2023

MULTIMODAL EMOTION ANALYSIS WITH FOCUSED ATTENTION

Siddhi Kiran Bajracharya

MULTIMODAL EMOTION ANALYSIS WITH FOCUSED ATTENTION

By

Siddhi Bajracharya

B.E., Tribhuvan University, Nepal, 2018

A Thesis Submitted in Partial Fulfillment of
the Requirements for the Degree of Master of Science

Department of Computer Science

Master of Science Program
In the Graduate School
The University of South Dakota
December 2023

The members of the Committee appointed to examine
the Thesis of Siddhi Kiran Bajracharya
find it satisfactory and recommend that it be accepted.

DocuSigned by:

KC Santosh

99944614C924467...

Chairperson

DocuSigned by:

Rodrigue Rizk

056A76355F814ED...

DocuSigned by:

Lu Baugh

05454119D36C468...

ABSTRACT

Emotion analysis, a subset of sentiment analysis, involves the study of a wide array of emotional indicators. In contrast to sentiment analysis, which restricts its focus to positive and negative sentiments, emotion analysis extends beyond these limitations to a diverse spectrum of emotional cues. Contemporary trends in emotion analysis lean toward multimodal approaches that leverage audiovisual and text modalities. However, implementing multimodal strategies introduces its own set of challenges, marked by a rise in model complexity and an expansion of parameters, thereby creating a need for a larger volume of data. This thesis responds to this challenge by proposing a robust model tailored for emotion recognition, specifically focusing on leveraging audio and text data. Our approach is centered on using audio spectrogram transformers (AST), and the powerful BERT language model to extract distinctive features from both auditory and textual modalities followed by feature fusion. Despite the absence of the visual component, employed by state-of-the-art (SOTA) methods, our model demonstrates comparable performance levels achieving an f1 score of 0.67 when benchmarked against existing standards on the IEMOCAP dataset [1] which consists of 12-hour audio recordings broken down into 5255 scripted and 4784 spontaneous turns, with each turn labeled by emotions such as anger, neutral, frustration, happy, and sad. In essence, We propose a fully attention-focused multimodal approach for effective emotion analysis for relatively smaller datasets leveraging lightweight data sources like audio and text highlighting the efficacy of our proposed model. For reproducibility, the code is available at 2AI Lab's GitHub repository: https://github.com/2ai-lab/multimodal-emotion.

DocuSigned by:

Thesis Advisor: KC Santosh

1D1EB20650034B9...

KC Santosh, Ph.D.

Acknowledgments

I want to express my deepest appreciation to my thesis advisor, Dr. KC Santosh, whose expert guidance, and steadfast support were crucial in completing this thesis. This work would not have come to fruition without his invaluable mentorship. I also want to recognize the contributions of the esteemed members of my thesis committee, Dr. Rodrigue Rizk and Dr. Lee Baugh. Their active engagement and help throughout this process have been greatly valued. Additionally, I extend my heartfelt thanks to the entire 2AI lab 2022 - 2023 team for their invaluable assistance and support during the thesis writing journey. Finally, I would like to thank the kind stranger on Reddit who quoted, “It’s better to reach the top and find out it wasn’t worth it than to never try at all”. Your collective contributions have been essential, and I am genuinely grateful for your involvement in this academic pursuit.

Dedication

I dedicate this thesis to my parents Tri Ratna and Jamuna Bajracharya. Thank you for everything. It is difficult to put into words the appreciation and gratitude I have towards you both. Thank you for everything; this achievement is as much yours as it is mine.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AI	Artificial Intelligence.
AST	Audio Spectrograma Transformers.
BERT	Bidirectional Encoder Representations from Transformers.
DNN	Deep Neural Network.
ANN	Artificial Neural Network.
TP	True Positive.
FP	False Positive.
FN	False Negative.
IEMOCAP	the Interactive Emotional Dyadic Motion Capture.
IO	Input Output
ADAM	Adaptive Momentum.
SGD	Stochastic Gradient Descent.
CNN	Convolutional Neural Network.
RNN	Recurrent Neural Network.
LSTM	Long Short Term Memory.
ViT	Visual Transformers.
NLP	Natural Language Processing.
GRU	Gated Recurrent Unit.
ReLU	Rectified Linear Unit.
LIWC	Linguistic Inquiry and Word Count .
LLM	Large Language Model.
SOTA	State Of The Art.
SVM	Support Vector Machines.

Chapter 1

Related work: Sentiment and emotion analysis

Summary: Emotion analysis and detection is a sub-domain in sentiment analysis. Sentiment analysis is where we try to measure the degree of positive and negative emotions in data. Emotion analysis (interchangeably also called sentiment analysis) is an application of AI where we go one step further and classify or predict emotions. Mostly sentiment and emotion analysis used to be related to text processing [2], [3], [4]. However, as we understand, emotions can be expressed not only through texts but also through spoken language or visual [5] cues—anger in a dialogue may not fully be captured only by textual data. Consider the sentence, "You like ice cream?". A rising intonation at the end may suggest that the person who asked the question is surprised that you like ice cream, but a neutral intonation may suggest that the person is just curious to know whether you like ice cream or not. The added tone to the sentence can provide more context to make more informed decisions.

Key topics: Motivation, goal, and contribution.

Organization: The chapter is structured as follows. In section 1.1, we will give a detailed explanation of the context and problem of our thesis work. After that, we will explain the motivation of our work in section 1.2. We will then discuss our research hypothesis and contribution in 1.3 and 1.4. Finally, section 1.5 describes the overall organization of the thesis.

1.1 Background

Emotion analysis has applications in diverse domains, such as health, business and marketing, and cybercrime. In the domain of health and medicine, emotion analysis can be used for mental health monitoring [6], [7], patient satisfaction feedback [8], and stress management. Significant research is also being conducted to minimize cybercrime and the risk of cybercrime. Intelligent algorithms are being used to detect cyberbullying [9] and cyber threats [10] on the internet. Although textual data have been dominantly used for emotion analysis, the addition of auditory data can significantly boost the performance of NLP models. Speech intonation, pitch, and other vocal cues can convey subtle emotional nuances that are often lost in text-based analysis. By incorporating auditory data, emotion analysis can gain a more comprehensive understanding of the emotional state of the speaker, leading to more accurate emotion classification. Auditory analysis can capture non-verbal cues such as laughter, sarcasm, and irony, which can significantly impact the overall sentiment of a conversation or utterance. These cues are often difficult to detect in text-based analysis, but they can provide valuable insights into the speaker's true intentions and emotions. The tone of voice, accent, and speech patterns can reveal information about the speaker's demographic characteristics, such as age, gender, and cultural background. This information can be useful for tailoring emotion analysis models to specific audiences and improving their accuracy in different contexts. Auditory cues can provide clues about the speaker's emotional state, which can be used to detect deception or manipulation. For example, changes in pitch, hesitation, or unusual vocal patterns may indicate that the speaker is being dishonest or trying to conceal their true intentions. Auditory analysis can help bridge cultural gaps in emotion analysis. Non-verbal cues and intonation patterns can convey cultural nuances that may not be reflected in the text, allowing emotion analysis models to better understand and interpret expressions from different cultures.

In recent years, researchers have become more interested in multimodal features to perform sentiment analysis [11], [12], [13]. Textual features along with audiovisual features have been widely used to demonstrate the robustness of multimodal representation learning for sentiment analysis. The current state-of-the-art method for emotion analysis leverages a cutting-edge approach known as Emformer[14], [15], which seamlessly integrates auditory, visual, and text data to enhance the accuracy and robustness of sentiment predictions. Emformer represents a pioneering step towards multimodal sentiment analysis, recognizing the importance of incorporating diverse modalities to capture a more comprehensive understanding of human expression. However, despite its advancements, the Emformer model is not without its limitations, as highlighted below:

Challenge of Large Video Data: One significant drawback lies in the handling of video data, which can be inherently voluminous. The storage and processing requirements for large video datasets pose practical challenges, potentially hindering the scalability and efficiency of the sentiment analysis system.

Unavailability of Pretrained Weights: Another notable limitation is the unavailability of pretrained weights for Emformer models. This absence of readily accessible pre-trained weights impedes the reproducibility of results for researchers and practitioners. Reproducing the exact conditions and outcomes of the original study becomes challenging without a standardized set of pre-trained weights.

1.2 Motivation

While multimodal techniques like Emformer have emerged as a promising approach to emotion analysis, it is essential to recognize its inherent limitations. The processing of large video datasets can pose significant computational challenges, while the scarcity of pre-trained models hinders reproducibility and collaboration. These limitations demand further research efforts to provide an alternative method to match SOTA performance. Our approach seeks to address these challenges by employing spectral representation for auditory signals and simplifying training algorithms for smaller datasets by eliminating the visual component. This strategy aims to enhance and foster the continued advancement of multimodal sentiment analysis. Our strategy focuses on two key aspects:

Spectral Representation for Auditory Signals: By employing spectral representation for auditory signals, we can effectively capture the emotional cues embedded within speech, reducing the computational burden while preserving relevant information.

Simplified Training Algorithms for Smaller Datasets: By eliminating the visual component during training, we can simplify the training algorithms and make them more efficient for smaller datasets.

1.3 Research hypothesis

Our research hypothesizes that the integration of audio spectrogram features and contextual information extracted from a BERT language model will significantly enhance emotion analysis accuracy compared to relying on either modality independently. Audio spectrograms effectively capture acoustic patterns in audio data, providing valuable insights into emotions and emotions conveyed through speech intonation, pitch variations, and pauses. BERT language models, on the other hand, excel at understanding the nuances of human language and extracting sentiment from text. Further, we consider a fully attention-focused architecture that is superior in performance for sequential data in comparison to convolutional methods. By seamlessly integrating these two modalities, we can harness the strengths of each to achieve superior emotion analysis performance. This hypothesis can be tested by conducting a comprehensive comparative analysis of emotion analysis accuracy using audio spectrogram features, the BERT language model, and the proposed multimodal approach.

1.4 Research contribution

Our research contributions are listed below:

1. Propose a purely attention-based architecture for multimodal feature fusion learning using Mel-filterbanks and BERT encoding for audio and text data respectively.

2. Propose a robust and lightweight model without the use of visual components for multimodal learning.

3. Propose an architecture that achieves comparable results to state-of-the-art methods without the need for pretraining.

1.5 Thesis outline

The thesis is organized into the following sections:

Chapter 1

In this chapter, we have established a foundational understanding of the research context through the introduction. The overarching theme of sentiment and emotion analysis, underscoring its significance in contemporary research, has been elucidated in the background. The motivation behind the study, serving as the driving force, has been presented, and the research hypothesis, outlining the key questions guiding the investigation, has been introduced. The chapter concludes with an overview of the organization of the thesis, providing readers with a roadmap for the subsequent chapters.

Chapter 2

In this chapter, we have surveyed the landscape of sentiment and emotion analysis, examining various methodologies. Lexicon-based methods have been introduced, and shallow and deep learning-based approaches have been explored, categorized into textual, acoustic, and multimodal methods. A research gap has been identified, underscoring the unique contributions of this research. Notably, a literature review covering both unimodal and multimodal sentiment and emotion recognition has been provided, assessing their strengths and limitations. This sets the groundwork for the subsequent chapters, offering a concise overview of existing methodologies and motivations for the novel contributions in this thesis.

Chapter 3

In this chapter, the foundational concepts and tools employed in the study have been laid out. Deep neural networks, encompassing their activation functions and artificial neural network architecture, have been discussed. The chapter delves into transformers, with an emphasis on the attention mechanism and the utilization of BERT. Additionally, convolutional neural networks are introduced, along with an overview of audio spectrograms, including Mel-frequency spectrograms and audio spectrogram transformers. Evaluation metrics, loss functions, and optimization techniques have also been presented, providing the theoretical underpinnings for the subsequent methodology.

Chapter 4

In this chapter, the research design has been outlined, providing details on the training and validation processes. The purpose and description of data collection have been elucidated, with a particular emphasis on the significance of a custom dataset. Preprocessing steps have been discussed to ensure data quality. The training setup has been defined, introducing the approach to unimodal learning and outlining the construction of a multimodal learning model through feature fusion.

Chapter 5

In this chapter, the findings of the research are presented. The examination begins with an exploration of unimodal approaches, followed by an in-depth analysis of the proposed multimodal model. Research limitations are candidly discussed, providing a transparent view of the study's constraints. Future work is proposed, offering potential directions for expanding and refining the research. Additionally, the chapter incorporates an ablation study.

Chapter 6

In this concluding chapter, the study's crucial findings are summarized, emphasizing the contributions made to the field of sentiment and emotion analysis.

Chapter 2

Related work: Sentiment and emotion analysis

Summary: In this chapter, we delve into the historical development of emotion analysis. We start with sentiment analysis and dive deeper into emotion analysis. We discuss lexicon-based and machine learning methods for sentiment analysis and study contemporary emotion analysis methods.

Key topics: Sentiment analysis, Emotion analysis.

Organization: The rest of the chapter is structured as follows: In section 2.1, a thorough review of lexicon-based methods for sentiment analysis is presented. In section 2.2, we reviewed related work so far that has been done by using shallow learning/deep learning methods for sentiment and emotion analysis. We also discuss contemporary emotion analysis techniques. Finally, in section 2.3 we explore and analyze the existing body of research to identify areas where further investigation is needed.

While sentiment analysis has a long-established history, the concept of emotion analysis is new. It is worthwhile to explore how sentiment analysis served as a foundational framework in the evolution of emotion recognition and analysis, representing a more advanced and nuanced approach. In 1964, Philip Stone et al.[16] the General Inquirer, a statistical tool that uses a dictionary of words and phrases to identify the sentiment of the text. It was used to analyze various text sources, including news articles, government documents, and medical records. This was one of the first attempts to extract sentiment from textual data. Up until the 1990s there were not any significant developments in sentiment analysis. The rise of the internet and social media in the 1990s and early 2000s led to a surge of interest in sentiment analysis. Businesses and organizations realized the potential of sentiment analysis to understand public opinion and behavior. WordNet [17], a lexical database that was published by George A. Miller in 1995 is still being used today by researchers and developers for sentiment analysis. In 1997, Hatzivassiloglou and McKeown [18]the use of semantic roles for sentiment analysis. Elkan [19] published a patent for text classification that included sentiment analysis as one of the possible class labels in 2001. Pang et al.[20] published a seminal paper on sentiment classification, introducing the notion of subjectivity and neutrality in sentiment analysis in 2002. Subsequent research developed into emotion analysis, where we now use various modalities to detect emotions.

2.1 Lexicon-based methods

Lexicon-based methods rely on a dictionary or lexicon of words associated with positive, negative, or neutral sentiment. The sentiment of a text is then determined by counting the number of positive, negative, and neutral words in the text. In 2006, Andrea Esuli and Fabrizio Sebastiani published their work SENTIWORDNET [21], which was a significant contribution to the field of sentiment analysis. SENTIWORDNET associated each WORDNET sysnet with three numerical scores Pos(s), Neg(s), and Obj(s) for positive, negative, and objective scores. In recent years, there has been a great deal of research on lexicon-based sentiment analysis. One of the most important developments has been the creation of more sophisticated sentiment lexicons. These lexicons include more domain-specific words and phrases [22], and they have also assigned more nuanced sentiment scores to each word or phrase. This has led to more accurate sentiment analysis results. Works such as [23], and [24] have developed domain-specific lexicons for financial and neural domains. Lexicon-based sentiment analysis is a straightforward and interpretable approach that relies on predefined sets of lexicons or dictionaries. The simplicity of these methods lies in their ability to be applied to various languages, given the availability of appropriate lexicon resources. Despite their universality, a significant challenge arises in the creation of high-quality lexicons, as developing comprehensive and accurate dictionaries requires considerable effort and linguistic expertise. While lexicon-based methods offer advantages in terms of simplicity and cross-language applicability, they exhibit limitations in handling nuanced language elements such as sarcasm and irony. The inherent rigidity of lexicons may lead to suboptimal performance when confronted with these complex linguistic constructs, as the fixed sets of predefined sentiments may struggle to capture the subtle, often contradictory, meanings inherent in sarcastic or ironic expressions. In summary, lexicon-based sentiment analysis provides a clear and understandable approach to analyzing sentiment in text, and its adaptability to multiple languages is a notable advantage. However, the quality of lexicons and the method's vulnerability to nuances like sarcasm and irony present challenges that researchers and practitioners must consider when employing these methods in sentiment analysis tasks.

2.2 Shallow and deep learning-based methods

2.2.1 Textual data

These methods use machine learning algorithms to learn to identify sentiment from text data. The algorithms are trained on a large dataset of labeled text, where each text has been manually assigned a sentiment label. Once trained, the algorithm can be used to predict the sentiment of new text data. Research by Rathi et al.[25] employs information gathered from microblogging platforms like Twitter and Facebook, utilizing a combination of Support Vector Machines (SVM) and decision trees for the categorization of expressed opinions into positive, negative, or neutral sentiments. Although traditional machine learning algorithms such as Naïve Bayes and support vector machines (SVMs) have been used for sentiment analysis [26], [25], research has shown that deep learning-based methods such as LSTMs, RNNs, and transformers outperform the nonsequential models [26], [27, 28, 29]. Wang Y et. al [28]. propose RNN-capsule, based on Recurrent Neural Network (RNN), where they demonstrate that RNN-based models outperform state-of-the-art sentiment classification models without a linguistic model. This demonstrates the superiority of deep learning models over traditional lexicon-based models. Research [30, 31, 32] suggests using hashtags and emojis in the text as labels to determine the emotions of the given text. The mentioned research shows the importance of additional dimensions to successfully detect and analyze emotions. In this case, the added dimensions are ASCII characters and not a different modality. In their work [30], the authors investigate the correspondence between emoticons, emoji, and hashtags that are associated with certain emotions. In another work [31], the authors propose a tool called Linguistic Inquiry and Word Count (LIWC), which is a word counting software that uses a dictionary of grammatical, psychological, and content word categories to efficiently classify texts along these verticals. In a different research paper [32], the authors employ self-labeling using the hashtags in Twitter posts that outperformed manually crafted WordNet. In 2019 [33], researchers used a graph convolutional neural network for emotion recognition in conversation. The authors identify two major context types: sequential and speaker-level context. Subsequently, they create graphs with nodes and edges that represent individual utterances and dependency between the speaker utterances respectively. Finally, the constructed graph is passed through a graph convolutional network. In their recent work [34], the authors proposed InstructERC, an emotion recognition tool that employed large language models. This work is currently an LLM-based SOTA model for the IEOMOCAP dataset.

2.2.2 Acoustic data

The sentiment analysis can also be done using spoken language. As we have already discussed in the background section of the thesis, acoustic speech signals can be useful in detecting nuances and emotional cues in spoken language. These acoustic signals can either be used in raw form or transformed form. Since acoustic signals are sequential data, it is common to see the use of sequential models [35], [36]. The research shows that the use of sequential models like Long Short-Term Memory (LSTM), a variant of RNN performs better than classical feed-forward deep neural networks. Concrete results using auditory modality are yet to be seen for sentiment analysis. However, there is research in the domain of multimodal learning where auditory modality is used along with other modalities to build speech models [15]. In 2018 in their paper [37], the authors introduce an approach called Speech SIMCLR that uses augmented raw audio and its spectrogram in combination with contrastive loss to maximize agreement between differently augmented samples.

2.2.3 Multimodal approach

Recent advancements and research in emotion analysis have shifted its focus to multimodal representation and learning. Research shows robust performance in multimodal learning [38]. In their work, Stappen et al. [12] use a multimodal approach by combining audiovisual, language, and biological signal modalities to analyze emotional and physiological-based stress. The scope of sentiment analysis has broadened beyond text-based modalities to include multimodal analysis. CLAP [39], another significant contribution, presents a two-stage framework for multimodal sentiment analysis utilizing language-audio modality. The methodology involves employing HuBERT and RoBERTa to extract both auditory and textual encodings. These encodings are subsequently fused and input into transformer models. In a different paper [40], the authors design a custom feature extractor that extracts embeddings from Mel-spectrograms and visual data. Subsequently, they fuse the extracted features, fuse them, and pass them to a multi-headed transformer network. The feature extractor is based on the triplet loss function. The EmoCaps [15] model proposed in 2022, utilizes audio, text, and visual modalities. The text features are extracted using BERT, and the auditory features are extracted using an open-source Library called OpenSmile [41]. Then the authors use Emformer (a transformer-based architecture) to extract subsequent embeddings create a fusion of features and finally pass it through a feed-forward network.

2.3 Research gap

The rapid research on different transformer-based models has given rise to an architecture of transformers called audio spectrogram transformers (AST), which has not yet been used for emotion recognition. The authors of the work AST [42] claim that the architecture is superior to CNN-based architectures in analyzing audio spectrograms. More studies by [43], [44], [45] also show promising results using spectral analysis rather than analysis of raw audio signals. In our work, we aim to employ AST to process the speech dataset for emotion recognition. Together with encoded BERT embeddings, we propose a multimodal approach using text and audio datasets. In the later section of the thesis, we demonstrate that our proposed architecture generates comparable performance to the SOTA emotion recognition architectures. We also compare our proposed model with other SOTA algorithms that utilize only audio and text modality.

Chapter 3

Materials

Summary: In this chapter, we give a detailed overview of the foundational blocks for the proposed model. We discuss foundational blocks of deep learning such as neurons, artificial neural networks, convolutional neural networks, and transformers.

Key topics: Deep neural network, Transformers, CNN.

Organization: The rest of the chapter is structured as follows: In section 3.1.1 we neurons and activation functions. Then in 3.1.2, we describe the artificial neural network. And then in 3.1.3 and 3.1.4, we shed light on important topics such as loss functions and optimizers. In further sections 3.2, 3.3, we discuss convolutional neural networks and transformers. Finally, in 3.4, we discuss some important evaluation metrics for classification.

3.1 Basics of neural network

3.1.1 Neurons and activation functions

We employ a deep neural network for the classification of input signals and texts, and in this section, we will provide a brief overview of deep learning. Deep learning is a machine learning approach founded on artificial neural networks (ANNs), which aim to mimic the functioning of the human brain. Neurons, the fundamental units of a neural network, receive inputs and produce outputs through mathematical computations. These artificial neurons are comparable to the neurons found in the human brain. Following the mathematical operations, the output undergoes activation through a function, leading to the activation or firing of the neuron if a specific threshold is attained.

The inputs will be the numeric values x1, x2, and so on. w1, w2, and so on are the weights of the neurons that tell us how important that input is:

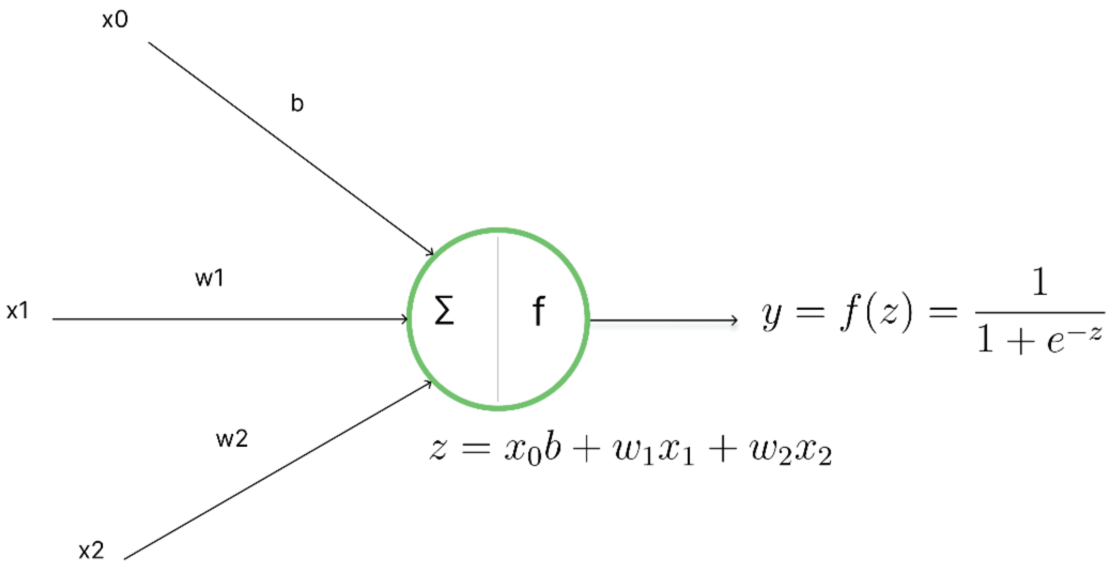

x0
b
w1
x1
Σ
f
$y = f(z) = \frac{1}{1+e^{-z}}$
w2
$z = x_0 b + w_1 x_1 + w_2 x_2$
x2

Figure 3.1: Basic structure of neuron.

$$z\tau = x_0 * \ b\tau + x_1 * \ w_1 + x_2 * \ w_2 + \ldots \tau + x_n * \ w_n,\tau \tag{3.1}$$

We pass the output through a function called the activation function which decides whether the neuron should be fired or not: y=f(z).

Activation functions introduce non-linearity to models, enabling them to effectively capture highly nonlinear data when enough neuron layers are present. Some common activation functions [46, 47, 48, 49] are shown in Figure 3.2. In theory, the described neurons can be combined in diverse configurations to approximate any type of function, forming what is known as neural networks. However, these neural networks necessitate a larger volume of data for training compared to the shallow learning techniques discussed in earlier sections.

After generating the output, it is essential to calculate the loss, which serves as a measure of the disparity between the real values and the predicted values. The selection of the loss function is contingent on the nature of the problems being addressed. Generally, a loss function L is denoted by $loss\tau = \ L\,(y, y_{target})$. Once the loss for the network is calculated, a mathematical algorithm called backpropagation is used.

Backpropagation is a mathematical method utilized for training neural networks, and it entails calculating the gradient of the loss function concerning the network's weights and biases. This gradient is then employed to modify the weights and biases, to minimize the loss function. To compute the gradient of the loss concerning the output at a given point, partial derivatives are employed:

$$\frac{\partial L\tau}{\partial z\tau} = \frac{\partial L}{\partial y} \cdot \frac{\partial y\tau}{\partial z\tau},^{\tau} \tag{3.2}$$

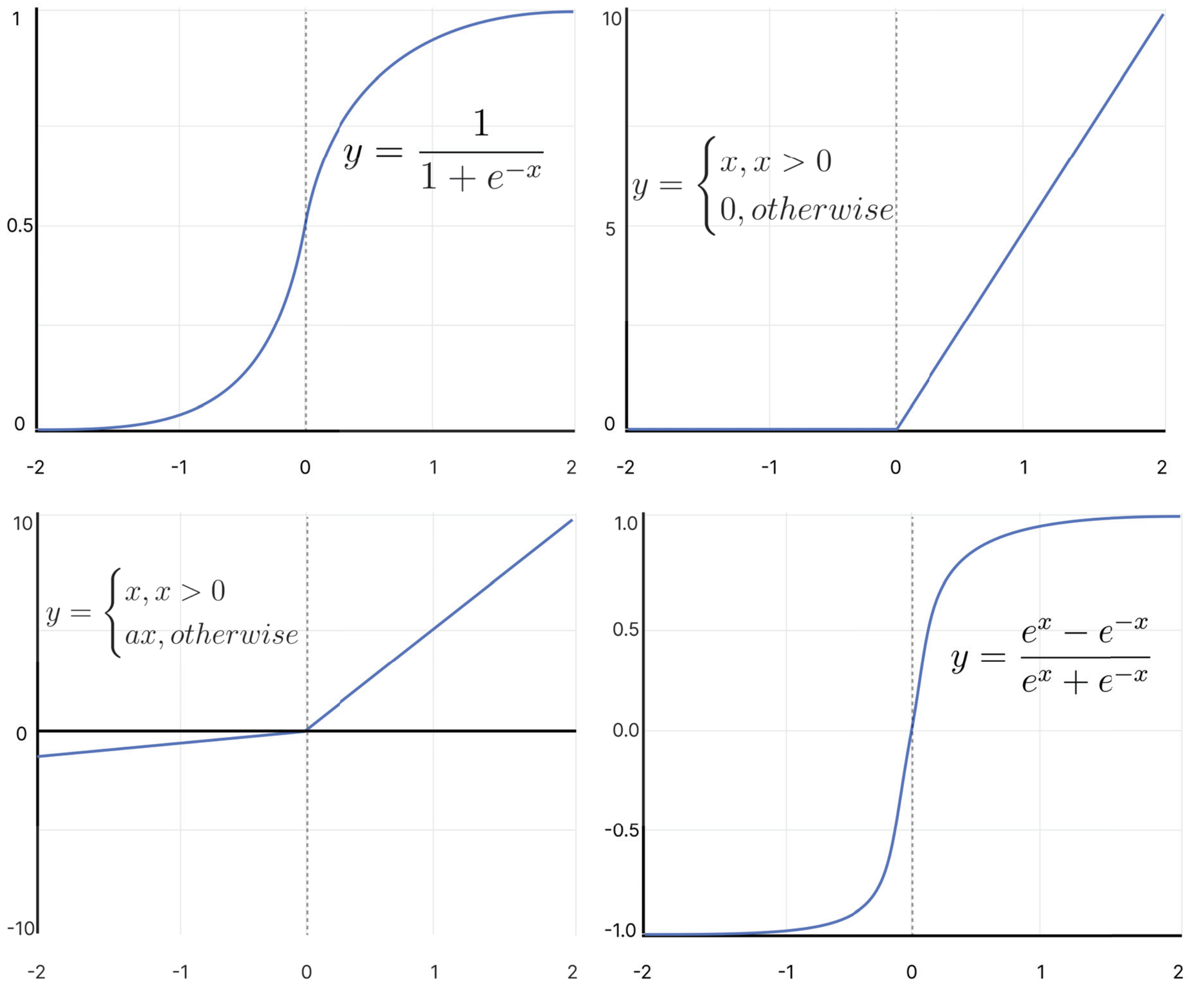


Figure 3.2: Sigmoid (upper-left), ReLU (upper-right), Leaky ReLU (lower-left), Tanh (lower-right) activation functions.

Using this gradient, we can update the weights at the nth layer, $w_{nnew} = w_n - \alpha.\frac{\partial L}{\partial z}.h_{n-1}$, where α is the learning rate and y_{n-1} is the output from the last layer. The non-trainable parameter α plays a crucial role in determining the learning speed of algorithms. If set too high, it may result in non-optimal solutions, while setting it too low can lead to extremely slow convergence. Therefore, it is advisable to experiment with various learning rates for different problems. In this method, the entire weight of the neural network is re-calibrated, and this process is iteratively repeated until the desired level of loss is achieved.

Sigmoid function

The sigmoid function produces output values within the range of 0 to 1, making it suitable for binary classification where the output is either 0 or 1. While the sigmoid function is a classic example of an activation function, it is prone to a challenge known as the vanishing gradient problem[1].

ReLU

ReLU (Rectified Linear Unit) produces an output equal to the input when the input is positive. For negative values, the output is zero. Since it assigns zero for negative inputs, certain neurons become inactive, leading to a simplification of computation. This not only reduces computational complexity but also aids in preventing overfitting and enhancing generalization.

Leaky ReLU

Leaky ReLU is a modified form of ReLU. Rather than discarding negative inputs, it produces an output with a specified magnitude. Typically, this value is small and regulated by a constant (see formula in Figure 3.2). Both the variants (ReLU and leaky ReLU) do not suffer from the vanishing gradient to the same extent as the sigmoid function. Because of this reason, this activation function has been employed in the fully connected layers of our architecture.

Tanh function

The hyperbolic tangent function, known as tanh, generates output values ranging from -1 to 1 for any given input. This characteristic proves valuable in tasks that demand a distinct differentiation between positive and negative inputs. Tanh activation is commonly applied in tasks such as recurrent neural networks (RNN), natural language processing (NLP), and speech recognition [50].

3.1.2 Artificial neural network

An artificial neural network (ANN) in deep learning consists of many such neurons connected (as shown in Figure 3.3). The initial layer is termed the input layer, while the concluding layer is known as the output layer. The intermediary layers are referred to as hidden layers. The quantity of these layers and the neurons within them is directly linked to the complexity of the model. Such neural networks are also recognized as universal function generators.

These neural networks can effectively address problems such as classification and regression like traditional algorithms, often with minimal tuning. Various configurations can be applied for classification and regression tasks. For instance, a single output for regression can predict a continuous value. In the case of binary classification, a single probabilistic value can still suffice. Moreover, multiple output nodes can be employed to represent different classes.

[1] a problem in deep neural networks where gradients become extremely small during training, making it challenging for the network to learn effectively.

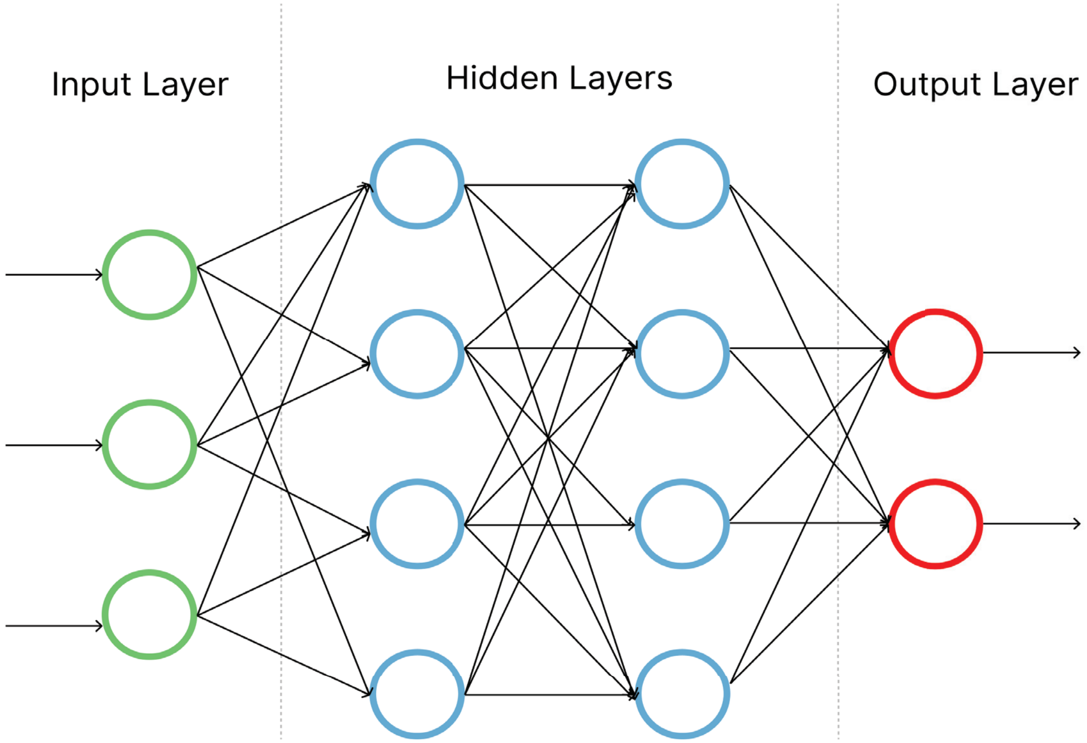


Figure 3.3: ANN architecture.

Yet, certain considerations demand caution when employing deep learning models. These algorithms operate as black boxes, making it exceptionally challenging to elucidate the reasoning behind the results, resulting in a loss of explainability. Deep learning models necessitate a substantially larger amount of data compared to traditional machine learning models. Overfitting these deep learning models is a common risk. Additionally, there is no universal rule dictating the ideal number of layers or neurons required for a specific task.

3.1.3 Loss function

A loss function (interchangeably also known as cost function or objective function) is a measure of error in the predicted and ground truth values. Its significance lies in guiding deep learning algorithms to minimize this error during the iterative process of backpropagation. The selection of an appropriate loss function is contingent upon the nature of the problem at hand. In our specific application, we opt for a loss function known as cross-entropy loss, which draws inspiration from information theory. Cross-entropy loss penalizes the model more when it is confident about incorrect predictions and rewards it when it is confident about correct ones. This characteristic makes cross-entropy loss a valuable metric for training models in classification tasks. We especially use a variant of cross-entropy loss called categorical crossentropy loss for multiclass classification. The categorical cross-entropy loss L is defined as:

$$L\tau(p,\tau y) = -\sum_{i=1}^{N}\Bigg) y_i.log\tau(p_i)\ ,\tau \tag{3.3}$$

where y is the ground truth vector, p is the predicted probability distribution produced by the model, and N is the number of classes. The loss is computed for each prediction and averaged during the training. The logarithm penalizes confidently wrong predictions than the less confident predictions as the logarithm yields large negative numbers when the prediction probability reaches close to zero. The negative sign in the equation ensures that the loss is minimized during the training process.

3.1.4 Optimizer

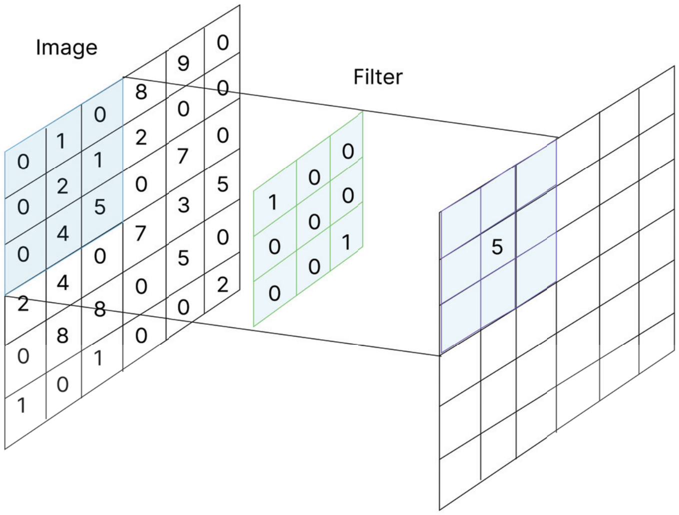
In addition to defining the loss function, it's crucial to specify the learning process in deep learning. While deep neural networks are learned by iteratively calculating and adjusting gradients, the optimization algorithm dictates how the model minimizes the loss function. One widely used optimizer is the stochastic gradient descent (SGD) [51] algorithm, which straightforwardly updates the weights to minimize the loss. The term "stochastic" refers to the randomness introduced by the algorithm, as it selectively and randomly samples a subset of data during each iteration to compute gradients and update the model parameters. For our architecture, we use an optimizer called Adam optimizer [52]. It stands for ADAptive Moment estimation. It is an adaptive learning rate optimizer that learns from historical gradients, in contrast to selecting gradients randomly. The Adam optimizer requires initialization of two moving average variables m (first momentum) and v (second momentum). To update the weights w of the model, the following equations are used:

$$m_t = \beta_1.m_{t-1} + (1+\beta_1)\,.g_t and v_t = \beta_2.v_{t-1} + (1+\beta_2)\,.g_t^2\,,\tau \tag{3.4}$$

Then corrections are calculated as $\delta m_t = \frac{m_t}{1-\beta_1^t}$, $\delta v_t = \frac{v_t}{1-\beta_2^t}$. Finally, the weights are updated as $w_t + 1 = w_t - \alpha.\frac{\delta m_t}{\sqrt{\delta v_t}+}$, where $t\tau$is the time step (or iteration), $\alpha\tau$is the learning rate, β_1 and β_2 control exponential decay rates for first and second moments, m_t and v_t are first and second moment estimates at iteration t, δm_t and δv_t are the correction terms, g_t is the gradient of the objective function concerning weights w, and is a small constant to avoid zero division.

3.2 Convolutional neural network

Spectral features alone can serve as a basis for emotion classification by employing Convolutional Neural Networks (CNNs) or ConvNets [53]. CNNs, a prevalent architecture in Artificial Neural Networks, are extensively utilized in computer vision applications like object detection, classification, segmentation, and identification. Their effectiveness in computer vision tasks stems from



$$y = Image \circledast Filter$$

$$y = \begin{matrix} 0 * 1 + 1 * 0 + 0 * 0 + \\ 0 * 0 + 2 * 0 + 1 * 0 + \\ 0 * 0 \quad 4 * 0 + 5 * 1 \end{matrix}$$

$$y = \begin{matrix} 0 + 0 + 0 + \\ 0 + 0 + 0 + \\ 0 + 0 + 5 \end{matrix}$$

$$y = 5$$

Figure 3.4: Convolution process.

their ability to capture spatial and temporal dependencies in an image through convolutional filters. Additionally, CNNs lead to a substantial reduction in trainable parameters, resulting in a notable improvement in performance metrics. The convolution process is illustrated in Figure 3.4. The term convolution comes from the convolution operation in signal processing [54], [55]. Mathematically, the convolution operation is given by $(x \tau\!* h)(t) = \int_{-\infty}^{\infty)} x(\ )h(t\tau-\ )$. In signal processing, $x(t)$ and $h(t)$ are input signals which are analogous to the image and the filter in the CNN process. * (usually asterisk or circled asterisk) denotes the convolution operation and is the integration variable representing time shift represents the amount of pooling in CNN (more about pooling in next sections).

Convolution layer

The convolution layer is the core block of CNN. In this layer, a filter (or feature detector) is sided over the image to calculate the dot product (see Figure 3.4) to give feature maps. These are the learnable parameters that learn the local features of the image.

Pooling layers

Pooling layers involve a filter moving across the image, generating a single output per stride. These layers summarize the features present in that specific location without incorporating any trainable parameters. The primary purpose of pooling layers is to decrease the number of trainable parameters within the feature maps.

Flattening

The flattening layers transform the 2-dimensional arrays of feature maps from the max pool layers into a single-dimensional array. This alteration is necessary to feed the data into the model, particularly a fully connected Artificial Neural Network (ANN).

Fully connected layers

Fully Connected Layers are the ANN that we already discussed in section 3.1.2.

3.3 Transformers

To extract text encoding, we will employ BERT encoders, a widely used transformer architecture. Transformers are neural network architectures designed for addressing sequence-to-sequence natural language processing (NLP) challenges. In sequence-to-sequence tasks, the model takes a sequence (e.g., a sequence of words) as input and produces a sequence (e.g., another sequence of words) as output.

Transformers rely on self-attention layers, making them particularly effective in dealing with long-range dependencies. To illustrate, consider the task of translating a sentence from French to English. This task demands that, regardless of the sentence's length, each word should have context and relevance to the words preceding it. Traditional architectures for sequence analysis, such as RNNs, LSTMs, and GRUs, struggle to handle larger sentences in comparison to transformers [50]. Transformers can effectively handle sequences of any length with sufficient computational power. Additionally, transformers do not necessitate the input to be in a sequential format, enabling parallelism—a challenge in models like RNN and its derivatives. The architecture of the transformer is depicted in the Figure 3.5 [50].

3.3.1 Attention mechanism

The transformer architecture utilizes the self-attention mechanism for handling sequential data. It comprises an encoder and a decoder, each composed of multiple layers. The self-attention mechanism enables transformers to grasp relationships and dependencies among each element in a sequence, irrespective of their positions. This capability empowers transformers to excel in natural language processing (NLP) tasks, enabling them to capture the context of lengthy and complex sentences. Below, we outline some essential components of transformers.

Self-attention mechanism

Self-attention, also referred to as scaled dot-product attention, serves as the fundamental element in transformers, enabling the model to assign significance to crucial segments of the input

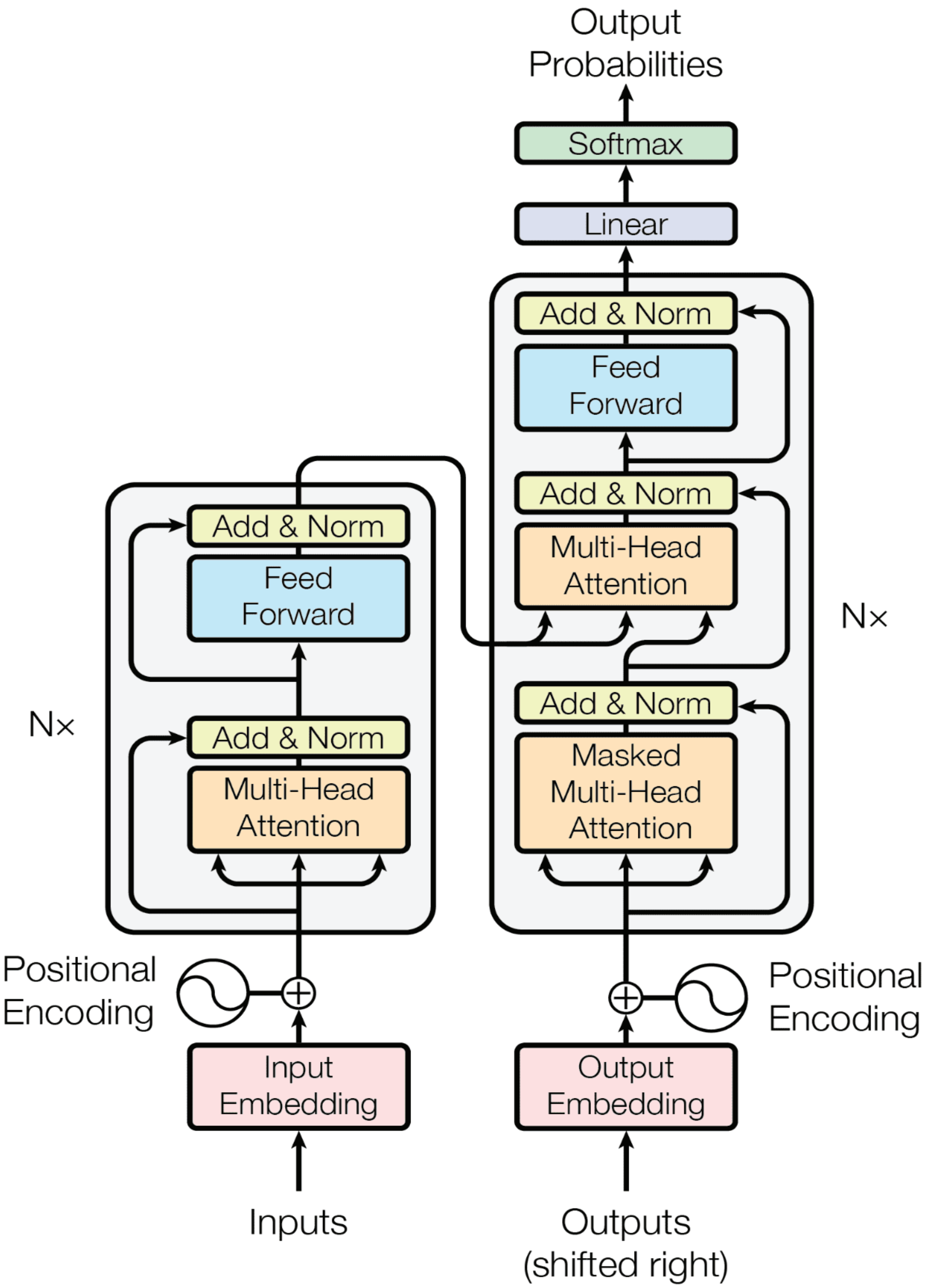


Figure 3.5: Transformer architecture.

sequence. This mechanism empowers the model to recognize and comprehend intricate relationships within sequential data. It involves three trainable parameter vectors known as query (Q), key (K), and value (V). Query represents the token's[2] importance in relation to other tokens in the sentence, the key vector helps to identify the relationship between the current token with all other tokens, and the value vector contains the information about the content of the token. Mathematically, the attention for a given token at position i is calculated as:

$$Attention(Q,\tau K,\tau V\tau)_i = softmax(\frac{Q_i.K^T}{\sqrt{d_k}}).V,\tau \tag{3.5}$$

where d_k is the dimension of the key vector. Because the output of the dot product can be a very large value, it is scaled by dividing the dot product with the square root of d_k.

Multi-head attention

In the original paper [50], the authors use eight attention heads (Figure 3.6 shows the stacking of multiple attention heads), enabling transformers to improve performance by acquiring diverse representations of K, Q, and V. Consequently, transformers can learn even more intricate dependencies. Furthermore, because these learners operate independently, parallelization becomes feasible.

3.3.2 Encoder-Decoder architecture

Transformers are commonly observed in an encoder-decoder architecture, although they can also be employed in either only-encoder or only-decoder configurations. In a machine translation scenario, the encoder takes the source language, and the decoder generates the translation. However, for our proposed framework, we exclusively utilize the encoder segment of the transformer. An architecture like BERT can be used to extract encoding from the text. Fortunately, pre-trained BERT models are readily accessible in the English language. Since our proposed model does not require us to generate texts, the decoder portion of the Transformer (see right side of figure 3.5 is not necessary.

3.4 Evaluation metrics

After the models have been trained, it is crucial to verify whether they are performing according to expectations during testing. The underperformance of the models may be caused due to various reasons, such as underfitting or overfitting [56, 57], poor model choice or poor feature selection/engineering [58]. The easiest and most comprehensive tool to visualize the performance of a binary classifier is through a confusion matrix. The confusion matrix is a special type of contingency table (as shown in Figure 3.7), where the rows and columns represent classes of

[2] a ”token” refers to a unit of text that a model can work with.

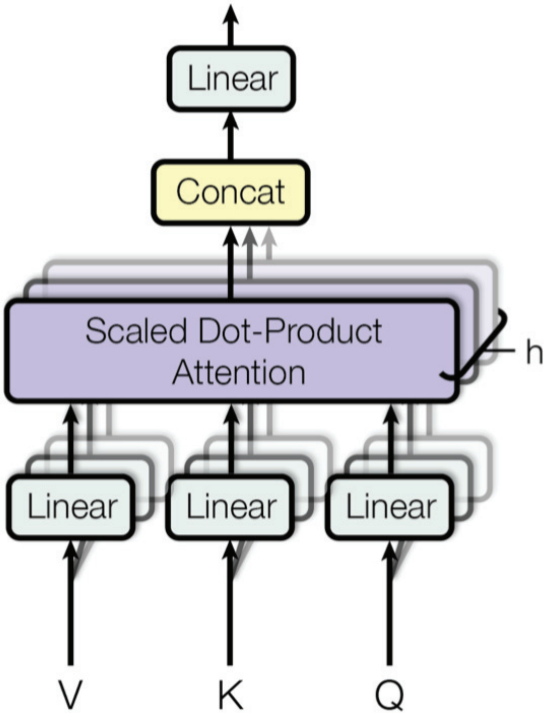


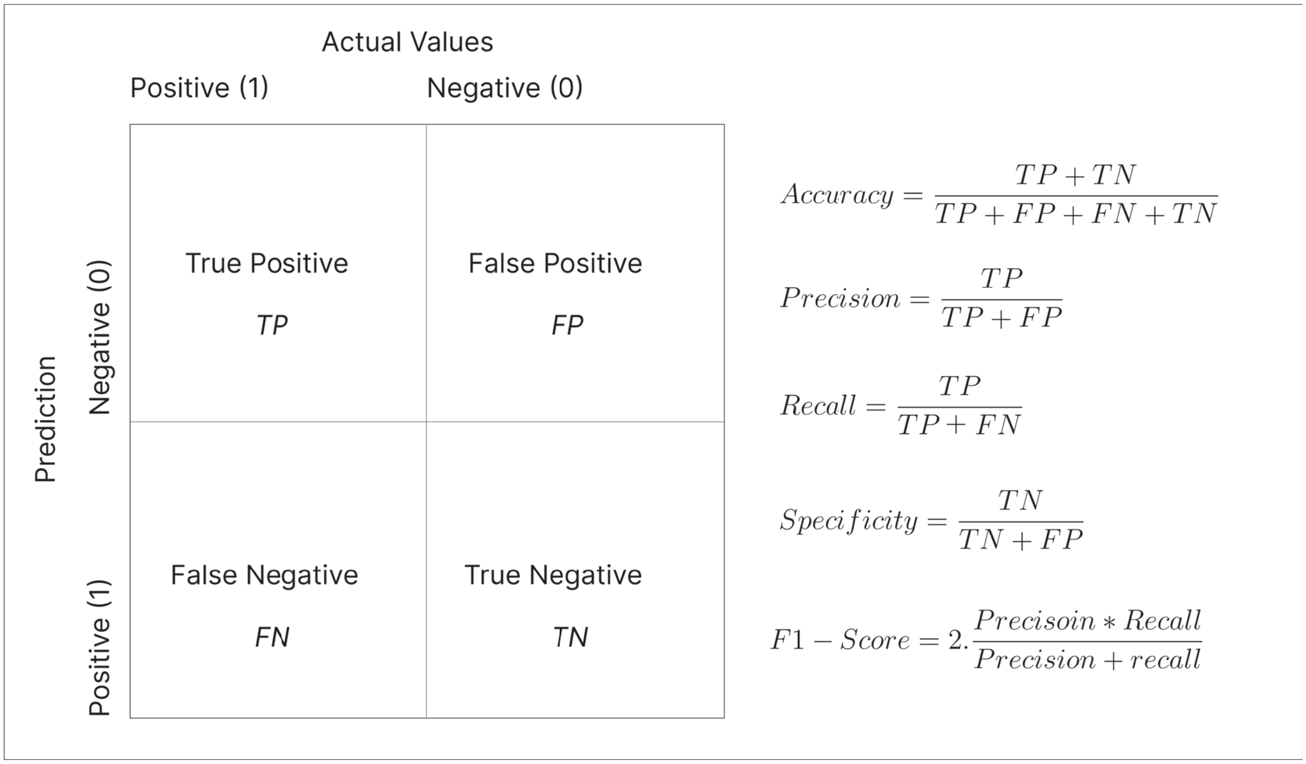
Figure 3.6: Multi-head attention.

predicted and ground truth values. An example of a confusion matrix for a binary classification is shown in Figure 3.7. For N classes, the matrix would be an NXN matrix. Some important information that can be extracted from the confusion matrix are:

- True Positive: Number of instances correctly classified as positive.
- True Negative: Number of instances correctly negative instances.
- False Negative: Number of instances incorrectly classified as negative.
- False Positive: Number of instances incorrectly classified as positive.

Figure 3.7 shows the confusion matrix for binary classification problems and derived evaluation metrics. Using these elements, we can calculate various evaluation metrics:

- Accuracy: It is the overall correctness of the classification model.
- Precision: Precision, also known as positive predictive value, represents the ratio of correctly predicted positive instances by the model out of all instances predicted as positive. It is a suitable metric when the goal is to minimize the occurrence of false positives.



	Positive (1)	Negative (0)
Negative (0)	True Positive *TP*	False Positive *FP*
Positive (1)	False Negative *FN*	True Negative *TN*

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$F1 - Score = 2.\frac{Precisoin * Recall}{Precision + recall}$$

Figure 3.7: Confusion matrix for binary classification.

- Recall (Sensitivity or True Positive Rate): Recall, or sensitivity, measures the accuracy of the model in identifying actual positive instances correctly. It is applicable in situations where the emphasis is on avoiding false negatives.
- Specificity (True Negative Rate): Specificity is the metric used to measure the ability of the model to correctly identify negative instances.
- F1 Score: The F1 score is the harmonic means of recall and precision. By incorporating both metrics, the F1 score offers a balanced assessment as recall and precision exhibit an inverse relationship.

To be consistent with the benchmark results for the IEMOCAP dataset [1], we have used weighted F1 scores as our primary evaluation metrics. The weighted F1 score is calculated using the following formula:

$$F1score = \sum_{i=1}^{n} Weight_i * F1score_i, \tag{3.6}$$

where N is the number of classes, $Weight_i$ is the weight of each class based on their distribution, and $F1score_i$ is the individual F1 score calculated for each class using the equation shown in Figure 3.7.

3.5 Summary

In this chapter, we explore the fundamental concepts of deep learning and their application to multimodal analysis tasks. We begin by introducing deep neural networks, a powerful class of machine learning models inspired by the structure of the human brain. We delve into the intricacies of transformers, a specialized architecture that excels at natural language processing tasks. Next, we investigate convolutional neural networks, particularly suited for image recognition applications. To effectively analyze audio data, we introduce audio spectrograms, a visual representation of the audio signal. We discuss Mel frequency spectrograms, a variant tailored to human perception. To process audio spectrograms efficiently, we introduce audio spectrogram transformers, a specialized transformer architecture. Finally, we explore evaluation metrics and tools for assessing the performance of machine learning models. Throughout the chapter, we emphasize the theoretical side of these concepts, showcasing their ability to tackle complex auditory and textual processing tasks. We also discuss the categorical cross-entropy and Adam optimizer used in the proposed model architecture.

Chapter 4

Multimodal focused attention architecture

Summary: In this chapter, we will delve into the proposed model, exploring its architecture, components, and underlying principles. We will dissect the model's layers, unraveling the intricate mechanisms that enable it to learn and make predictions. Along the way, we will shed light on the model's strengths and limitations, providing insights into its suitability for various applications.

Key topics: Focused attention, Dataset, Preprocessing.

Organization: The chapter is structured as follows: dataset and preprocessing techniques are discussed in 4.1 used in 4.2. Then in 4.3 we discuss basic setup for experimentation. Further in 4.4 and 4.5 we describe the unimodal and proposed model, exploring its architecture, components, and underlying principles. Finally, we also discuss how the ablation study is conducted in 4.6.

4.1 Dataset

The dataset we use for this extension is the Interactive Emotional Dyadic Motion Capture (IMEOCAP) dataset [1]. The dataset stands as a rich resource in the field of emotional communication research. Compiled at the SAIL lab at the University of South Carolina, this database is a curated collection of approximately 12 hours of audiovisual content. Encompassing video footage, speech recordings, facial motion capture data, and text transcriptions, the dataset is designed to offer a comprehensive understanding of human emotional expression. The dataset offers a diverse range of modalities for analysis, including motion capture face information, speech recordings, videos, head movement, head angle information, and dialog transcriptions. Notably, it provides alignment at various linguistic levels, such as word, syllable, and phoneme. The IEMOCAP database comprises dyadic sessions where actors participate in both improvised and scripted scenarios.

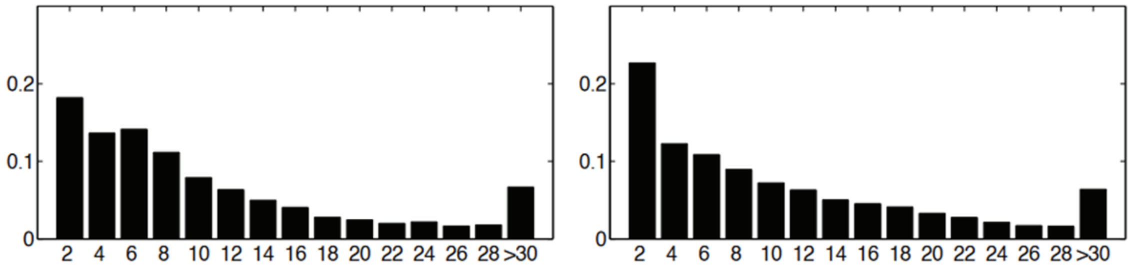


Figure 4.1: Multi-Histogram showing the number of words per turn in percentage for the scripted and spontaneous session (left to right) [1].

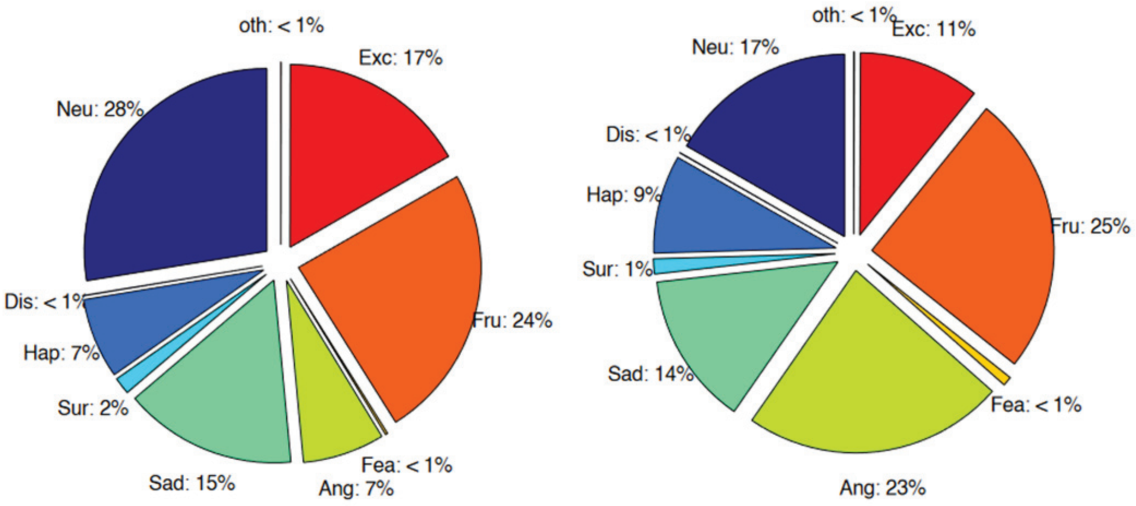


Figure 4.2: Distribution of emotion category in scripted and spontaneous sessions (from left to right) [1].

These scenarios are carefully selected to elicit a diverse range of emotional responses, offering a detailed exploration of human expression. What sets this database apart is not just its extensive content but also its integration of various modalities, including speech, facial expressions, and body language. The database's value is further elevated by meticulous annotations from multiple annotators. Emotions are categorized into labels like anger, happiness, sadness, and neutrality, while additional dimensional labels such as valence, activation, and dominance provide a more nuanced understanding of the emotional dynamics captured in the dataset. For professionals and researchers in the field of multimodal and expressive human communication, the IEMOCAP database proves to be an indispensable resource. With its comprehensive motion capture data, utilization of interactive scenarios for genuine emotional responses, and considerable size, the dataset stands out as a noteworthy contribution to the available resources in the research community. This collection of data holds the potential to drive ongoing investigations and progress in the understanding and modeling of human emotional communication. Interested parties can obtain and download the dataset through their website[1] by making a request. Two datasets are available: one includes visual recordings, while the other does not. To simplify reproducibility and for demonstration purposes, we opt to utilize the dataset that exclusively contains auditory recordings.

The 12-hour audiovisual recordings have been divided into sentences or turns, resulting in 10,039 segments. Among these, 5,255 turns are scripted, and 4,784 turns are from spontaneous sessions. The annotations include categorical labels for emotions (such as anger, happiness, and sadness) and continuous value-based annotations (valence, activation, and dominance). In this chapter, our focus is on emotion classification using categorical labels. The average duration of the turns is 4.5 seconds. The number of words per turn is shown in Figure 4.1.

The data distribution for each emotion category is generally well-balanced, except for surprise and fear. The distribution patterns for scripted and spontaneous sessions are depicted in Figure 4.2. The charts indicate a higher proportion of individuals expressing positive emotions like happiness and excitement compared to negative emotions such as sadness and anger. Fear is the least prevalent emotion category, accounting for less than one percent of the total emotion categories in both scripted and spontaneous sessions.

4.2 Preprocessing

Preprocessing is a very important step for building deep learning models. For our approach, we under-sampled the majority classes as imbalanced classes can introduce bias in classification. To achieve this, we use under-sampling functions from the imblearn library [59]. Following the undersampling process, approximately 1100 samples remained for each class. Most of the related work only uses major emotions like happiness, sadness, neutrality, anger, and frustration. Thus, in total, we have 5500 records after undersampling. It is not necessary to convert these categorical classes to numerical labels, but we map the mentioned classes to numerical values from 0 to 4 respectively for reproducibility. The average length of sentences after undersampling is illustrated in Table 4.1. Since we are using pre-trained encoders from BERT, it was not necessary to perform preprocessing for the text data. The audio samples also did not require preprocessing before extracting the Mel frequency banks.

4.2.1 Extracting spectral features from audio

Research [60, 61, 62, 63, 64] show that the use of spectrograms significantly aids in waveform analysis and recognition. The spectrograms can capture minute changes in frequencies. This consequently allows to detection of emotional cues that may not be apparent from the raw signals. The extracted spectral features may be used in combination with CNNS [63]. Recently, advance-

[1] https://sail.usc.edu/iemocap/index.html

Table 4.1: Average length of sentences in each class after undersampling.

Label	Average length
happiness (0)	54.80
Sadness (1)	58.06
Neutral (2)	54.10
Anger (3)	63.50
Frustration (4)	71.49

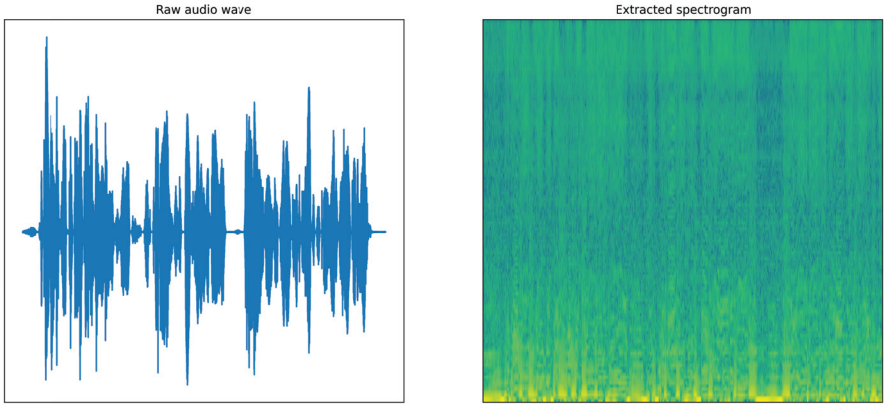


Figure 4.3: Conversion of raw audio (left) to spectral representation (right).

ments in visual transformers (VIT), research have also shown that audio spectrogram transformers (AST) [42],[65] are better able to capture both emotional cues and temporal information.

We are accustomed to seeing audio in the form of waveform (see left of Figure 4.3). The waveform shows a change in the signal's amplitude (y-axis) over time (x-axis). In contrast, spectrograms illustrate the alterations in frequency over time, using the x-axis for time, the y-axis for frequency, and brightness or color to signify amplitude. Spectrograms provide a detailed view of the frequency components of a signal across time, making them valuable for tasks such as audio processing and speech recognition, offering a more comprehensive perspective compared to waveform. In waveforms, slight variations in higher frequency ranges are typically less noticeable compared to equivalent changes in lower frequency ranges. However, spectrograms depict frequency changes consistently across all frequency ranges, providing a spectral representation that aligns with human perception of waveforms. In our approach, we use Mel-filter banks as a feature and pass it to subsequent layers (with and without fusing it to other features).

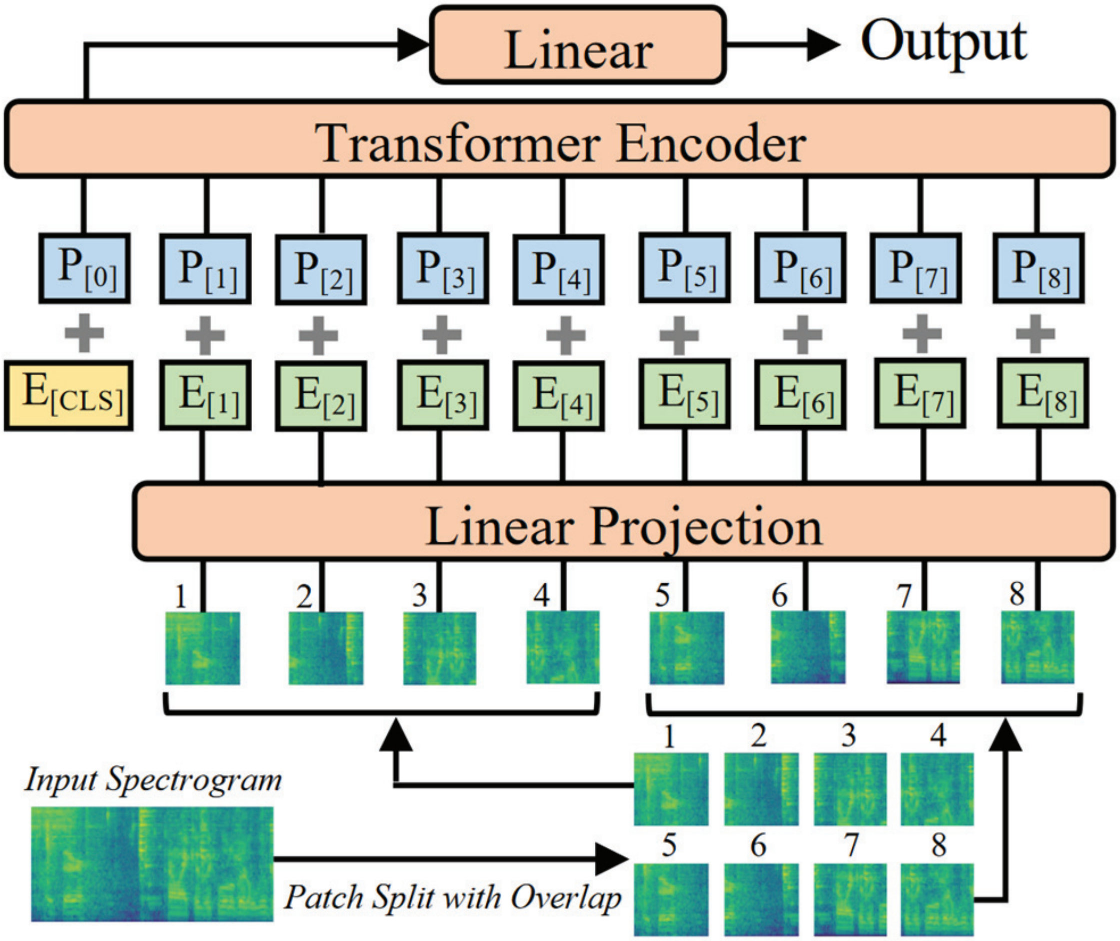


Figure 4.4: Audiospectrogram transformer architecture for extracting features from spectrograms.

Mel frequency spectrograms and audio spectrogram transformer

Mel frequency spectrograms are one of the methods used to create spectral representation from audio waves. Let's consider an audio waveform $x(t)$, where t ranges from 0 to t seconds. First, we apply the hamming window function h(t) to the waveform, to obtain the windowed segments $x_h : x_h\,(t) = x\tau(t) * h\tau(t)$. The hamming window is defined as $h\tau(t) = 0.54 - 0.46cos\tau\left(\frac{2\pi t}{N-1}\right),\big)$where $0\ <=\ t\ <=\ N\tau - 1$, $N\tau$is the window length. Other window functions may also be used, but the Hamming window function is a standard one for speech-processing tasks. The windowed segments are divided into frames with some time duration and shift. Then, the discrete Fourier Transform (DFT) of each window is calculated to obtain the magnitude spectrum $X(f)$, where f is the frequency index: $X\tau(f) = DFTx_h\,(t)$. It is necessary to apply the window function before performing the Fourier transformation to address the assumption made by the Fourier transform (like data being infinite) and to reduce spectral leakage. Then we apply a Mel-filterbank to $X(f)$ to obtain Mel-filterbank energies E_m.

$$E_m = \sum\big)H_m\,(f)\,.\tau|X\tau(f)|^2\,,\tau \tag{4.1}$$

where H_m are triangular filters that are applied to approximate the frequency of audio signals to the frequency response of the human auditory system. The parameters of this filter are determined based on the Mel Scale, which reflects the nonlinear relationship between frequency and perceived pitch. The audio can be converted to the Mel scale from the decibel scale and vice versa using the following relationship:

$$m\tau = 2595log_{10}\left(1\right) + \frac{f\tau}{700} \quad andf\tau = 700\left(10^{m/2595} - 1\right),\tau \tag{4.2}$$

where m and f are Mel and Hertz scale.

The original paper [42] describes an audio spectrogram transformer (AST) as a convolution-free and purely attention-based model. The model architecture is shown in Figure 4.4. The AST utilizes Mel-filter bank features described in Mel frequency spectrograms. Instead of relying on CNN to process spectral representation, the AST architecture relies on the sequential information provided by the spectral features. The AST architecture is also independent of the input sequence length. The input spectrogram is divided into a sequence of N 16X16 overlapping patches. Embeddings are generated from the linear projection of these patches using transformer-inspired architecture. The output of this transformer can be used for downstream tasks and thus is used in our proposed model.

4.2.2 Extracting text encoding using BERT

Bidirectional Encoder Representations from Transformers (BERT) [66, 67] is an architecture based on transformers by Google AI. BERT is specifically designed for pre-training deep bidirectional representations by leveraging unlabeled text, taking into account both left-to-right and right-to-left context across all layers. As a result, the pre-trained BERT model can attain state-of-the-art performance in diverse tasks, such as question answering and language inference, with minimal adjustments to the task-specific architecture. Often, only the inclusion of a single output layer is needed for fine-tuning. This training approach is commonly referred to as transfer learning.

To load the dataset, we utilized PyTorch's Dataset and Dataloader classes. During dataset loading, we extract the Mel frequency-banks from the raw audio signals using the ASTFeature-Extractor module. We limit the length of the signals to 1024 and produce 128 Mel frequency bins. We also normalize the Mel features using the mean and standard deviation of -4.2677393 and 4.5689974. These are the default parameters for the feature extractor class. The sampling rate is set to 16Khz which is also the sampling frequency of the input audio signals. The spectrogram that is derived has dimensions of [1024, 128].

For the text transcripts, we use the pre-trained BERT tokenizer. Little to no preprocessing is required before this step. We use the uncased BERT encoder, which is trained on large unlabelled English dataset. The maximum limit for the encoder is chosen to be 50 which is close to the average length of the texts. Both preprocessing (for audio and text) is done on the fly using the Dataset and Dataloader classes.

4.3 Experimental setup

The implementation of the proposed model is done using the PyTorch lightning framework, which is one of the popular deep learning frameworks based on Python. We utilize a pre-trained BERT-base-uncased encoder and a pre-trained AST model from the Hugging face repository. We chose Lightning over vanilla PyTorch due to its high modularity and reproducibility. We also use PyTorch's audio spectrogram extractor to extract mel-filterbanks. To train the model we use a single-node HPC[2].

4.4 Unimodal learning

Our experiments involve unimodal learning with both audio and text datasets. For the first experiment using the audio dataset, we utilize custom CNN models, where we extract features from the spectral data using specialized CNN filters. We use two convolution layers followed by maxpool, activation, and normalization layers. These features undergo flattening and are then processed through a feedforward network. In a separate experiment, we employ pre-trained AST models for audio. In this strategy, we extract Mel-filter banks with a sampling rate set to 16Khz and 128 Mel-bins. The max length of the audio signals is clipped to 1024 and the output is normalized. The resulting filter banks are of the shape (1024, 128). This is the input for AST transformer architecture initialized with a finetuned audioset dataset. We modify the model by

[2] Computation was performed on Lawrence Supercomputer at the University of South Dakota awarded by NSF.1626516.

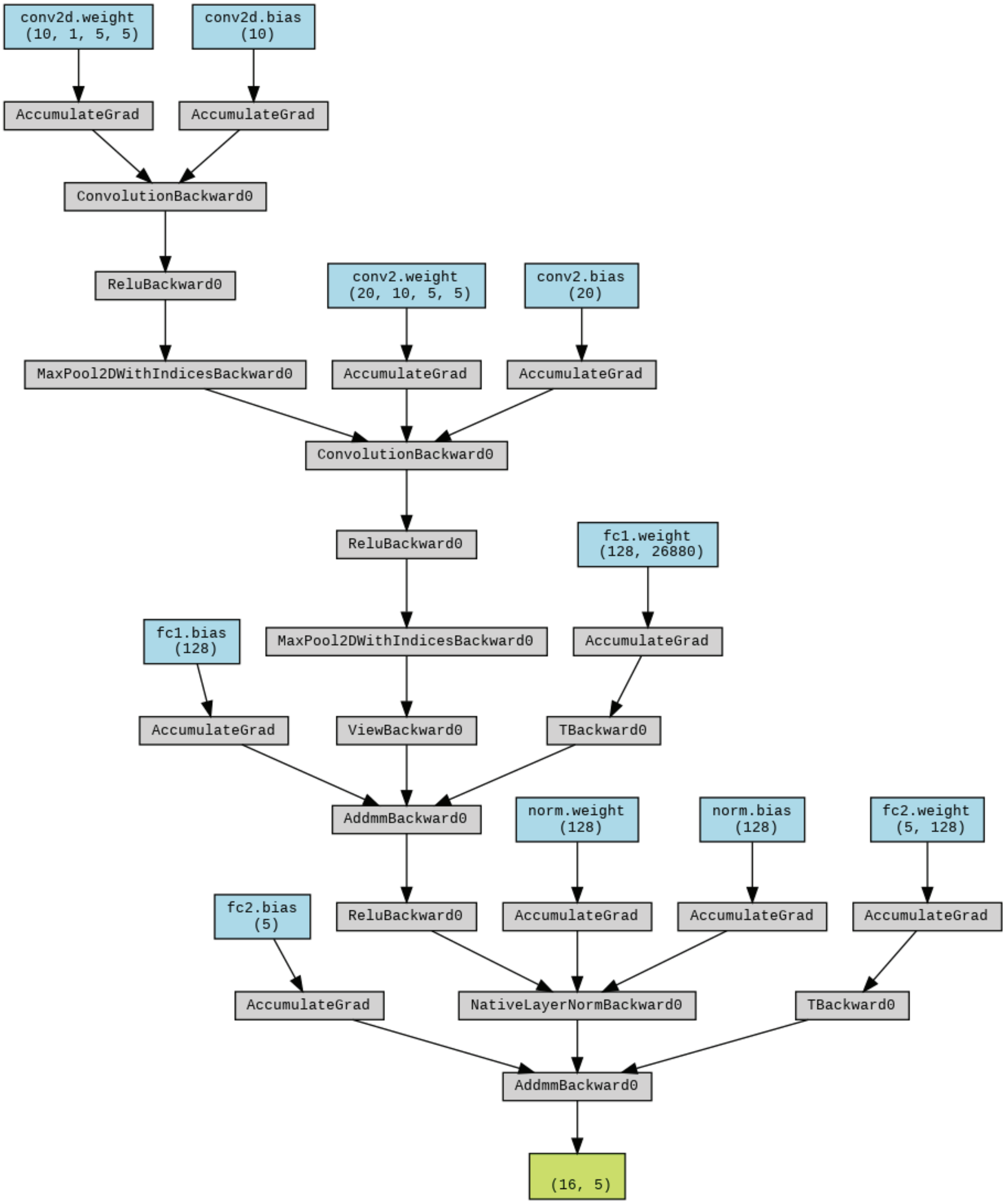


Figure 4.5: 12 Audiospectrogram model architecture using CNN (only trainable parameters).

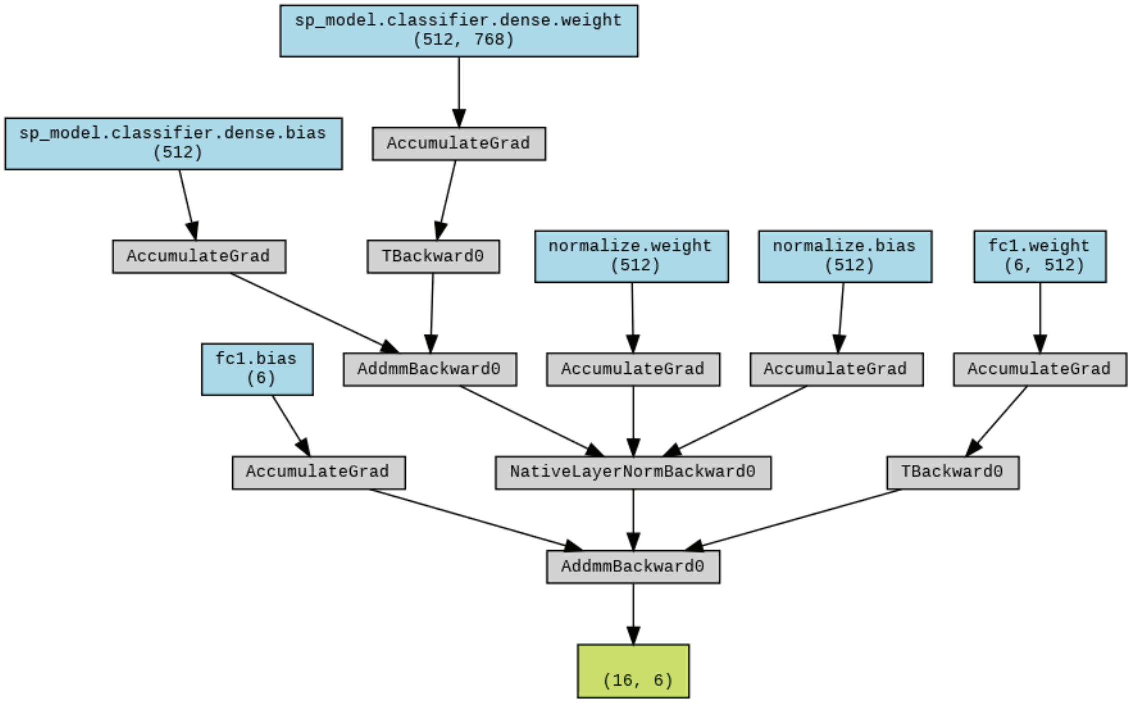


Figure 4.6: Audio spectrogram model architecture using AST (only trainable parameters).

replacing the last layer and introducing a custom linear layer that aligns with the total number of emotion classes, tailoring it to our specific classification needs.

For textual modality, we use the pre-trained BERT encoder. The inputs for the BERT encoder model are the tokenized sentences. BERT provides the weights for both the last hidden layer and the pooled layer in its outputs. The pooled layer, often favored for classification tasks, offers a lower-dimensional representation and is considered a general overview of the input sequence. This reduced dimensionality and summarization make it a commonly employed choice for classification purposes. The output for pre-trained BERT is 764 encoded vectors. The pooled layers are passed through the feed-forward network with five outputs.

The trainable parameter for the considered text model is shown in Figure 4.7. Compared to other architecture, this model has only 8.6K trainable parameters. This is because we leverage transfer learning techniques to extract the vector embeddings of the sentences. Thus, the need for training an encoder is eliminated.

4.5 Focused attention architecture

We propose a multimodal learning model using audio spectrogram transformers and the BERT language model. The architecture is shown in Figure 4.8. The overall architecture is implemented

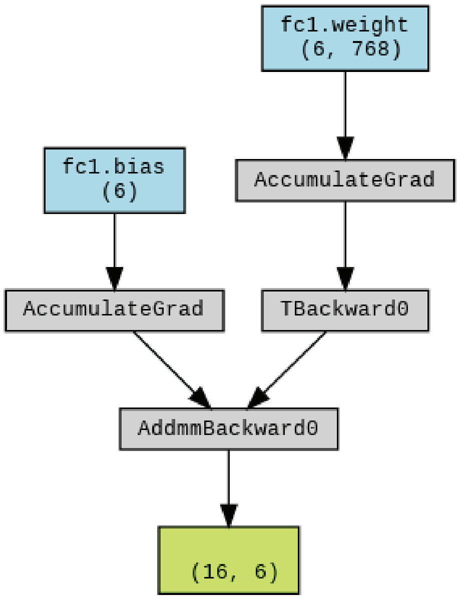


Figure 4.7: Language model architecture (only trainable parameters).

using PyTorch Lightning. We have taken advantage of pre-trained BERT and pre-trained AST classifiers from Huggingface. To process raw audio input, we have employed Torchaudio and for reading textual data, we have used Python's basic IO functions. The input for the architecture is 16Khz raw audio (in .wav format) and their respective text transcripts. The audio is converted to spectral representation, i.e., mel-filterbanks. The audio spectrogram is a visual representation of the audio data, showing the frequency and intensity of the sound waves over time. It is created by performing a Fourier transform on the audio data. Then the audio feature extractor extracts features from the audio spectrogram. These features are numerical quantities that represent the characteristics of the audio. The audio transformer transforms the audio features into a format that is suitable for the machine learning algorithm that will be used. The feature from audio is fused together with encoded text (using BERT). This helps the machine learning algorithm to learn more complex relationships between the data and the target variable. Then a fully connected combines all the input features into a single output. The output of the fully connected layer is then fed to a classifier, which predicts one of the target emotions: neutral state, frustration, happiness, sadness, and anger.

The detailed feed-forward architecture of the proposed model is shown in Figure 4.8 and Figure 4.9. The feature extracted from the spectrogram undergoes processing through linear layers comprising 128 neurons each. Simultaneously, the encoded textual information undergoes a similar transformation through a linear layer consisting of 128 neurons. The reason behind the reduction of layers to 128 neurons is to balance the contribution of each modality. The extracted audio spectral features $F_{spec} = [fs_1, fs_2, \ldots, fs_{128}]$ and text features $F_{text} = [ft_1, ft_2, \ldots, ft_{128}]$ are fused or concatenated and passed through an additional linear layer containing 256 neurons.

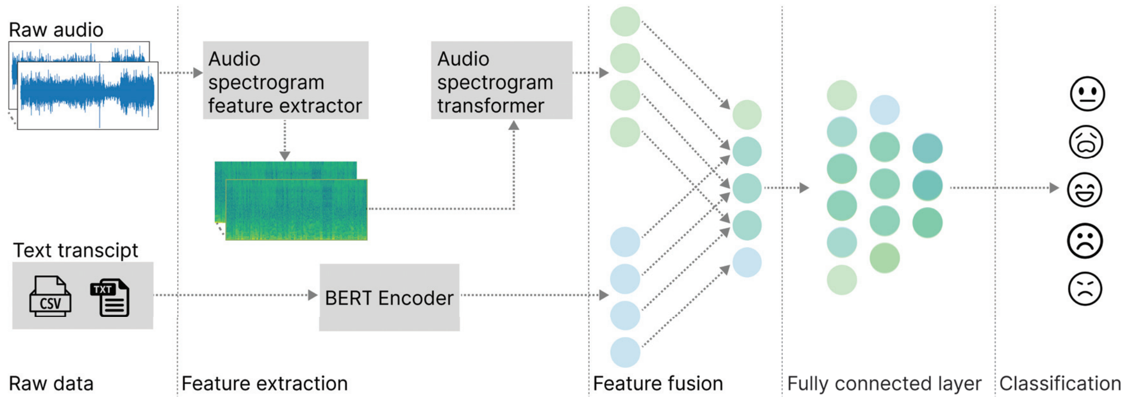


Figure 4.8: Proposed multimodal learning architecture.

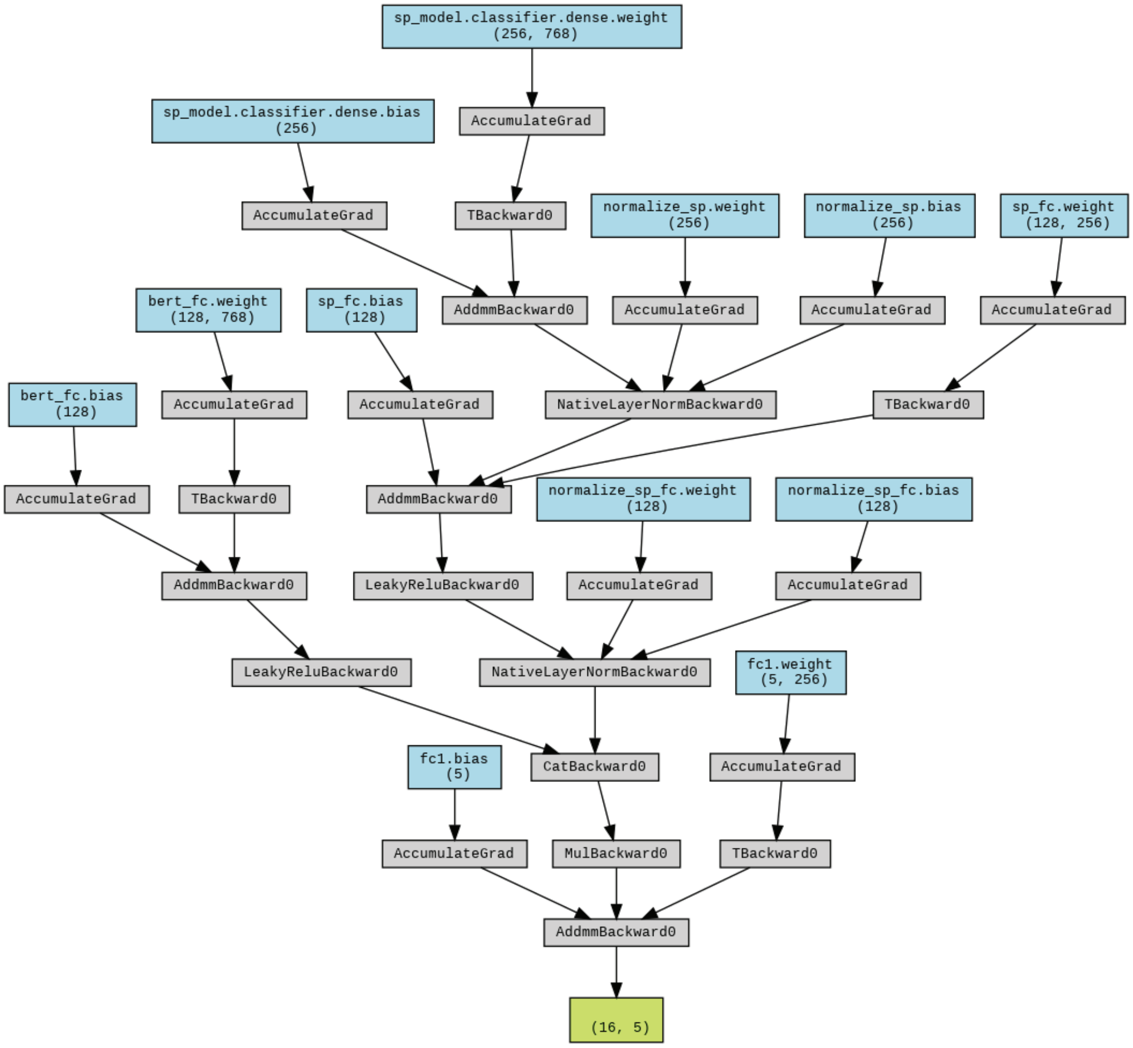


Figure 4.9: Multimodal architecture using text and audio modality (only trainable parameters).

Table 4.2: Number of trainable and non-trainable parameters for each model considered for experimentation.

Model architecture	No. of non-trainable parameters	No. of trainable parameters
Audio (Custom CNN)	0	3.4 M
Audio (AST)	86.2 M	398 K
Text (BERT)	109 M	3.8 K
Multimodal (AST+BERT)	195 M	331 K

i.e., The concatenated features $F_{\mathcal{T}}= F_{spec} + F_{text}$ is the fused features. Finally, an output layer with 5 classes is introduced to complete the model. Table 4.2 lists the number of trainable and non-trainable parameters for each of the models considered for experiments.

4.6 Ablation study

We perform an ablation study and investigate each of the model architectures listed in Table 4.2. We have tried to keep the models as consistent as possible for the ablation study. In evaluating the models, we compute the average F1 score and loss metrics across all experiments for each model. The goal of this study is to analyze each model architecture individually and investigate the results of combining diverse architectures. The IEMOCAP dataset is employed for this purpose.

Chapter 5

Results and discussion

Summary: In this chapter, we present the results for unimodal and proposed focused attention architecture. We discuss and disect the results obtained in ablation study as well as compare our proposed architecture with benchmark models on IEMOCAP dataset. Each model is cross-validated 5 times with an 80:20 split ratio for training and validation. We display the mean and standard deviation of the F1 score for each run. The evaluation of our models is based on the weighted F1 score, aligning with the metric utilized by the IEMOCAP benchmark.

Key topics: Results, Discussion, Analysis.

Organization: The chapter is organized as follows: In 5.1 and 5.2 unimodal and proposed focused attention architecture is discussed. Then deliberation on research limitation and future work is presented in 5.3 and 5.4 respectively. A summary of F1 evaluation metrics is shown in Table 5.1.

5.1 Unimodal approaches

The results of the experiments for all the architectures are presented in table 5.1 At first look, both the audio models show promising results that significantly score higher than the SOTA architecture [15] The state-of-the-art (SOTA) architecture achieves a notable score of 71.77 on the weighted F1 score metric, while our audio-based architecture surpasses 80 on the F1 score metric. Despite these seemingly impressive results, a more in-depth examination reveals that the audio model is generating predictions with a lack of confidence. The validation set does not conclusively indicate whether the model is overfitting, as both the validation loss and F1 score show improvement over time. However, there are concerns that the model may not generalize well to real-world datasets, indicated by the escalating loss function for these audio models. This problem is more evident in custom CNN-based architecture than in the AST-based model. This problem is also illustrated in Figure 5.1.

To delve into the underlying issues, we can scrutinize the loss function equation (see Figure 5.1. This function involves taking the logarithm of the prediction probabilities. As the predicted

Table 5.1: Average length of sentences in each class after undersampling.

Model architecture	Average F1 score	Standard deviation of F1 scores
Audio (CNN)	0.83	0.009
Audio (AST)	0.82	0.005
Text (BERT)	0.42	0.010
Multimodal (AST+BERT)	0.67	0.001

probability approaches zero, the loss increases substantially. This suggests that while the models are enhancing performance metrics, they are simultaneously making fewer assured predictions as the iterations progress. Another reason for such performance could be the high number of trainable parameters for a relatively smaller training dataset. The custom CNN model with 3.4 million trainable parameters has the highest F1 score, but its loss function exponentially degrades compared to the AST model with just 398 K parameters. In contrast, BERT-based models exhibit greater stability, although their performance, as indicated by the F1 score metric, is comparatively lower. Both the validation F1 score, and validation loss follow a typical trend, yet the F1 score hovers just above 40, a significant deviation from benchmark results. This subdued F1 score can be attributed to the smaller number of trainable parameters, which stands at a mere 3.8K. This reduction in trainable parameters is achieved through the implementation of transfer learning techniques. Specifically, the last layer of the pre-trained BERT model is removed, and a linear layer is subsequently added to the output of the pre-trained BERT encoder. The observed normal behavior in the training process of these BERT-based models, marked by stable trends in validation metrics, is a positive indication. The strategy of leveraging transfer learning not only reduces the overall number of parameters but also contributes to the model's stability during training.

5.2 Focused attention model

The proposed multimodal architecture with auditory and textual modality offers a fine tradeoff between the independent audio-based and text-based models. Both the F1 score and loss metrics for validation sets improve throughout the iterations. The evaluation metric is comparable to the benchmark result achieving .67 on the F1 score metric. This puts the proposed model in the top 10 of the benchmark results using auditory and textual modality. The comparison results are shown in Table 5.2.

From Table 5.2, we can see that the proposed model is still behind the SOTA architectures. However, our model achieves comparable results (places third) among the audio-textual model architectures. Through our research, we can demonstrate that attention-based architectures used in parallel with spectrographic features can achieve comparable results with SOTA architectures. We also demonstrate the superior performance of convolution-free models over classical CNN

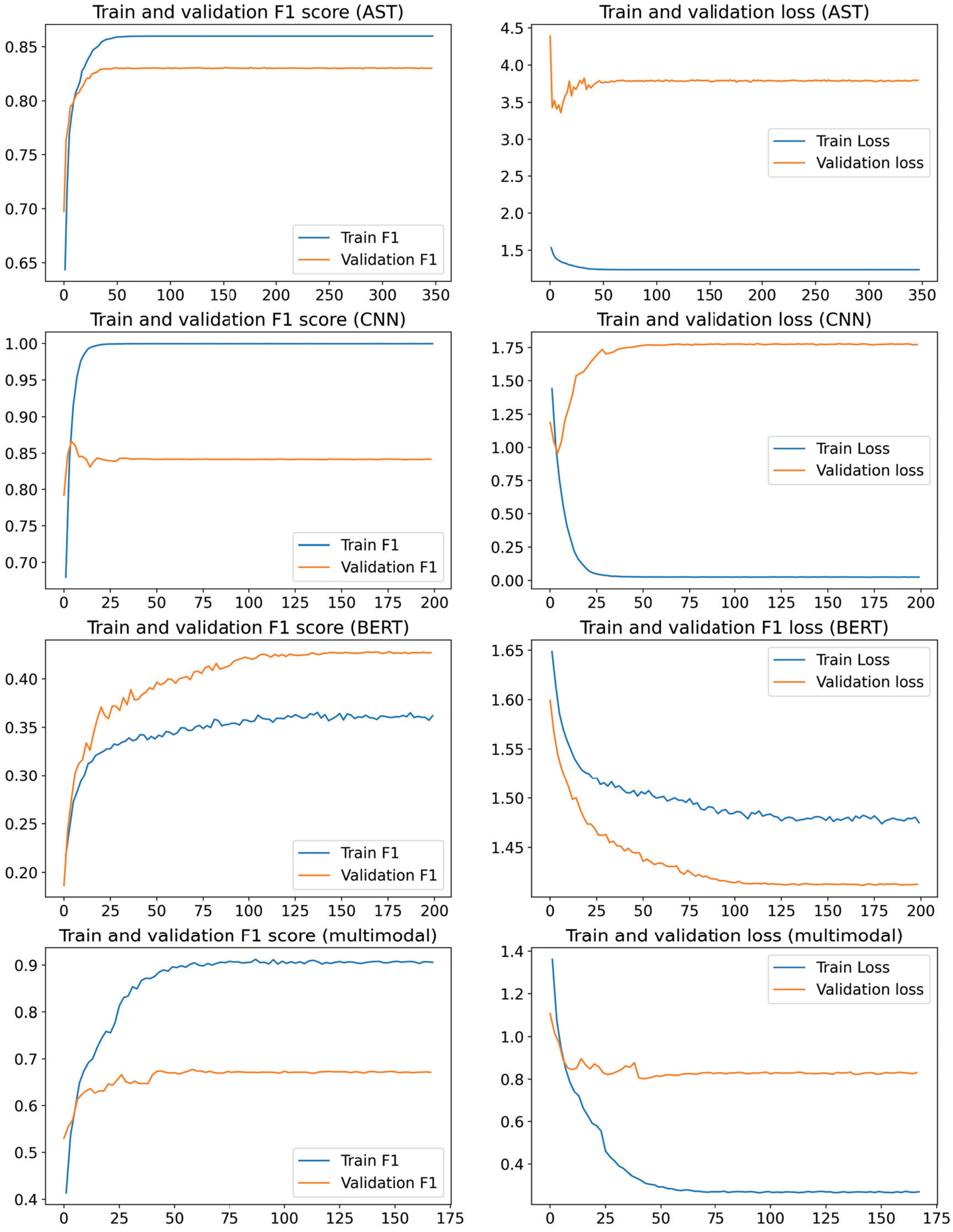


Figure 5.1: Validation and training F1 score and loss (averaged results from cross-validation of 5 experiments).

Table 5.2: Average length of sentences in each class after undersampling.

Model	F1 score	Audio + Text
EmoCaps [15]	71.77	71.39
InstructERC [34]	71.39	-
CFN-ESA [68]	71.04	68.46
UniMSE [69]	70.66	-
M2Fnet [40]	69.86	66.32
Proposed model	67.78	67.78

methods for spectrogram analysis for emotion recognition in speech. Additionally, our research demonstrates that multimodal learning exhibits increased resilience and produces predictions with higher confidence levels.

5.3 Research limitation

The study utilizes openly accessible datasets, leading to a lack of control over both the quality and quantity of the data. Our findings suggest that while intricate models can effectively capture nuanced emotional cues, they exhibit sensitivity to overfitting. Conversely, simpler models may struggle to accurately capture concealed emotional nuances. The necessity for additional data sources, preferably from primary sources, is crucial for fine-tuning the model. Transfer learning expedites and improves model convergence, but it's essential to consider that pre-trained models were designed for comparable purposes. A larger dataset and the capability to train a custom network, rather than depending on pre-trained models that may not align with the specific problem statement, could lead to enhanced model performance. As trained facial marker models were not accessible, their integration into our architecture was not feasible. Developing a custom model would have resulted in architectures with millions of parameters, an impractical approach due to the constraints posed by the dataset's limitations. The inclusion of video datasets also poses similar problems. Our investigation exclusively centered on the IEMOCAP dataset, which is specific to the English language. Consequently, we were unable to assess whether the model would exhibit comparable performance across various languages and datasets.

5.4 Future work

Most state-of-the-art (SOTA) models optimize their performance by utilizing tailor-made datasets. Likewise, our approach stands to gain substantial advantages through fine-tuning with a custom dataset sourced from primary or secondary channels. Moreover, the expanded dataset will enable the integration of additional modalities into our methodology. Additionally, incorporating datasets featuring diverse languages and scenarios is vital to ensuring the resilience of the architecture, representing an intriguing avenue for further research. The architecture can also be tuned by experimenting with different numbers of fully connected layers, and other pre-trained models. To tackle the problem of data availability, we can apply techniques such as active learning [70]. Using active learning, we do not require a lot of data and we also eliminate the need for expensive annotators.

Chapter 6

Conclusion

In conclusion, our thesis introduces a novel multimodal attention focused approach to emotion classification. This innovative method leverages fully attention-based mechanisms, integrating spectral and text embeddings to enhance the model's understanding of emotional content. The architecture is intentionally crafted for the development of robust models, especially tailored for scenarios with smaller datasets, eliminating the necessity for fine-tuning datasets. The integration of audio spectrogram transformers and BERT encoders in our methodology facilitates the extraction of meaningful features from both audio and text data, contributing to a comprehensive understanding of emotional nuances.

Our study culminates in the presentation of results using the F1 score metric, a robust measure of precision and recall. Notably, our approach achieves performance levels comparable to well-established state-of-the-art (SOTA) models in the field of emotion classification. This validation underscores the efficacy and competitiveness of our proposed multimodal framework, positioning it as a noteworthy contribution to the evolving landscape of sentiment and emotion analysis.

References

[1] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. 2007.

[2] Tetsuya Nasukawa and Jeonghee Yi. Sentiment analysis: Capturing favorability using natural language processing. pages 70–77. Association for Computing Machinery, 2003.

[3] M Shakeel, S Faizullah, T Alghamidi, and I Khan. Language independent sentiment analysis. pages 1–5. IEEE Computer Society, 2 2020.

[4] S Pham and B Kieu. Sentiment analysis for vietnamese. pages 152–157. IEEE Computer Society, 10 2010.

[5] Nikita Jain, Vedika Gupta, Shubham Shubham, Agam Madan, Ankit Chaudhary, and KC Santosh. Understanding cartoon emotion using integrated deep neural network on large dataset. *Neural Comput. Appl.*, 34(24):21481–21501, 2022.

[6] Meishu Song, Andreas Triantafyllopoulos, Zijiang Yang, Hiroki Takeuchi, Toru Nakamura, Akifumi Kishi, Tetsuro Ishizawa, Kazuhiro Yoshiuchi, Xin Jing, Vincent Karas, Zhonghao Zhao, Kun Qian, Bin Hu, Björn W Schuller, and Yoshiharu Yamamoto. Daily mental health monitoring from speech: A real-world japanese dataset and multitask learning analysis. pages 1–5. IEEE, 6 2023.

[7] Saad Awadh Alanazi, Ayesha Khaliq, Fahad Ahmad, Nasser Alshammari, Iftikhar Hussain, Muhammad Azam Zia, Madallah Alruwaili, Alanazi Rayan, Ahmed Alsayat, and Salman Afsar. Public's mental health monitoring via sentimental analysis of financial text using machine learning techniques. *International Journal of Environmental Research and Public Health*, 19:9695, 8 2022.

[8] Jesus Serrano-Guerrero, Mohammad Bani-Doumi, Francisco P. Romero, and Jose A. Olivas. Understanding what patients think about hospitals: A deep learning approach for detecting emotions in patient opinions. *Artificial Intelligence in Medicine*, 128:102298, 6 2022.

[9] Balvinder Singh Gambhir, Jatin Habibkar, Anjesh Sohrot, and Rashmi Dhumal. *Cybercrime Detection Using Live Sentiment Analysis*, pages 409–419. 2022.

[10] Shaurjya Mandal, Banani Saha, and Rishov Nag. *Exploiting Aspect-Classified Sentiments for Cyber-Crime Analysis and Hack Prediction*, pages 200–212. 2020.

[11] Rada Mihalcea. Multimodal sentiment analysis. page 1. Association for Computational Linguistics, 2012.

[12] Lukas Stappen, Alice Baird, Lukas Christ, Lea Schumann, Benjamin Sertolli, Eva-Maria Meßner, Erik Cambria, Guoying Zhao, and Björn W. Schuller. The muse 2021 multimodal sentiment analysis challenge. pages 5–14. ACM, 10 2021.

[13] A Agarwal, A Yadav, and D Vishwakarma. Multimodal sentiment analysis via rnn variants. pages 19–23. IEEE Computer Society, 5 2019.

[14] Yangyang Shi, Yongqiang Wang, Chunyang Wu, Ching-Feng Yeh, Julian Chan, Frank Zhang, Duc Le, and Mike Seltzer. Emformer: Efficient memory transformer based acoustic model for low latency streaming speech recognition. pages 6783–6787, 2021.

[15] Zaijing Li, Fengxiao Tang, Ming Zhao, and Yusen Zhu. Emocaps: Emotion capsule based model for conversational emotion recognition, 2022.

[16] Philip J Stone, Dexter C Dunphy, Marshall S Smith, and Donald M Ogilvie. *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press, 1966.

[17] George A. Miller. Wordnet. *Communications of the ACM*, 38:39–41, 11 1995.

[18] Vasileios Hatzivassiloglou and Kathleen R McKeown. Predicting the difficulty of identifying word sense ambiguity. pages 123–130, 1997.

[19] Charles Elkan. Text classification system with bayes networks, 2001.

[20] Bo Pang, Lillian Lee, and Shiva Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. pages 79–86, 2002.

[21] F Sebastiani, A Esuli of the 5th international conference on . . . , and undefined 2006. Sentiwordnet: A publicly available lexical resource for opinion mining. *researchgate.netF Sebastiani, A EsuliProceedings of the 5th international conference on language resources, 2006•researchgate.net*, 2016.

[22] Gulsen Demiroz, Berrin Yanikoglu, Dilek Tapucu, and Yucel Saygin. Learning domain-specific polarity lexicons. pages 674–679. IEEE, 12 2012.

[23] Savas Yildirim, Dhanya Jothimani, Can Kavaklioglu, and Ayse Basar Bener. Building domain-specific lexicons: An application to financial news. pages 23–26. IEEE, 8 2019.

[24] Oscar Araque, Marco Guerini, Carlo Strapparava, and Carlos A. Iglesias. Neural domain adaptation of sentiment lexicons. pages 105–110. IEEE, 10 2017.

[25] Megha Rathi, Aditya Malik, Daksh Varshney, Rachita Sharma, and Sarthak Mendiratta. Sentiment analysis of tweets using machine learning approach. pages 1–3. IEEE, 8 2018.

[26] Lepeng Wang. Text sentiment analysis method based on support vector machine and long short-term memory network. pages 87–91. ACM, 5 2023.

[27] Qiao Liu, Haibin Zhang, Yifu Zeng, Ziqi Huang, and Zufeng Wu. Content attention model for aspect based sentiment analysis. pages 1023–1032. ACM Press, 2018.

[28] Yequan Wang, Aixin Sun, Jialong Han, Ying Liu, and Xiaoyan Zhu. Sentiment analysis by capsules. pages 1165–1174. ACM Press, 2018.

[29] Hitkul Jangid, Shivangi Singhal, Rajiv Ratn Shah, and Roger Zimmermann. Aspect-based financial sentiment analysis using deep learning. pages 1961–1966. ACM Press, 2018.

[30] Jared Suttles and Nancy Ide. Distant supervision for emotion classification with discrete binary values. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 7817 LNCS:121–136, 2013.

[31] Cindy K. Chung and James W. Pennebaker. Linguistic inquiry and word count (liwc): Pronounced "luke,". and other useful facts. *Applied Natural Language Processing: Identification, Investigation and Resolution*, pages 206–229, 2011.

[32] Saif Mohammad. emotional tweets. pages 246–255. Association for Computational Linguistics, 11 2012.

[33] Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation, 2019.

[34] Shanglin Lei, Guanting Dong, Xiaoping Wang, Keheng Wang, and Sirui Wang. Instructerc: Reforming emotion recognition in conversation with a retrieval multi-task llms framework, 2023.

[35] Albert Zeyer, Patrick Doetsch, Paul Voigtlaender, Ralf Schlüter, and Hermann Ney. A comprehensive study of deep bidirectional lstm rnns for acoustic modeling in speech recognition. pages 2462–2466, 2017.

[36] Has¸im Has¸im Sak, Andrew Senior, Kanishka Rao, and Beaufays Google. Fast and accurate recurrent neural network acoustic models for speech recognition.

[37] Dongwei Jiang, Wubo Li, Miao Cao, Wei Zou, and Xiangang Li. Speech simclr: Combining contrastive and reconstruction objective for self-supervised speech representation learning, 2021.

[38] KC Santosh and Sameer Antani. Guest editorial multimodal learning in medical imaging informatics, 2023.

[39] Tianqi Zhao, Ming Kong, Tian Liang, Qiang Zhu, Kun Kuang, and Fei Wu. Clap: Contrastive language-audio pre-training model for multi-modal sentiment analysis. pages 622–626. ACM, 6 2023.

[40] Vishal Chudasama, Purbayan Kar, Ashish Gudmalwar, Nirmesh Shah, Pankaj Wasnik, and Naoyuki Onoe. M2fnet: Multi-modal fusion network for emotion recognition in conversation. *ArXive*, 2022.

[41] Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile: The munich versatile and fast open-source audio feature extractor. pages 1459–1462. Association for Computing Machinery, 2010.

[42] Yuan Gong, Yu An Chung, and James Glass. Ast: Audio spectrogram transformer. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 1:56–60, 4 2021.

[43] Himadri Mukherjee, Sk Md Obaidullah, KC Santosh, Santanu Phadikar, and Kaushik Roy. Line spectral frequency-based features and extreme learning machine for voice activity detection from audio signal. *International Journal of Speech Technology*, 21, 2018.

[44] KC Santosh. Speech processing in healthcare: Can we integrate? *Intelligent Speech Signal Processing*, pages 1–4, 1 2019.

[45] Rafia Sharmin Alice, Laurent Wendling, and KC Santosh. 2d respiratory sound analysis to detect lung abnormalities. volume 1704 CCIS, 2023.

[46] Adrien Cauchy. Mémoire sur les fonctions dont les dérivées sont périodiques. *Journal de l'École Polytechnique*, 17:133–200, 1829.

[47] Andrew L. Maas Awni Y. Hannun and Andrew Y Ng. Rectifier nonlinearities improve neural network acoustic models. *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, page 3, 2013.

[48] Pierre-Simon Laplace. Théorie analytique des probabilités. *Courcier*, 1812.

[49] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 807–814, 2010.

[50] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. volume 30. Curran Associates, Inc., 2017.

[51] Jack Kiefer and Jacob Wolfowitz. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, pages 462–466, 1952.

[52] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

[53] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86:2278–2324, 1998.

[54] Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of shape recognition. *Proceedings of the IEEE*, 68:826–834, 1980.

[55] Yann LeCun, Bernhard Boser, John S Denker, David Henderson, Richard E Howard, Warren Hubbard, and Lawrence D Jackel. Handwritten digit recognition with a back-propagation network. *Advances in neural information processing systems*, pages 396–404, 1990.

[56] Johnson Kolluri, Vinay Kumar Kotte, M S B Phridviraj, and Shaik Razia. Reducing overfitting problem in machine learning using novel l1/4 regularization method. pages 934–938, 2020.

[57] Haotian Zhang, Lin Zhang, and Yuan Jiang. Overfitting and underfitting analysis for deep learning based end-to-end communication systems. pages 1–6, 2019.

[58] Joseph Paul Cohen, Tianshi Cao, Joseph D. Viviano, Chin Wei Huang, Michael Fralick, Marzyeh Ghassemi, Muhammad Mamdani, Russell Greiner, and Yoshua Bengio. Problems in the deployment of machine-learned models in health care. *CMAJ*, 193:E1391–E1394, 9 2021.

[59] Guillaume Lemaître, Fernando Nogueira, and Christos K Aridas. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18:1–5, 2017.

[60] Alan V Oppenheim. Speech spectrograms using the fast fourier transform. *IEEE Spectrum*, 7:57–62, 1970.

[61] Shirali Kadyrov, Cemil Turan, Altynbek Amirzhanov, and Cemal Ozdemir. Speaker recognition from spectrogram images. pages 1–4, 2021.

[62] Evaggelos Spyrou, Rozalia Nikopoulou, Ioannis Vernikos, and Phivos Mylonas. Emotion recognition from speech using the bag-of-visual words on audio segment spectrograms. *Technologies*, 7, 2019.

[63] Mustaqeem and Soonil Kwon. A cnn-assisted enhanced audio signal processing for speech emotion recognition. *Sensors*, 20, 2020.

[64] Siddhi Bajracharya, Rodrigue Rizk, and KC Santosh. Deep spectral features to detect atrial fibrillation using single-lead ecg signals. 2023.

[65] Yuan Gong, Cheng-I Lai, Yu-An Chung, and James Glass. Ssast: Self-supervised audio spectrogram transformer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36:10699–10709, 6 2022.

[66] Rohit Kumar Kaliyar. A multi-layer bidirectional transformer encoder for pre-trained word embedding: A survey of bert. pages 336–340. IEEE, 1 2020.

[67] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[68] Jiang Li, Yingjian Liu, Xiaoping Wang, and Zhigang Zeng. Cfn-esa: A cross-modal fusion network with emotion-shift awareness for dialogue emotion recognition, 2023.

[69] Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu, Yuchuan Wu, and Yongbin Li. Unimse: Towards unified multimodal sentiment analysis and emotion recognition, 2022.

[70] KC Santosh and Suprim Nakarmi. *Active Learning to Minimize the Possible Risk of Future Epidemics*. Springer Nature Singapore, 2023.

Biographical Notes

Siddhi Kiran Bajracharya is currently pursuing his graduate studies in the Department of Computer Science at the University of South Dakota (USD). His academic journey includes obtaining a Bachelor of Engineering in computer engineering from Tribhuvan University, Nepal. With a research emphasis on deep learning and computer vision, Siddhi has contributed to the field by presenting a conference paper at the Institute of Electrical and Electronics Engineers (IEEE) Conference on Artificial Intelligence in Santa Clara in 2023. The paper introduces an approach utilizing deep spectral features for the classification of atrial fibrillation, a specific type of cardiac arrhythmia, in single-lead ECG signals.

During his time at USD, Siddhi has taken on roles as a graduate research assistant and a teaching assistant. Before his academic pursuits, he accrued valuable industry experience spanning four years, specializing in machine learning, data science, and software engineering. His professional background includes working with two companies based in Nepal. Siddhi's comprehensive educational background, research contributions, and industry experience collectively shape his expertise in computer science and technology.