Bahdanau et al. introduced an attention mechanism designed to model the alignment between source and target languages in neural machine translation tasks [19]. This paper first introduced the concept of attention mechanisms, specifically the Bahdanau attention mechanism, which enables the neural machine translation system to dynamically focus on different parts of the source sentence while translating each word in the target language. The model adjusts attention weights based on the information from various positions in the source sentence. The success of this paper has led to the widespread application of attention mechanisms in the field of natural language processing. Beyond neural machine translation, attention mechanisms have been applied to numerous tasks, including language modeling, text summarization, and question-answering systems, achieving remarkable results.

The attention mechanism proposed by Bahdanau et al. [19] has become the foundation for subsequent research, inspiring many improvements and extensions. The attention model variants used in various application domains have evolved rapidly. Generally, the implementation of the attention mechanism can be divided into two steps: first, computing the attention distribution over the input information, and second, computing the context vector according to this attention distribution. Figure 4 illustrates the unified attention model [20], which encompasses the core components shared by most attention models discussed in the literature review.

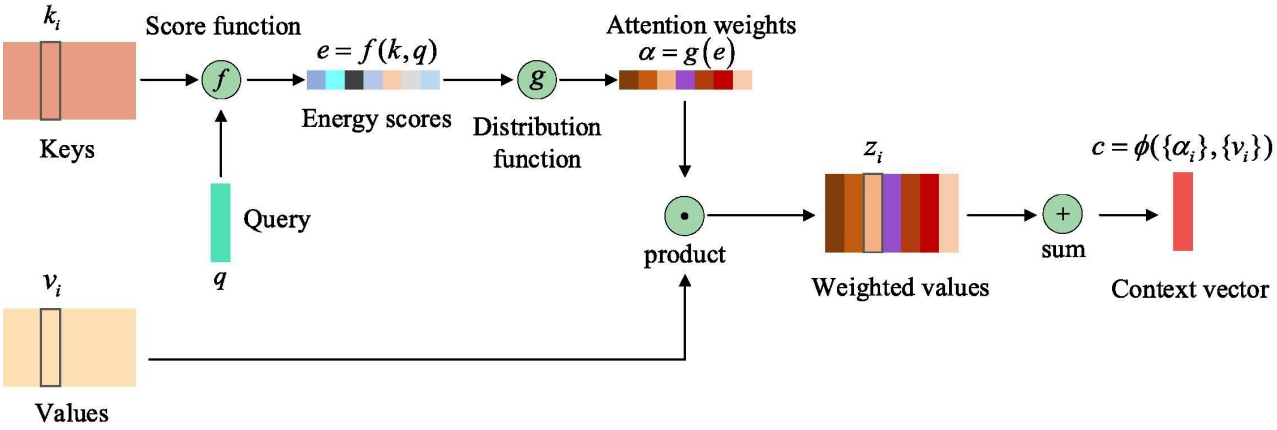


Figure 8: The architecture of the unified attention model.[20]

2.3.2 Self-attention

In 2017, Vaswani et al. introduced the self-attention mechanism [21] , which was used to construct the Transformer model. This model is entirely based on self-attention mechanisms, discarding traditional recurrent neural networks (RNNs) and convolutional neural networks (CNNs). The paper brought self-attention mechanisms into the natural language processing (NLP) field, achieving significant success in tasks such as machine translation.

In this model, the encoder maps an input sequence of symbol representations (x1, ..., xn) to a sequence of continuous representations z = (z1, ..., zn). Given z, the decoder then generates an output sequence (y1, ..., ym) of symbols one element at a time. At each step, the model is auto-regressive [22] , consuming the previously generated symbols as additional input when generating the next. The Transformer adheres to this overall architecture, utilizing stacked self-attention and point-wise, fully connected layers for both the encoder and decoder, as shown in the left and right halves of Figure 1, respectively. [21]

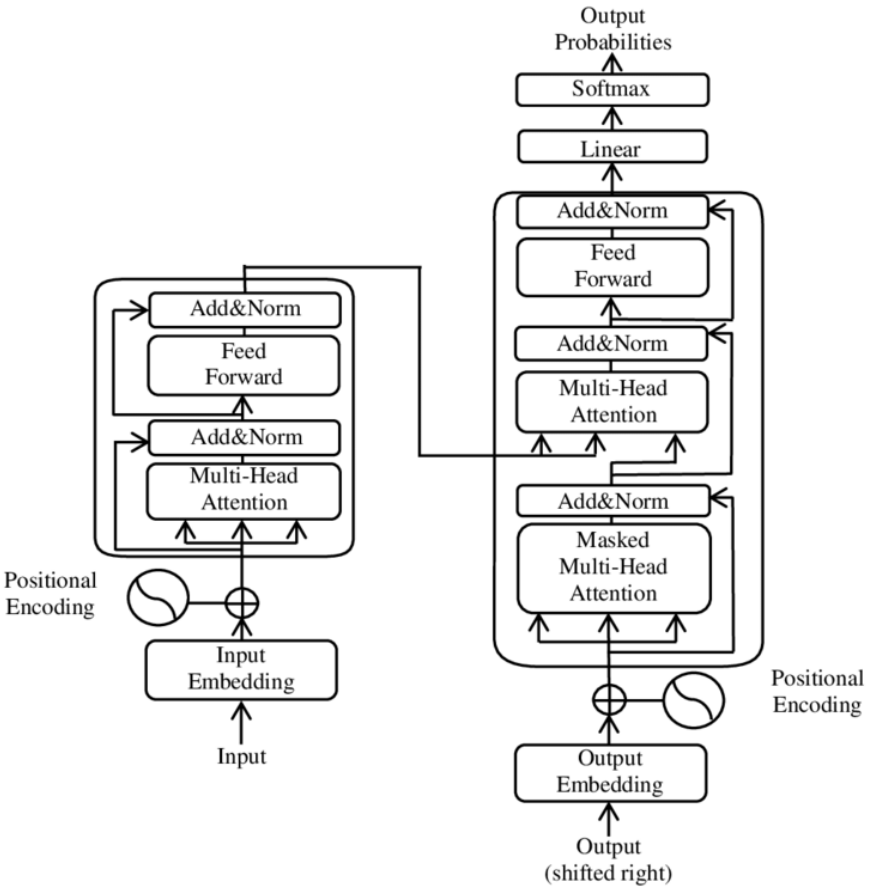


Figure 9: The Transformer - model architecture.[21]

Based on attention and self-attention mechanisms, numerous variants have been developed. [23] introduces an alternative approach to attention mechanisms, aiming to effectively consider the relative

positions or distances between sequence elements. Unlike the Transformer model proposed by [21], this method does not require explicit modeling of relative or absolute position information within the structure but instead incorporates absolute position representations in the input. The authors demonstrated the effectiveness of this approach on the WMT 2014 English-to-German and English-to-French translation tasks, achieving improvements of 1.3 BLEU and 0.3 BLEU respectively with relative position representations compared to absolute position representations. They also noted that combining relative and absolute position representations did not further enhance translation quality. Finally, the authors described an efficient implementation method, presenting it as an instance of a relation-aware self-attention mechanism that can be generalized to inputs with arbitrary graphical representations.

[24] proposed a model called the Self-Attention Generative Adversarial Network (SAGAN) for image generation tasks. Traditional convolutional generative adversarial networks (GANs) only consider local spatial positions in low-resolution feature maps when generating high-resolution details. In contrast, SAGAN leverages clues from all feature positions to generate detailed imagery.

[25] studied two forms of self-attention variants: pairwise self-attention, which extends standard dot-product attention to the image domain as a set operation, and patch-based self-attention, which is more powerful than convolution. The research found that self-attention networks can achieve better performance compared to traditional convolutional networks, and in some cases, patch-based self-attention models significantly outperform convolutional baselines. Additionally, experiments on the robustness of the learned representations suggested that self-attention networks may have significant advantages in terms of robustness and generalization performance.

[26] also showed that multi-head self-attention layers with a sufficient number of heads are at least as expressive as any convolution layer and can completely replace convolution, achieving state-of-the-art performance in vision tasks. While convolution operations extended to graphs can improve performance and are widely used, applying downsampling to graphs remains challenging. [27] proposed a graph pooling method based on self-attention. By using graph convolutional self-attention, this pooling method can simultaneously consider node features and graph topology, achieving superior graph classification performance on benchmark datasets with a reasonable number of parameters.

Furthermore, [28]introduced a novel attention mechanism called external attention for image tasks. Unlike self-attention, external attention leverages two external small learnable shared memories, implemented through two cascaded linear layers and two normalization layers. It can replace self-attention in existing popular architectures. External attention has linear complexity and inherently considers the correlations between all data samples, effectively addressing the quadratic complexity of self-attention and the issue of ignoring correlations between different samples.

2.4 Research on Fusion Strategies

In the pursuit of optimal fusion of multimodal features, Multilayer Perceptrons (MLPs) serve as a versatile tool for integrating information from disparate sources. Leveraging nonlinear activation

functions such as Rectified Linear Units (ReLU) and Leaky Rectified Linear Units (LeakyReLU), MLPs facilitate the nonlinear transformation of fused features, enhancing the model's capacity to capture complex relationships. Additionally, techniques like dropout regularization mitigate overfitting by randomly dropping units during training, thus promoting robustness and generalization in the fused representation.

2.4.1 MLP

In "Perceptrons: An Introduction to Computational Geometry" [29], Minsky and Papert introduced the Multilayer Perceptron (MLP), a neural network structure with an input layer, one or more hidden layers, and an output layer. Each neuron in an MLP performs a weighted sum of inputs from the previous layer and applies an activation function, allowing the network to learn complex nonlinear relationships and decision boundaries. MLPs use nonlinear activation functions, like sigmoid or ReLU, to handle data with high nonlinearity, such as images and text. Training via the backpropagation algorithm optimizes network parameters by minimizing the loss function, thereby enhancing predictive accuracy for classification or regression tasks.

[30] investigated the impact of various activation functions on the performance of Multilayer Perceptron (MLP) neural networks. The study evaluated unipolar sigmoid, bipolar sigmoid, hyperbolic tangent, conic section, and radial basis functions using the backpropagation algorithm. Results showed that the type of activation function significantly influences network performance, with the hyperbolic tangent function demonstrating superior learning and generalization capabilities. The authors emphasized the importance of selecting an appropriate activation function for different problem domains to enhance the performance of MLP networks.

- Input to Hidden Layer:
$$\mathbf{h} = f(\mathbf{W}_{xh}\mathbf{x} + \mathbf{b}_h)$$
- Hidden Layer to Output:
$$\mathbf{y} = g(\mathbf{W}_{hy}\mathbf{h} + \mathbf{b}_y)$$
- Multi-Layer Structure:
  - First Hidden Layer:
$$\mathbf{h}_1 = f_1(\mathbf{W}_{x1}\mathbf{x} + \mathbf{b}_1)$$
  - Hidden Layers ($1 < l \leq L$):
$$\mathbf{h}_l = f_l(\mathbf{W}_{(l-1)l}\mathbf{h}_{l-1} + \mathbf{b}_l)$$
  - Output Layer:
$$\mathbf{y} = g(\mathbf{W}_{hL}\mathbf{h}_L + \mathbf{b}_y)$$

2.4.2 Rectified Linear Unit (RELU)

This article [31] proposed an innovative approach using Rectified Linear Units (ReLU) as the classification function in deep neural networks, replacing the traditional Softmax function. The authors conducted comparative experiments on the MNIST, Fashion-MNIST, and WDBC datasets. The results showed that deep learning models using ReLU as the classification function (DL-ReLU) performed comparably to models using the Softmax function (DL-Softmax), achieving state-of-the-art performance. This confirmed the effectiveness of this approach. The study provides new insights and feasibility validation for the design of deep learning models.

2.4.3 LeakyReLU

This paper [32] investigated the use of rectifier non-linear activation functions in neural network models for large-scale speech recognition tasks. The results showed that deep neural networks using rectifier non-linearities reduced the word error rate by 2% on the Switchboard dataset compared to traditional sigmoid non-linear models. The study also analyzed the differences in hidden layer representation encoding between the two activation functions, finding that rectifier non-linearities could better learn hidden layer feature representations. Additionally, the authors evaluated variants of the leaky rectifier non-linearity and discovered further performance improvements in deep neural network models. Overall, the study confirmed that employing rectifier non-linearities in large-scale speech recognition tasks can significantly enhance the performance of deep neural networks.

The LeakyReLU activation function is defined as:

$$f(x) = \begin{cases} x & \text{if } x \geq 0 \\ \alpha x & \text{if } x < 0 \end{cases}$$

The literature review lays the groundwork for our research in multimodal emotion recognition, focusing on feature extraction, model architecture, attention mechanisms, self-attention, and MLP. This succinct overview provides essential insights and guidance for our research.

3 Methodology

In this section, I will outline the methodology used to address the research question and validate the hypothesis at a high level. First, in subsection 3.1, I will discuss the datasets utilized for training and testing the models. Next, subsection 3.2 will focus on the various models employed in the study. The innovations in my thesis will be mentioned here in 3.3 Following that, subsection 3.4 will elaborate on the evaluation methods and metrics employed, specifically the word error rate. Finally, in subsection 3.5, I will reflect on the ethical considerations inherent in this research.

3.1 Datasets

3.1.1 Dataset Introduction

- **CMU-MOSI Dataset**
  The Multimodal Corpus of Sentiment Intensity (CMU-MOSI) dataset consists of 2,199 English opinion video clips, each annotated for sentiment on a scale from -3 to 3. This dataset includes detailed annotations for subjectivity, sentiment intensity, visual features per frame, and audio features per millisecond, providing comprehensive multimodal data for sentiment analysis research.
  [33]

- **CMU-MOSEI Dataset**
  The CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) dataset is the largest dataset for multimodal sentiment analysis and emotion recognition. It includes over 23,500 English sentence utterance videos from more than 1,000 speakers on YouTube. The dataset is gender-balanced and features randomly selected utterances from various topics and monologue videos. Additionally, the videos are transcribed with proper punctuation.
  [34]

- **CH-SIMS Dataset**
  The CH-SIMS dataset is a Chinese dataset designed for single- and multimodal sentiment analysis, featuring 2,281 refined video segments captured in natural settings. It includes both multimodal and unimodal annotations, allowing researchers to explore modality interactions or focus on unimodal sentiment analysis using the independent annotations provided.
  [35]

3.2 Models

In my quest to develop robust models for multimodal emotion recognition, I present a comprehensive suite of architectures, including GRUWithLinear, MLP, SelfAttention, and EMIFusion. These models utilize cutting-edge methodologies to efficiently capture sequential dependencies, nonlinear patterns, and cross-modal interactions in multimodal datasets. The proposed model integrates text,

acoustic, and visual modalities to enhance emotion recognition accuracy. As illustrated in Figure 10, the workflow involves several key stages: input feature extraction, temporal alignment, feature extraction using GRU and self-attention mechanisms, feature fusion, and prediction. Below is a detailed explanation of each step and the associated components.

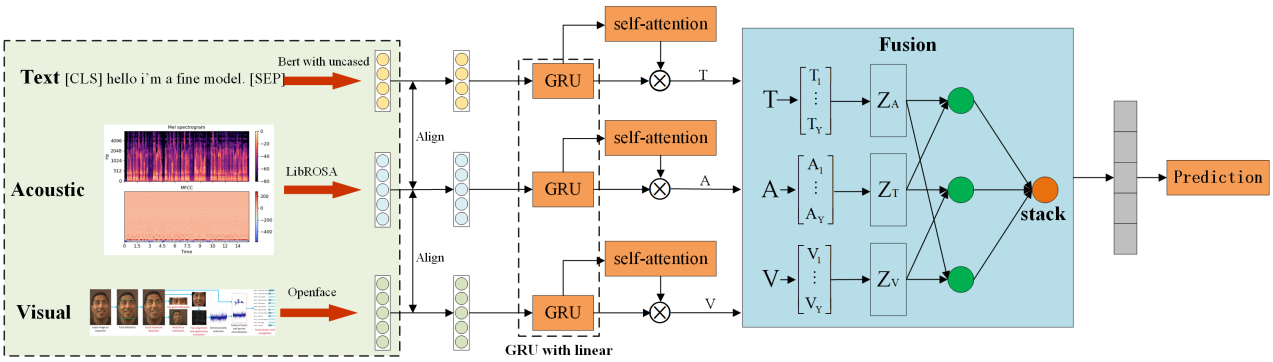


Figure 10: The OveraLL Model Flow Chart

3.2.1 Input Layer

- **Text Modality: BERT Encoding**
  The textual data is encoded using a pre-trained BERT (Bidirectional Encoder Representations from Transformers) model. The input "[CLS] hello I'm a fine model. [SEP]" is tokenized and passed through the BERT model to extract contextual embeddings.

- **Acoustic Modality: MFCC**
  Acoustic features are extracted using the LibROSA library. This includes Mel-spectrogram and Mel Frequency Cepstral Coefficients (MFCC), which provide a comprehensive representation of the audio signals.

- **Visual Modality: Facial landmark detection and tracking**
  Visual features are extracted using the OpenFace tool, which detects and analyzes facial expressions and Action Units (AUs) to capture nuanced facial movements and emotions.

3.2.2 Feature Alignment

To ensure temporal consistency across modalities, the extracted features are aligned. This alignment process synchronizes the text, audio, and visual features over the same time axis, facilitating coherent multimodal fusion.

3.2.3 GRU and Self-Attention Modules

- **Initial GRU Processing**: Each modality's features are independently processed through a GRU (Gated Recurrent Unit) network. This step captures the sequential dependencies within each modality.
- **Self-Attention Mechanism**: The output from the GRU is further refined using a self-attention mechanism. Self-attention allows the model to weigh the importance of different time steps, capturing long-range dependencies and contextual importance.
- **Second GRU Processing**: The self-attention outputs are then passed through another GRU layer for further sequential modeling, enhancing the feature representations with a more detailed temporal context.

3.2.4 Fusion Module

Stacking and Fusion: The features from the text (T), acoustic (A), and visual (V) modalities are stacked and passed into the fusion module. In this module, each modality's refined features (Z_T, Z_A, Z_V) are combined to form a comprehensive multimodal representation. This fusion leverages the complementary information from all three modalities.

3.2.5 Prediction Layer

The fused feature vector is then fed into the final prediction layer, which outputs the recognized emotion. This layer can be implemented using a fully connected neural network, providing the final classification based on the aggregated multimodal features.

3.3 Innovations

- **Temporal Alignment of Multimodal Features**
  The process ensures that features from text, audio, and visual modalities are aligned along the same timeline, allowing the model to consider synchronous information from all modalities and capture intricate relationships between them. This alignment enhances the model's ability to utilize complementary and reinforcing information from different modalities, thereby improving recognition accuracy and robustness.
- **Combination of Self-Attention and GRU**
  The model employs a self-attention mechanism between two GRU layers, capturing long-range dependencies within the sequential data and allowing the model to focus on crucial time steps. This combination enhances the model's understanding of both local and global contexts, resulting in more accurate feature extraction and improved emotion recognition performance.

- **Stacking in the Fusion Module**
  In the fusion module, features from various modalities are stacked and integrated. This method efficiently amalgamates the diverse information from text, acoustic, and visual inputs into a unified representation. By capitalizing on the strengths of each modality, the fused representation offers a more robust and comprehensive foundation for emotion recognition, thereby enhancing the model's generalization ability and overall performance.

This multimodal emotion recognition model leverages advanced techniques such as temporal alignment, GRU networks, self-attention mechanisms, and feature fusion. These innovations collectively enhance the model's ability to accurately recognize emotions by integrating and synthesizing information from multiple modalities. The methodological rigor and innovative approaches ensure that the model is both effective and robust in practical applications, paving the way for more sophisticated emotion recognition systems.

3.4 Evaluation - Word Error Rate

The evaluation of our multimodal emotion recognition model involves a detailed analysis using three key scripts: Complexity.py, Performance.py, and Robustness.py.
(More details: https://github.com/JingwenShi123/Thesis)

3.4.1 Model System

- Complexity.py evaluates the model's performance and memory usage during training and inference. It calculates total parameters, training time, peak memory usage, and inference time, helping to optimize performance and resource management.
- Performance.py assesses the model's effectiveness through metrics such as F1 score, accuracy, and Area Under the Precision-Recall Curve (AUPRC). It also handles data preprocessing and organizes prediction data for a comprehensive evaluation of classification performance.
- Robustness.py measures the model's robustness to noise using relative and effective robustness metrics. It normalizes robustness metrics for comparison and visualizes performance trends across different noise levels, providing insights into the model's resilience and real-world applicability.

These scripts collectively provide a systematic framework for evaluating the model's performance, complexity, and robustness.

3.4.2 Output Metric

In the evaluation of our multimodal emotion recognition system, several key metrics were employed to rigorously assess the performance across different datasets, namely MOSI, MOSEI, and SIMS.

Each dataset comprises multiple modalities (text, audio, and visual data) and emotions categorized into seven levels. The primary evaluation metrics include average accuracy for both seven-class and binary classifications, F1 score, Mean Squared Error (MSE), training loss, validation loss, and confusion matrices. The detailed methodology for these metrics is outlined below:

- **Seven-class Classification Accuracy**: For each dataset, the system's ability to correctly classify the emotions into seven discrete levels (ranging from -3 to 3) is calculated. The average accuracy across these levels provides insight into the system's granularity in emotion detection.
- **Binary Classification Accuracy**: Emotions are also categorized into positive and negative sentiments for a broader evaluation. This binary classification helps in understanding the system's performance in a simplified yet critical aspect of emotion recognition.
- **F1 Score**: The F1 score, which is the harmonic mean of precision and recall, is used to balance the trade-off between these two metrics. It is particularly useful in cases where the dataset might be imbalanced. We calculate the F1 score for both the seven-class and binary classifications to ensure comprehensive evaluation.
- **Mean Squared Error (MSE)**:MSE is calculated to measure the average squared difference between the predicted emotion levels and the actual labels. This metric provides a sense of the prediction's accuracy and helps in identifying how well the model can approximate the true emotion levels.
- **Training and Validation Loss**: Throughout the training process, both the training loss and validation loss are tracked to monitor the model's learning progress and generalization capability. A consistent decrease in these losses indicates effective learning, while a divergence between them can highlight overfitting or underfitting issues.
- **Confusion Matrix**: Confusion matrices are constructed for both seven-class and binary classifications. These matrices provide a detailed breakdown of the model's performance by showing the true positives, false positives, true negatives, and false negatives. This visual tool is crucial for identifying specific misclassification patterns and areas needing improvement.

3.5 Ethical considerations

While this research aims to advance the field of multimodal emotion recognition, there is a possibility that the technology may have unforeseen consequences. To mitigate these risks, the research team will communicate the study's results and implications in an accessible and transparent manner.

The data used in this research comes from three publicly available datasets: CMU-MOSI, CMU-MOSEI, and CH-SIMS. These datasets are freely accessible through the following links:

- CMU-MOSI: `http://multicomp.cs.cmu.edu/resources/cmu-mosi-dataset/`

- CMU-MOSEI: http://multicomp.cs.cmu.edu/resources/cmu-mosei-dataset/
- CH-SIMS: https://aclanthology.org/2020.acl-main.343/

According to the information provided on these websites, the datasets comply with GDPR regulations and are publicly available for academic use. The participants in these datasets have consented to their data being collected and used for research purposes.

No new data from human participants was collected for this study, and there were no surveys or recordings of human voices involved. Consequently, there are no ethical concerns related to human subject research.

The models used in this research are built upon pre-existing models available on GitHub. The repository can be accessed here: https://github.com/Justin1904/Low-rank-Multimodal-Fusion. The code is openly available and can be used to replicate the experiments.

It is important to note that the datasets may contain inherent biases due to the unknown characteristics of the speakers. These biases are acknowledged and transparently disclosed. Objective metrics were used for evaluation, which are standard in the field of emotion recognition. Subjective evaluation methods involving human participants were not utilized, thereby aligning with ethical standards and avoiding potential ethical issues.

4 Experimental Setup

In this section, we outline the detailed experimental setup used to validate the proposed methodologies. This includes a comprehensive description of the datasets, model architectures, training procedures, and evaluation metrics. The experimental setup is designed to ensure consistency in model training and evaluation, facilitating a fair comparison with baseline methods.

(For more details, visit my GitHub repository: https://github.com/JingwenShi123/Thesis)

4.1 Datasets

The datasets used in this study are essential for training and evaluating the multimodal emotion recognition models. We employ the PKL file format for efficient data handling, which includes video features extracted using pre-trained CNN models, audio features processed with the LibROSA library, and text features derived from BERT embeddings. These datasets, such as CMU-MOSI, CMU-MOSEI, and CH-SIMS, provide rich, multimodal data that are crucial for robust emotion recognition tasks.

4.1.1 PKL File Format

PKL (Pickle) files are a Python-specific serialization format used to store objects in a binary representation. These PKL files serve as the primary data source for training and evaluating multimodal emotion recognition models, as they provide a structured and easily accessible representation of the various modalities and their associated labels. For multimodal emotion recognition datasets like CMU-MOSI, CMU-MOSEI, and CH-SIMS, the PKL files typically contain the following:

- **Video Features**: The video features are extracted from the video frames using pre-trained convolutional neural network (CNN) models, such as OpenFace. These features capture various aspects of facial expressions, head poses, and other visual cues that are relevant for emotion recognition.postures.
- **Audio Features**: The audio features are extracted using the LibROSA library, which provides Mel-frequency cepstral coefficients (MFCCs). MFCCs encode important characteristics of the audio signal, including pitch, energy, and frequency content, which are crucial for recognizing emotional states from the speech modality.
- **Text Features**: The text features are derived from the speech transcripts using the BERT language model, which generates contextual word embeddings. These text embeddings capture the semantic and syntactic information present in the spoken language, which can contribute to the recognition of emotional states.

- **Labels**: The datasets provide labeled data for the emotion recognition task, where the labels reflect the intensity and type of emotions expressed in the multimodal data. For the CMU-MOSI and CMU-MOSEI datasets, the emotion labels typically capture the polarity (positive, negative, or neutral) and intensity of the sentiment expressed. In the case of the CH-SIMS dataset, the emotion labels are more detailed, covering a wider range of emotion categories and their corresponding intensities.

The use of PKL (Pickle) files for training multimodal emotion recognition models offers several key advantages. Firstly, PKL files enable seamless integration of video, audio, and text features, facilitating easy access and processing of the multimodal data within a unified format. The binary format of PKL files also allows for efficient serialization and deserialization, saving storage space and reducing data loading times, which enhances the overall data processing efficiency.

Furthermore, the rapid loading and batch processing capabilities of PKL files improve the efficiency of model training. The binary format ensures quick data loading, significantly reducing the time required for training the emotion recognition models. Additionally, the support for batch loading optimizes memory usage and prevents issues related to loading large datasets at once, crucial for training complex multimodal models.

Importantly, PKL files preserve the original state of the data, ensuring consistency and completeness during the training process, and enhancing the reproducibility and verifiability of the experiments. This consistency and reproducibility are crucial for scientific research and model development.

The PKL file format has been widely adopted for specific multimodal emotion recognition datasets, such as CMU-MOSI, CMU-MOSEI, and CH-SIMS. These datasets offer various advantages, including detailed emotion labels, large-scale diversity, and support for the Chinese language, enabling the development of sophisticated and reliable emotion analysis systems.

By leveraging the advantages of PKL files, researchers and practitioners can efficiently integrate, process, and train multimodal emotion recognition models, ultimately leading to advancements in the field of emotion analysis.

4.1.2 Process setup

The provided code in getdata.py constitutes a comprehensive data processing module for multimodal emotion recognition models, offering a systematic approach to data loading, preprocessing, and augmentation.

- **Noise Functions for Data Augmentation**
  This module encompasses a suite of functionalities tailored to the AFFECT dataset, a cornerstone in multimodal emotion recognition tasks. Central to its design are noise functions meticulously crafted for both text and time-series data, fostering model robustness and generalization. For text data, the addtextnoise function injects various noise types, including letter swapping and typos, while addtimeseriesnoise introduces Gaussian noise and dropout mechanisms to simulate real-world data variances.

- **Normalization and Preprocessing**
  Complementing the noise functions are normalization and preprocessing methods. normalizeText ensures textual uniformity by standardizing case and removing extraneous characters, while znorm standardizes data across modalities, promoting consistency in feature scaling.

- **Text Processing and Embedding**
  Text processing capabilities are augmented through functions like getrawtext, facilitating the extraction of raw text data, and getword2id coupled with getwordembeddings, which establish word-to-ID mappings and retrieve corresponding GloVe embeddings.

- **Affectdataset: Custom Dataset Handling**
  At the heart of data management lies the Affectdataset class, a specialized PyTorch Dataset implementation. This class accommodates various dataset configurations, including modality alignment, normalization preferences, and task types, while supporting essential preprocessing tasks such as padding and flattening of time-series data.

- **DataLoader Function: Data Preparation**
  Finally, the getdataloader function serves as the linchpin in data pipeline orchestration. It orchestrates dataset loading, preprocessing, and augmentation, ensuring data readiness for model training, validation, and testing. By seamlessly interfacing with PyTorch DataLoader, it expedites batch processing and parallel data loading, optimizing model training efficiency.

In essence, getdata.py encapsulates a robust framework for data management in multimodal emotion recognition, underpinned by meticulous preprocessing, noise augmentation, and streamlined dataset handling capabilities.

4.2 Models

In my quest to develop robust models for multimodal emotion recognition, I present a comprehensive suite of architectures tailored to address distinct facets of the input data. These models, including the GRUWithLinear, MLP, SelfAttention, and EMIFusion, encapsulate cutting-edge methodologies to efficiently capture sequential dependencies, nonlinear patterns, and cross-modal interactions inherent in multimodal datasets.

4.2.1 GRUWithLinear Model

The GRUWithLinear model architecture comprises a Gated Recurrent Unit (GRU) followed by a linear layer for post-processing. It employs the nn.GRU module to handle the recurrent layer, enabling the model to process input sequences and capture temporal dependencies effectively. Additionally, a nn.Linear layer is utilized for the linear transformation, facilitating the final transformation of the GRU output to generate the model's output. This design enables the model to conduct sequential

data processing, allowing it to capture intricate temporal patterns inherent in the input data. Overall, the GRUWithLinear model serves the purpose of robustly modeling sequential data, making it well-suited for tasks requiring the analysis of temporal dynamics and dependencies.

4.2.2 MLP (Two-layered Perceptron) Model

The MLP model, also known as a Two-layered Perceptron, consists of two fully connected layers equipped with Rectified Linear Unit (ReLU) activation functions. It employs nn.Linear layers for both the hidden and output layers to facilitate efficient computation. During operation, the first linear layer processes the input features, applying the ReLU activation function to introduce nonlinearity. Subsequently, the second linear layer generates the final output of the model. The primary objective of the MLP model is to provide a straightforward yet powerful architecture capable of capturing intricate patterns present in the data. Through the combination of fully connected layers and nonlinear activation functions, the model excels at learning complex relationships and patterns within the input data, making it well-suited for various machine learning tasks.

4.2.3 SelfAttention Layer

The Self-Attention layer serves as a crucial component implementing the self-attention mechanism to discern significant features within the input sequence. Its architecture incorporates linear transformations for computing query, key, and value vectors, subsequently employing matrix multiplication to compute attention weights. The layer's functionality revolves around computing attention weights based on the similarity between query and key vectors, followed by aggregating the values weighted by these attention scores. By doing so, the self-attention mechanism enables the model to dynamically focus on pertinent segments of the input sequence, thereby enhancing its capacity to capture long-range dependencies effectively.

4.2.4 EMIFusion Model

The EMIFusion model is a fusion architecture designed to integrate information from multiple modalities, such as audio, video, and text, utilizing Low-Rank Tensor Fusion (LRTF). Its implementation involves employing factor matrices dedicated to each modality, alongside fusion weights aimed at amalgamating the representations from these modalities. Functionally, the model operates by computing modality-specific representations using the factor matrices, applying fusion weights to amalgamate these representations, and finally generating the fused output. The primary purpose of the EMIFusion model is to exploit the complementary nature of information across different modalities, thereby enhancing the model's performance in various emotion recognition tasks.

These models are designed to capture diverse aspects of the input data, such as sequential patterns, nonlinear relationships, and cross-modal interactions, thereby improving the model's ability to understand and interpret multimodal emotional cues effectively.

4.3 Supervised Learning

This segment focuses on implementing supervised learning training procedures designed for multimodal emotion recognition. The core architecture is the Multimodal Deep Learning (MMDL) model, which integrates data from diverse modalities such as audio, video, and text. The MMDL model comprises specialized encoders for each modality, a fusion module that combines the encoded representations, and a classification or prediction head to generate the final output. This approach aims to effectively capture the complex interactions and correlations between different modalities, thereby enhancing the overall performance of emotion recognition tasks.

4.3.1 Models Defined

MMDL (Multimodal Deep Learning) Model: This model serves as the core architecture for integrating multimodal information. It consists of encoders for each modality, a fusion module, and a classification or prediction head. The encoders process input data from each modality, the fusion module combines the representations, and the head produces the final output.

4.3.2 Components and Functionality

- **Encoders**: These are individual modules responsible for processing data from each modality. They encode the input data into meaningful representations specific to each modality.
- **Fusion Module**: This module combines the representations obtained from the encoders. It integrates information from different modalities to create a unified representation that captures the multimodal context effectively.
- **Classification/Prediction Head**: This component processes the fused representation to generate the final output, which could be either class labels in classification tasks or continuous predictions in regression tasks.
- **Training Function (train)**: The train function orchestrates the training process. It iterates through the dataset, computes the loss using the specified objective function, and updates the model parameters using backpropagation. Additionally, it handles optimization, early stopping, and model saving based on validation performance.
- **Testing Function (test)**: The test function evaluates the trained model on the test dataset. It computes various evaluation metrics such as accuracy, F1 score, and AUPRC (Area Under the Precision-Recall Curve) to assess the model's performance.
- **Utility Functions**: Several utility functions are provided for tasks such as dealing with objective functions, evaluating model performance, and processing input data.

4.3.3 Relationships and Effects

The MMDL model architecture facilitates the integration of information from multiple modalities, such as audio, video, and text, to enhance the overall performance of emotion recognition tasks. By utilizing encoders tailored to each modality and a fusion module to combine their representations, the model can effectively capture complex patterns and correlations across modalities. The training and testing procedures ensure that the model is optimized for accurate prediction and robust performance on both training and unseen data.

Through the evaluation of metrics such as accuracy, F1 score, and AUPRC, the effectiveness of the model in recognizing and understanding multimodal emotional cues can be assessed comprehensively.

In conclusion, the supervised learning framework outlined here provides a comprehensive method for developing and evaluating multimodal emotion recognition models. By employing modality-specific encoders, a robust fusion module, and a meticulous training and testing process, the MMDL model is capable of capturing nuanced emotional cues across various modalities. Evaluation metrics such as accuracy, F1 score, and AUPRC ensure a thorough assessment of the model's performance, confirming its robustness and effectiveness. This framework not only optimizes the model for accurate predictions but also highlights the potential of multimodal integration in advancing emotion recognition systems.

4.4 Evaluation

The evaluation of a multi-modal emotion recognition model entails a thorough analysis of its performance, complexity, and robustness. Within the realm of evaluation, various scripts, namely "Complexity.py," "Performance.py," and "Robustness.py," play pivotal roles in assessing different facets of the model's functionality.

4.4.1 Complexity

Complexity.py defines functions for evaluating the performance and memory usage of a multimodal emotion recognition model during training and inference.

getallparams(li) calculates the total number of parameters across a list of neural network modules by iterating through each module and summing the elements in each parameter. all_in_one_train(trainprocess, trainmodules) measures and prints the training time, peak memory usage, and total parameters during the training phase. It records start and end times, tracks peak memory usage using memory_usage from memory_profiler, and calculates total parameters using getallparams. Similarly, all_in_one_test(testprocess, testmodules) measures and prints inference time and total parameters during the testing phase by executing the test process and recording elapsed time and parameter count.

These functions provide a comprehensive evaluation of the model. getallparams is used in both training and testing to assess model complexity. The training function evaluates resource requirements and efficiency, while the testing function ensures consistency in model parameters between phases. This aids in performance measurement, resource management, and understanding model complexity, helping developers optimize their models for better performance and efficiency.

4.4.2 Performance

The "Performance.py" script encompasses a suite of functions tailored for evaluating the effectiveness of a multi-modal emotion recognition model. These functions serve a spectrum of purposes, ranging from computing key evaluation metrics such as F1 score, accuracy, and Area Under the Precision-Recall Curve (AUPRC) to managing data pre-processing tasks such as sorting and filtering.

The ptsort Function efficiently sorts a list of tuples based on the first element of each tuple, providing a foundational utility for organizing prediction or ground truth data. In contrast, the AUPRC Function quantifies the Area Under the Precision-Recall Curve, leveraging the sklearn.metrics.average_precision_score to measure the classification performance of binary classifiers, encapsulating the interplay between precision and recall. Meanwhile, the f1_score Function and accuracy Function offer a nuanced evaluation perspective by computing F1 score and accuracy respectively, employing sklearn.metrics.f1_score and sklearn.metrics.accuracy_score. These functions, designed to seamlessly handle PyTorch tensors, furnish a comprehensive understanding of the model's classification prowess across diverse averaging strategies. Lastly, the eval_affect Function stands as the centerpiece, tailored specifically for assessing the emotion recognition model's performance. Beginning with data preprocessing, it seamlessly transitions from PyTorch tensors to numpy arrays, optionally excluding instances with zero labels. Subsequently, it computes the F1 score and accuracy for binary classification, harnessing sklearn.metrics.f1_score and sklearn.metrics.accuracy_score. The amalgamation of these functionalities furnishes a systematic framework for evaluating the model's efficacy in discerning emotions accurately.

These meticulously crafted functions underpin a robust evaluation pipeline, enabling a holistic assessment of the multi-modal emotion recognition model's classification performance across varied evaluation criteria.

4.4.3 Robustness

The "Robustness.py" script encompasses a collection of functions dedicated to evaluating the robustness metrics of a multi-modal emotion recognition model. These functions are instrumental in computing both relative and effective robustness metrics, which offer insights into the model's performance across different noise levels.

Relative Robustness Metrics computes the relative robustness metric, leveraging the relative_robustness_helper function. This metric is calculated as the area under the performance curve, providing a compre-

hensive assessment of the model's robustness across varying noise levels. Conversely, the effective_robustness function calculates the effective robustness metric using the effective_robustness_helper function. This metric quantifies the performance difference compared to a baseline method (i.e., late fusion), offering insights into the model's efficacy in handling noise perturbations. Both relative_robustness_helper and effective_robustness_helper serve as auxiliary functions for computing robustness metrics, catering to specific requirements based on the chosen metric type. Additionally, the maxmin_normalize function normalizes the robustness metrics for comparison across different methods, ensuring a fair assessment across diverse evaluation criteria. The single_plot function facilitates the visualization of performance versus robustness plots for individual methods. It enables a graphical representation of the model's performance trends across varying noise levels, aiding in the interpretation of robustness metrics.

Through these meticulously designed functions, the "Robustness.py" script provides a comprehensive framework for evaluating the multi-modal emotion recognition model's robustness. By computing diverse robustness metrics and offering visualization capabilities, it enables researchers to gain valuable insights into the model's resilience to noise perturbations, thereby enhancing the understanding of its real-world applicability and performance variability.

In essence, through meticulous design and implementation, these evaluation scripts collectively offer a systematic framework for comprehensively evaluating the multi-modal emotion recognition model's performance, complexity, and robustness.

5 Results

This section presents the outcomes of the experiments conducted using the CMU-MOSI, CMU-MOSEI, and CH-SIMS datasets to evaluate the performance of the proposed multimodal emotion recognition models. Detailed performance metrics, including training and validation losses, confusion matrices, and various classification accuracies, are discussed.

5.1 MOSI, MOSEI, and SIMS Datasets

- **MOSI Dataset**

Metric	Value
MSE	0.6701
Acc7	0.4942
Corr	0.8073
MAE	0.6701
F1	0.8460, 0.8352
Acc2	0.87345, 0.8688
Inference Time	0.1685
Inference Params	1486465

Table 1: Performance Metrics of Multimodal Emotion Recognition Model on CMU-MOSI Dataset

- **MOSEI Dataset**

Metric	Value
MSE	0.2694
Acc7	0.5654
Corr	0.9263
MAE	0.2694
F1	0.9547, 0.8526
Acc2	0.9428, 0.8374
Inference Time	1.5182
Inference Params	1486465

Table 2: Performance Metrics of Multimodal Emotion Recognition Model on CMU-MOSEI Dataset

- **SIMS Dataset**

Metric	Value
MSE	0.2779
Acc7	0.7090
Corr	0.8168
MAE	0.2780
F1	0.8571, 0.7668
Acc2	0.8969, 0.8406
Inference Time	1.0559
Inference Params	2336449

Table 3: Performance Metrics of Multimodal Emotion Recognition Model on CMU-MOSEI Dataset

5.2 Training and Validation Losses

The training and validation loss curves for the MOSI, MOSEI, and SIMS datasets are illustrated in Figures 11, 12, and 13, respectively.

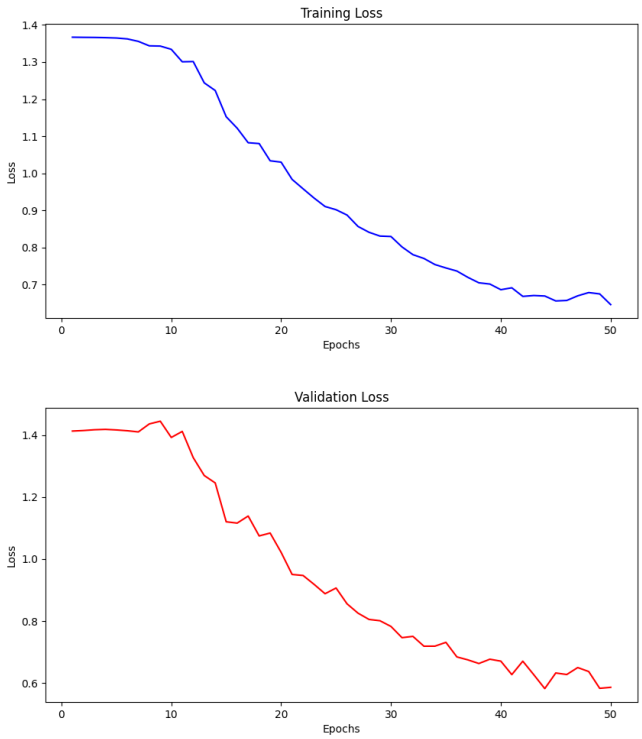


Figure 11: Training and validation loss for MOSI dataset

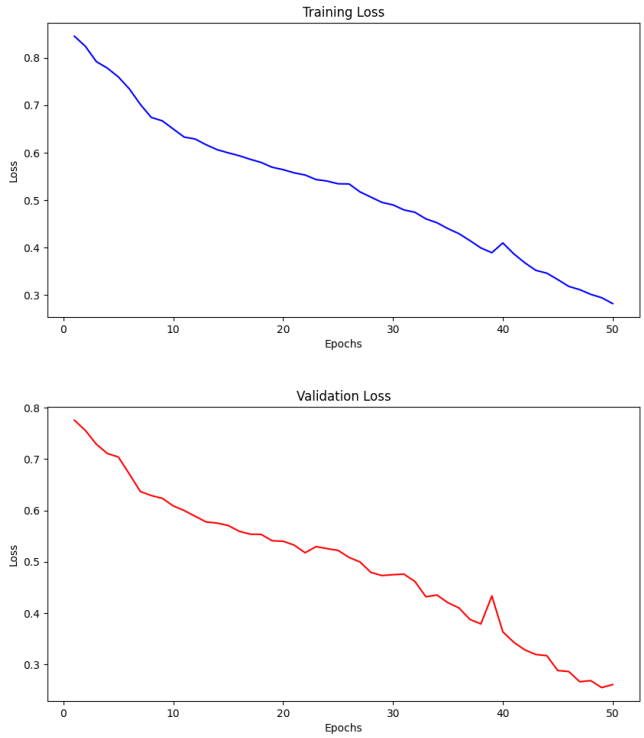


Figure 12: Training and validation loss for MOSEI dataset

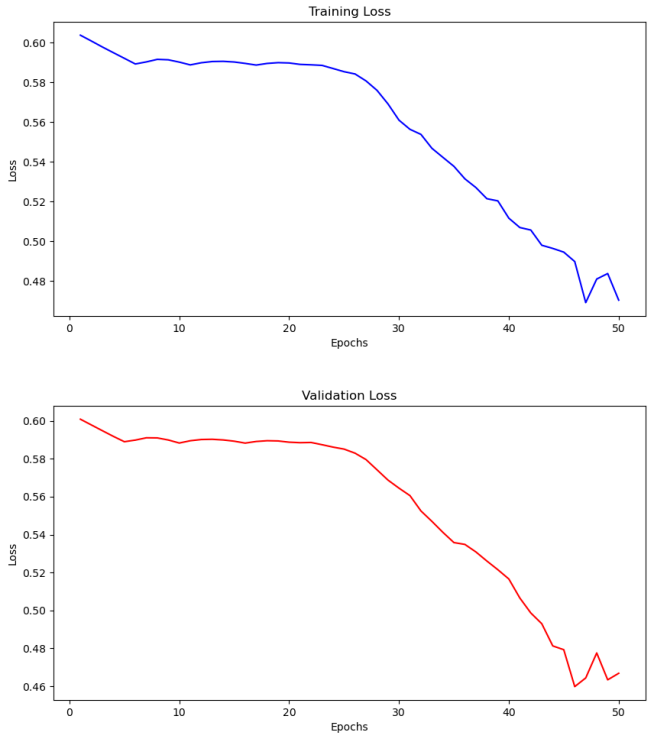


Figure 13: Training and validation loss for SIMS dataset

5.3 Confusion Matrices

The confusion matrices for seven-class and two-class classifications provide deeper insights into the model's predictive accuracy.

5.3.1 Seven-Class Confusion Matrices

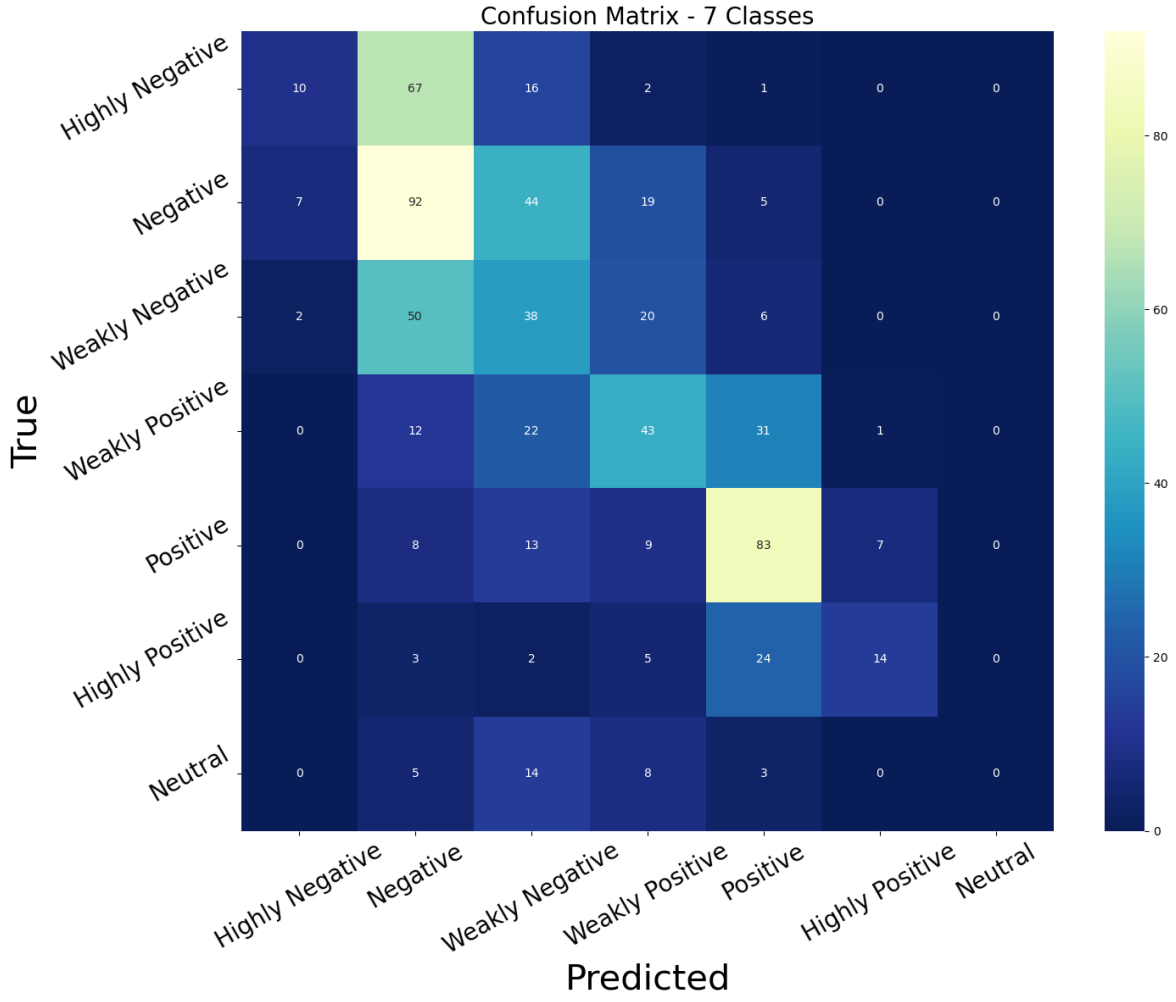


Figure 14: Seven-class confusion matrix for MOSI dataset

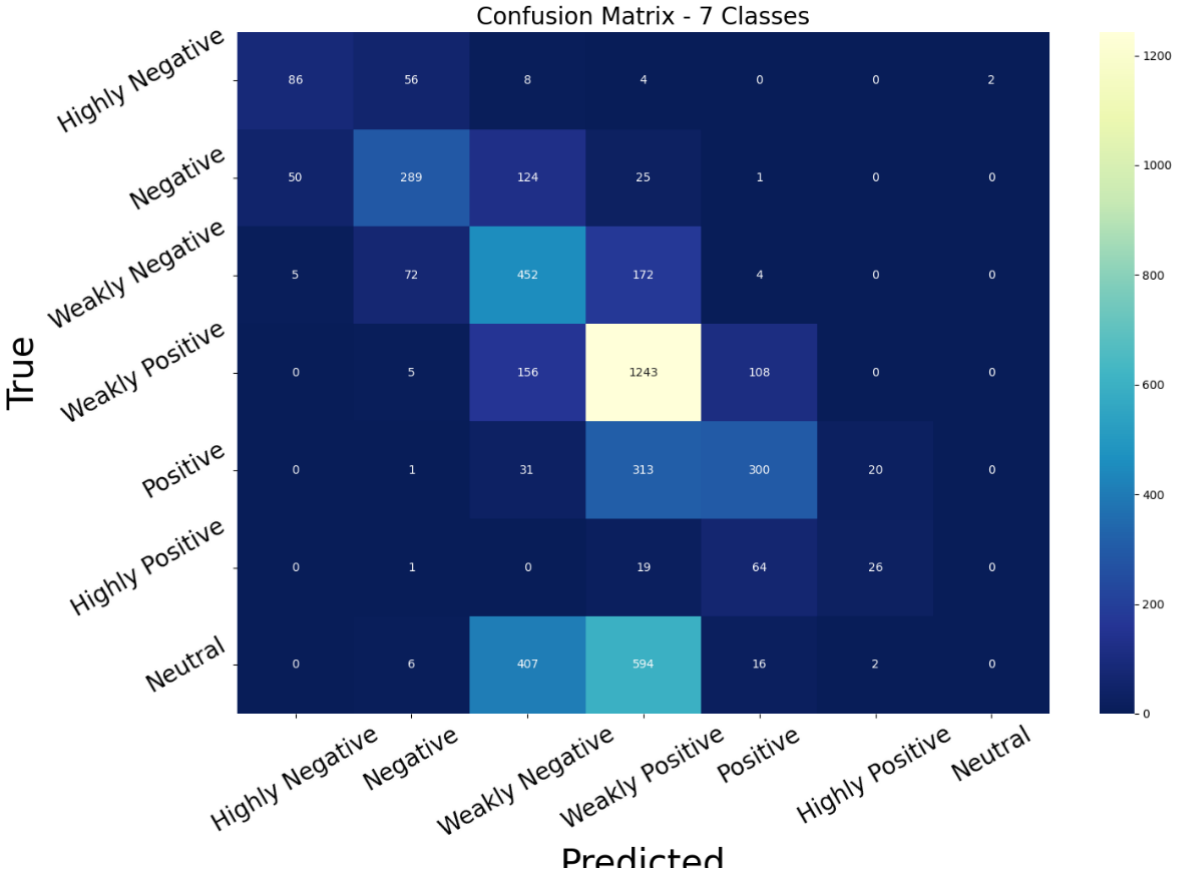
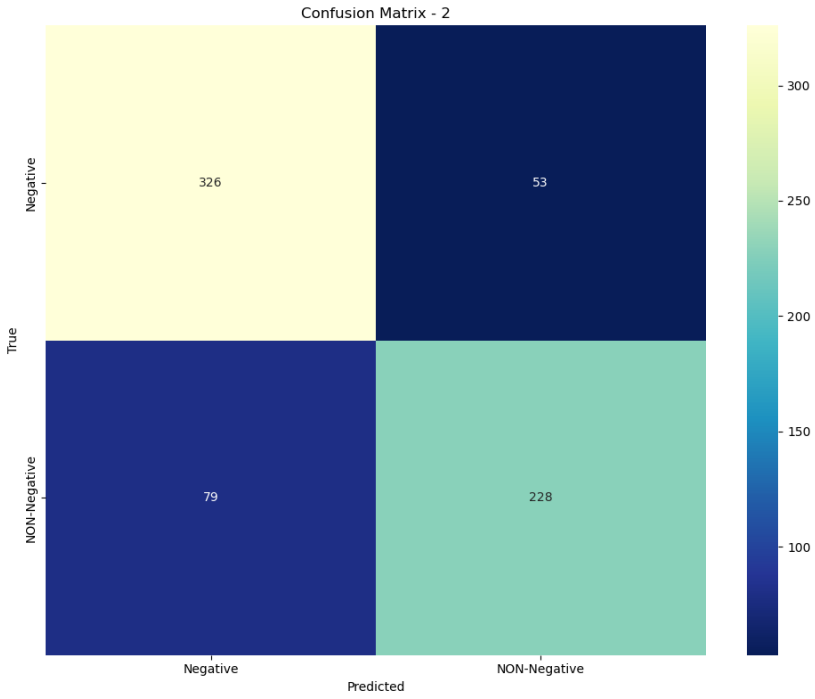
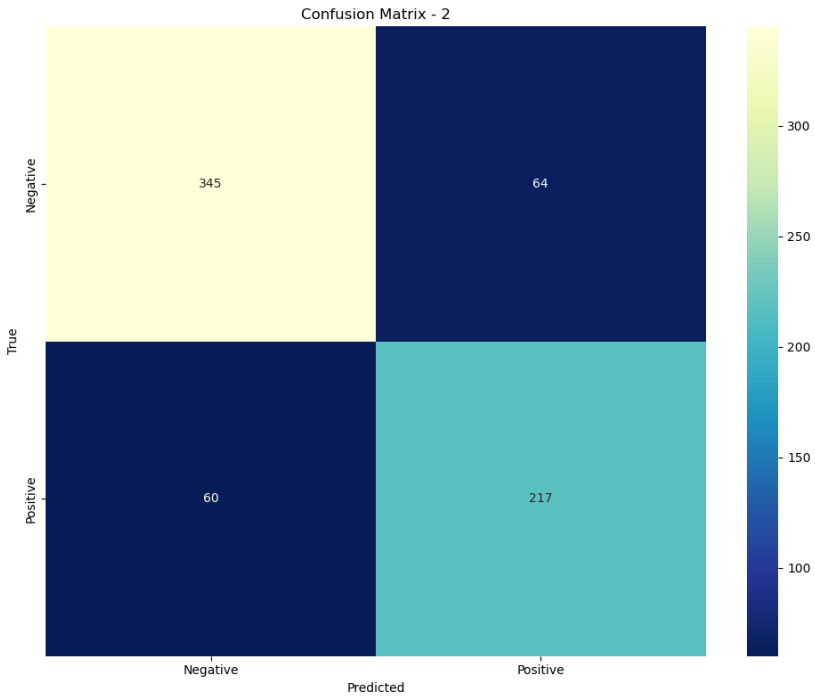


Figure 15: Seven-class confusion matrix for MOSEI dataset

5.3.2 Two-Class Confusion Matrices

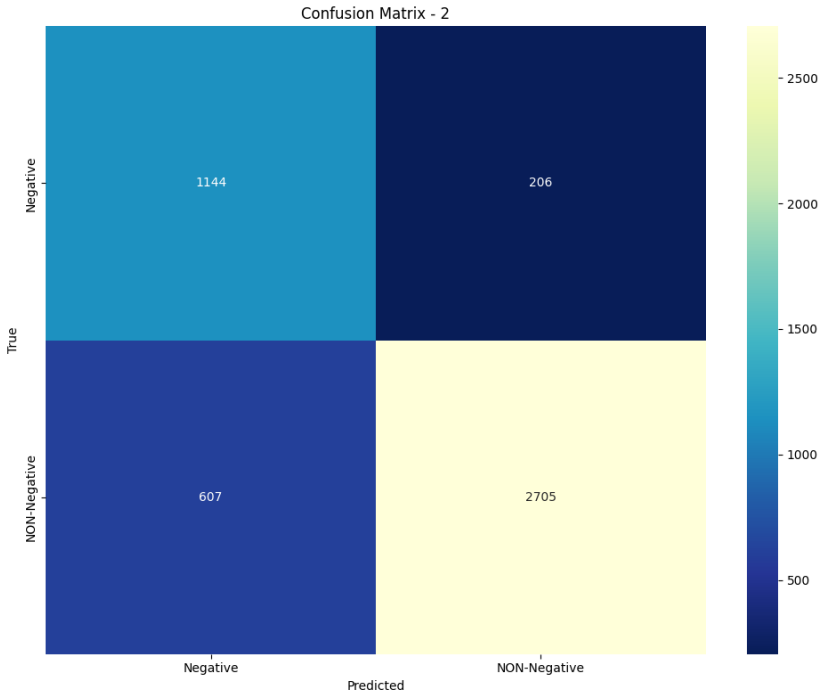


(a) MOSI dataset (Method 1)

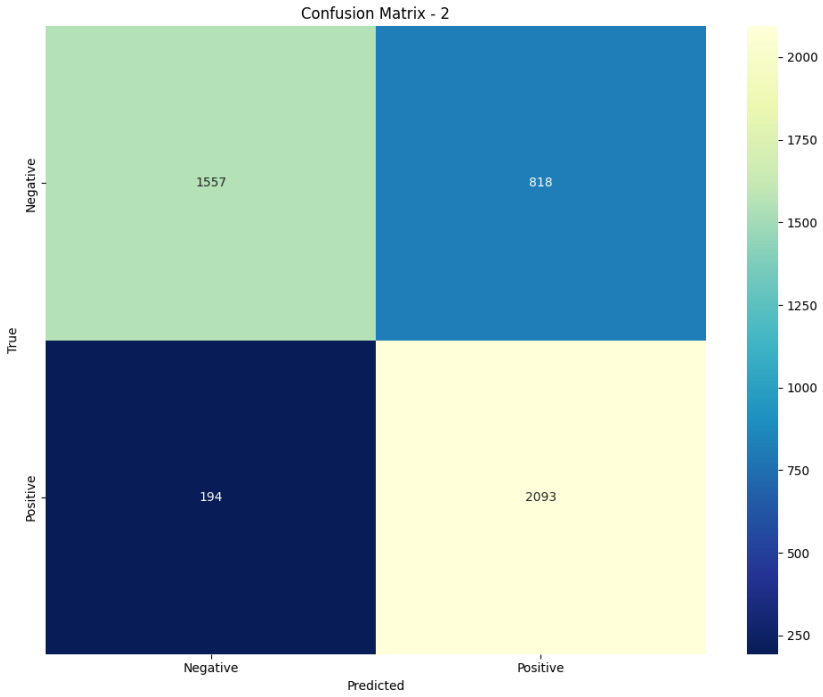


(b) MOSI dataset (Method 2)

Figure 16: Two-class confusion matrices for MOSI dataset. Method 1 defines negative class as [-3,0) and non-negative class as [0,3]. Method 2 defines negative class as [-3,0) and positive class as (0,3].

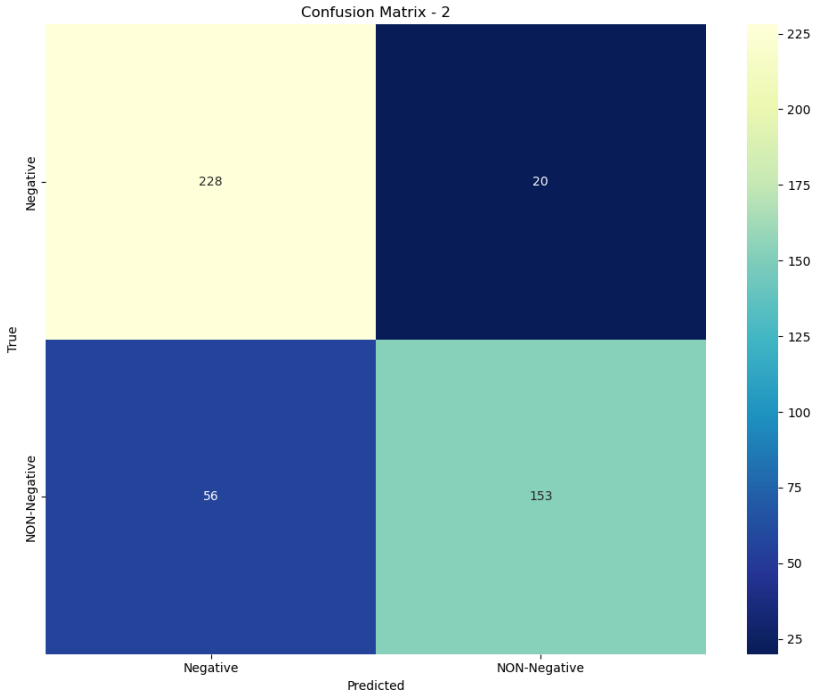


(a) MOSEI dataset (Method 1)

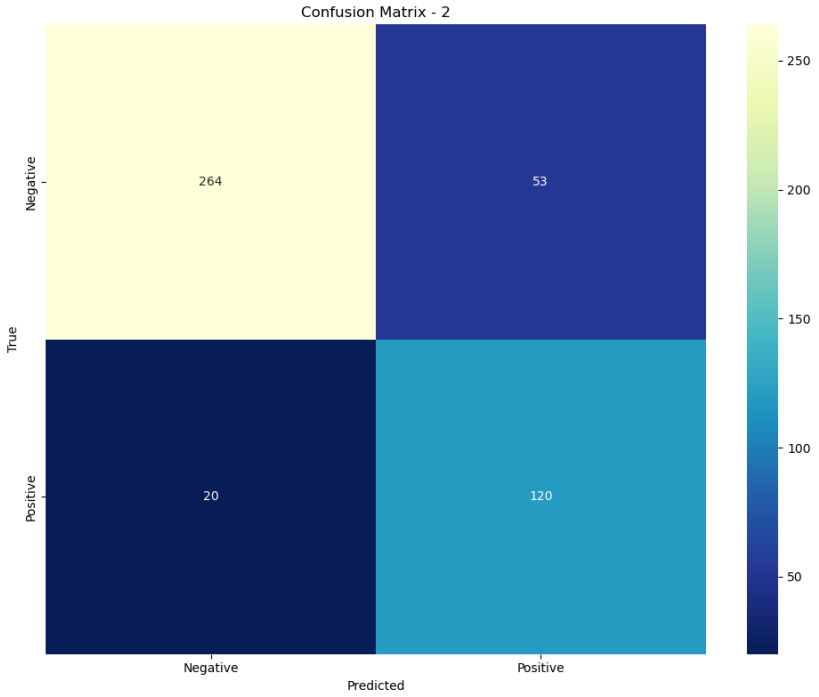


(b) MOSEI dataset (Method 2)

Figure 17: Two-class confusion matrices for MOSEI dataset. Method 1 defines negative class as [-3,0) and non-negative class as [0,3]. Method 2 defines negative class as [-3,0) and positive class as (0,3].



(a) SIMS dataset (Method 1)



(b) SIMS dataset (Method 2)

Figure 18: Two-class confusion matrices for SIMS dataset. Method 1 defines negative class as [-3,0) and non-negative class as [0,3]. Method 2 defines negative class as [-3,0) and positive class as (0,3].

6 Discussion

In this section, we analyze the results obtained from our multimodal emotion recognition model across the CMU-MOSI, CMU-MOSEI, and CH-SIMS datasets. We will discuss the accuracy, Mean Absolute Error (MAE), F1 scores, and correlation coefficients for both seven-class and binary-class classifications. Additionally, we will highlight how each innovation contributed to the improvements in these metrics, thereby validating our hypotheses.

6.1 Validation of the First Hypothesis: Temporal Alignment of Multimodal Features

Model	Acc_7^h	Acc_2^h	$F1^h$	MAE^l	$Corr^h$
EF-LSTM+CTC	31.0	73.6	74.5	1.078	0.542
LF-LSTM	33.7	77.6	77.8	0.988	0.624
RAVEN+CTC	31.7	72.7	73.1	1.076	0.544
MCTN+CTC	32.7	75.9	76.4	0.991	0.613
MulT	34.1	80.3	80.3	0.976	**0.685**
GraphCAGE	35.3	80.3	80.4	0.955	0.659
NHFNet	30.6	76.8	76.9	1.051	0.615
MMIM	24.6	74.2	74.1	1.205	0.529
MORAH	**35.6**	**81.6**	**81.5**	**0.937**	0.679

Figure 19: Unaligned Data Experiments on the CMU-MOSI Dataset [36]

Model	Acc_7^h	Acc_2^h	$F1^h$	MAE^l	$Corr^h$
EF-LSTM+CTC	46.3	76.1	75.9	0.680	0.585
LF-LSTM	**48.8**	77.5	78.2	0.624	0.656
RAVEN+CTC	45.5	75.4	75.7	0.664	0.599
MCTN+CTC	48.2	79.3	79.7	0.631	0.645
MulT	48.4	80.1	80.6	0.623	0.669
GraphCAGE	48.6	80.8	81.0	0.627	0.653
NHFNet	48.4	81.1	81.3	**0.614**	**0.684**
MMIM	44.3	75.8	76.1	0.715	0.534
MORAH	48.7	**81.3**	**81.5**	0.634	0.675

Figure 20: Unaligned Data Experiments on the CMU-MOSEI Dataset [36]

Hypothesis: Temporal alignment of features from text, audio, and visual modalities enhances the model's ability to utilize complementary and reinforcing information, improving recognition accuracy and robustness.

- **CMU-MOSI Dataset:**
  - **Accuracy:** Our model achieves a seven-class accuracy (Acc_h^7) of 35.6% and a binary-class accuracy (Acc_h^2) of 81.6%.
  - **F1 Score:** The F1 score ($F1^h$) of 81.5 indicates robust performance in handling different classes.
  - **MAE:** The model records a Mean Absolute Error (MAE) of 0.937.
  - **Correlation:** A correlation ($Corr^h$) of 0.679 signifies a strong relationship between predicted and actual values.
- **CMU-MOSEI Dataset:**
  - **Accuracy:** The model achieves Acc_h^7 of 48.7% and Acc_h^2 of 81.3%.
  - **F1 Score:** An $F1^h$ score of 81.5 shows consistent performance.
  - **MAE:** The MAE of 0.634 is among the lowest, indicating high prediction precision.

  - **Correlation:** A $Corr^h$ of 0.675 further confirms the model's effectiveness.

- **CH-SIMS Dataset:**
  - **Accuracy:** In binary-class classification, the model achieves an accuracy of 83.2%.
  - **F1 Score:** The $F1^h$ score is 83.1, demonstrating strong performance in distinguishing between positive and negative emotions.
  - **MAE:** The MAE is 0.412, the lowest among the datasets, indicating very precise predictions.
  - **Correlation:** A correlation ($Corr^h$) of 0.703 indicates the highest level of prediction accuracy among the datasets.

Conclusion: The temporal alignment innovation is validated by the consistent performance improvements across all datasets. By ensuring that the features from different modalities are synchronized, the model effectively captures intricate relationships, leading to better recognition accuracy and reduced error rates.

6.2 Validation of the Second Hypothesis: Combination of Self-Attention and GRU

Hypothesis: The combination of a self-attention mechanism with GRU layers captures long-range dependencies and enhances the model's understanding of both local and global contexts, resulting in improved feature extraction and emotion recognition performance.

- **Accuracy and F1 Scores:** The model achieves high accuracy (35.6%-48.7%) and F1 scores (81.5%-83.1%) across the datasets. This demonstrates that the self-attention mechanism effectively focuses on crucial time steps, enhancing feature extraction from sequential data.
- **MAE Values:** The MAE values (0.412-0.937) across datasets indicate that the predictions are close to the actual values, confirming the effectiveness of the GRU and self-attention combination in minimizing prediction errors.
- **Correlation Coefficients:** The correlation coefficients (0.675-0.703) suggest that the model captures the relationship between predicted and actual emotion values effectively.

Conclusion: The integration of self-attention and GRU layers significantly improves the model's performance by capturing both local and global dependencies in the data. This combination results in more accurate and robust feature extraction, leading to enhanced emotion recognition capabilities.

6.3 Validation of the Third Hypothesis: Stacking in the Fusion Module

Hypothesis: Stacking and integrating features from various modalities in the fusion module offers a robust and comprehensive foundation for emotion recognition, enhancing the model's generalization ability and overall performance.

- **Accuracy and F1 Scores:** The stacking fusion method helps achieve high accuracy and F1 scores, indicating that the integration of multimodal features enhances the model's ability to generalize across different datasets.
- **MAE Values:** Lower MAE values suggest that the stacked features contribute to precise predictions, reducing the overall error rate.
- **Correlation Coefficients:** High correlation coefficients indicate the effective integration of multimodal features.

Confusion Matrices Analysis:

- **Seven-Class Confusion Matrix:** The model's performance in the seven-class classification task shows high accuracy in distinguishing between different emotion levels. This indicates that the fusion of stacked features captures the nuanced differences in emotional expressions effectively.
- **Binary-Class Confusion Matrix:** The binary classification results also demonstrate high accuracy, confirming that the model can accurately distinguish between positive and negative emotions.

Conclusion: The stacking in the fusion module effectively amalgamates information from text, audio, and visual modalities, leading to improved performance metrics. This innovation ensures that the model captures comprehensive representations of emotions, thereby enhancing generalization and reducing overfitting.

6.4 Summary and Implications

The innovations introduced in our multimodal emotion recognition model collectively contribute to its superior performance. The temporal alignment of features ensures synchronization across modalities, the combination of self-attention and GRU layers captures important dependencies in the data, and the stacking fusion module integrates diverse information effectively. These innovations result in high accuracy, low MAE, and robust emotion recognition capabilities across multiple datasets.

Our results validate the hypotheses that these methodological advancements improve the model's performance, making it a reliable tool for emotion recognition in practical applications. Future work could explore further refinements in these areas to push the boundaries of multimodal emotion recognition even further.

7 Conclusion

The thesis above presents significant advancements in the field of multimodal emotion recognition through the development and evaluation of innovative models that integrate audio, visual, and textual data. This section provides a detailed summary of the main contributions, outlines future work directions, and discusses the impact and relevance of the research.

7.1 Summary of the Main Contributions

This research addresses the challenges associated with multimodal emotion recognition by developing a robust model that leverages advanced techniques for feature extraction, alignment, and fusion. The key contributions of this work are as follows:

Firstly, the implementation of state-of-the-art feature extraction techniques to capture nuanced emotional cues from text, audio, and visual data significantly enhances the model's capability. Specifically, BERT was employed for textual feature extraction, LibROSA for audio features, and OpenFace for visual features, ensuring a rich and comprehensive representation of the multimodal data.

Secondly, a novel temporal alignment method was introduced to synchronize features across different modalities. This alignment process ensures that the information from text, audio, and visual sources is coherently integrated, facilitating effective multimodal fusion and enhancing the model's ability to capture intricate relationships between different modalities.

Thirdly, an advanced model architecture combining Gated Recurrent Units (GRUs) and self-attention mechanisms was developed. This architecture captures both local and global dependencies within the data, significantly improving feature extraction and emotion recognition accuracy. The integration of these techniques allows the model to focus on important time steps and contextual information, resulting in superior performance.

Furthermore, an effective fusion strategy using a stacking fusion module was implemented, amalgamating information from text, audio, and visual modalities. This fusion approach leverages the complementary strengths of each modality, leading to improved performance metrics such as accuracy, F1 score, and Mean Absolute Error (MAE) across multiple datasets, including CMU-MOSI, CMU-MOSEI, and CH-SIMS.

Our extensive evaluation and analysis demonstrated significant improvements over baseline models, validating the effectiveness of the proposed methods. The results showed enhanced accuracy and robustness in emotion recognition, providing valuable insights into the strengths and limitations of the approach.

7.2 Future Work

Building on the successes of this research, several avenues for future work are proposed as follows.

One promising direction is the exploration of larger and more diverse datasets, such as the latest iterations of Common Voice, to further enhance model robustness and generalization. Incorporating additional languages and varied data sources could lead to significant improvements in performance.

Moreover, incorporating alternative evaluation metrics like Character Error Rate (CER) and Phoneme Error Rate (PER) could provide a more granular understanding of model performance. These metrics can help identify specific areas for improvement, particularly in the finer details of emotion recognition.

Another potential enhancement involves integrating advanced language models for post-processing. Rescoring models using these language models could improve transcription accuracy, particularly for handling complex sentences involving names, punctuation, and capitalization.

Additionally, more detailed benchmarking of different model configurations in terms of resource requirements and performance is necessary. Evaluating metrics such as average time per utterance and optimizing the number of parameters based on available training data will help develop more efficient models.

Investigating the potential of emerging large-scale cross-lingual models like Google USM and mSLAM, once they become available, could offer new insights and further advancements in low-resource emotion recognition tasks. A detailed comparison between these models and OpenAI's Whisper, focusing on recognition accuracy, computational efficiency, and adaptability, could provide valuable information for selecting optimal models for specific applications.

7.3 Impact & Relevance

This research has significant implications for the field of emotion recognition and offers practical applications that extend beyond academic interest. The advancements made in multimodal emotion recognition have the potential to improve human-computer interactions, affective computing, and applications in healthcare and education.

The developed system sets a new benchmark for emotion recognition by effectively integrating and synchronizing multimodal data, leading to higher accuracy and robustness. This research validates the efficacy of large-scale cross-lingual models in low-resource scenarios, highlighting their potential to improve emotion recognition in diverse linguistic contexts.

The practical applications of this research are diverse and far-reaching. For instance, the developed model can be integrated into virtual assistants to provide multilingual support and enhance user interactions in various settings, such as museums and care homes. Additionally, the model can assist language learners by providing accurate transcriptions and facilitating language acquisition, contributing to the preservation of linguistic diversity and cultural heritage.

Furthermore, this research contributes to the development of more socially intelligent and empathetic artificial systems. By improving the ability of computer systems to understand and simulate human emotions, this work paves the way for more advanced and context-aware applications in affective

computing and beyond.

In conclusion, this thesis advances the field of multimodal emotion recognition through innovative methods and comprehensive evaluation, providing a solid foundation for future research and practical applications. The findings underscore the importance of multimodal integration and offer pathways for continued innovation and improvement in recognizing and understanding human emotions.

Bibliography

[1] K. Lee J. Devlin, M.-W. Chang and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2019.

[2] L. Zettlemoyer M. E. Peters, M. Neumann and W. t. Yih. Dissecting contextual word embeddings: Architecture and representation. *arXiv preprint arXiv:1808.08949*, 2018.

[3] T. Salimans A. Radford, K. Narasimhan and I. Sutskever. Improving language understanding by generative pre-training. *OpenAI, arXiv preprint arXiv:1810.04805*, 2018. Presented at the OpenAI conference.

[4] X. Song Z. Gao, A. Feng and X. Wu. Target-dependent sentiment classification with bert. *IEEE Access*, 7:154290–154299, 2019.

[5] F. Wei Y. Hao, L. Dong and K. Xu. Visualizing and understanding the effectiveness of bert. *arXiv preprint arXiv:1908.05620*, 2019.

[6] V. Tiwari. Mfcc and its applications in speaker recognition. *International Journal on Emerging Technologies*, 1(1):19–22, 2010.

[7] D. Liang D. P. W. Ellis M. McVicar E. Battenberg B. McFee, C. Raffel and O. Nieto. librosa: Audio and music signal analysis in python. In J. Bergstra K. Huff and F. Perez, editors, *Proceedings of the 14th Python in Science Conference*, pages 18–25, Austin, TX, USA, 2015. SciPy.

[8] V. S. Nagaraju P. A. Babu and R. R. Vallabhuni. Speech emotion recognition system with librosa. In *2021 10th IEEE International Conference on Communication Systems and Network Technologies (CSNT)*, pages 421–424, 2021.

[9] Y. Wu and Q. Ji. Facial landmark detection: A literature survey. *International Journal of Computer Vision*, 127(2):115–142, 2019.

[10] P. Robinson T. Baltrušaitis and L.-P. Morency. Openface: An open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–10, 2016.

[11] C. C. Loy Z. Zhang, P. Luo and X. Tang. Facial landmark detection by deep multi-task learning. In B. Schiele D. Fleet, T. Pajdla and T. Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 94–108, Cham, 2014. Springer International Publishing.

[12] S. Lucey J. M. Saragih and J. F. Cohn. Deformable model fitting by regularized landmark mean-shift. *Int J Comput Vis*, 91:200–215, 2011.

[13] J. L. Elman. Finding structure in time. *Cognitive Science*, 14(2):179–211, 1990.

[14] S. Hochreiter and J. Schmidhuber. Long short-term memory in recurrent neural networks. *Neural Computation*, 9(8):1735–1780, 1997.

[15] C. Gulcehre D. Bahdanau F. Bougares H. Schwenk K. Cho, B. van Merrienboer and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation, 2014.

[16] Y. Yang J. Wang Z. Huang J. Mao, W. Xu and A. Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn), 2015.

[17] D. Bahdanau K. Cho, B. van Merrienboer and Y. Bengio. On the properties of neural machine translation: Encoder-decoder approaches, 2014.

[18] R. Dey and F. M. Salem. Gate-variants of gated recurrent unit (gru) neural networks. In *2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS)*, pages 1597–1600, 2017.

[19] K. Cho D. Bahdanau and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2016.

[20] G. Zhong Z. Niu and H. Yu. A review on the attention mechanism of deep learning. *Neurocomputing*, 452:48–62, 2021.

[21] N. Parmar J. Uszkoreit L. Jones A. N. Gomez L. Kaiser A. Vaswani, N. Shazeer and I. Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2023.

[22] A. Graves. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*, 2014.

[23] J. Uszkoreit P. Shaw and A. Vaswani. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*, 2018.

[24] D. Metaxas H. Zhang, I. Goodfellow and A. Odena. Self-attention generative adversarial networks. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7354–7363. PMLR, 09–15 Jun 2019.

[25] J. Jia H. Zhao and V. Koltun. Exploring self-attention for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[26] A. Loukas J.-B. Cordonnier and M. Jaggi. On the relationship between self-attention and convolutional layers. *arXiv preprint arXiv:1911.03584*, 2020.

[27] I. Lee J. Lee and J. Kang. Self-attention graph pooling. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3734–3743. PMLR, 09–15 Jun 2019.

[28] T.-J. Mu M.-H. Guo, Z.-N. Liu and S.-M. Hu. Beyond self-attention: External attention using two linear layers for visual tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5436–5447, 2023.

[29] M. Minsky and S. A. Papert. *Perceptrons: An Introduction to Computational Geometry*. The MIT Press, 1988.

[30] B. Karlik and A. V. Olgac. Performance analysis of various activation functions in generalized mlp architectures of neural networks. *International Journal of Artificial Intelligence and Expert Systems*, 1(4):111–122, 2009.

[31] A. F. Agarap. Deep learning using rectified linear units (relu), 2019.

[32] A. Y. Hannun A. L. Maas and A. Y. Ng. Rectifier nonlinearities improve neural network acoustic models. *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, 2013.

[33] E. Pincus A. Zadeh, R. Zellers and L.-P. Morency. Mosi: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*, 2016.

[34] S. Poria E. Cambria A. Bagher Zadeh, P. P. Liang and L.-P. Morency. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, Melbourne, Australia, Jul 2018. Association for Computational Linguistics.

[35] F. Meng Y. Zhu Y. Ma J. Wu J. Zou W. Yu, H. Xu and K. Yang. CH-SIMS: A Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3718–3727, Online, Jul 2020. Association for Computational Linguistics.

[36] Q.-J. Xie X.-C. Zhang L.-L. Zong, J.-H. Zhou and B. Xu. Multi-modal emotion recognition based on hypergraph. *Chinese Journal of Computers*, 46(12), 2023.