Deep Regression of Social Signals in Dyadic Scenarios

Author: Ítalo Vidal Lucero
Director: Dr. Sergio Escalera Guerrero
Co-director: Dr. Julio Jacques Junior
Co-director: Cristina Palmero Cantariño

Abstract

The purpose of this project is to design a general system for emotion recognition through social signals in dyadic using deep learning methods using raw data from audio, video and text transcriptions from publicly available database records. The automatic emotion recognition problem has increased the attention in the scientific community considering the multi applications for emotion detection but also to design more accurate and complex empathic machines. During this project are proposed alternatives for utterance representation of multi-modal data generated from text, audio and video, in order to improve the state of the art system for emotion recognition based on deep learning networks. The proposed framework is based in IEMOCAP database but it has a general scope for any multi-modal database. The performance of this system outperforms the state of the art method and delivers an informative analysis concerning the utterance representation quality. Finally, the conclusions of this work are exposed along with potential future lines of work related to emotion recognition systems and emotion representations.

This work was funded by the National Agency for Research and Development (ANID) / Scholarship Program / BECAS CHILE MAGISTER AREAS PRIORITARIAS/2018 - 730190091

Acknowledgements

In the first place, I would like to express my sincere thanks to my Director Sergio Escalera and my two Co-directors Julio Jacques Junior and Cristina Palmero for all the time dedicated, guidance throughout all this process and for the opportunity to collaborate with your research team, which has made this project a unique experience.

I would also like to thank all my family who at a distance have been an indispensable support for the realization of this project.

Contents

List of Figures

List of Tables

1 Introduction

Human communication is one of the key drivers in humanity's evolution that makes possible all our social interactions. Consequently, human communication helps to define us as social living beings as it has shaped our perception of the world making us a unique specie in the animal kingdom with an advanced communication system and, at the same time, an efficient way to preserve knowledge through time.

Nowadays, with the promising advances in new technologies, there is also increasing attention to reveal the mechanism that supports our communication system and be able to explain, in detail, how it works. In particular, this is the case of many efforts emerging from Artificial Intelligence studies in the seek of a better understanding of human communication that allows us to replicate it in a natural and fluid manner. However, the task increases its complexity since the human communication system relies on a highly complex compound of communication sources, commonly called modalities, that work in parallel to generate a multi-modal broadcast while interact simultaneously to the extend of producing a message. Additionally, the source modalities can proceed from very different natures like are text, audio, and visual sources, and their interrelationship is responsible for the message's abstract representation. In this scenario, deep neural networks models have shown promising results for features representations that outperform standard machine learning algorithms.

Emotion analysis has is strictly related with human communication since it allows us to express emotional states using multi-modal sources to enrich the language and increasing the efficiency in human interactions. This interest has derived in the deep analysis of emotional effects over the human communication system using automatic detection of emotions. This work is framed in the automatic methods for emotion recognition in dyadic interactions to provide a better understanding of the relationship between communication and emotions.

1.1 Motivation

The dynamic effects of emotional states during a conversation has been proven by Richards et al. [34] to be an *interactive* phenomenon between participants not restricted to the isolated self-state of individuals due to the emotional influence between speaker and listener. This means that the emotional state of a subject is conditioned by the state of his counterpart, which can modify or mirror his emotions.

As a consequence, the correct detection of emotions in human communication opens several fields of application, being Affective Computing one of them. As it might be expected, the first goal in Affective Computing is the development of efficient emotion recognition methodologies to reduce the prediction error in automatic systems, so they could be applied in a variety of usage contexts such as customer behavior analysis and health, security, among others.

In the Affective Computing field, emotion recognition is the first milestone that must be fulfilled as the cornerstone for more ambitious goals. In this scenario, there is a growing interest in Artificial Intelligence to provide tools for the development of more accurate and emphatic machines not limited to predict emotions but also to recognize the emotional state of an interlocutor and emulate a more effective response in terms of the three modalities, those are text, audio and video. Therefore, emotion recognition plays a key role in this continuous process of discovery and is the primer aim in the present project, since the proper application of new techniques in emotion recognition can guide us in a more informed understanding of human communication.

1.2 Context

As previously mentioned, Affective Computing is a growing area of investigation, and many research groups across the globe have been created to address this topic under a diversity of approaches. One of them is the case of the current collaboration between the Center of Computer Vision (CVC) at Universitat Autonoma de Barcelona (UAB) and the Human Pose Recovery and Behavior Analysis group (HuPBA) at Universitat de Barcelona (UB) who are actively contributing with new lines of research in human behavior and communication, based on different technologies to produce valuable knowledge in a diversity of social scenarios. One of these scenarios is the dyadic interactions where the focus is the interrelations restricted to a pair of individuals which is the line of research in this work.

In terms of the theoretical foundation, in the first place, it is important to define the conceptual environment where human communication is developed. For this reason, in this section we specify the set of required framework and elements present in the human communication process. In the research line of work related to the area of social communication based on multi-modal data, it is commonly defined as a starting point, the process representation of human communication. Here, the most generally accepted model is the one proposed by Shannon [38] which is depicted in **Figure 1**.

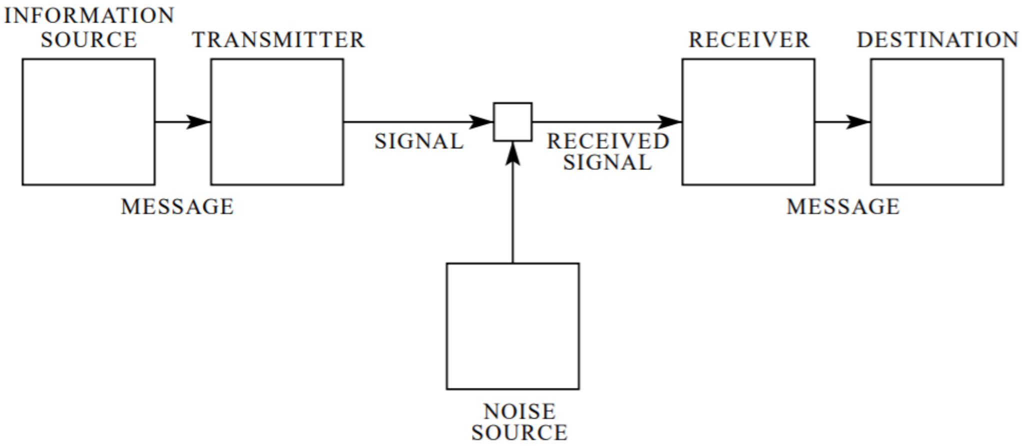


Figure 1: Shannon's Communication model [38].

In Shannon's general representation, five unique factors that intervene in human communication are introduced. A brief description of each element is provided as follow:

1. Information Source: Where the message or sequence of messages are created. The nature of the message can take multiple forms as in the case of audiovisual data, for instance: a sequence of letters to form a text, a combination of sound waves from a radio, a function to describe a video in color televison, and others.

2. Transmitter: Consisting of a physical means to encode the source message into a signal able to be transmitted through a proper channel.

3. Channel: Define as the medium needed to propagate the message between the source origin, the information source, to the destination.

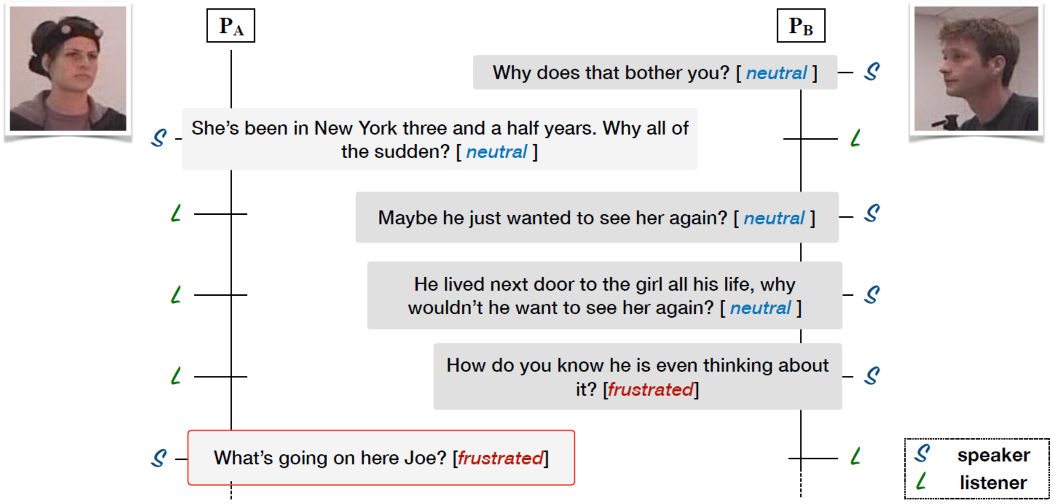
4. Receiver: Responsible to decode the message and transform it into a proper signal. It performs the transmitter inverse operations.

5. Destination: The objective self of the generated message.

Another distinction in Shannon's model is the classification type of communication system that can be discrete, continuous, or mixed. To clarify, the nature of both the message and the signal determines the type of communication. In particular, it is called a discrete system when the message and signal are represented with discrete input values or symbols, as in the case of letters in a text; continuous, in the event of message and signal presented as a continuous function of values, like in the example of television; or mixed, when both types of representations are present in the message and signal, as in the case of Pulse-code Modulation (PCM) systems for speech transmission.

On the other side, Shannon's model also introduces the presence of a Noise Source which modifies the signal transmitted. As result, any effort in a posterior analysis must take into account the detection and mitigation of this undesirable effect especially when it comes to audiovisual data, e.g. in the processing step but also in a prediction phase. In the particular event of audiovisual emotion recognition task, a potential noise source is the emotion classification ground truth data, since many databases contain manual labeling from individual experts which adds a bias, except when the database is adjusted employing consensus from a group of experts to reduce this noise.

In the second place, it is also needed a clear definition of the scope of application in the present work that is focused on dyadic interactions between individuals and the emotions that can be obtained from their multi-modal codification. In fact, as it will be discussed, human interaction is an important factor to be considered in the emotion recognition problem in dyadic scenarios. As described in **Figure 2**, the emotion dynamics in a dyadic interaction have a relevant role, as they can influence the emotional state of the counterpart during a conversation. In this case, utterances from Person B, which can be defined as a unit of speech bound by breathes or pauses [10], can modify the emotional state of Person A, represented by its utterances.



PA
PB
Why does that bother you? [neutral]
She's been in New York three and a half years. Why all of the sudden? [neutral]
Maybe he just wanted to see her again? [neutral]
He lived next door to the girl all his life, why wouldn't he want to see her again? [neutral]
How do you know he is even thinking about it? [frustrated]
What's going on here Joe? [frustrated]
S speaker
L listener

Figure 2: Emotional dynamics in a dyadic scenario [20].

These components must be considered in the model representation of social interactions from multi-modal sources. Social interactions are the starting point for the emotion analysis in dyadic scenarios and their proper design between the two individuals has to combine the right data preparation to produce a suitable approximation of the human communication phenomenon in order to build an accurate emotion recognition framework.

1.3 Objectives of this work

Having explained the importance of human emotion recognition from multimodal data and the existing gap for further improvement in automatic emotion recognition, the present work has defined the following objectives to be fulfilled:

1. Investigate the public state-of-the-art databases used in dyadic scenarios.
2. Perform an utterance feature extraction pipeline from raw modalities in the study.
3. Reproduce a state-of-the-art multi-modal system in terms of performance in emotion recognition.
4. Analyze and propose a state-of-the-art feature representation at the utterance level for text modality.
5. Analyze and propose a joint optimization of modalities for feature representation at the utterance level.
6. Perform an ablation study to compare modalities importance at the utterance level.

To be precise, objectives 1, 2 and 3 aim to reproduce the current framework in the state of the art considering: (a) the independent deep feature representation of the text, audio, and video modalities, plus (b) the correct alignment of these features as input for the multi-modal system proposed to retrieve temporal meaning.

In contrast, objectives 4 and 5 are set to enhance the state-of-the-art feature representation considering a uni-modal and global approaches based on current improvements suggested by the literature.

Finally, objective 6 is defined to retrieve more in-depth conclusions about the contribution in the performance of all the possible combinations of modalities considering at the same time the complexity associated.

1.4 Reader's guide

The document is structured in the following way:

1. **State of the art**: Review of the bibliography more applicable to modality feature representations and recurrent networks systems for emotion recognition in dyadic scenarios.
2. **Analysis and Proposal**: Description of the proposal considering the characteristics of the database selected and objectives for emotion recognition.

   (a) **Problem Specification**: An in-depth analysis of the emotion recognition requirements and database selection.

   (b) **Proposal**: Description of the methodology to obtain baseline features representations with multi-modal data and alternatives to improve them.

3. **Experiments**: Description, results and analysis of the experiments.

   (a) **Description of the experiments**: Experimental set up used along the experimentation.

   (b) **Results and Analysis**: Results of the experiments together with a comparison with baseline results in terms of performance.

   (c) **Discussion**: Overall discussion of the results.

4. **Conclusions and Future Work**: Conclusion of the work and recommended guidelines for future work in the area.

2 State of the Art

In the introduction was mentioned the advances in emotion recognition with deep neural networks to produce deep features representations and also as a part of relevant works based on Recurrent Neural Networks (RNN) [12, 14] to improve context-awareness of utterance in speech. There is a vast literature associated with emotion recognition methods and also with deep learning techniques for feature extraction, the present work is focused on the studies with an application on emotion recognition in dyadic scenarios. Furthermore, as most of the works made in this field have employed a recursive deep learning architecture over time to retrieve context temporal information, the present will be focused on this type of method based on Recurrent Neural Networks (RNN) architectures.

For this purpose, the section begins with a review of the most known databases collected in different social interaction contexts including the description of the type of data stored and the labeling process and scheme employed. Database selection is an important decision since any automatic algorithm, including deep neural networks, learns from examples and the final performance in the emotion recognition system will depend on the database application.

Next, for each of the three modalities to analyze, these are text, audio, and visual sources, there is a review about the more commonly used techniques to represent them and also, some modern approaches to investigate their potential applicability to the problem under study. After the feature representation discussion, and within the context of multi-modal techniques for emotion recognition based on audiovisual representations, a complete survey of multimodal methodologies are analyzed to determine the more suitable for the present project considering the objectives previously defined.

2.1 Multi-modal databases in dyadic interactions

Multiple databases contain emotions under speech scenarios, however, the focus on this work is the subset of databases that have audiovisual records and also speech transcriptions to used them in multi-modal systems. Following, there is a description of the most relevant multi-modal databases of the literature that fulfill these requirements. Considering all the characteristics, advantages and disadvantages, it will be decided which is the most suitable for emotion recognition in dyadic scenarios.

2.1.1 MOSI database

MOSI, Multimodal Opinion-level Sentiment Intensity [48] corpus is a very popular corpus and it was introduced as the first dataset for sentiment and subjectivity analysis focused on online videos collected from review of a variety of topics extracted from the Youtube webpage [1]. MOSI contains 93 original videos from 89 distinct speakers and a classification scheme at utterance level. In fact, the annotation provided comes fine-grained at frame level for videos, per transcription for opinion in texts, and per milliseconds for audio features. The main drawback of MOSI is the lack of emotion annotations since it was designed for sentiment analysis using different levels of intensity.

2.1.2 SEMAINE database

SEMAINE, the Sustained Emotionally colored Machine-human Interaction using Nonverbal Expression [22] database, another very well known database generated for human-agent simulated interactions. SEMAINE's main distinction is the used of a different methodology to express emotion by two latent factors named valance and arousal that can be annotated in a continuum emotional space following the Circumplex model of Affect introduced by Russell [32] and presented in **Figure 3**. From this project, it was born another initiative called audio/Visual Emotion Challenge (AVEC) [35, 36] which, in 2016 [45], also includes physiological signals for emotion and depression recognition. AVEC contains 24 videos from SEMAINE, role-played by humans, and annotations are provided every 0.2 seconds.

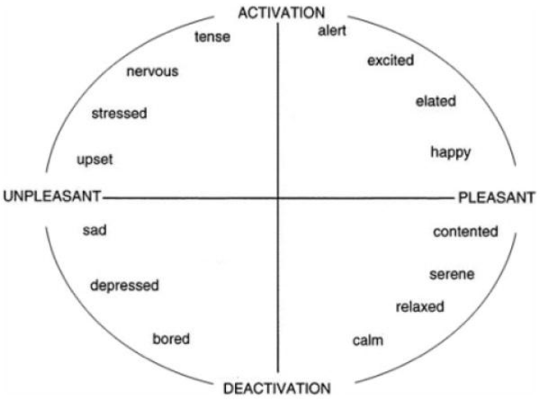


Figure 3: Circumplex model of affect for emotion continuous representation [32].

2.1.3 IEMOCAP database

IEMOCAP, the Interactive Emotional Dyadic Motion Capture database [2], is probably the most well-accepted database in emotion recognition in dyadic scenarios. The database contains records from 10 actors split into pairs for two-way dyadic interactions into five sessions which are 151 videos interplays. IEMOCAP contains labels for anger, happiness, sadness, excitement, frustration, fear, surprise, neutral and other, and annotations at utterance level using a majority vote consensus for the final evaluation.

[1] https://www.youtube.com/

Figure 4: Video frame extracted from IEMOCAP database [2].

2.2 Deep feature extraction

The second step for an automatic emotion recognition system is the construction of a proper end-to-end information process flow to generate feature descriptions of each modality at the utterance level, those are text, audio and video. Naive models that used directly raw data have worse performance in emotion recognition, thus, it is imperative the right design of feature extraction methods to obtain better abstract representation for each unit of speech represented by the utterances. These methods are called Deep feature extraction models since they used deep neural architecture to retrieve the utterance representations.

For this purpose, the widely accepted procedure is the feature extraction using the penultimate layers of deep neural networks which are previously trained with the same dataset or with a more general public one. This method has been proven as a good abstract representation for text [24], audio [1], images [19] and video [44], using the penultimate layers of deep neural networks with a fixed number of neurons that will define the new descriptor dimensions. Therefore, it is suitable for any of the three modalities analyzed in this work.

In the case of text, the penultimate layers of deep neural networks have been used as a descriptor for text classification problems [24]. In audio sources, it has been applied to improve speech classifications from raw audio data [1]. Finally, in video analysis, the feature extraction in image classification using pre-trained data [21] or from neural networks trained from scratch [19], which later inspire other methods for action recognition in sports that can be used as general video extractors, which is the case of C3D architecture able to learn spatiotemporal features [44].

The following sections describe the most influential methods in the state of the art for feature extraction along with the advantages and main drawbacks, for each type of source in multi-modal emotion recognition with audiovisual data.

2.2.1 Text modality

Text representation has been studied using multiple approaches from Natural Language Processing literature. The main goal is to produce a language model from text represented by sequences of characters and symbols. In this sense, language models can be classified into 2 types of general schemes: context-free models and context-based models.

In context-free models, the basic representation comes from a 1-dimension vector for each utterance using their whole used vocabulary as an encoding system, the so-called one-hot encoding. One-dimensional Convolutional Neural Networks have been used in this scenario for sentence classification to represent more complex abstract descriptors from words for text classification [49] and for text to speech conversion [41]. The main problems with this approach are the sparseness derived from the potential high dimensional vocabulary size employed, and the lack of context awareness since words are coded independently of the other words in the same utterance.

A better approach is the context-based models where the surrounding words defined a contextual meaning that is reflected in the utterance representation. At the same time, context-based models are able to reduce the text dimensionality as they compress utterance into a hidden latent representation from a dimension space smaller than the vocabulary size which translates into a reduction in training time. The latent representation is called word embedding and it has been proved a good estimator even from pre-trained models obtained from a general purpose using their outputs embeddings as enhanced inputs for language model representation [33]. Also, the complexity of the model can be reduced if are used pre-trained representations since these can be ”frozen”, which means that they do not change during training.

The most popular pre-trained word embeddings in context-based models are Word2Vec [23] and Glove [30]. The main difference between both frameworks is how they compute the surrounding context of each word. In the case of Word2Vec, the context is exclusively defined from the immediately contiguous words while, in the case of Glove, the context reflects the frequency of co-occurrences of words in the corpus used for training, highlighting the importance of the probability distribution of co-occurrences over just the closest neighbors.

A powerful advantage of word embeddings is the ability to learn implicit relationships between words using abstract concepts without any previous supervised information and perform arithmetic operations between them. This can be observed in **Figure 5** through the dimensionality reduction algorithm PCA [29], as word embeddings from countries and their capital cities are automatically assorted in two-dimensional space describing their relationship in a low dimension.

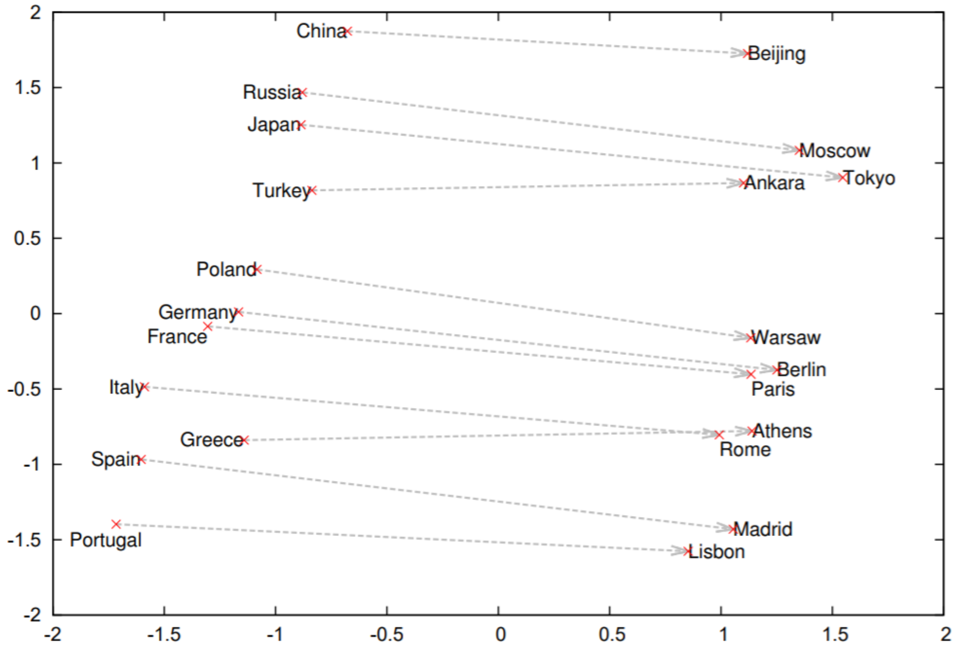


Figure 5: Two-dimensional PCA representation of countries and their capital cities using Word2Vec embeddings [23].

Moreover, recent researches have built word embeddings inspired by transformers architectures [47], which is the case of BERT: Bidirectional encoder Representations from transformers [5] pre-trained with the English Wikipedia database [2] as corpus almost 110 millions of parameters in the Base version, which even for pre-trained usage implies a huge computational effort compare to Word2Vec or Glove. BERT also introduced certain improvements as the masked language model [42] that improves the word embeddings hiding a percentage of words during training. But, it is also possible to use a basic bidirectional encoder as an approximation of BERT, in other words, a transformer with a shallow architecture, fewer parameters, and trained from scratch.

The transformer architecture is depicted in **Figure 6**, there are two main components in grouped with dashed lines: the encoder in red, and Decoder in blue. Since the transformer are primarily used for translation task, the encoder captures the meaning of a given a sequence of words within the context of the sentences, and the Decoder outputs the translated sequences of words into the new language. In particular, the Decoder contains a Positional encoding and Input embedding to extract context, followed by a transformer block (in grey) which weights the words in the sentences using a Multi-Head Attention layer and a Fed-Forward layer, which finally produce a vector representation for each input word.

[2] https://en.wikipedia.org/wiki/Wikipedia:Database_download

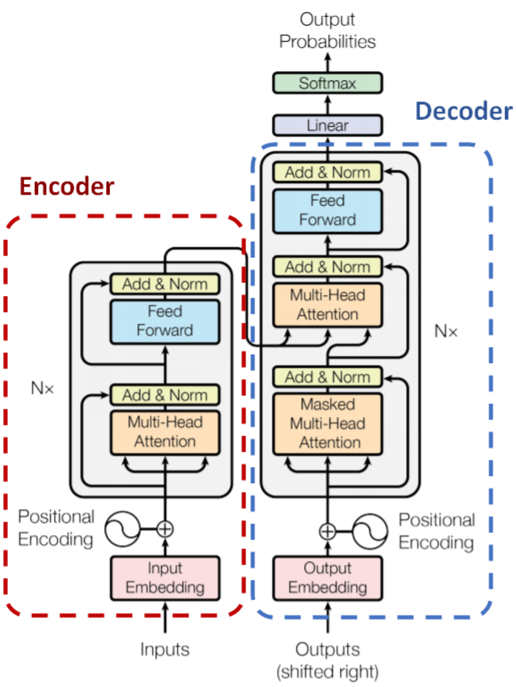


Figure 6: Transformer architecture [47].

2.2.2 Audio modality

The intrinsic complexity of audio modalities is increased with the bias introduced by the high variability in an individual's tone and pace, influenced by social and demographic factors like age, gender, or even precedence. In general, methods for speech analysis can be classified into two groups according to the type of information to consider as input [28]. On one side, there is explicit or linguistic information related to differently articulated patterns from the speaker and, on the other side, implicit or paralinguistic information associated exclusively to variations in the pronunciation of linguistic patterns, that is, low-level descriptors and/or statistics extracted from these descriptors. The main disadvantage in explicit or linguistic information models is that they are not able to provide an independent language model due to the variability from speakers previously mentioned.

The most common process to deal with speech data is to split audios into sequences of smalls segments, usually using a window length of 30-200 milliseconds, after which they can be processed with a proper method to extract the final audio features.

One group of works for audio feature extraction is based on hand-crafted features. For this purpose, a well known tool software is OpenSMILE [7], an open source software that provides high dimensional vectors from Speech processing and Music Information retrieval communities [3]. The configuration setup

[3] https://www.audeering.com/opensmile/

IS13_ComParE is widely accepted since it was the method employed for feature extraction in the INTERSPEECH 2013 ComParE Vocalization Challenge [37] where the output is a 6373 dimensional vector. This configuration generates low-level descriptors such as CHROMA and CENS features, perceptual linear predictive cepstral coefficients, linear predictive coefficients, fundamental frequency and Mel Frequency Cepstral Coefficient (MFCC) which are particularly informative since stress the importance of Fourier transformations over the human audible range of frequencies. Mel Frequency Cepstral Coefficients have been used in different speech tasks such as voice and speaker recognition using deep neural networks [25, 9]. Finally, low-level descriptors have an important role in emotion recognition since there is a strong correlation between several statistical measures and the speaker's emotional state [4]. Here high pitch and fast speaking pace can be a sign of anger, or low variance in pitch and slow speaking pace can denote sadness, are some of the many examples.

There is another method for audio feature extraction based on spectrograms and Convolutional Neural Networks. In general, spectrograms are used in phonetics and speech analysis through the identification of spoken words phonetically. However, the most important works in speech analysis perform low-level descriptors extraction for the high number of informative statistical measures covered.

2.2.3 Visual modality

Visual modalities are also indicators of emotional states as they can identify expressions like a frown or a smiling face. Additionally, previous studies provided have found that facial expressions can be described as universal indicators of human emotions [6]. Also as in the case of audio data, the process starts with the utterance record split using 8-frames as video window length following the procedure from [44].

A standard approach for video feature extraction is the use of 2-Dimensional Convolutional Neural Networks sequentially over segments of video frames [39]. Modern studies have improved the performance of visual features by means of the exploration of a 3-Dimensional Convolutional Neural Network that has the ability to learn spatiotemporal representation at once. This is the case of Learning Spatiotemporal Features with 3D Convolutional Networks (C3D) by Tran et al. [44], based on 3-Dimensional Convolutional Neural Network, was a pioneer in this line for applying this type of architecture over videos for action recognition in sport clips from Youtube website. It combines several 3D-CNN filters connected to 3D-MaxPooling that ends with a flatten layer to make predictions over annotated videos. **Figure 7** third layer visualization of C3D architecture that is able to detect moving body parts based on 3D-CNN. On the right side, there are sequences of frames representing video sport clips while on the left side are shown in color the activations in the third layer.

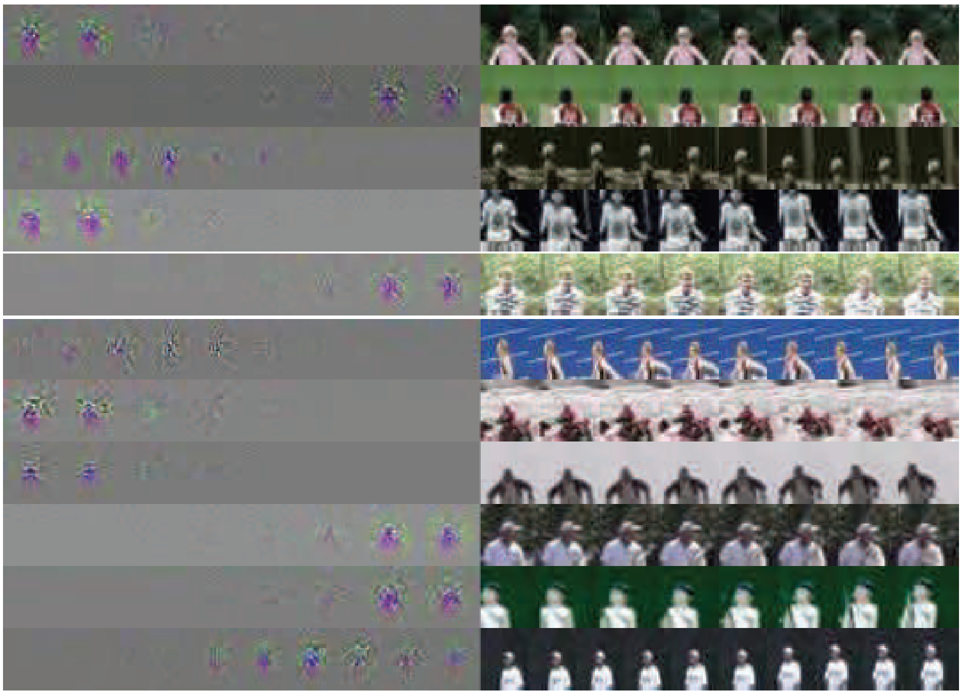


Figure 7: C3D layer visualization under moving body parts detection [44].

2.3 Multi-modal systems

The feature extraction process using deep neural networks is the first step in almost every model proposed for emotion recognition in dyadic interactions. At the same time, these models can be classified between context-based or context-free methods considering if they include the temporal information from the previous utterance or not as relevant signals to define the speaker's emotional state. This is a reasonable assumption since utterances are semantically dependent on each other within a conversation, thus the previous utterance contains relevant information that affects the next ones. Further models also are capable to treat this context separately for each speaker to model their intra and also inter dependencies since it has been demonstrated that emotional dynamics in a conversation is an interpersonal phenomenon, rather than a one-directional phenomenon [34].

2.3.1 Context-free based

Within the context-free models are grouped the first researches using deep neural networks for sentence classification. The Convolutional Neural Networks for Sentence Classification (CNN) [16] architecture was a baseline effort to represent utterances based on word embeddings vectors from Word2Vec [23] pre-trained models. The CNN [16] applies several filters with different length size with 100 units each to proceed with a flatten layer before the classification output. Here, each sentence is treated independently of the rest during the conversation therefore, it is considered a context-free method. In **Figure 8** is presented the work made by Kim [16] based on a CNN architecture with the word embeddings layer, 1-D CNN with max-pooling and Fully connected layer for classification.

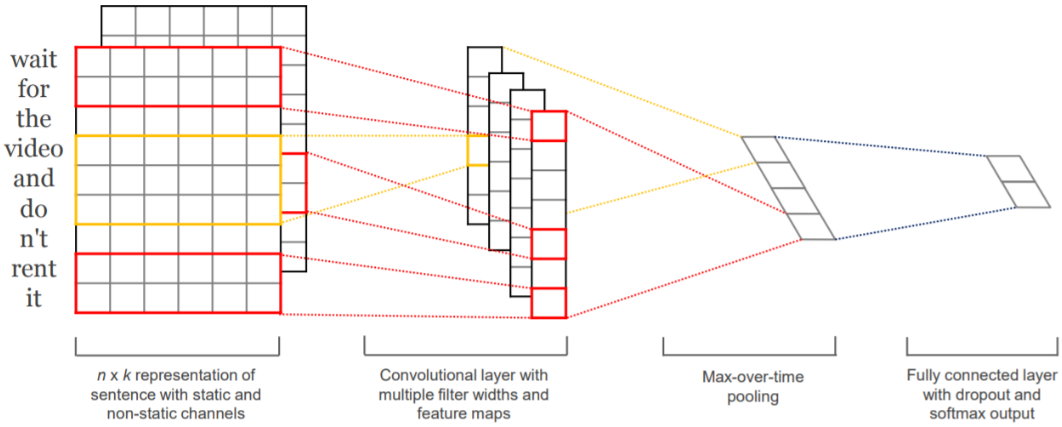

wait
for
the
video
and
do
n't
rent
it
n x k representation of sentence with static and non-static channels
Convolutional layer with multiple filter widths and feature maps
Max-over-time pooling
Fully connected layer with dropout and softmax output

Figure 8: CNN architecture for sentence classification [16].

2.3.2 Temporal based

Context-based multi-modal models rely on the importance of the temporal awareness of utterance to define the emotional state of the speaker. Deep architectures based on Recurrent Neural Networks [12, 14] have made a breakthrough in emotion recognition prediction since they are effective to retrieve information from past utterances. However, RNN presents an optimization issue when they face long-term utterances due to the vanished gradient problem that decreases its effectiveness [13].

Long Short-Term Memory (LSTM) [13] network introduced by Hochreiter and Schmidhuber is a variant of RNN that is less sensitive to the vanished gradient problem by means of a series of gates that are responsible to preserve information from the previous utterance improving the predictions compared with naive RNNs. In **Figure 9** can be observed the principal components of an LSTM cell consist of the input sequence x_t, the cell state c and the hidden state h that serves as information regularized and output, respectively. Inside the LSTM cell can be found an input gate i, an output gate o, and a forget gate f. As their names indicate the forget gate is in charge to delete previous information not relevant for the new states, the input gate decides how much new information must be updated in the cell state c and the output gate acts as a filter for the hidden state h [43].

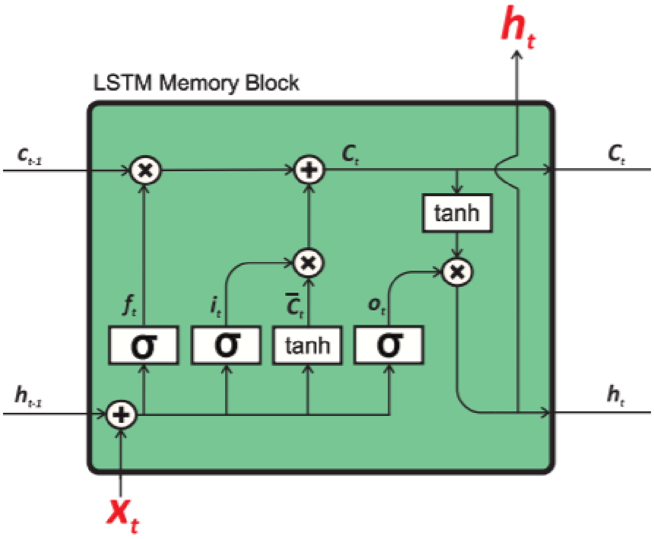


Figure 9: Long Short-Term Memory (LSTM) architecture [43].

Gated Recurrent Unit (GRU) [3] is another variant in RNN architectures introduced by Cho et al.. The idea behind GRU is to simplify the filtering neural networks inside the cell with fewer gates in order to make more efficient the backpropagation process during training, thus GRUs reduce the complexity of the RNN. **Figure 10** shows the similarities in the GRU architecture that only has two gates to control and process the temporal information: the reset gate r and the update gate z, where the former is in charge of deleting information from previous steps to keep relevant data in the current state h while the latter, decides the amount of information much be updated in the output using the new information.

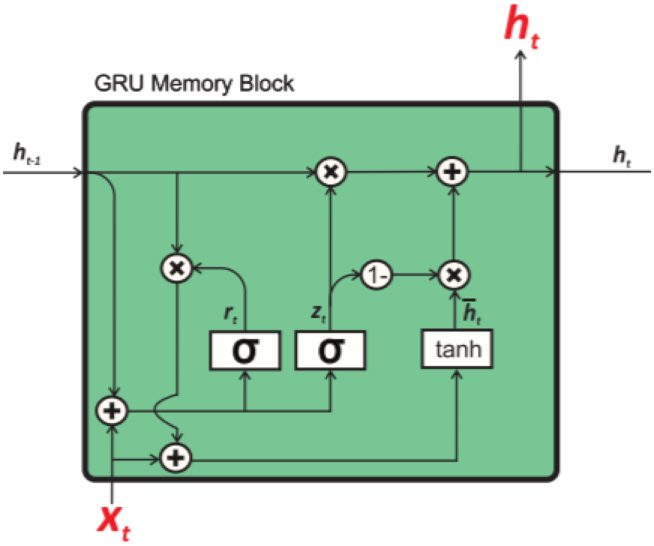


Figure 10: Gated Recurrent Unit (GRU) architecture [43].

There is a well-known method called bi-directional LSTM (bc-LSTM) [31] that has employed RNN in its framework. This method is considered context-based as it takes advantage of RNN architecture to capture temporal information in the emotion recognition given an utterance. The bi-directional LSTM (bc-LSTM) [31] contains a hierarchical fusion for multi-modal data and as its name suggests, it extracts contextual features using unimodal LSTM that later are concatenated and connected into a final LSTM layer to perform the classification prediction. There is also a variant for bc-LSTM called bc-LSTM with Attention (bc-LSTM+Att) that includes an attention mechanism over the output of the regular bc-LSTM in order to improve the context. Although bc-LSTM and bc-LSTM+Att preserve information from past utterance, the main issue in these methods is that they are not able to reflect inter-speaker dependencies, also important in emotion recognition.

In general, the temporal models inspired by RNN produce good results as they create a proper context based on historical records. Nevertheless, there is a second context representation linked to the inter-speaker context that it is necessary to distinguish since there are multiple studies that show their influence over their counterpart during a conversation.

2.3.3 Memory based

Another family of methods is the memory-based models that preserve historical information by means of dynamically updated memory cells. One effort in memory-based models is Memnet [40] that generates memory representation for each historical utterance, employing word embedding vectors, to describe the current utterance that is used in the classification layer. As consequence, the memories from past utterance lose sequential information.

On the other side is the Conversational Memory Network (CMN) [10] that improves the Memnet descriptors keeping historical and sequential information through two independent GRUs [3] for each of the two speakers. These two GRU are merged using memory cells and a series of hops designed to refine the final output prediction.

A recent approach called DialogueRNN [20], inspired in memory cell attention from CMN, has proposed a design to tackle the main problem from past models related to the lack of a discriminative treatment for different individuals with the intent of adding inter-speaker information to consider their emotional state dependencies. **Figure 11** displays the DialogueRNN architecture where the multimodal inputs utterance u_t are fed on each time step t into different GRU cells to capture their temporal context. In the middle with the green font is presented the Global GRU which serves as a contextual pivot for Person A and Person B GRUs with blue fonts at top and bottom, which at the same time fed the emotion GRU to perform the emotion prediction. The three major components to describe emotions using GRUs are:

1. Party state: that can be the speaker or listener state depending on the turn during the conversation. The party state is meant to represent the current individual state taking into account the previous party state and the current multi-modal features plus the context weighted with an attention mechanism that keeps track of past global state using a memory cell.

2. Global state: it is defined to update the current global state processing the past global state and party state and, the current multi-modal features plus the party state.

3. Emotion state: receives as input the previous emotional state and the current party state to finally make the prediction.

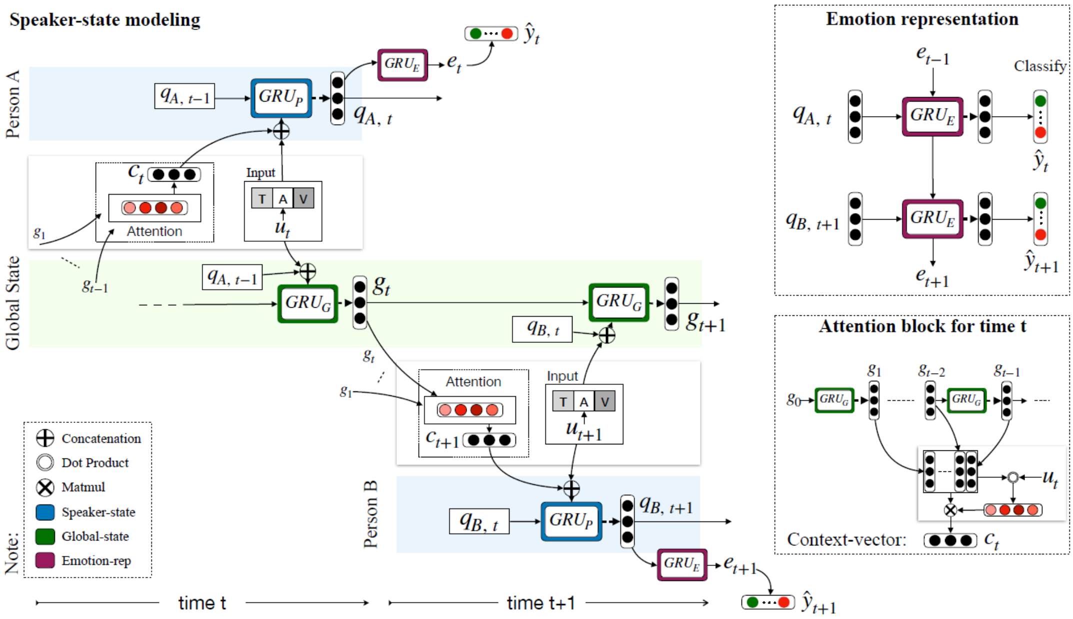


Figure 11: DialogueRNN architecture [20].

In general, there are several variants in the RNN architecture to test in the emotion recognition system as the addition of peepholes that are connections to the cell state to refine the outputs, but the experiments show no evidence of improvement compares with more simplistic RNNs [8].

It is important to mention that in the most relevant models, those are CMN [10] and DialogueRNN [20], the input features for multi-modal data come from the same pre-process that ends with the mere concatenation of uni-modal features before the utilization of the actual recurrent model on it. This means that alternative methods can be explored to merge uni-modal features.

2.4 State of the art summary

In light of the evidence obtained during the review of the available literature referred to resources and current methods for emotion recognition, some key ideas were taken to design the proposal work in this master thesis.

One important consideration of the data resources is the lack of public annotated data for emotion recognition purposes in dyadic scenarios at the utterance level. Databases like MOSI and SEMAINE contains data at a low level that can be summarized to represent the utterance's emotion. Moreover, MOSI was designed for sentiment analysis and SEMAINE uses a four-dimensional evaluation based on valence and arousal latent factors that can be translated into emotions but it is not a straight valuation. In summary, IEMOCAP is the more suitable dataset for the present project scope using utterance and evaluation for the six principal emotions.

It was interesting that most relevant models, like DialogueRNN and CMN, rely on pre-trained word embeddings as an early representation of utterances without the proper fine-tuning concerning the particular scope of dyadic interactions. For instance, Word2Vec [23] was trained with every English Wikipedia entry and this could not be the optimal corpus for the specific task of emotion recognition. Making the embedding weights part of the trainable matrix is an intermediate alternative but increases the complexity of the model that in the case of BERT [5] Base it would represent 110 million additional parameters. It is crucial to analyze a different prospect scheme to provide a closer approximation without compromising complexity.

In the case of audio modality, hand-crafted features from low-level descriptors combined with deep neural networks produce better results in emotion recognition tasks [26]. In the opposite side are text and video modalities, where the state of the art methods are exclusively built from deep neural networks with no hand-craft support [20, 10]. Nevertheless, the common multi-modal utterance representation ends as the simple concatenation to serve as entry-level input for any multi-modal system [20, 10]. This point will be analyzed in order to explore other methods to merge uni-modal features.

To summarize, the learnings from the state of the art provide a critical discussion for further experiments that are the core of this master thesis. Detailed proposals will be provided in the next section to tackle all the points addressed.

3 Analysis and Proposal

This section presents the main issues to deal with an emotion recognition system and a set of specific solutions based on earlier references. First, the selected database is explored and associated with the most suitable pre-processing procedure according to the characteristics of each modality and preceding results. Then, architectures to develop features representations as part of the experiment section are presented.

3.1 Problem specification

In this section are presented the arguments to justify the database selection to serve as input for the emotion recognition in dyadic scenarios. The particular characteristics of the database chosen will determine the proper pre-processing methods to execute which in turn affect the posterior feature extraction models that limit the scope of action for the proposal of this project.

Nevertheless, the proposed system for emotion recognition is design to be for general purpose, this means, it can be applied to any multi-modal databases with audio, video and text transcription records.

3.1.1 Data selection

Database selection is the first decision in order to reach the thesis objectives. Thus, considering the advantages and disadvantages of different public databases it has been decided that IEMOCAP is more adequate than MOSI or SEMAINE. The decision is based on the social interplays addressed in the videos focused on social interactions and, considering the available labels that are aligned with emotion recognition requirements, in other terms, annotated labels are not just an intensity evaluation for polarity sentiment analysis as in MOSI database, and they do not need an early transformation from latent factors as in the case of SEMAINE. However, the final system design is not restricted to this specific dataset since it can be used in any database with multi-modal records. From this point, the pre-processing steps will follow the most accepted state of the art approaches.

It is important to clarify that it will not be used the whole corpus from IEMOCAP but a subset related to the six principal emotions to make a fair comparison with the state of the art models. Therefore, from here and on, the data referenced in this work will be restricted to the emotions under study.

IEMOCAP database is divided into six Sessions of interactions roleplayed in pairs by ten actors as participants using pre-defined scripts taken from social scenarios and also a series of makers on the face, head and hands. Actors follow pre-defined scripts and also improvised social scenarios to induce certain emotional states. It contains approximately twelve hours of records distributed in 151 videos with a variable number of utterances each and during the recording, the two actors remain seated to keep all the gestures in the camera frame with 3 meters of distance between them. It was used one camera and one microphone per participant where the latter is merged to produce a unique audio source. Concerning the evaluation method, they used six evaluators to classify each utterance within the set of emotional states, following a majority consensus to obtain the final ground truth labels.

In **Figure 12** is presented the distribution of utterance length that shows a variable duration at the utterance level, which is the unit for evaluation purposes. Longest utterance speech has a duration of 34.11 seconds but the majority are concentrated in short utterance where 93.32% of the total is below or equal than 10 seconds long.

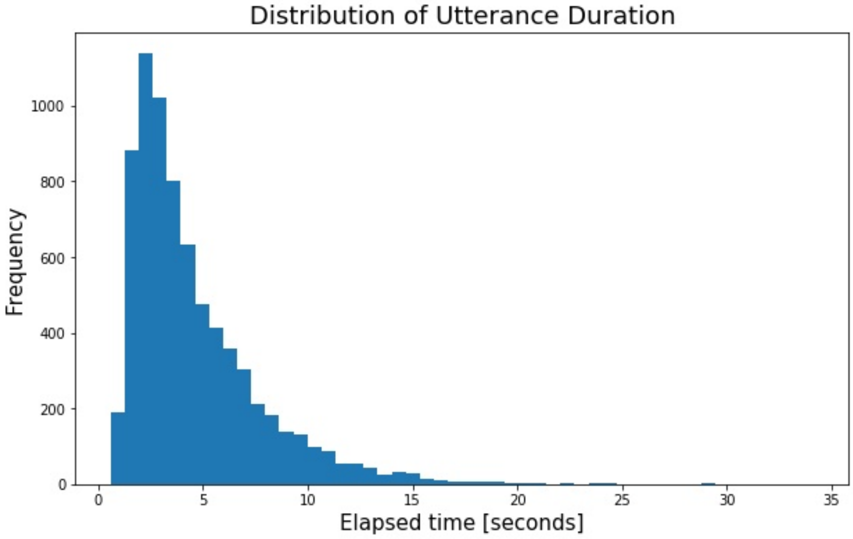


Figure 12: Utterance duration distribution in IEMOCAP database.

With respect to the number of words, 82.57% has less or equal than 20 words per utterance. **Figure 13** helps to explain the reasonably correlated distribution patterns between time and word length. It is important to say that the maximum number of words per utterance is 100 which can be used to defined the maximum sequence length necessary for the text vector representation to define the amount of padding.

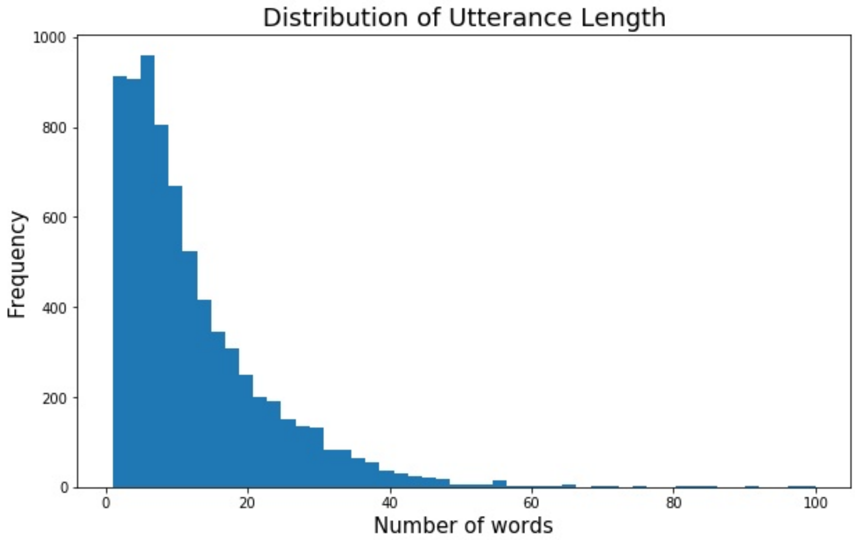


Figure 13: Utterance length distribution of words in IEMOCAP database.

Emotions have a degree of variance in the distribution over the corpus, been neutral and frustrated emotions the ones with more number of instances as is shown in **Figure 14**. A particular case is happy which has the lowest participation in the corpus utterance.

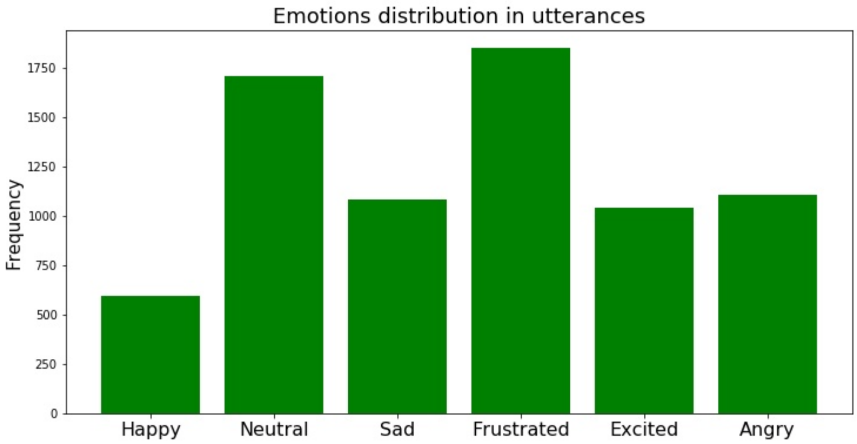


Figure 14: Emotion distribution in IEMOCAP database using six emotions.

In the gender distribution case, there is a balanced number of instances for each gender, as shown in **Figure 15**, which is also reasonable since there was an equal number of men and women but it is important to discard any sign of influence that implies an undesirable bias over gender.

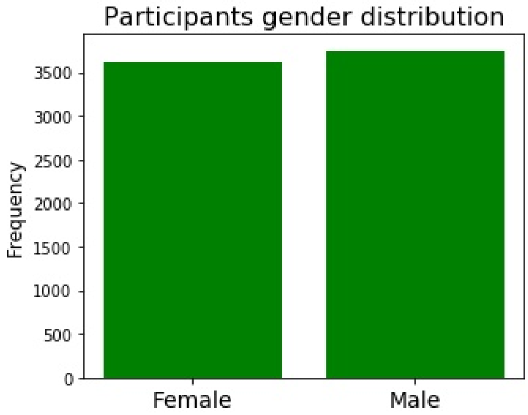


Figure 15: Gender distribution in IEMOCAP database using six emotions.

3.1.2 Evaluation protocol

Actors from IEMOCAP are separated in pairs to perform each session separately, this means, the 10 actors are paired to be in only one session. All of them perform similar scripts and improvisations from different social scenarios. **Figure 16** describe the distribution of utterance through all the five sessions. Taking that information, above all the split protocol followed by previous works using IEMOCAP, it has been defined that the first four sessions will be part of the training and validation datasets, leaving the complete session 5 only for test purposes.

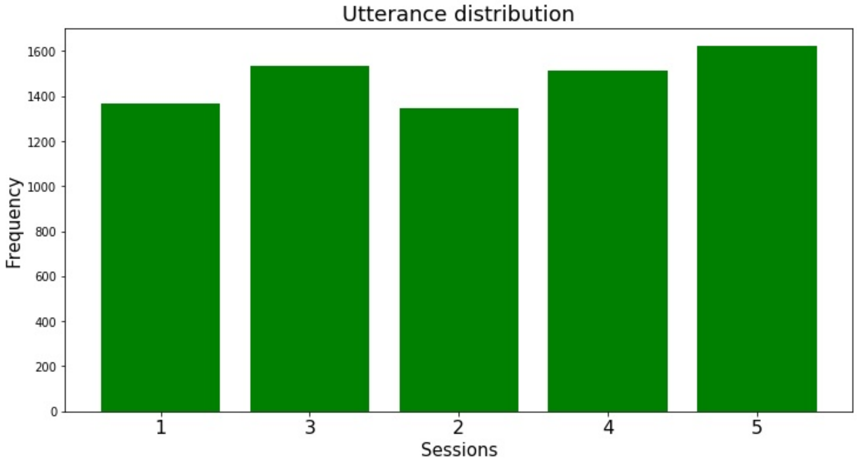


Figure 16: Utterance distribution over Sessions in IEMOCAP database using six emotions.

Have a completely person independent test set allows obtaining a more realistic evaluation of the generalization model performance and comparison. **Table 1** presents the resulting partition following the evaluation protocol with the number of utterances and videos contained in training plus validation set, and test set. In this way, the proportion of data used for training plus validation will be 78.02% which is similar to the standard 80%.

Partition	**Utterance Count**	**Dialogue Count**
Train + Validation	5757	120
Test	1622	31
Total	**7379**	**151**

Table 1: IEMOCAP dataset split protocol.

This protocol will be followed either for the feature extraction and the multi-modal training. In the case of feature extraction, it will be randomly separated 20% over *Train + Validation* partition to validate the performance, but in the case of multi-modal training, the evaluation protocol proposed in DialogueRNN does not assign a validation set and it keeps a monitor over the test set to select the best model based on the lowest categorical cross-entropy loss function. This work considers the same protocol for comparison purposes.

3.2 Proposal

In this section are specified the procedures and protocols adopted on each project stage, starting from uni-modal raw data transformation, to continue with the training of independent feature extraction models for each modality. Later are described the available methods to merge these uni-modal descriptors into a single utterance multi-modal vector. This final utterance representation will be the early step to determine the input for Recurrent Neural Network (RNN) [12] systems for emotion recognition.

Over the list of RNN systems shown in **Table 4**, the DialogueRNN system outperforms the rest of the methods on every emotion and over the global average weighted of accuracy and average weighted F1-score (defined as the average Recall and average F1-score *weighted* by the number of instance in each emotion category), thus this method is defined as the baseline to be exploited. In general lines, the pre-processing and feature extraction models are compiled from scratch following their procedures described from the literature [20, 10], and the implementation for DialogueRNN [20] is used from its authors Deep Cognition and Language Research lab (DeCLaRe) [4] from Singapore University of Technology and Design, accessed from their GitHub repository [5].

[4] https://declare-lab.net/

[5] https://github.com/declare-lab/conv-emotion

3.2.1 Pre-processing

In this section are listed the series of sequential procedures employed for modal pre-processing purposes starting from the raw data published in the IEMOCAP corpus. As the three modalities come from very disparate sources and the expected outputs are also different, the particular procedures are described for each source based on the specifics described in [20, 10] or in the works [44] that inspire them that were mentioned in **Section 2**.

3.2.1.1 Text data

All utterances have their manually annotated transcription using the utterance identification and also information about speaker turn and timestamps to extract the starts and endings in milliseconds. This information is critical for the sequence extraction as they define the utterance duration and speaker turn. There are 2 main steps for text pre-processing:

1. **Tokenization:** In the case of text, the procedure starts by encoding each utterance into an ordered sequence of words or tokens that represent the minimal unit of language in this textual representation.

2. **Padding:** Each utterance has a sequence of tokens with different lengths. The common procedure to handle it is by padding them using a maximum length allowed. In the previous section, it was found that the maximum number of words in IEMOCAP utterances is 100 words which is a reasonable limit for padding.

3.2.1.2 Audio data

IEMOCAP provides the audio data in separate files that contain both speaker's microphone records. The audios are stored in *wav* with a sample rate of 44100 Hz and stereo using 2 channels. These files are split into overlapped segments and, using the OpenSMILE toolkit, the handcrafted features from Low-Level Descriptors (LLD) are extracted in the form of a 6373-dimensional vector. The following steps must be executed to process IEMOCAP raw data:

1. **Conversion to Mono channel:** Since audio files come in 2 different channels, that is the stereo format, it is necessary to merge them into a single channel or mono to simplify the extraction of audio segments.

2. **Window Splits:** All utterances are split using a 100 milliseconds window length and an overlapping of 33.3 milliseconds following the procedure explained in DialogueRNN work [20] to produce the *Audio Segments*.

3. **LLD Extraction:** *Audio Segments* are the inputs to proceed with the Low-Level Descriptors extraction using OpenSMILE toolkit under the configuration IS13-ComParE which generates 6373 features for each segment as is described in **Figure 17**.

4. **Normalization:** LLD are normalized using Z-normalization to centered the features with zero means and unit variance.

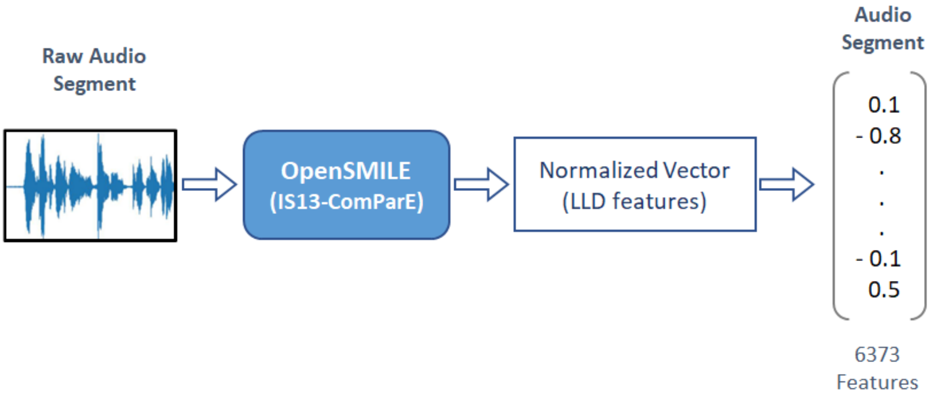


Figure 17: Text pre-processing using OpenSMILE toolkit to produce 6373 features segments.

3.2.1.3 Visual data

Video records are presented at 29.97 frame rate and full size of 720x480 pixels, it is important to notice that both speaker's recordings are merged into the same video file using a black background as can be seen in **Figure 4**, thus, regions with relevant information are a static subset of video frames and must be cropped individually. Furthermore, since it is necessary to extract the video of the current speaker only, the file segmentation has to follow the transcriptions participant identification. The following steps are performed to obtain the segments or "bins" from IEMOCAP raw video data:

1. **Video cropped:** Using the timestamps and the current speaker identification, the videos are cropped using a size of 346x236 switching between the speaker on the right and left side. **Figure 18** shows how the current speaker is selected using a blue bounding box to extract each frame. The output of this process is a video file for each utterance.

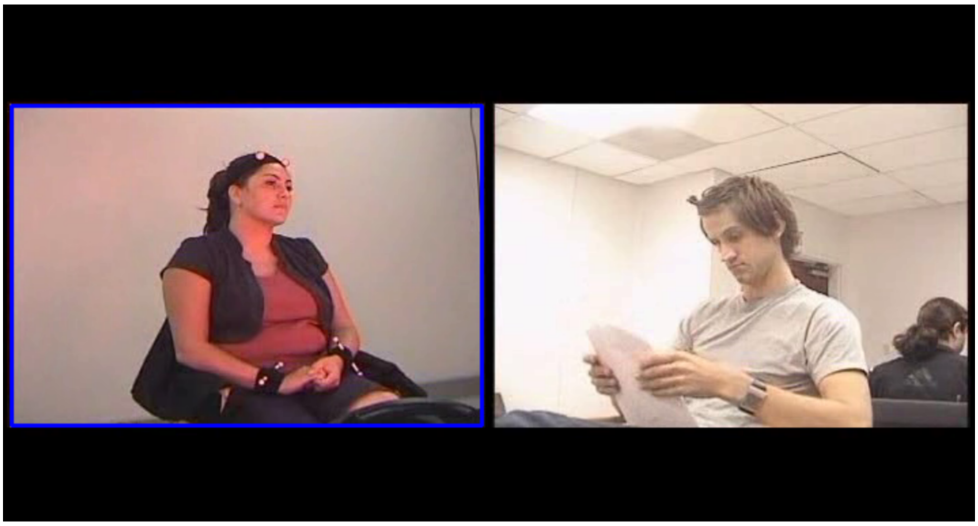


Figure 18: raw video from IEMOCAP, with both speakers, and illustration of the video cropping proceeding.

2. **Bins Extraction:** The standard procedure to split each video utterance is a series of *bins* with a length of 8 frames without overlap. This produces a 4D temporal representation with dimensions of 236x346x8x3 which represents the height, width, temporal depth and the three RGB channels, according to the "channels-last" convention. This format is described in Convolutional 3D Network (C3D) [44] work as required for the later application of 3D-CNN.

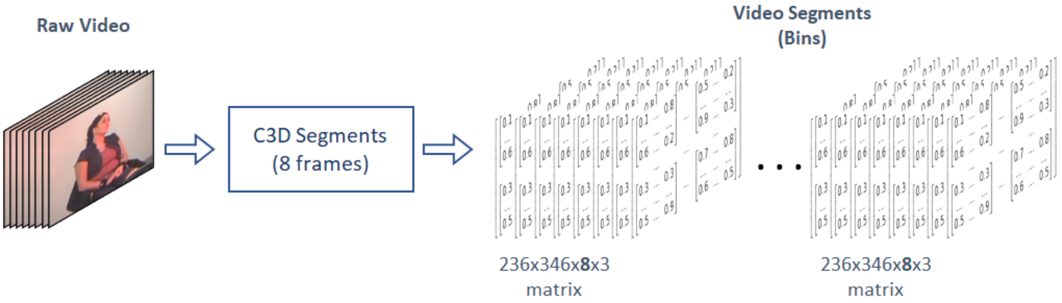


Figure 19: Video pre-processing to produce C3D temporal segments.

3.2.2 Independent feature extraction

During feature extraction, each modality is trained using an independent model to predict the six selected emotions. During inference, the learned model's weights are frozen and the activations from the dense layer that precedes the emotion predictions are used as feature extractor representation. These penultimate dense layers of each modality architecture have 100 neurons which generate a vector of 100-dimensions per utterance. Thus, given the three feature representations for text, audio and video with t_u, a_u and $v_u \in \mathbb{R}^{100}$, the proposed method in DialogueRNN and CMN, is the concatenation to produce the final utterance representation $u \in \mathbb{R}^{300}$ as in equation 1.

$$u = [t_u;\ a_u;\ v_u] \quad \forall\, t_u, a_u, v_u \in \mathbb{R}^{100} \tag{1}$$

3.2.2.1 Text features

For text representation, two approaches have been developed based on CNN [49] and transformer encoder [47]. Text based on 1D-CNN is the baseline proposed in the state of the art feature extraction method for emotion recognition, on the other hand, transformer architecture is a novel approach mainly used in translation tasks but its capabilities can be extended for text feature representation.

1. **Text based on CNN:** The baseline neural network model for text takes the padded sequences of tokens as inputs and using an Embedding Matrix fitted with the IEMOCAP vocabulary to extract the word representations from Glove [30] pre-trained version for 300-Dimensional vectors which is the most used word embedding.

   The embedding sequences are processed using 1-D Convolutional Neural Networks [49] to extract local context between surrounding words, which is an n-gram representation of language. In parallel, are applied three different convolutions with distinct filter sizes of 3, 4 and 5 words to varying the context range to capture during training. At the same time, each convolution contains 50 filter maps. The convolution filters are followed by a max-pooling layer with a size of 2 with the Rectified Linear activation Unit (ReLU) [27] introduced by Nair and Hinton. The output of the max-pooling layer is flattened and linked to a Dense Layer with 100 neurons that makes the *Scores* for text posterior representation, therefore, the textual vector is defined as $t_u \in \mathbb{R}^{100}$. The output layer contains six emotions labels for training purposes with Softmax activation to produce pseudo-probabilities. **Figure 20** describe the textual feature extraction architecture following [20] work.

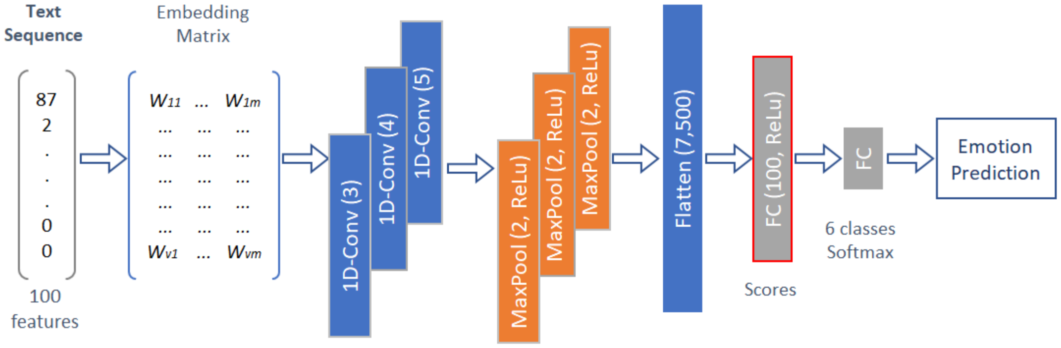


Figure 20: Text feature architecture based on CNN using 1D-CNNs, MP and FC layers.

To avoid overfitting and the covariate shift problem during training were added to the baseline architecture two types of layers commonly used in deep neural networks. The first is Batch Normalization (BN) [15] layer to avoid the internal covariate shift problem using normalization over previous activations at batch level and the second, Dropout [11] that randomly drops a percentage of neurons on each iteration offering a regularization effect. For simplicity, these layers are omitted in all the presented architecture schemes but represent well-probed and widely accepted mechanisms.

2. **Text based on transformer encoder:** An alternative approach to describe sequences of words is through a transformer encoder [47]. In this architecture is important to notice that there is no need for a pre-trained word embedding since the transformer trains its word representation using a Position Encoding and Embedding layers [47], where the former gives context based on the position of the word in the sentence and the latter retrieves context based on the meaning of the surrounding words in the sentence, this type of layer is also known as Token and Positional Embedding. This representation is fed into the encoder block where is applied a set of attention mechanisms, also called Multi-Head Attention, to weight relevant words given an input, and a fully-connected or Fed Forward layer that generates a vector for each word. All this is represented by the encoder block in **Figure 21**. After the encoder block, it is attached a Global Max-Pooling that summarizes each filter maps and the regular Dense layer with 100 neurons as *Scores* for text representation $t_u \in \mathbb{R}^{100}$. Then, the outputs to predict are the six emotions at the utterance level.

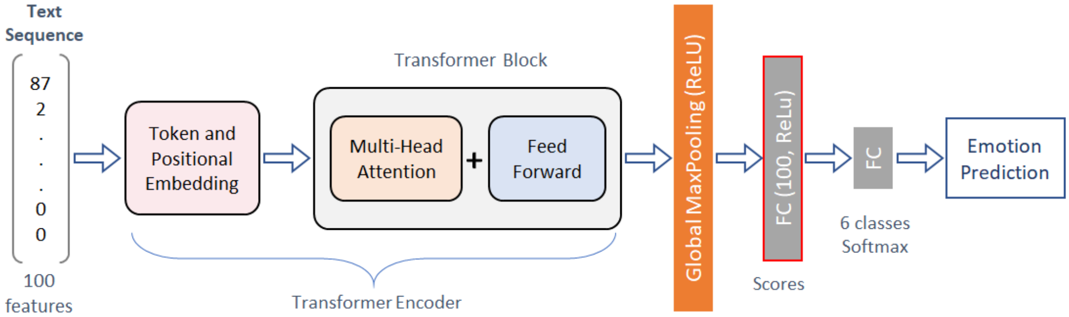


Figure 21: Text feature architecture based on transformer encoder. Input: 100-D vector.

3.2.2.2 Audio features

The normalized LLD features extracted with the OpenSMILE toolkit are fed into the audio neural network [10]. These features have local context information since they were calculated from a temporal window speech, the next step is to reduce the high dimensionality from 6373 features to 100 using a Dense layer with 100 neurons that precedes the final Dense layer with Softmax used to predict the six emotions. **Figure 22** presents the audio feature extraction model that generates the vector $a_u \in \mathbb{R}^{100}$ from the penultimate Dense layer activations. Since there are multiple audio segments per utterance, the final representation at the utterance level is calculated using the mean of all the activations made from the penultimate Dense layer. For simplicity, Dropout and Batch Normalization layers are omitted and their details can be found in the Experiments in **Section 4**.

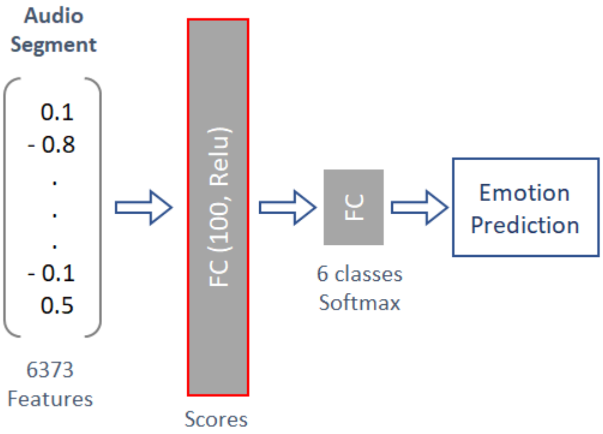


Figure 22: Audio feature architecture using a FC layer.

3.2.2.3 Visual features

For video feature architecture the logic follows the same as in the case of 2D-CNN over images but adding an extra dimension to represent temporal information. In this sense, the video bins extracted in the pre-processing step with dimensions of 346x236x8x3 representing width, height, depth and channels, are fed into a neural network and applied 128 3D-Convolution filters with a size of 5. The output is connected to a 3D-Max-Pooling layer where, according to the C3D work on 3D-CNN [44], the best filter maps size is 3. The output is fed to a Dense layer that contains 100 neurons to be used in the inference step to extract the video representation v_u. As in the case of audio segments, there are multiple video bins per utterance, then the average over all the bin activations is computed to obtain the final utterance vector $v_u \in \mathbb{R}^{100}$. **Figure 23** describe the video feature architecture.

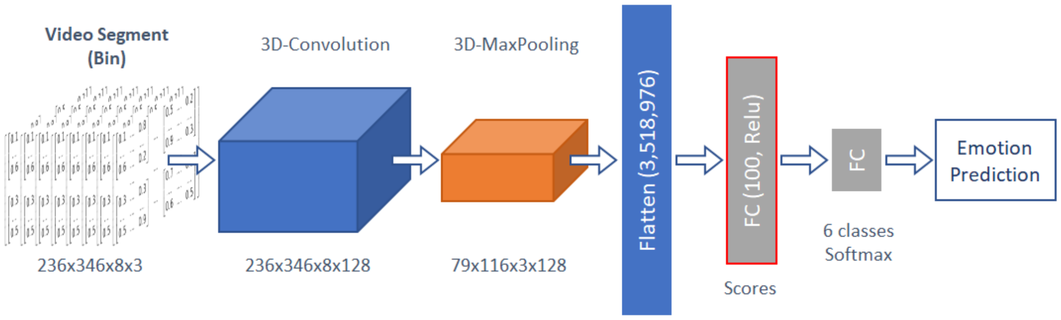


Figure 23: Video feature architecture using 3D-CNN, MP and FC layers. Input: 236x346x8x3 vector.

3.2.3 Joint optimization of features

A different proposal to obtain the utterance vector u is the joint optimization of multiple modalities with a single neural network able to receive multiple inputs. This multi-input network will have a Dense layer with 100 neurons per input modality to represent the ”Scores” before the output layer that makes the predictions. In this way, the utterance representation will be optimized taking into account the simultaneous contribution of different input modalities in a supervised manner.

The main issue with this approach is the number of instances per utterance in each modality. For text, there is only one instance per utterance but, for audio and video, one utterance has multiple and disparate numbers of audio segments and video bins as inputs. A simple solution is oversampling the text inputs using a fixed number of instances. In particular, the text examples are repeated while the audio and video sequences are selected using a linearly spaced sampled method preserving the temporal order. The number of samples is determined considering the distribution of video bins per utterance. **Figure 24** shows that most of the utterances have a low number of video bins, in fact, 85.65% of utterances contain 20 video bins or less. Therefore, the sampling method will use 20 instances at the utterance level.

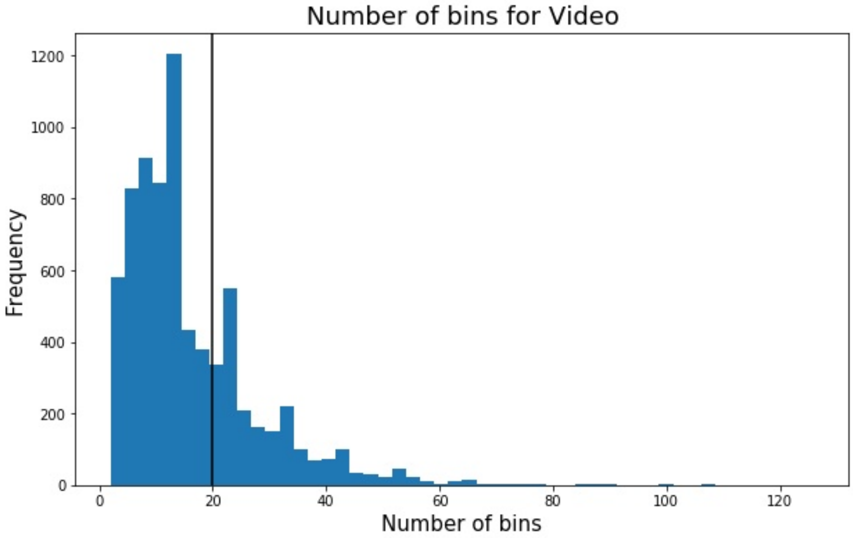


Figure 24: Histogram of number of bins files per video utterance in IEMOCAP database using six emotions.

In this work are proposed the joint optimization of all the three modalities Text + Audio + Video to obtain the utterance vector $u \in \mathbb{R}^{300}$, and the joint optimization using only Text + Audio which generates a modified version with fewer dimensions of $u \in \mathbb{R}^{200}$. In **Figure 25** is presented the feature extraction model using multi-modal inputs of text, audio and video. The uni-modal feature extraction models presented with blue boxes are the neural networks described in **Sections 3.2.2.1, 3.2.2.2 and 3.2.2.3**. The difference with the feature extraction models used here is that the output layer and "Score" layer from uni-modal versions are deleted and replaced by a new Dense layer with 300 units for each modality. The Network continues with a Flatten layer to concatenate the different modalities and the new "Score" Dense layer with 100 neurons per modality. As usual, the last Dense layer makes the predictions using Softmax activation over the six emotions analyzed. It is important to notice that in the case of text, the joint optimization architecture can be used with the text based on CNN or transformer without loss of generality.

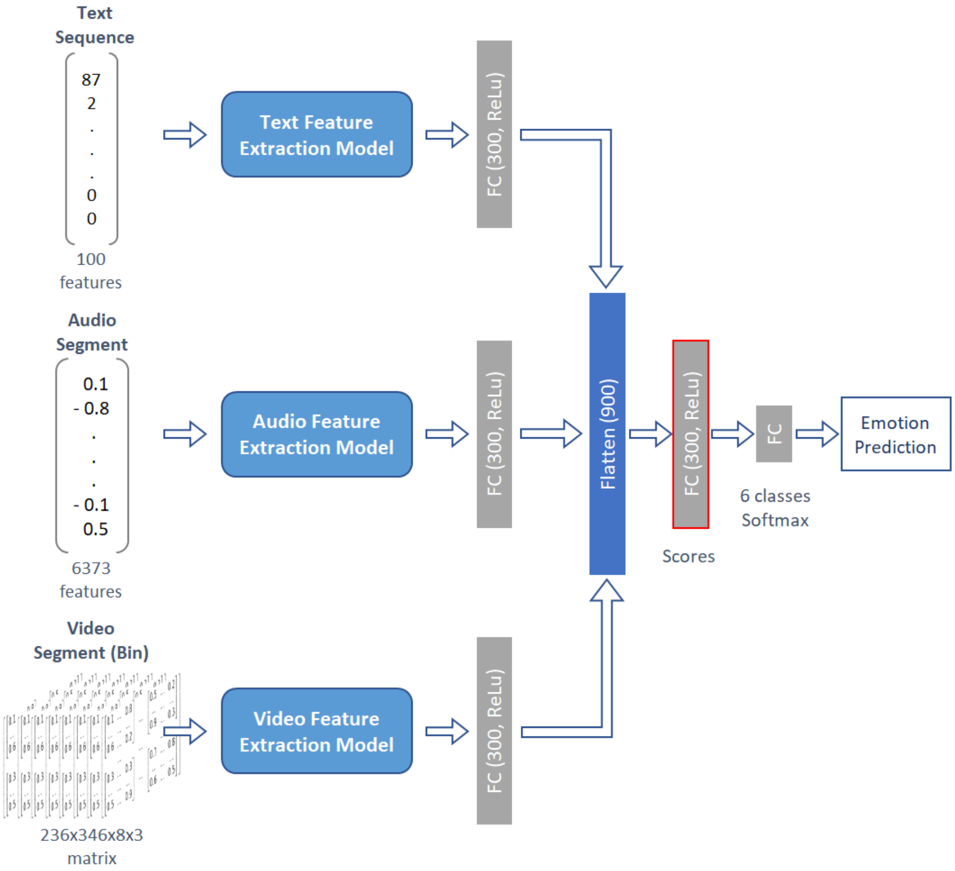


Figure 25: Joint optimization architecture using multi-inputs of text, audio and video feature extraction models.

Figure 26 is a modified version of the previous model meant to be a lightweight variation merging the two uni-modal feature extraction architectures of text and audio that have fewer trainable parameters compared to the heavy video architecture based on 3D-CNN.

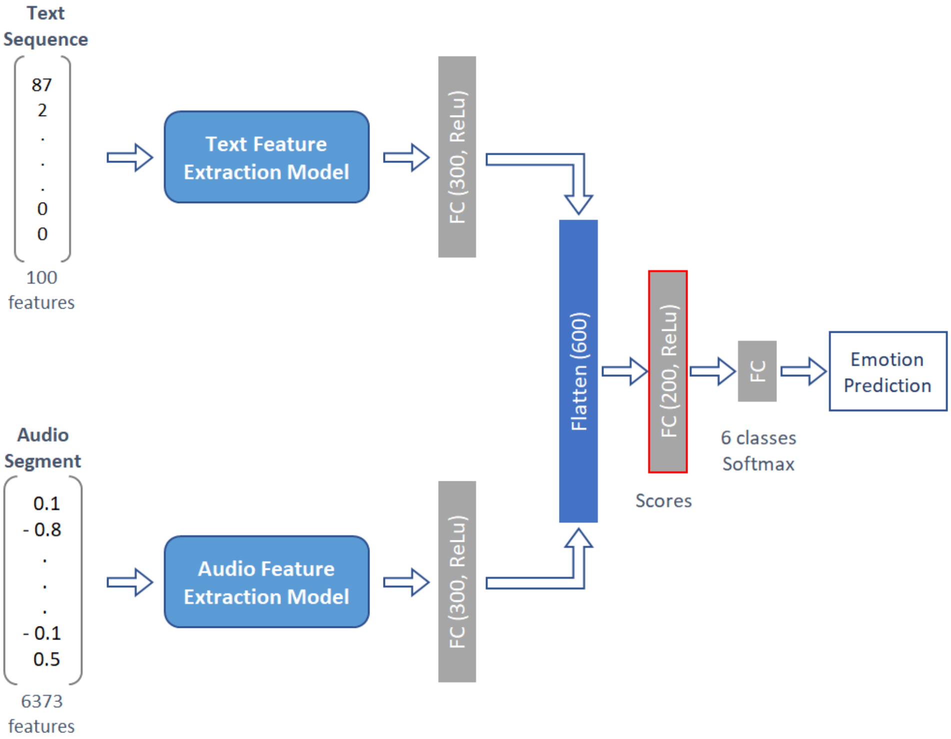


Figure 26: Joint optimization architecture using multi-inputs of text and audio feature extraction models.

4 Experiments

The experiment section is designed to provide in detail description of the methods aimed in the proposal **section 3.2** along with the comparative analysis of their results obtained in each set of experiments to generate a framework for discussion. In **section 4.1**, are presented the experiments for independent feature extraction models, those are uni-modal, and for joint optimization of features models, with bi and tri-modalities to be analyzed. Then, in **section 4.2** are tested multi-modal emotion regression models using tri-modal utterance representations obtained from **section 4.1** to perform a proper comparison with the baseline model and state of the art methods presented in **section 4.2.1**. A qualitative study of tri-modal utterance representation is provided in **section 4.2.2** to complement the benchmark analysis and **section 4.2.3** describe an extensive comparison of possible combinations of uni, bi and tri-modal regression models to test their generalization capabilities to open the discussion about the project discoveries in **section 4.3**.

4.1 Features

This section delivers a comprehensive explanation of settings and results used across different methods for independent feature extraction and joint optimization, both with the aim to generate a multi-modal utterance representation. Models are analyzed under the performance metrics over the test set and the specifics details of architecture and training experiments are provided in **section A** and **section C**, respectively.

4.1.1 Independent feature extraction

In this set of models, each modality is treated independently to provide the uni-modal descriptor. For each uni-modal experiment, the best design is selected based on the validation categorical cross-entropy loss function to avoid overfitting. The multi-modal utterance representation detailed in **section 3.2.2** will be generated through the concatenation of these uni-modal features.

4.1.1.1 Text features

The text based on CNNs was built following the next arrange of layers:

- Embedding Matrix: Glove representation with 300-Dimensions.
- 1D-CNN: 3 Convolutions layers of sizes of 3, 4 and 5 with 50 filter maps each.
- Batch Normalization: over each convolutional layer.

- Dropout of 20%.
- MaxPooling: size of 2 for each layer.
- Flatten and concatenation: of each MaxPooling layer.
- Dense layer: with 100 neurons used as feature extractor.
- Dropout of 20%.
- Dense layer: with softmax for the 6 emotions.

The model was compiled using categorical cross-entropy loss function, Stochastic Gradient Descend [18] optimizer during 24 epochs, batch size of 128 and an *annealing* learning rate with an initial value of 0.03 halved 2 every 4 epochs.

On the other side the text based on transformer architecture was built following the next arrange of layers:

- Token and Positional Embedding block: with 300 dimensions for the token representation and using linear position of the word as positional encoding.
- Transformer block: with 5 heads for the multi-head attention layer and 32 units in the fed-forward layer.
- Dropout of 20%.
- Global Average Pooling.
- Batch Normalization: over the Global Average Pooling layer.
- Dense layer: with 100 neurons used as feature extractor.
- Dense layer: with softmax for the 6 emotions.

The model was also compiled using categorical cross-entropy loss function, Adam [17] optimizer during 16 epochs with a batch size of 64 and an *annealing* learning rate with an initial value of 0.0003 divided every 4 epochs.

4.1.1.2 Audio features

The audio modality was built following DialogueRNN baseline with the addition of Batch Normalization and Dropout to improve the generalization:

- Dense layer: with 100 neurons used as feature extractor.
- Batch Normalization: over the Dense layer.
- Dropout of 20%.
- Dense layer: with softmax for the 6 emotions.

The model was compiled with categorical cross-entropy loss function using Stochastic Gradient Descend [18] optimizer during 16 epochs with a batch size of 128 and an *annealing* learning rate with an initial value of 0.003 halved every 4 epochs.

4.1.1.3 Video features

The video modality follows the baseline describe in DialogueRNN with 3D-CNN as base method:

- 3D-CNN: with 128 neurons, kernel size of (5, 5, 5).
- 3D-MaxPooling: with kernel size of (3, 3, 3).
- Dropout of 20%.
- Flatten layer.
- Dense layer: with 100 neurons used as feature extractor.
- Dropout of 20%.
- Dense layer: with softmax for the 6 emotions.

The model was compiled using Stochastic Gradient Descend [18] optimizer during 8 epochs with a batch size of 8 and *annealing* learning rate with an initial value of 0.003 halved every 4 epochs. In this case, the 3D-CNN has a heave architecture in terms of GPU memory consumption that has limited the batch size to 8 samples.

4.1.1.4 Performance

Table 2 present the results obtained for different uni-modal utterance representations. All the models produce as output the same 100-dimensions to be combined. Video has the highest number of parameters and at the same time the lowest performance in terms of average weighted accuracy. The most interesting result is the performance of text based on transformer that improves the scores from text based on CNN in 8.08 percentage points over the average weighted accuracy and in 8.62 percentage points over the average weighted F1-score. Moreover, text based on transformer contains less trainable parameters thus, has a lower complexity compared to text based on CNN.

Modalities	Output Dimension	Number of Parameters	Average (w)	
			Accuracy	F1-score
Text based on CNN	100	1 895 056	42.11	41.16
Text based on Transformers	100	1 407 138	50.19	49.79
Audio	100	638 206	24.93	9.97
Video	100	351 946 434	18.61	18.19

Table 2: Comparison feature extraction models for uni-modal utterance representation.

4.1.2 Joint optimization of features

The second group of models was designed to exploit the interdependencies between different modalities using multi-input neural networks as an alternative to the simple concatenation. The final utterance representation is obtained directly using the activations from the penultimate Dense layer.

Notice that the results from **Section 4.1.1** has proven the better performance for text feature extraction based on transformer encoder over text baseline, therefore, text based on transformer method is chosen for the incremental modeling using the joint optimization procedure of this section.

4.1.2.1 Text and audio features

The two modalities follow the same scheme presented for the independent versions but now they are represented as brunches within the neural network and combined to predict together the utterance emotion.

The brunches for text based on transformer and audio were designed as follows:

1) Text brunch based on transformer:

- Token and Positional Embedding block: with 300 dimensions for the token representation and using linear position of the word as positional encoding.
- Transformer block: with 5 heads for the multi-head attention layer and 32 units in the fed-forward layer.
- Global Average Pooling.
- Dropout of 20%.

2) Audio brunch:

- Batch Normalization: over features.
- Dense layer: with 300 neurons to reduce dimensionality.

- Batch Normalization: over the Dense layer.
- Dropout of 20%.

Then, the two brunches end with equally length vectors of 300-dimensions each that are combined adding, among others, the Dense layer responsible for the feature extraction through 200 neurons (100 per modality) to finally predict utterance emotions. The next layers were added to the neural network:

- Concatenation of branches 1) and 2).
- Batch Normalization: over the concatenation.
- Dense layer: with 100 neurons used as feature extractor.
- Dropout of 20%.
- Dense layer: with softmax for the 6 emotions.

The model was compiled using Adam [17] optimizer during 16 epochs with a batch size of 128 and *annealing* learning rate with an initial value of 0.00001 halved every 4 epochs.

4.1.2.2 Text, audio and video features

This is the extended version of the previous joint optimization where is added a new brunch with video modality to combine these three sources conjointly. However, during the experiments, it was not possible to fit this architecture in the GPU with 24 Gb of memory available thus, it was decided to reduce the number of filters in the 3D-CNN to reduce exponentially the number of parameters needed. In this way, the number of 3D-CNN filters in the video was modified from 128 to 64.

The general scheme follows the same 2 brunches for text based on transformer and audio, and now it is included the third modified brunch for video layers:

3) Video:

- 3D-CNN: with 64 neurons, kernel size of (5, 5, 5).
- 3D-MaxPooling: with kernel size of (3, 3, 3).
- Dropout of 20%.
- Flatten layer.
- Dense layer: with 300 neurons to reduce dimensionality.
- Dropout of 20%.

– Dense layer: with softmax for the 6 emotions.

The three brunches end with equally length vectors of 300-dimensions each that are combined adding, among others, the Dense layer responsible for the feature extraction through 300 neurons (100 per modality) to finally predict utterance emotions. The next layers were added to the neural network:

– Concatenation of branches 1), 2) and 3).

– Batch Normalization: over the concatenation.

– Dense layer: with 100 neurons used as feature extractor.

– Dropout of 20%.

– Dense layer: with softmax for the 6 emotions.

The model was compiled using Adam [17] optimizer during 16 epochs with a batch size of 12 and *annealing* learning rate with initial value 0.00001 halved every 4 epochs. In this case, the architecture size has limited the batch size. In this case, the 3D-CNN has a heave architecture in terms of GPU memory consumption that has limited the batch size to 12 samples.

4.1.2.3 Performance

The comparison between the two versions of optimized modalities is not direct since the output for bi-modal methods delivers a smaller vector representation with 200-dimensions while the tri-modal generates a 300-dimensions vector, as is shown int **Table 3**. Moreover, the latter has a bigger structure explained for the video 3D-CNN architecture. Despite these considerations, the bi-modal method outperforms the tri-modal one in 10.32 percentage points over the average weighted accuracy and in 11.60 percentage points over the average weighted F1-score. However, if this bi-modal optimized joint of features is compared with the uni-models results from **Table 2**, the text based on transformer design has slightly higher performance with 0.45 percentage points over average weighted accuracy and 0.49 percentage points over average weighted F1-score, and all this, with 59.06% less trainable parameters with information from only one source: text.

Modalities	**Output Dimension**	**Number of Parameters**	**Average (w)**	
			Accuracy	**F1-score**
Text based on Transformers + Audio	200	3 438 530	49.73	49.30
Text based on Transformers + Audio + Video	300	531 433 302	39.41	37.70

Table 3: Comparison joint optimization of feature models for bi and tri-modal utterance representation.

4.2 Multi-modal emotion regression

The obtained multi-modal utterance representations are the input for the emotion regression model with temporal context. DialogueRNN [20] is designed as the baseline framework to receive these representations including the three modalities: text, audio and video. In this section, the DialogueRNN [20] model is analyzed through the first proposed model: the concatenation of the three feature extraction models following the implementation of its work, which represents the **Baseline** for further comparisons in this work. The other two multimodals alternatives to be compared against this baseline are built with text based on transformer, considering the results from **Table 2** where it obtains the highest performance metrics over CNNs. In particular, the second proposed model will also use the simple concatenation of text based on **Transformer** (instead of CNN), audio and video, and the third proposed model will be a **Joint Optimization** of these three modalities, rather than using the simple concatenation.

In terms of architecture design, all the utterance representations will be fed into the DialogueRNN [20] network using the bi-directional version with attention which performs the best overall results according to the state of the art benchmark presented in **Table 4**. The Gated Recurrent Units (GRUs) outputs proposed by their authors are:

- Party State GRU: represented by a vector space of 150-dimensions.
- Global State GRU: represented by a vector space of 150-dimensions.
- Emotion State GRU: represented by a vector space of 100-dimensions.

The model was compiled using Adam optimizer and categorical cross-entropy loss function, during 60 epochs with a batch size of 30 instances and a learning rate of 0.0001. During training, it was monitored the test categorical cross-entropy loss function to select the best model along with its weighted average Accuracy and F1-score.

4.2.1 Benchmark comparison

In order to compare further implementations, one of the previous recurrent systems must be selected as the baseline model. The decision is made based on the performance metrics of Accuracy and F1-Score. To this end, the DialogueRNN [20] work has provided a complete performance benchmark of methods for multimodal data in dyadic scenarios explained in **section 2.3**. **Table 4** is a complete summary of the methods using the IEMOCAP database. It can be observed that DialogueRNN [20] outperforms the state of the art frameworks especially with the bi-directional + attention variant of the base model. This is a reasonable conclusion taking into account that DialogueRNN [20] is an incremental recurrent system inspired by previous advances in this line of research proposed in the literature. DialogueRNN [20] obtains the best metrics performance in accuracy and F1-score over each of the six emotion under evaluation and over the global evaluation taking the weighted average accuracy and weighted average F1-score considering all the emotions and their mass in the test set.

Methods	**Happy**		**Sad**		**Neutral**		**Angry**		**Excited**		**Frustrated**		**Average(w)**	
	Acc.	**F1**	**Acc.**	**F1**	**Acc.**	**F1**	**Acc.**	**F1**	**Acc.**	**F1**	**Acc.**	**F1**	**Acc.**	**F1**
CNN [16]	27.77	29.86	57.14	53.83	34.33	40.14	61.17	52.44	46.15	50.09	62.99	55.75	48.92	48.18
memnet [40]	25.72	33.53	55.53	61.77	58.12	52.84	59.32	55.39	51.50	58.30	67.20	59.00	55.72	55.10
c-LSTM [31]	29.17	34.43	57.14	60.87	54.17	51.81	57.06	56.73	51.17	57.95	67.19	58.92	55.21	54.95
c-LSTM+Att [31]	30.56	35.63	56.73	62.90	57.55	53.00	59.41	59.24	52.84	58.85	65.88	59.41	56.32	56.19
CMN [10]	25.00	30.38	55.92	62.41	52.86	52.39	61.76	59.83	55.52	60.25	71.13	60.69	56.56	56.13
DialogueRNN [20]	31.25	33.83	66.12	69.83	63.02	57.76	61.76	62.50	61.54	64.45	59.58	59.46	59.33	59.89
DialogueRNN l [20]	35.42	35.54	65.71	69.85	55.73	55.30	62.94	61.85	59.20	62.21	63.52	59.38	58.66	58.76
BiDialogueRNN [20]	32.64	36.15	71.02	74.04	60.47	56.16	62.94	63.88	56.52	62.02	65.62	**61.73**	60.32	60.28
DialogueRNN+Att [20]	28.47	**36.61**	65.31	72.40	62.50	57.21	67.65	**65.71**	70.90	68.61	61.68	60.80	61.80	61.51
BiDialogueRNN+Att [20]	25.69	33.18	75.10	**78.80**	58.59	**59.21**	64.71	65.28	80.27	**71.86**	61.15	58.91	63.40	**62.75**

Table 4: State of the art methods with tri-modal utterance representations for emotion recognition in dyadic scenarios [20]. Bold font denotes best performance for F1-score.

In **Table 5** are presented the performance evaluation over test set using the 3 multi-modal methods described in **section 4.2**, those are **Baseline**, **Transformers** and **Joint Optimization**. At emotion level, Joint Optimization obtains the best results in accuracy and F1-score in happy, neutral, excited and frustrated emotions while Transformer achieves the best results in sad and angry, except for accuracy in angry where the Baseline reaches the best accuracy with 74.12%. In the overall, the lowest results come from the Baseline method with a score of 30.02% average weighted accuracy and 25.42% average weighted F1-score; Joint Optimization gets the highest overall average weighted accuracy with 42.91% and Tranformers multi-modal method, the best overall average weighted F1-score with 39.26%. Compared with the published benchmark performances of multi-modal methods presented in **Table 4**, the three methods proposed have lower performance in overall average weighted accuracy and F1-score, but since the Baseline proposed method is inspired in the BiDialogueRNN+Att [20] implementation, the difference could be explained with differences in the utterance representation.

Methods	**Happy**		**Sad**		**Neutral**		**Angry**		**Excited**		**Frustrated**		**Average(w)**	
	Acc.	**F1**	**Acc.**	**F1**	**Acc.**	**F1**	**Acc.**	**F1**	**Acc.**	**F1**	**Acc.**	**F1**	**Acc.**	**F1**
Baseline	32.87	29.84	15.10	21.70	55.73	37.06	74.12	44.37	11.71	18.09	7.35	11.72	30.02	25.42
Transformer	29.37	35.00	18.78	**28.40**	77.60	48.06	55.88	**46.12**	27.42	41.52	28.35	34.12	41.37	**39.26**
Joint Optimization	37.76	**43.72**	7.76	14.18	85.68	**48.28**	0.59	1.17	34.45	**48.13**	49.87	**49.54**	42.91	38.06

Table 5: Comparison of methods with tri-modal utterance representations using DialogueRNN. Bold font denotes best performance for F1-score.

4.2.2 Qualitative study

The performance metrics in the three proposed multi-modal methods show the global evaluation of their capabilities to be used in the emotion recognition task using a temporal context, however, it is important to analyze how the underlying structure of these multi-modal features actually produces the utterance vector in a high-dimensional space to perform the emotion regression. To this end, is used the t-distributed Stochastic Neighbor Embedding t-SNE [46] visualization method for dimensionality reduction where closer points, in a high-dimensional space, are represented by means of t-student distributions that reproduce their local neighborhood. Through two-dimensional t-SNE [46], it is possible to perform a qualitative comparison rather than a global classification score.

In **Figure 27** is presented the two-dimensional t-SNE [46] for the **Baseline** multi-modal feature space from original 300-dimensions. Here, are depicted some small clusters for neutral, angry and happy emotions. There are some bigger clusters with two or tri emotions but the middle area shows a mix of emotions without a clear order in the space, which is a sign that the emotions do not have distinctive distributions in the vector space, which undermine the emotion classification in further methods that use these representations as inputs.

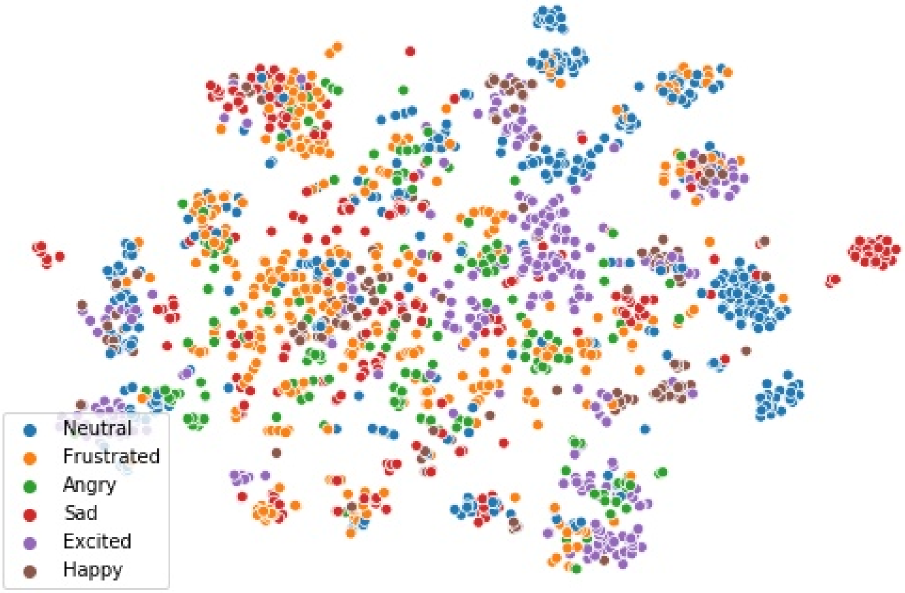


Figure 27: Two-dimensional t-SNE utterance representation of text based on CNN, audio and video.

The second proposed multi-modal method, the **Transformers**, is represented by two-dimensional t-SNE [46] in **Figure 28** where each emotion is cluster into multiple small clusters similarly than in the case of **Baseline**. The main difference here is the middle area where there is still a mix of emotions but with a more distinctive separation, thus, the emotions have more similar distributions in high-dimensional space.

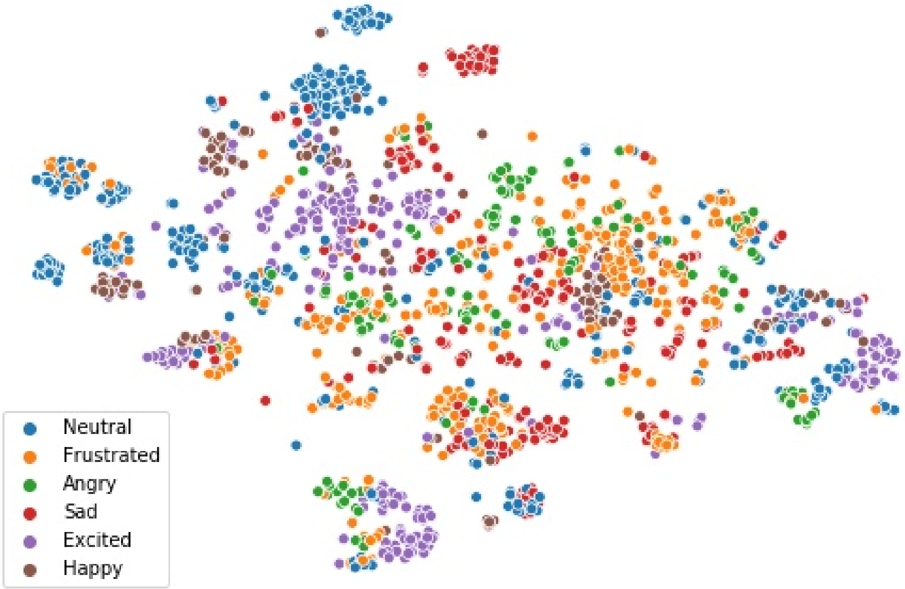


Figure 28: Two-dimensional t-SNE utterance representation of text based on transformer, audio and video.

In the third case using the **Joint Optimization** of features, the patterns for two-dimensional t-SNE [46] in **Figure 29** are different than the previous cases, in fact, the emotions representations show more compact distributions meaning that the emotions are more concentrated in certain areas rather than small clusters in all the space. This is particularly significant since allows to make more precise comparisons between emotions where in general, the negative emotions like frustrated, angry and sad are concentrated in the lower right side of the plot while positive emotions like excited and happy are concentrated in the opposite side, in the upper left side of the plot. In other words, the **Joint Optimization** is able to capture hidden relationships between emotions without previous supervision.

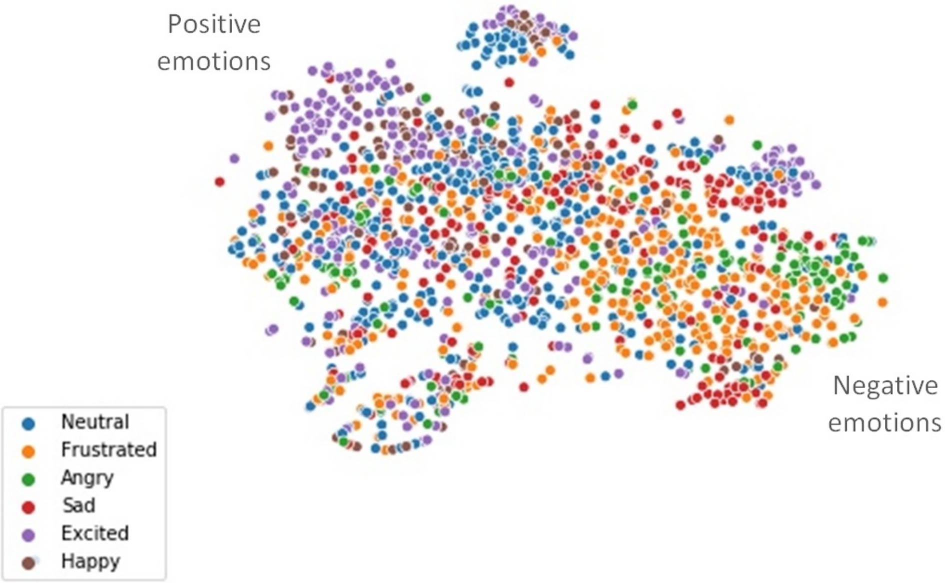


Figure 29: Two-dimensional t-SNE utterance representation of joint optimization of text based on CNN, audio and video.

In summary, these discoveries add a different perspective for the quality evaluation of the utterance representations where the **Joint Optimization** highlights with more informative representations of the emotional relationships.

Further t-SNE [46] dimensionality reduction representations for other modalities can be found in **section B**.

4.2.3 Ablation study

This section presents a more in detail contrast of different models starting from the three proposed methods for multi-modal utterance representation: **Baseline**, **Transformer** and **Joint Optimization**. Since DialogueRNN [20] can be modified to accept different input sized, it is feasible to test all the possible combinations of uni, bi and tri-modalities to obtain a more sensitive measuring of each modalities contribution in the emotion regression task in terms of overall average weighted and average weighted F1-score performance metrics over the test set.

In **Figure 6** are presented the different proposed methods described in **section 4.2** and their combination of modalities. In first place, text based on transformers has the best performance in average weighted accuracy with 61.22% and F1-score with 61.21% for uni-modal evaluations which is correlated with the results at utterance level presented in **section 4.1.1.4**.

As can be expected, the best bi-modal method is the concatenation of text based on transformers and audio with 64.30% and 64.13% for average weighted and average weighted F1-score, respectively, which also represent an improvement of 0.9 and 1.38 percentage points over the average weighted and average weighted F1-score, each, with respect to the state of the art bi-directional DialogueRNN with attention [20] performance presented in **Table 4**.

In third place, the tri-modal evaluations show the results already presented in **Table 5** where the best performance in average weighted accuracy is obtained with **Joint Optimization** with 42.91% and the best for average weighted F1-score is the **Transformers** with 39.26%. These tri-modal performance scores are worse than the best overall combination with text base on transformers and audio, which only use a bi-modal input. The difference can be explained considering that both tri-modal methods imply an exponential increase in complexity with respect to the number of parameters caused by the 3D-CNN architecture used in the video modality, therefore, there is a clear compromise between complexity and performance to be considered in a multi-modal design, where from an optimal point, the increase in complexity produces a reduction in performance.

Methods	**Modalities**	**Average (w)**	
		Accuracy	**F1-score**
Baseline	Text_{CNN}	56.66	57.35
	Audio	22.50	16.00
	Video	18.25	12.87
	Text_{CNN} + Audio	59.19	58.68
	Text_{CNN} + Video	28.11	25.56
	Audio + Video	21.21	14.65
	Text_{CNN} + Audio + Video	30.02	25.42
Transformer	Text_{TE}	61.22	61.21
	Text_{TE} + Audio	**64.30**	**64.13**
	Text_{TE} + Video	44.20	43.68
	Text_{TE} + Audio + Video	41.37	39.26
Joint Optimization	{ Text_{TE} + Audio }	59.80	60.02
	{ Text_{TE} + Audio + Video }	42.91	38.06

Table 6: Ablation study of modalities through different methods for utterance representations using DialogueRNN [20]. Bold font denotes best performance for Weighted average Accuracy and F1-Score.

4.3 Discussion

First, the analysis of performance metrics from the feature extraction methods for utterance representation highlights the importance of text over audio and video modalities. This result is reasonable since the text data is able to compress more informative signals in dyadic interactions compared with the highly sparse audio and video sources that need a strong pre-processing to summarize useful information for emotion recognition. In the case of transformer architecture as an alternative for text representation instead of CNN, it was proven the relevance of a good feature extractor model that enhances the utterance representations and consequently the emotion regression using temporal context.

In this line, the three multi-modal proposed methods show a related improvement when it was used text based on transformers instead of CNN. In particular, the concatenation with text transformer obtains the best average weighted F1-score and the joint optimization of features reaches the best average weighted accuracy. However, the similar achievement of those two methods in terms of performance classification scores is complemented with a qualitative analysis of the utterance representations in two-dimensional space employing t-SNE [46] dimensionality reduction visualization, where the Joint Optimization of features is able to retrieve more compact representations to group emotions using latent relationships among them where positive and negative emotions are distributed in opposite areas describing the intrinsic polarity in emotions. This can use in other areas for emotional analysis of utterance in low dimensions.

Finally, the comparison of multiple combinations of modalities in the ablation study show and improvement over the state of the art using DialogueRNN benchmark results in 0.9 and 1.38 percentage points over the average weighted and average weighted F1-score, respectively. This is interesting since the improvement is obtained using just text based on transformers and audio instead of the three modalities. This is a sign of undermining in performance due to the increase in variability explained by the increment in complexity represented by the number of parameters needed to exploit 3D-CNN architectures.

5 Conclusions and Future Work

This section matches the discoveries extracted in each step of the framework proposed for deep regression of social signals in dyadic scenarios focus on emotion recognition based on multi-modal sources. In **section 5.1** these findings are paired with the objectives defined for the master thesis project, and in **section 5.2** are depicted further lines of study and new perspectives to address the results obtained.

5.1 Conclusions

The purpose of this master thesis is to propose and test a new framework for deep regression of social signals in dyadic scenarios through the exploration of the state of the art methods to enhance multi-modal utterance representations for emotion recognition. For this aim, in **section 1.3** were presented a series of objectives to be fulfilled during this work:

The first three objectives were related to the correct implementation of state of the art methods as a baseline case for further improvements. The first objective defined as the analysis of conventional public databases used in dyadic scenarios analysis. In **section 2.1** were analyzed the three most relevant databases and in **section 3.1** were considered the best option for the problem considering the application and benchmark available for comparisons. In general, this project was analyzed under the IEMOCAP database but the framework can perfectly be applied to any database with multi-modal data. The second objective was the proper pipeline design for feature extraction from raw modalities which was treated in detail in **section 3.2.1** describing all the considerations needed and schemes for pre-processing multi-modal data. The third objective intended to reproduce the state of the art system which was defined in **section 3.2.2** and tested in **section 4.2.1** against an extensive benchmark of methods.

The fourth objective finds to compare the baseline text modality utterance representation based on CNN against a state of the art alternative based on transformer encoder which was presented in **section 3.2.2.1** and proven to be a major improvement in **section 4.1.1** in terms of performance metrics of accuracy and F1-score.

The fifth objective meant to compare the baseline multi-modal representation against a joint optimization of features using a neural network instead of mere concatenation. **Section 4.1.2.3** proves that the joint optimization of feature has a superior performance in term of average weighted accuracy but the most remarkable outcome is that this method is able to represent more compact emotion representations in and describe latent factors like the polarity of emotions as is explained in **section 4.2.2**.

Finally, the sixth objective pretends to make a modality importance evaluation by means of an ablation study. The results from **section 4.2.3** confirmed that text is the most relevant modality and its representation with transformer and audio outperforms the state of the art method using the three modalities.

In summary, all the proposed objectives have been fulfilled delivering informative insights for the emotion recognition system that can be improved considering the future lines of work.

5.2 Future Work

A conclusion was the importance of text modality as a feature to perform emotion regression. The ablation study shown an improvement when it is combined with audio and video however, these last two modalities have lower performance compared with text. Therefore, there is a space for future improvement to generate better audio and video feature representations considering the main issue to tackle which is the sparsity of this type of data. This can be achieved using different types of deep architectures like graph neural networks that have promising results compared with traditional approaches.

Secondly, this work opens new questions about how the emotion's relationships can be represented rather than just perform emotion recognition. In this work, the intrinsic emotion polarity was outlined using the resulting utterance feature representation but they can be improved to represent more complex associations between emotions to understand their latent associations.

References

[1] Bihong Zhang, Lei Xie, Yougen Yuan, Huaiping Ming, Dongyan Huang, and Mingli Song. Deep neural network derived bottleneck features for accurate audio classification. In *2016 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, pages 1–6, 2016.

[2] 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. *Lang. Resour. Evaluation*, 42(4):335–359, 2008. URL `http://dblp.uni-trier.de/db/journals/lre/lre42.html#BussoBLKMKCLN08`.

[3] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1179. URL `https://www.aclweb.org/anthology/D14-1179`.

[4] F. Dellaert, T. Polzin, and A. Waibel. Recognizing emotion in speech. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, volume 3, pages 1970–1973 vol.3, 1996.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL `http://arxiv.org/abs/1810.04805`.

[6] P Ekman. Facial expression and emotion. *The American psychologist*, 48(4): 384—392, April 1993. ISSN 0003-066X. doi: 10.1037//0003-066x.48.4.384. URL `https://doi.org/10.1037//0003-066x.48.4.384`.

[7] Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM International Conference on Multimedia*, MM '13, page 835–838, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450324045. doi: 10.1145/2502081.2502224. URL `https://doi.org/10.1145/2502081.2502224`.

[8] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber. Lstm: A search space odyssey. *IEEE Transactions on Neural Networks and Learning Systems*, 28(10):2222–2232, 2017.

[9] A. Gupta and H. Gupta. Applications of mfcc and vector quantization in speaker recognition. In *2013 International Conference on Intelligent Systems and Signal Processing (ISSP)*, pages 170–173, 2013.

[10] Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. Conversational memory network for emotion recognition in dyadic dialogue videos. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2122–2132, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1193. URL `https://www.aclweb.org/anthology/N18-1193`.

[11] Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *CoRR*, abs/1207.0580, 2012. URL `http://arxiv.org/abs/1207.0580`.

[12] Sepp Hochreiter. Untersuchungen zu dynamischen neuronalen netzen. 1991. URL `https://academic.microsoft.com/paper/194249466`.

[13] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, November 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.8.1735. URL `https://doi.org/10.1162/neco.1997.9.8.1735`.

[14] Sepp Hochreiter, Yoshua Bengio, and Paolo Frasconi. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies. In J. Kolen and S. Kremer, editors, *Field Guide to Dynamical Recurrent Networks*. IEEE Press, 2001.

[15] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR*, abs/1502.03167, 2015. URL `http://arxiv.org/abs/1502.03167`.

[16] Yoon Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1181. URL `https://www.aclweb.org/anthology/D14-1181`.

[17] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv e-prints*, art. arXiv:1412.6980, December 2014.

[18] Yann Lecun. *PhD thesis: Modeles connexionnistes de l'apprentissage (connectionist learning models)*. Universite P. et M. Curie (Paris 6), June 1987.

[19] B. Liu, X. Yu, P. Zhang, A. Yu, Q. Fu, and X. Wei. Supervised deep feature extraction for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 56(4):1909–1921, 2018.

[20] Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander F. Gelbukh, and Erik Cambria. Dialoguernn: An attentive RNN

for emotion detection in conversations. *CoRR*, abs/1811.00405, 2018. URL http://arxiv.org/abs/1811.00405.

[21] D. Marmanis, M. Datcu, T. Esch, and U. Stilla. Deep learning earth observation classification using imagenet pretrained networks. *IEEE Geoscience and Remote Sensing Letters*, 13(1):105–109, 2016.

[22] G. McKeown, M.F. Valstar, R. Cowie, Maja Pantic, and M. Schroeder. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE transactions on affective computing*, 3(1):5–17, 1 2012. ISSN 1949-3045. doi: 10.1109/T-AFFC.2011.20. eemcs-eprint-22960.

[23] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 3111–3119, Red Hook, NY, USA, 2013. Curran Associates Inc.

[24] Xiaoyue Feng Xiaojing Fan Zhili Pei Yu Xue Renchu Guan Mingyang Jiang, Yanchun Liang. Text classification based on deep belief network and softmax regression. *Neural Computing and Applications*, 29(1):61–70, 2016. ISSN 0941-0643. doi: 10.1007/s00521-016-2401-x.

[25] Lindasalwa Muda, Mumtaj Begam, and I. Elamvazuthi. Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques. *CoRR*, abs/1003.4083, 2010. URL http://arxiv.org/abs/1003.4083.

[26] Bhalaji Nagarajan and V. Ramana Murthy Oruganti. Deep learning as feature encoding for emotion recognition. *ArXiv*, abs/1810.12613, 2018.

[27] Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, page 807–814, Madison, WI, USA, 2010. Omnipress. ISBN 9781605589077.

[28] Michalis Papakostas, Evaggelos Spyrou, Theodoros Giannakopoulos, Giorgos Siantikos, Dimitrios Sgouropoulos, Phivos Mylonas, and Fillia Makedon. Deep visual attributes vs. hand-crafted audio features on multidomain speech emotion recognition. *Computation*, 5(4):26, Jun 2017. ISSN 2079-3197. doi: 10.3390/computation5020026. URL http://dx.doi.org/10.3390/computation5020026.

[29] Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

[30] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. URL http://www.aclweb.org/anthology/D14-1162.

[31] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 873–883, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1081. URL https://www.aclweb.org/anthology/P17-1081.

[32] JONATHAN POSNER, JAMES A. RUSSELL, and BRADLEY S. PETERSON. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and Psychopathology*, 17(3):715–734, 2005. doi: 10.1017/S0954579405050340.

[33] Ofir Press and Lior Wolf. Using the output embedding to improve language models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 157–163, Valencia, Spain, April 2017. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/E17-2025.

[34] Jane M. Richards, Emily A. Butler, and James J. Gross. Emotion regulation in romantic relationships: The cognitive consequences of concealing feelings. *Journal of Social and Personal Relationships*, 20(5):599–620, 2003. doi: 10.1177/02654075030205002. URL https://doi.org/10.1177/02654075030205002.

[35] Björn Schuller, Michel Valstar, Florian Eyben, Gary McKeown, Roddy Cowie, and Maja Pantic. Avec 2011-the first international audio/visual emotion challenge. In *Proceedings of the 4th International Conference on Affective Computing and Intelligent Interaction - Volume Part II*, ACII'11, page 415–424, Berlin, Heidelberg, 2011. Springer-Verlag. ISBN 9783642245701.

[36] Björn Schuller, Michel Valster, Florian Eyben, Roddy Cowie, and Maja Pantic. Avec 2012: The continuous audio/visual emotion challenge. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction*, ICMI '12, page 449–456, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450314671. doi: 10.1145/2388676.2388776. URL https://doi.org/10.1145/2388676.2388776.

[37] Björn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian

Eyben, Erik Marchi, Marcello Martillaro, Hugues Salamin, Anna Polychroniou, Fabio Valente, and Samuel Kim. The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism. In ISCA, editor, *Proceedings of the 14th Annual Conference of the International Speech Communication Association (INTERSPEECH 2013)*, pages 148–152, 2013. URL `http://www5.informatik.uni-erlangen.de/Forschung/Publikationen/2013/Schuller13-TI2.pdf`.

[38] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/j.1538-7305.1948.tb01338.x`.

[39] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *CoRR*, abs/1406.2199, 2014. URL `http://arxiv.org/abs/1406.2199`.

[40] Sainbayar Sukhbaatar, arthur szlam, Jason Weston, and Rob Fergus. End-to-end memory networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2440–2448. Curran Associates, Inc., 2015. URL `http://papers.nips.cc/paper/5846-end-to-end-memory-networks.pdf`.

[41] H. Tachibana, K. Uenoyama, and S. Aihara. Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4784–4788, 2018.

[42] Wilson L. Taylor. ”cloze procedure”: a new tool for measuring readability. *Journalism & Mass Communication Quarterly*, 30:415–433, 1953.

[43] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, and J. Ortega-Garcia. Exploring recurrent neural networks for on-line handwritten signature biometrics. *IEEE Access*, 6:5128–5138, 2018.

[44] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[45] Michel F. Valstar, Jonathan Gratch, Björn W. Schuller, Fabien Ringeval, Denis Lalanne, Mercedes Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic. AVEC 2016 - depression, mood, and emotion recognition workshop and challenge. *CoRR*, abs/1605.01600, 2016. URL `http://arxiv.org/abs/1605.01600`.

[46] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008. URL `http://www.jmlr.org/papers/v9/vandermaaten08a.html`.

[47] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017. URL http://arxiv.org/abs/1706.03762.

[48] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. MOSI: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *CoRR*, abs/1606.06259, 2016. URL http://arxiv.org/abs/1606.06259.

[49] Peng Zhou, Zhenyu Qi, Suncong Zheng, Jiaming Xu, Hongyun Bao, and Bo Xu. Text classification improved by integrating bidirectional LSTM with two-dimensional max pooling. *CoRR*, abs/1611.06639, 2016. URL http://arxiv.org/abs/1611.06639.

Appendices

A Neural Networks Architectures

This section presents the graph neural networks produces for each feature extraction models. Each models contains the type of layer used and the dimensions for input and output vectors.

A.1 Independent feature extraction models

A.1.1 Text based on CNN feature extraction model

In text based on CNN there are three parallel CNN with different sizes that are combined for utterance prediction.

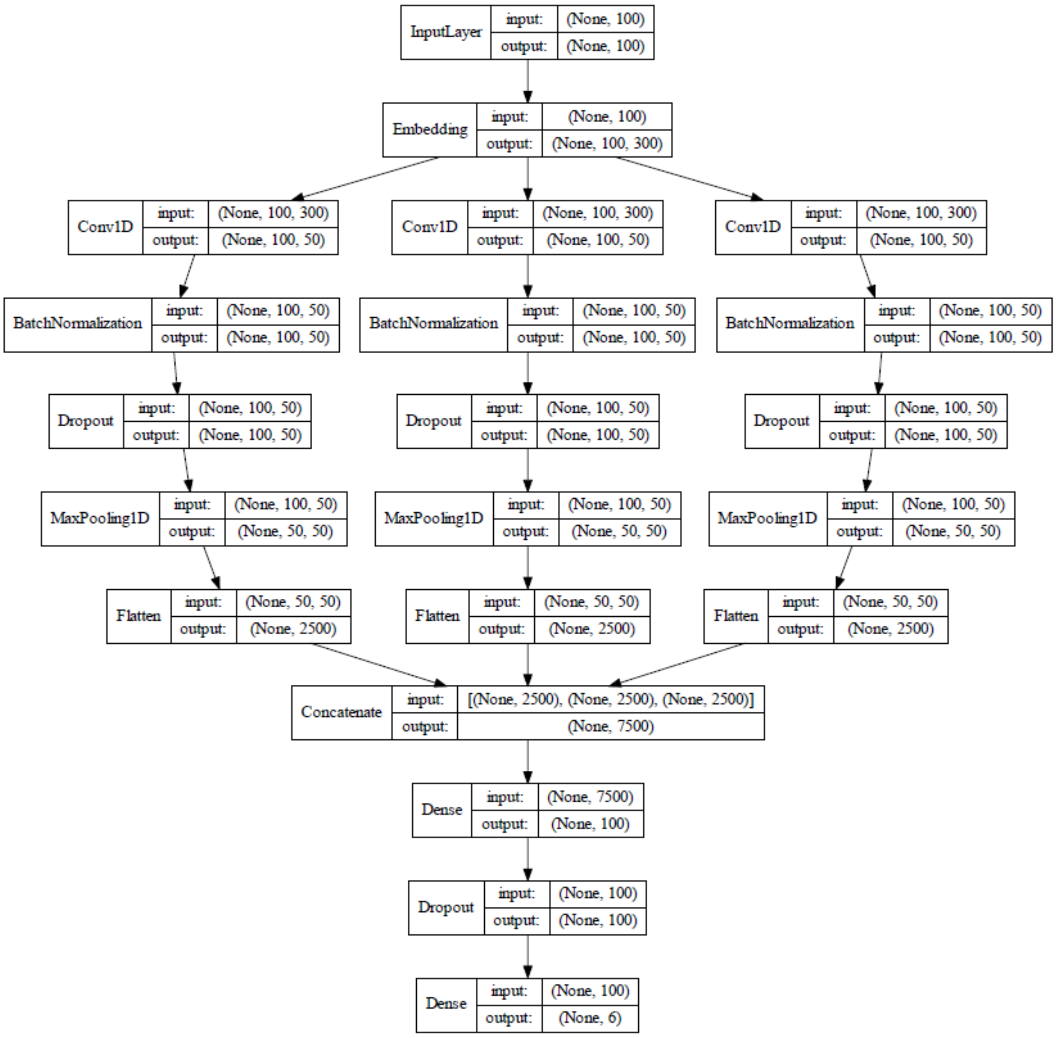

InputLayer | input: (None, 100) | output: (None, 100)
Embedding | input: (None, 100) | output: (None, 100, 300)
Conv1D | input: (None, 100, 300) | output: (None, 100, 50)
Conv1D | input: (None, 100, 300) | output: (None, 100, 50)
Conv1D | input: (None, 100, 300) | output: (None, 100, 50)
BatchNormalization | input: (None, 100, 50) | output: (None, 100, 50)
BatchNormalization | input: (None, 100, 50) | output: (None, 100, 50)
BatchNormalization | input: (None, 100, 50) | output: (None, 100, 50)
Dropout | input: (None, 100, 50) | output: (None, 100, 50)
Dropout | input: (None, 100, 50) | output: (None, 100, 50)
Dropout | input: (None, 100, 50) | output: (None, 100, 50)
MaxPooling1D | input: (None, 100, 50) | output: (None, 50, 50)
MaxPooling1D | input: (None, 100, 50) | output: (None, 50, 50)
MaxPooling1D | input: (None, 100, 50) | output: (None, 50, 50)
Flatten | input: (None, 50, 50) | output: (None, 2500)
Flatten | input: (None, 50, 50) | output: (None, 2500)
Flatten | input: (None, 50, 50) | output: (None, 2500)
Concatenate | input: [(None, 2500), (None, 2500), (None, 2500)] | output: (None, 7500)
Dense | input: (None, 7500) | output: (None, 100)
Dropout | input: (None, 100) | output: (None, 100)
Dense | input: (None, 100) | output: (None, 6)

Figure 30: Text based on CNN architecture.

A.1.2 Text based on transformer encoder feature extraction model

The transformer encoder is represented by the Token and Positional Embedding layer and the Transformer Block layer.

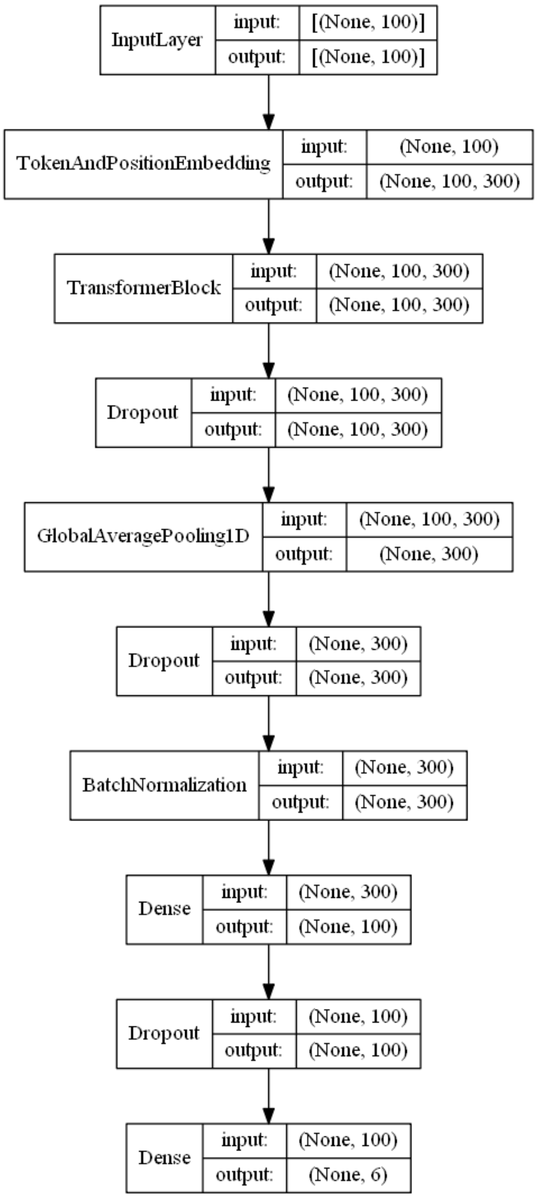


Figure 31: Text based on transformer architecture.

A.1.3 Audio feature extraction model

Audio architecture is a simple network with two fully-connected layers with 100 neurons and 6 neurons, respectively.

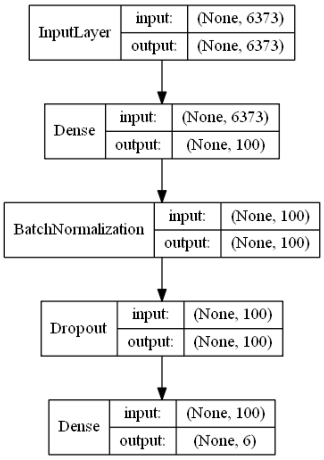


Figure 32: Audio architecture.

A.1.4 Video feature extraction model

The video architecture contains the biggest architecture en terms of number of trainable parameters caused by the used of 3D-CNN.

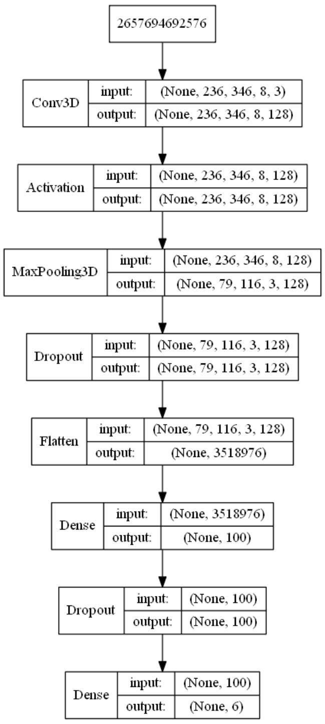

2657694692576
Conv3D
input: (None, 236, 346, 8, 3)
output: (None, 236, 346, 8, 128)
Activation
input: (None, 236, 346, 8, 128)
output: (None, 236, 346, 8, 128)
MaxPooling3D
input: (None, 236, 346, 8, 128)
output: (None, 79, 116, 3, 128)
Dropout
input: (None, 79, 116, 3, 128)
output: (None, 79, 116, 3, 128)
Flatten
input: (None, 79, 116, 3, 128)
output: (None, 3518976)
Dense
input: (None, 3518976)
output: (None, 100)
Dropout
input: (None, 100)
output: (None, 100)
Dense
input: (None, 100)
output: (None, 6)

Figure 33: Video architecture.

A.2 Joint optimization of features

Here are presented the feature extraction models for Joint Optimization of features combined in an early stage.

A.2.1 Text based on transformer and audio feature extraction model

Text in the left side and audio in the right side, are combined and by means of a dense layer of 200 neurons (100 per each modality), the features are obtained.

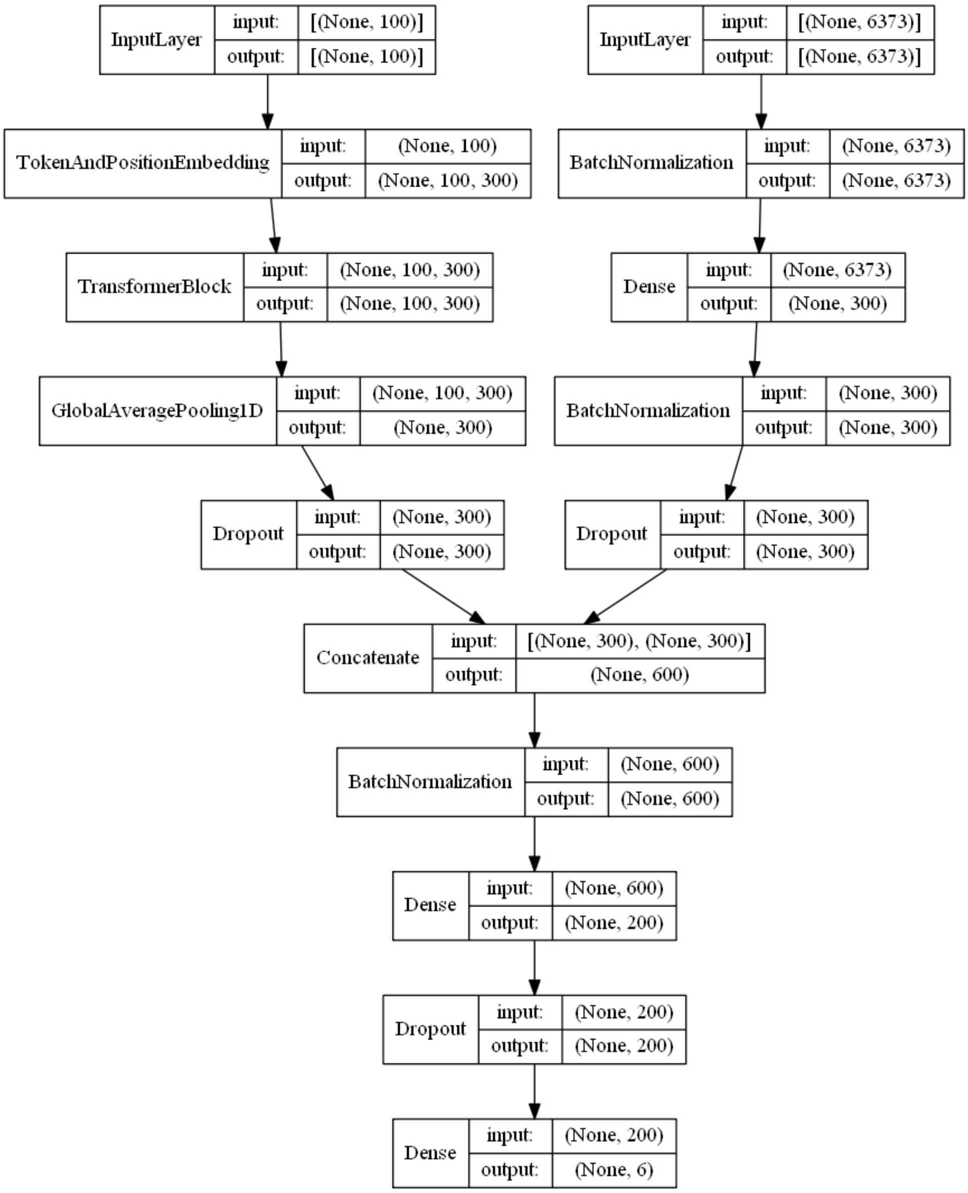


Figure 34: Joint optimization of text based on transformer and audio architecture.

A.2.2 Text bases on transformer, audio and video feature extraction model

This is the more complex network in terms of number of parameters. In the left is the video input architecture, in the middle the text based on transformer architecture and in the right side the audio architecture. All of them are combined to extract the 300-dimension features (100 dimensions per feature).

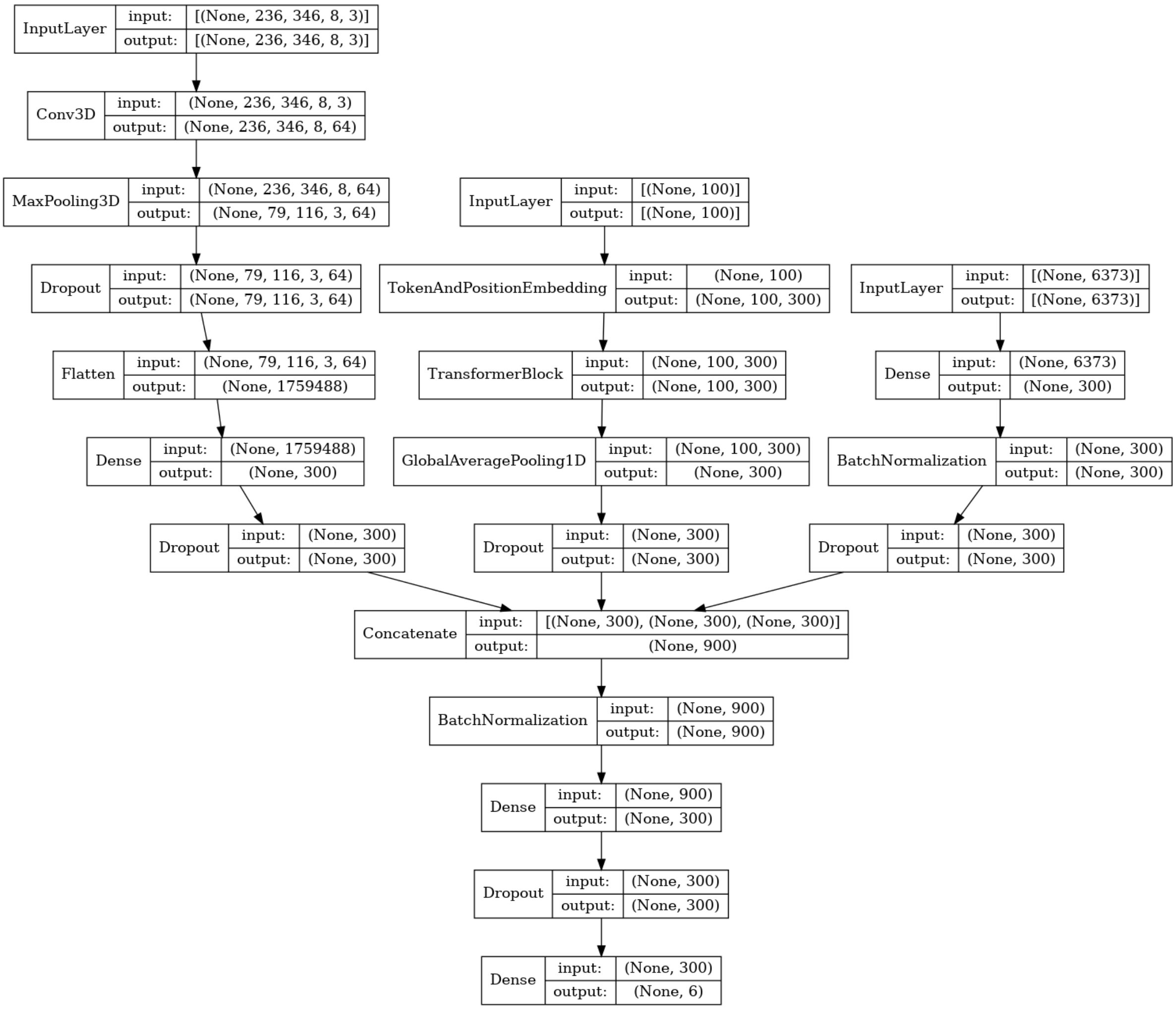

InputLayer | input: [(None, 236, 346, 8, 3)] | output: [(None, 236, 346, 8, 3)]
Conv3D | input: (None, 236, 346, 8, 3) | output: (None, 236, 346, 8, 64)
MaxPooling3D | input: (None, 236, 346, 8, 64) | output: (None, 79, 116, 3, 64)
Dropout | input: (None, 79, 116, 3, 64) | output: (None, 79, 116, 3, 64)
Flatten | input: (None, 79, 116, 3, 64) | output: (None, 1759488)
Dense | input: (None, 1759488) | output: (None, 300)
Dropout | input: (None, 300) | output: (None, 300)
InputLayer | input: [(None, 100)] | output: [(None, 100)]
TokenAndPositionEmbedding | input: (None, 100) | output: (None, 100, 300)
TransformerBlock | input: (None, 100, 300) | output: (None, 100, 300)
GlobalAveragePooling1D | input: (None, 100, 300) | output: (None, 300)
Dropout | input: (None, 300) | output: (None, 300)
InputLayer | input: [(None, 6373)] | output: [(None, 6373)]
Dense | input: (None, 6373) | output: (None, 300)
BatchNormalization | input: (None, 300) | output: (None, 300)
Dropout | input: (None, 300) | output: (None, 300)
Concatenate | input: [(None, 300), (None, 300), (None, 300)] | output: (None, 900)
BatchNormalization | input: (None, 900) | output: (None, 900)
Dense | input: (None, 900) | output: (None, 300)
Dropout | input: (None, 300) | output: (None, 300)
Dense | input: (None, 300) | output: (None, 6)

Figure 35: Joint optimization of text based on transformer, audio and video architecture.

B Qualitative study

B.1 Dimensionality Reduction visualization

B.1.1 Text based on CNN t-SNE representation

Dimensionality reduction for text based on CNN.

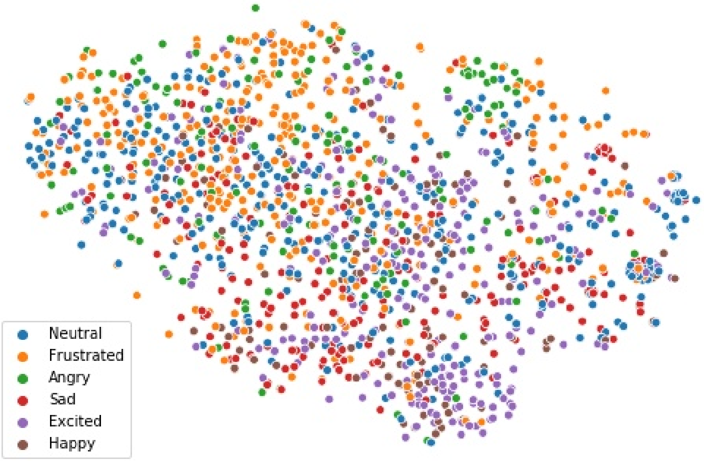


Figure 36: Two-dimensional t-SNE utterance representation of text based on CNN.

B.1.2 Audio t-SNE representation

Audio shows a particular patter that can be justify by similar distribution of vectors.

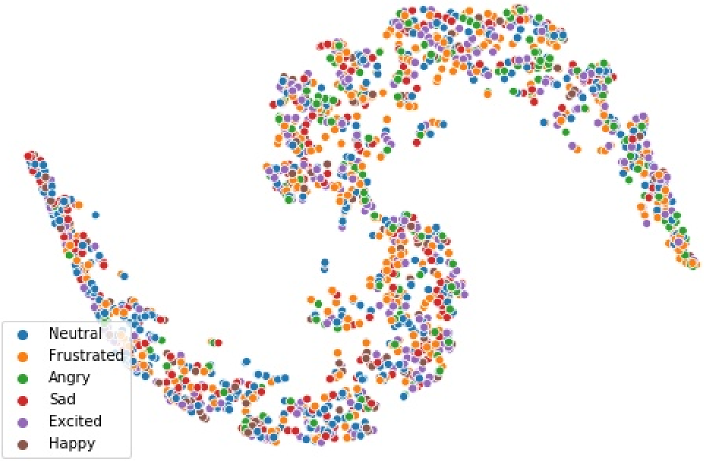


Figure 37: Two-dimensional t-SNE utterance representation of audio.

B.1.3 Video t-SNE representation

Video presents very sparse clusters with small number of samples each.

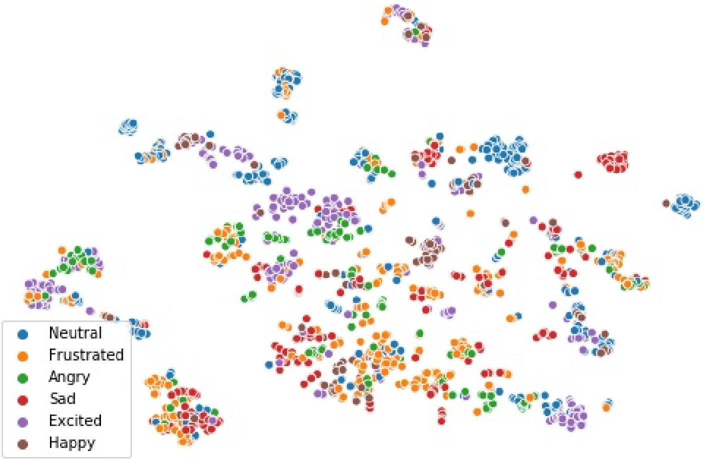


Figure 38: Two-dimensional t-SNE utterance representation of video.

B.1.4 Text based on transformer t-SNE representation

Similar to text based on CNN, transformer present a mixed patter with sparse representations.

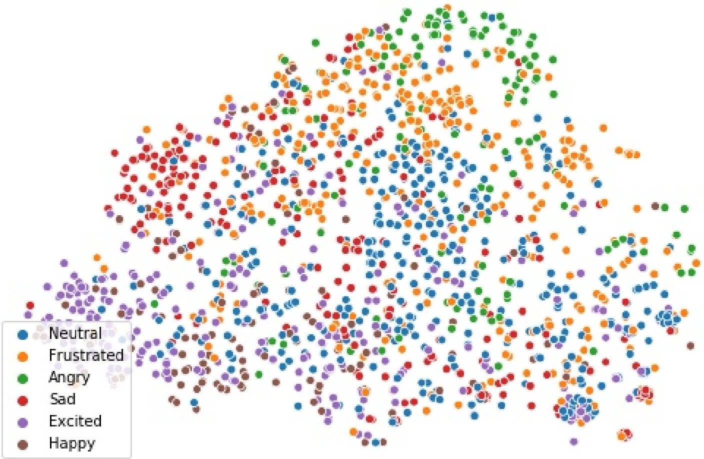


Figure 39: Two-dimensional t-SNE utterance representation of text based on transformer.

B.1.5 Joint optimization of text based on transformer and audio t-SNE representation

The joint optimization of text based on transformers and audio split emotions in more stable representations.

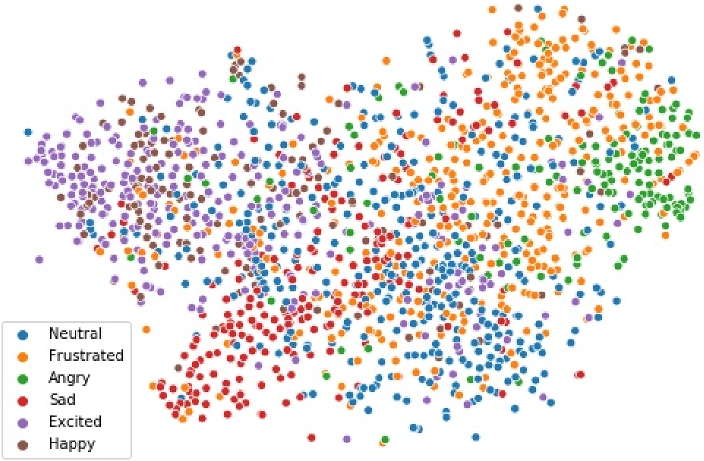


Figure 40: Two-dimensional t-SNE utterance representation of joint optimization of text based on transformer and audio.

C Training Performance

C.1 Independent feature extraction models training

This section presents the training performance over train and validation sets for independent feature extraction models.

C.1.1 Text based on CNN feature extraction training

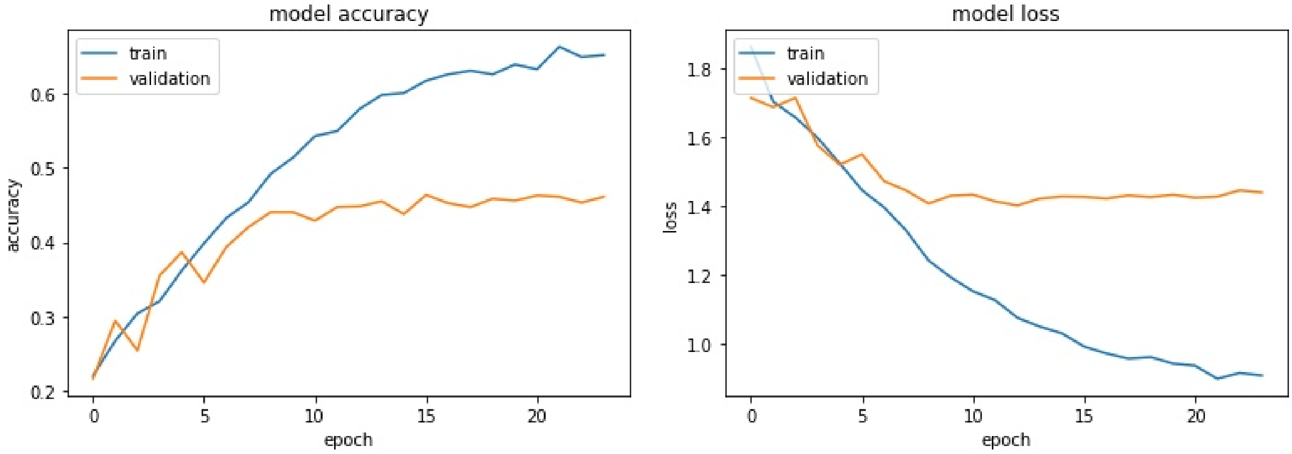


Figure 41: Accuracy and Loss performance during training in text based on CNN feature extraction model.

C.1.2 Text based on transformer feature extraction training

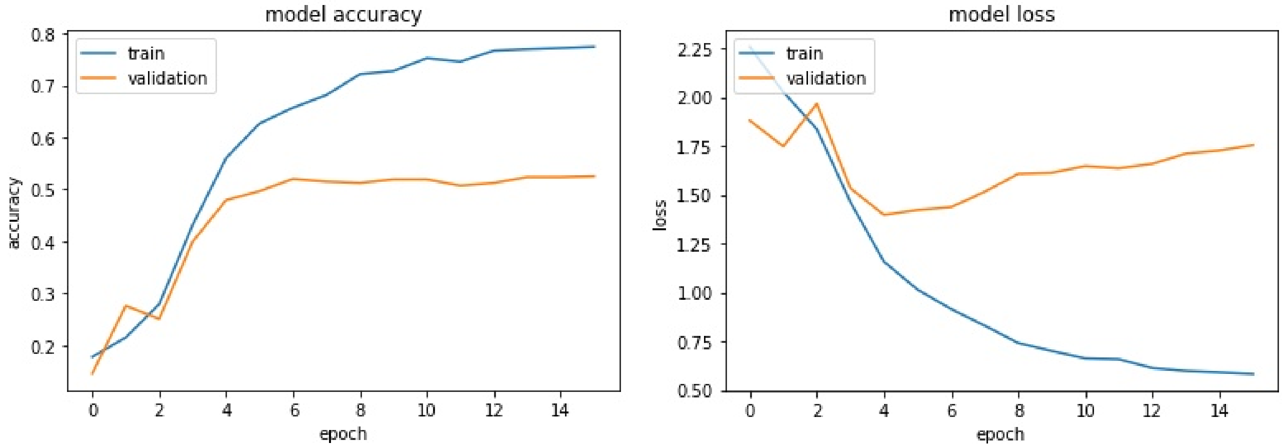


Figure 42: Accuracy and Loss performance during training in text based on transformer feature extraction model.

C.1.3 Audio feature extraction training

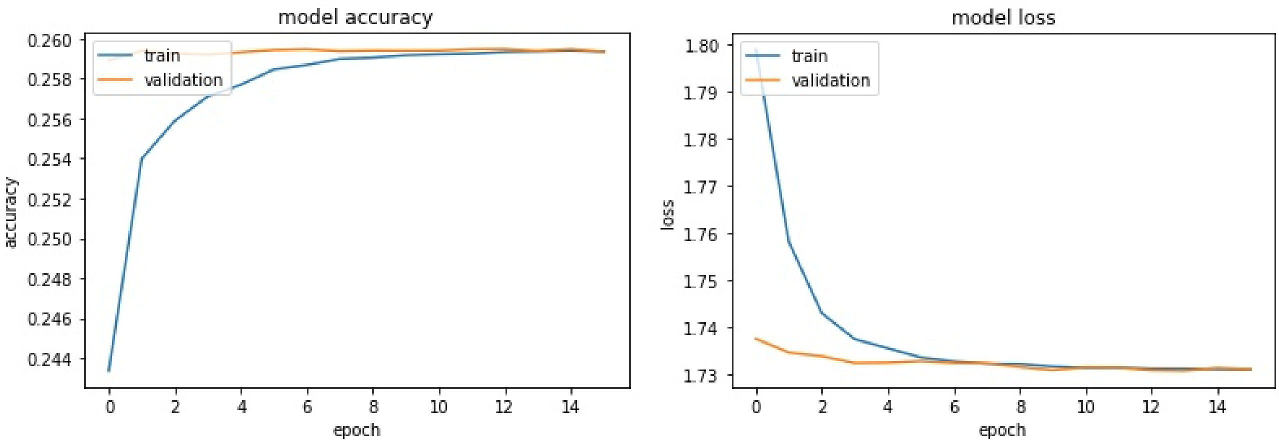


Figure 43: Accuracy and Loss performance during training in audio feature extraction model.

C.1.4 Video feature extraction training

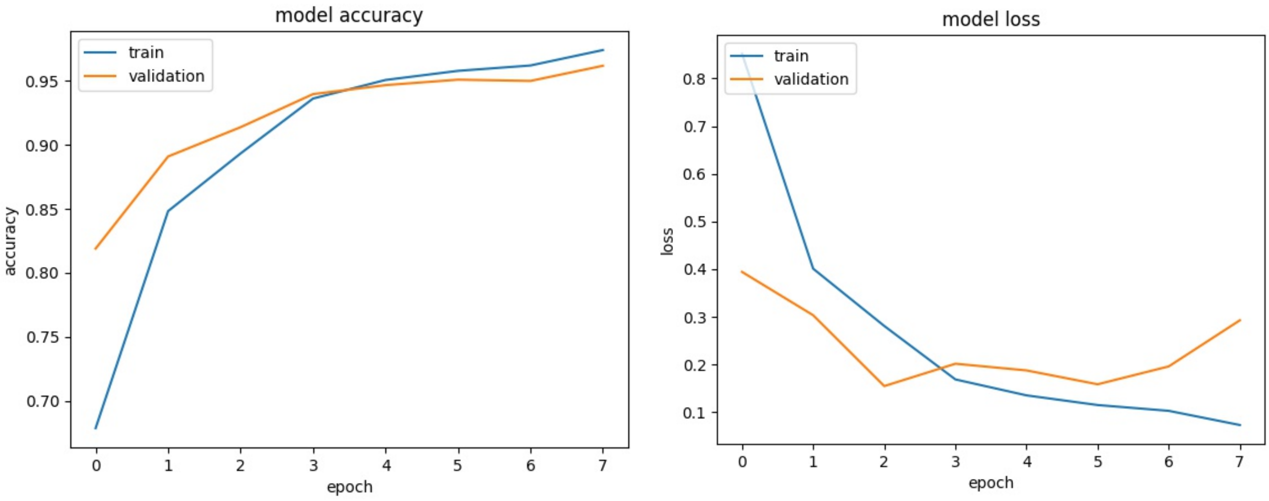


Figure 44: Accuracy and Loss performance during training in video feature extraction model.

C.2 Joint optimization of features models training

This section presents the training performance over train and validation sets for joint optimization of features extraction models.

C.2.1 Text based on CNN transformer and audio training

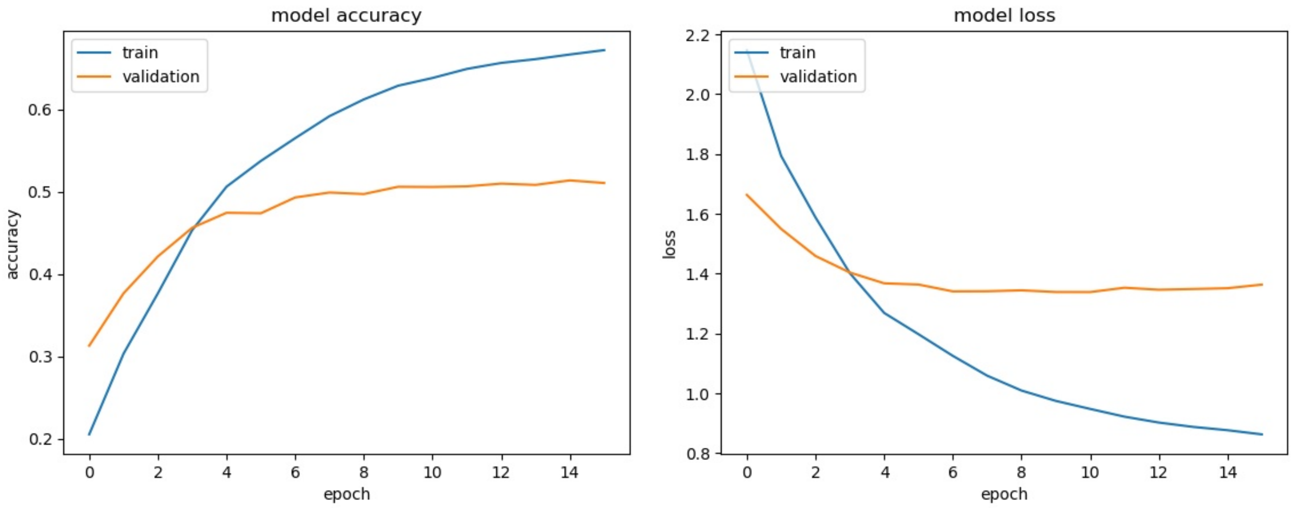


Figure 45: Accuracy and Loss performance during training in text based on transformer and audio feature extraction model.

C.2.2 Text based on CNN transformer, audio and video training

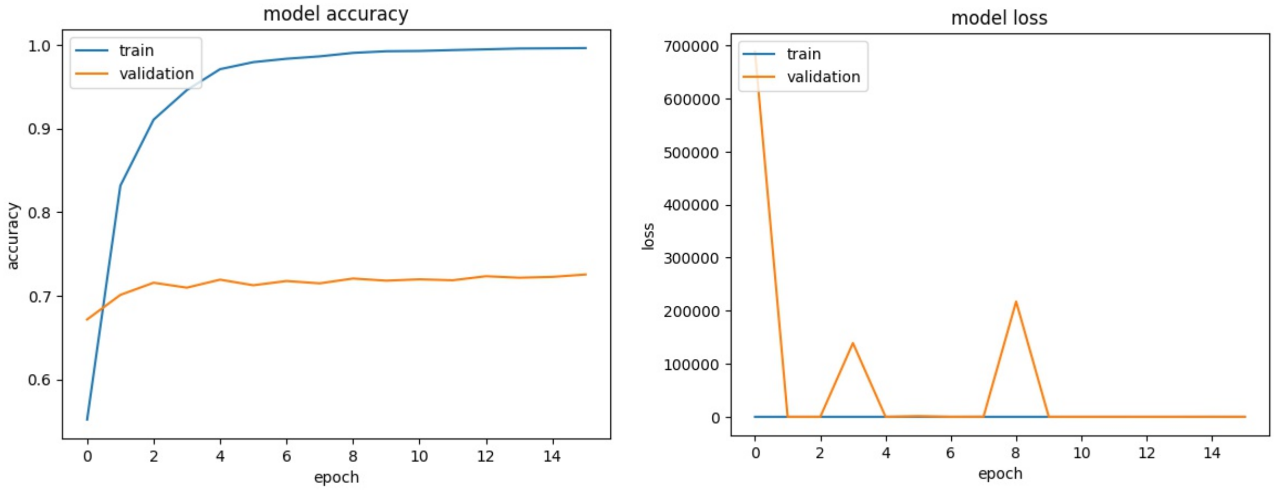


Figure 46: Accuracy and Loss performance during training in text based on transformer, audio and video feature extraction model.

C.3 Multi-Modal Emotion Recognition models training

Here are presented the performance during training using only test set as monitoring to select the best model in terms of the lowest categorical cross-entropy loss function.

C.3.1 Baseline training

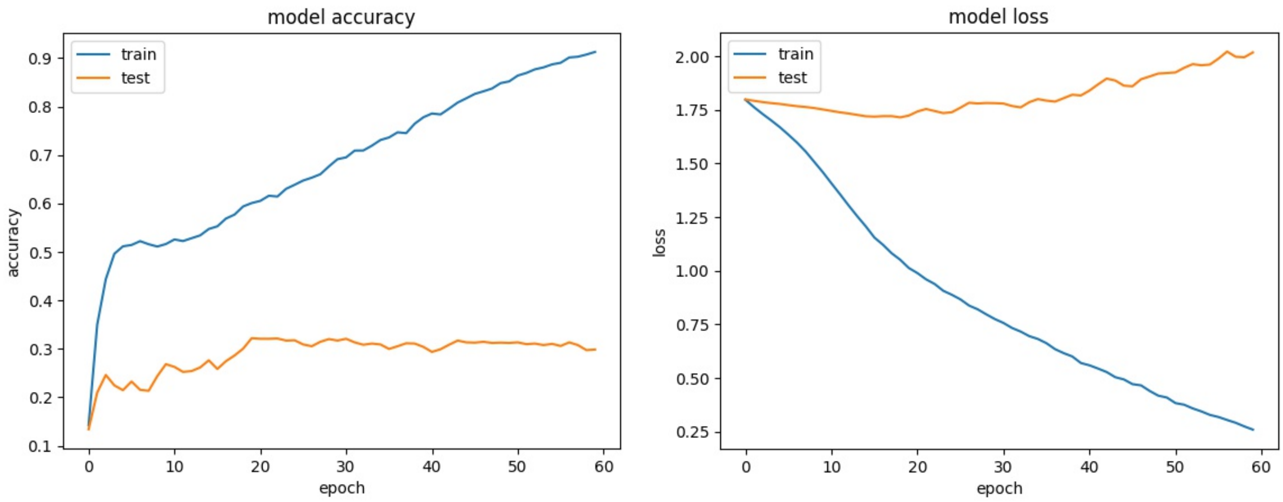


Figure 47: Accuracy and Loss performance during training with text based on CNN, audio and video in DialogueRNN.

C.3.2 Multi-modal with Transformer training

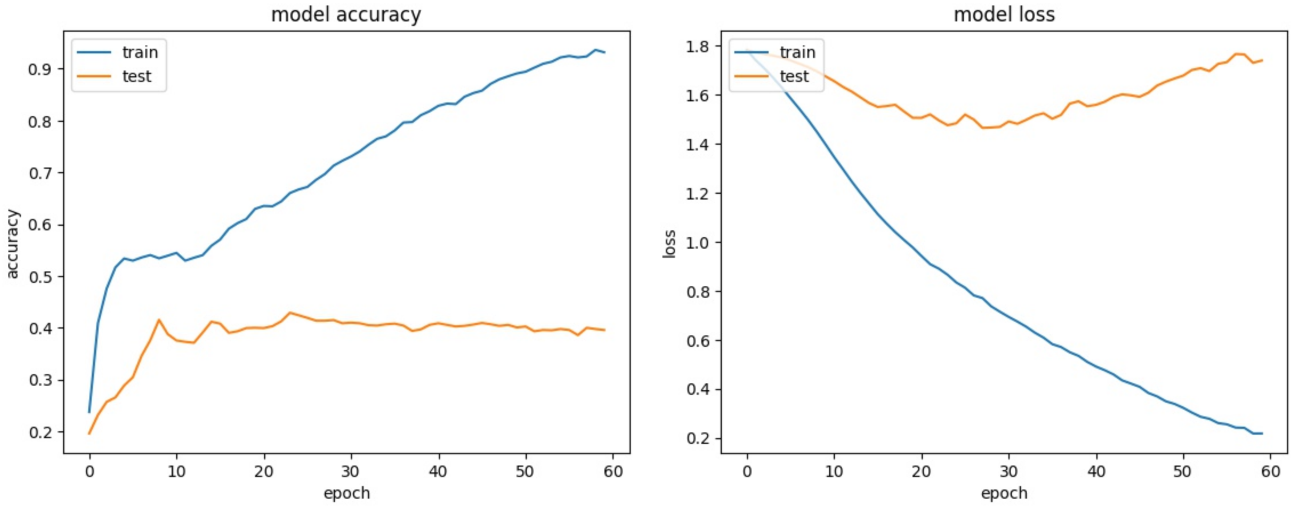


Figure 48: Accuracy and Loss performance during training with text based on transformers, audio and video in DialogueRNN.

C.3.3 Multi-modal Joint Optimization with Transformers training

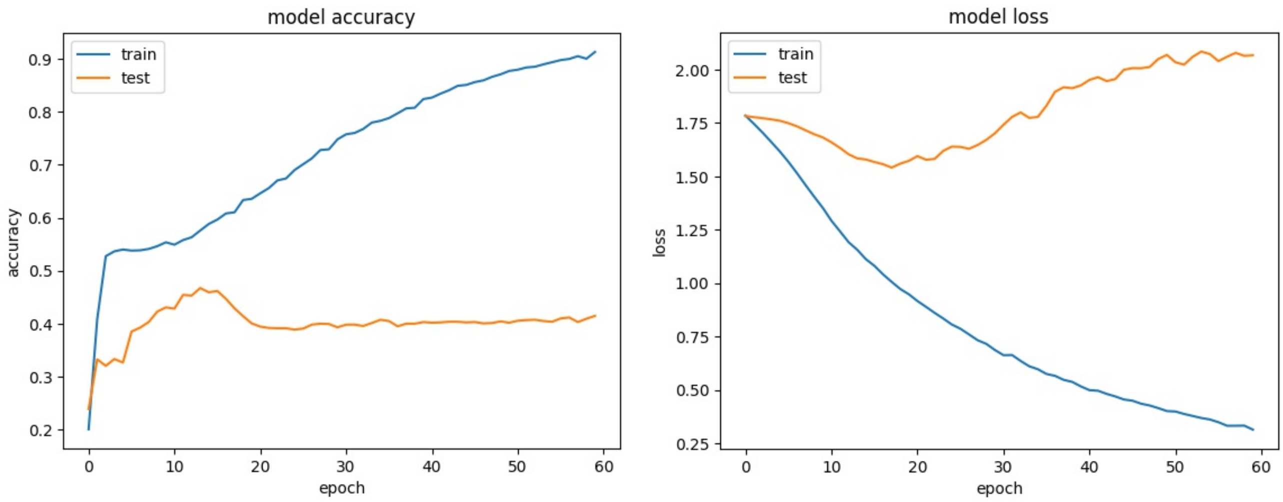


Figure 49: Accuracy and Loss performance during training with joint optimization of text based on transformers, audio and video in DialogueRNN.

C.4 Ablation study training

C.4.1 Text based on CNN DialogueRNN training

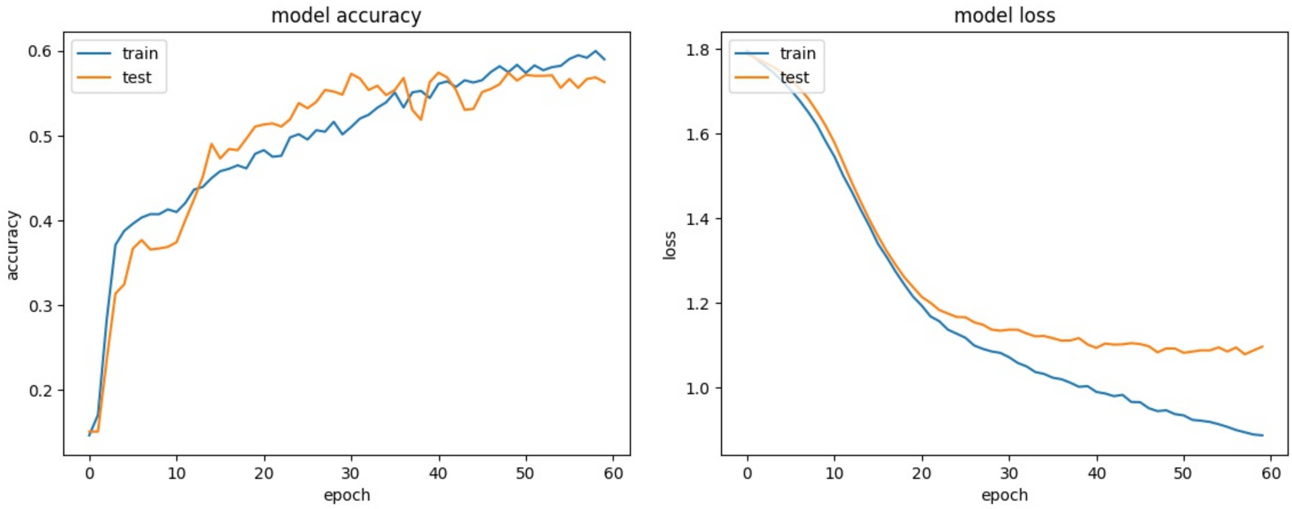


Figure 50: Accuracy and Loss performance during training with text based on CNN in DialogueRNN.

C.4.2 Audio DialogueRNN training

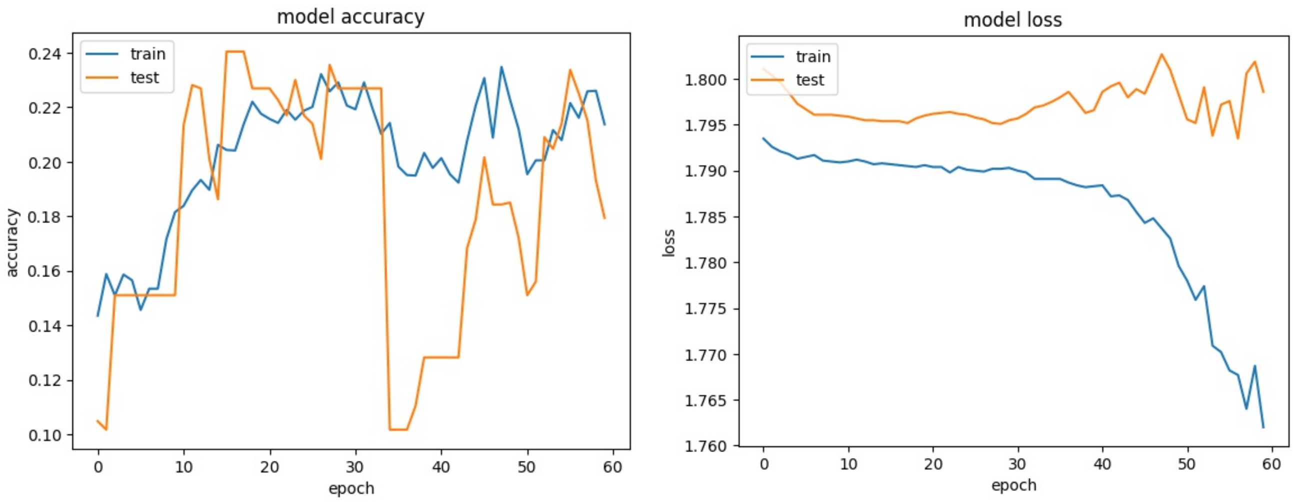


Figure 51: Accuracy and Loss performance during training with audio in DialogueRNN.

C.4.3 Video DialogueRNN training

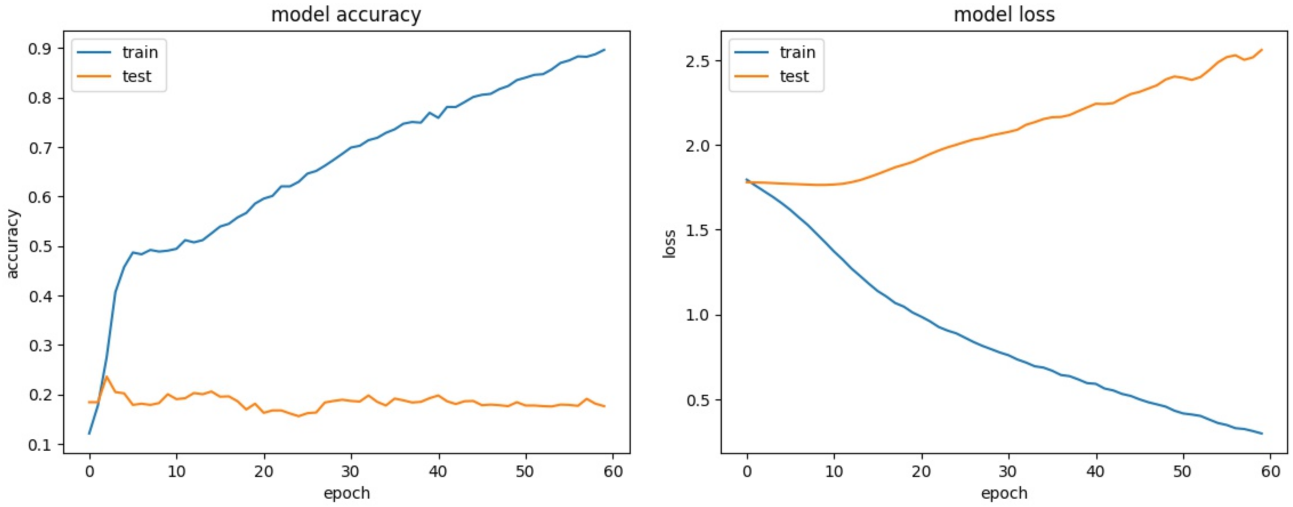


Figure 52: Accuracy and Loss performance during training with video in DialogueRNN.

C.4.4 Text based on CNN and audio DialogueRNN training

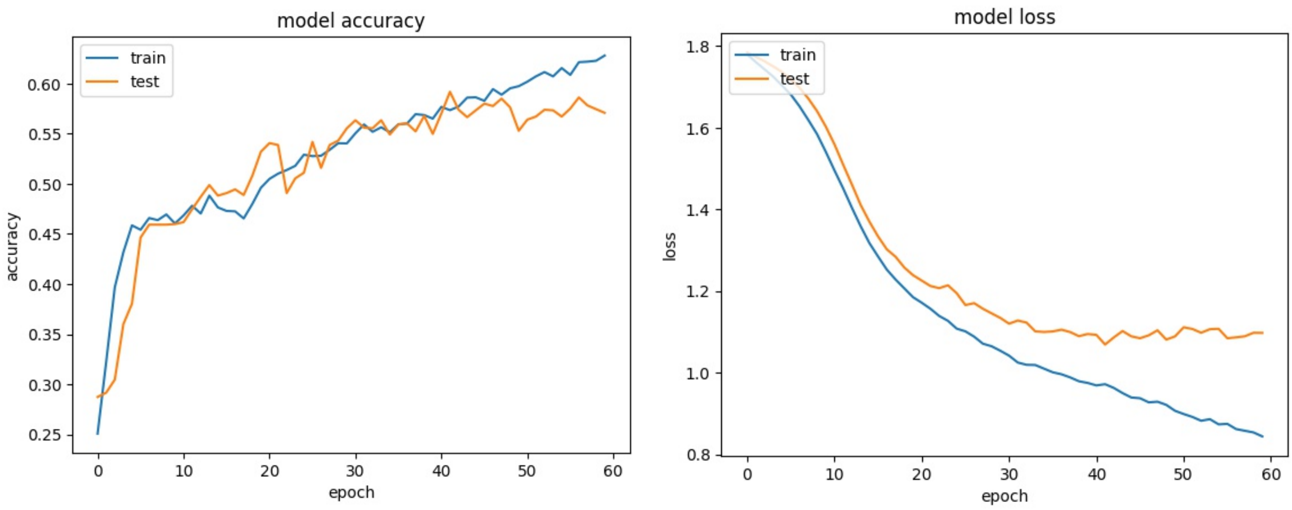


Figure 53: Accuracy and Loss performance during training with text based on CNN and audio in DialogueRNN.

C.4.5 Text based on CNN and video DialogueRNN training

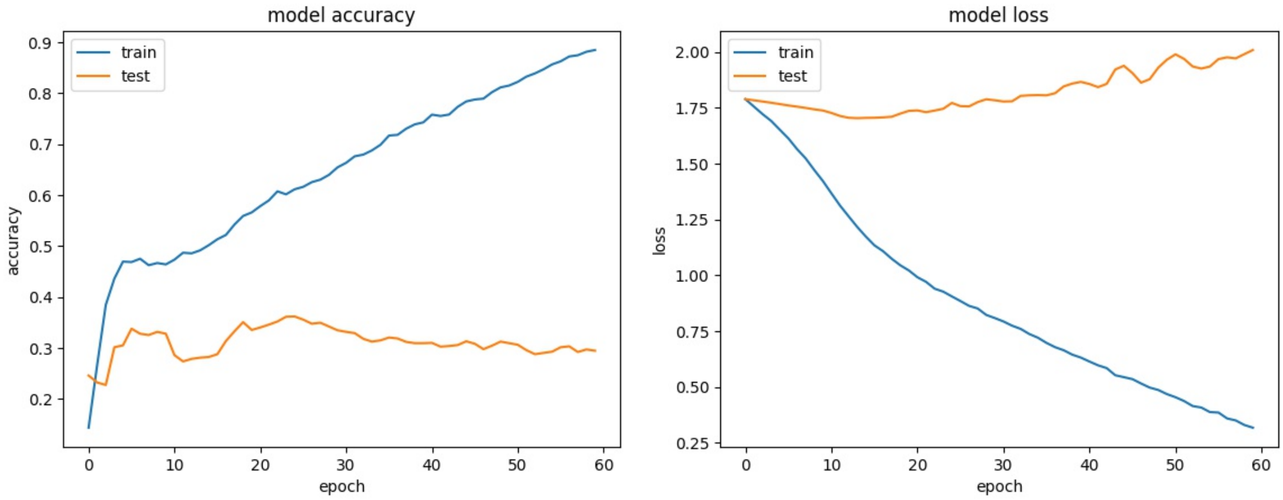


Figure 54: Accuracy and Loss performance during training with text based on CNN and video in DialogueRNN.

C.4.6 Audio and video DialogueRNN training

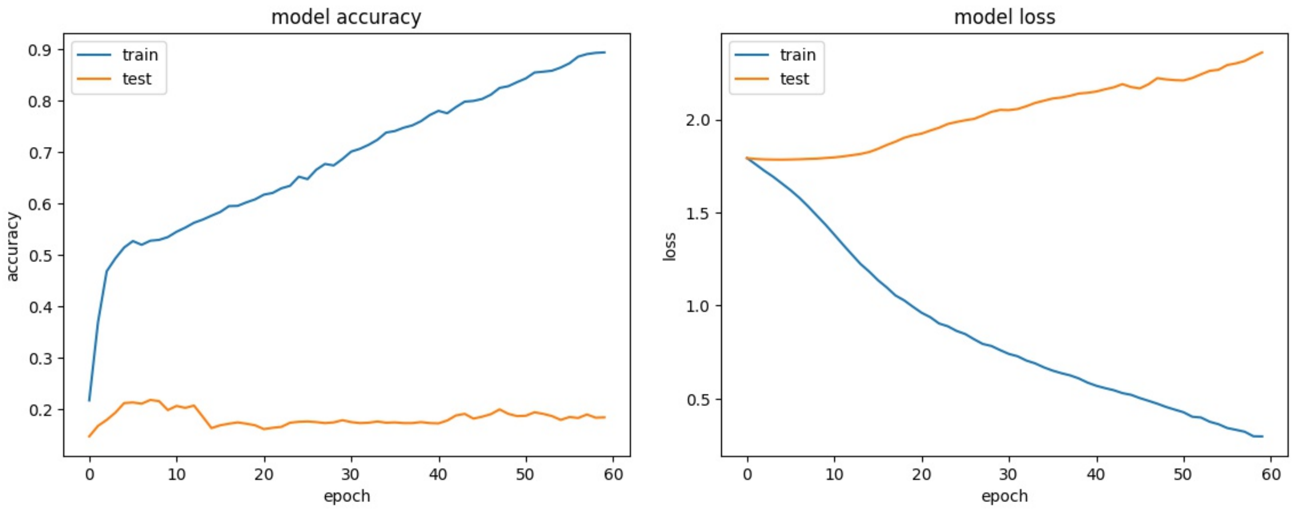


Figure 55: Accuracy and Loss performance during training with audio and video in DialogueRNN.

C.4.7 Text based on transformer encoder DialogueRNN training

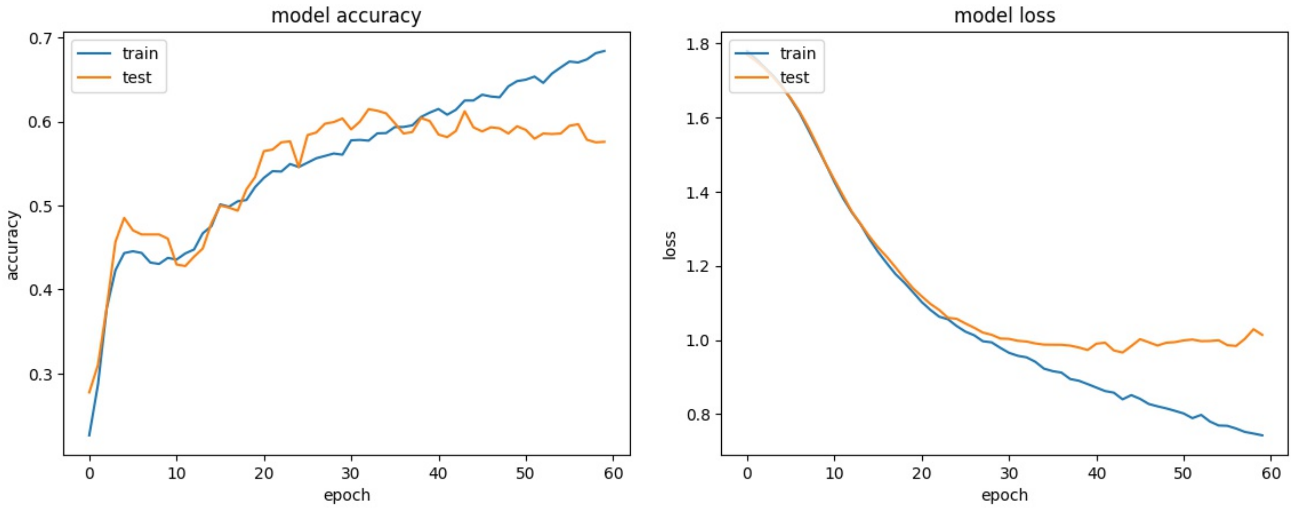


Figure 56: Accuracy and Loss performance during training with text based on transformers in DialogueRNN.

C.4.8 Text based on transformer encoder and audio DialogueRNN training

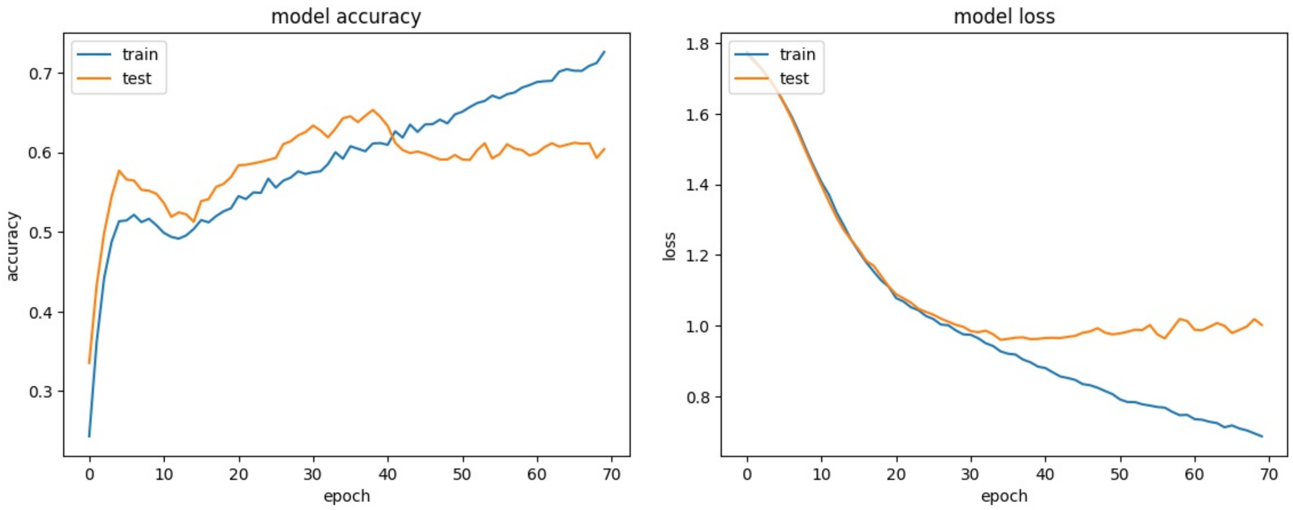


Figure 57: Accuracy and Loss performance during training with text based on transformers and audio in DialogueRNN.

C.4.9 Text based on transformer encoder and video DialogueRNN training

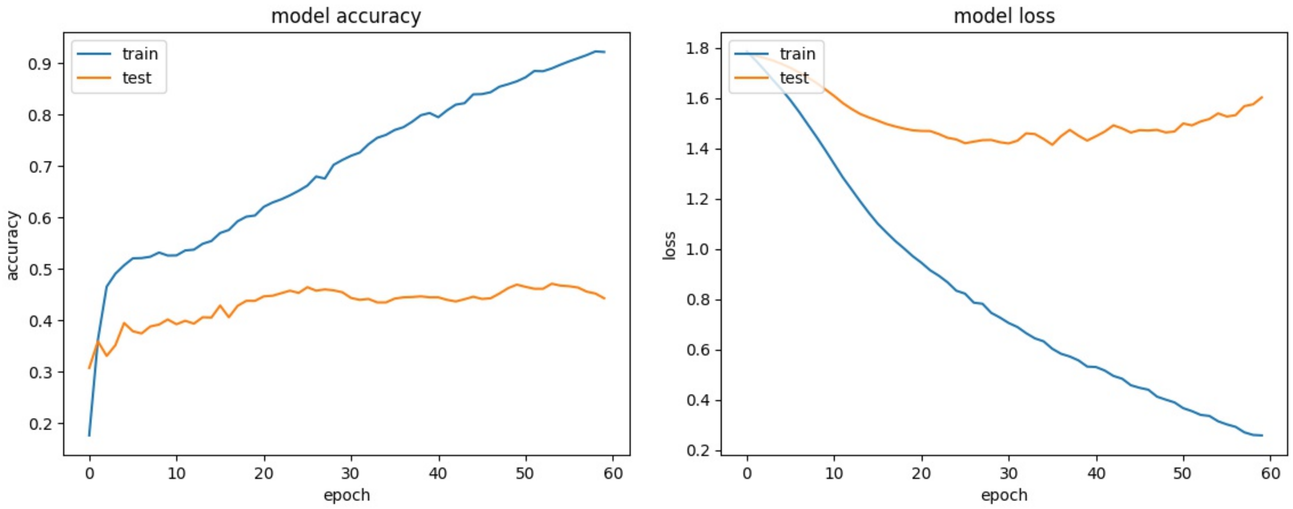


Figure 58: Accuracy and Loss performance during training with text based on transformers and video in DialogueRNN.

C.4.10 Joint optimization of text based on transformer encoder and audio DialogueRNN training

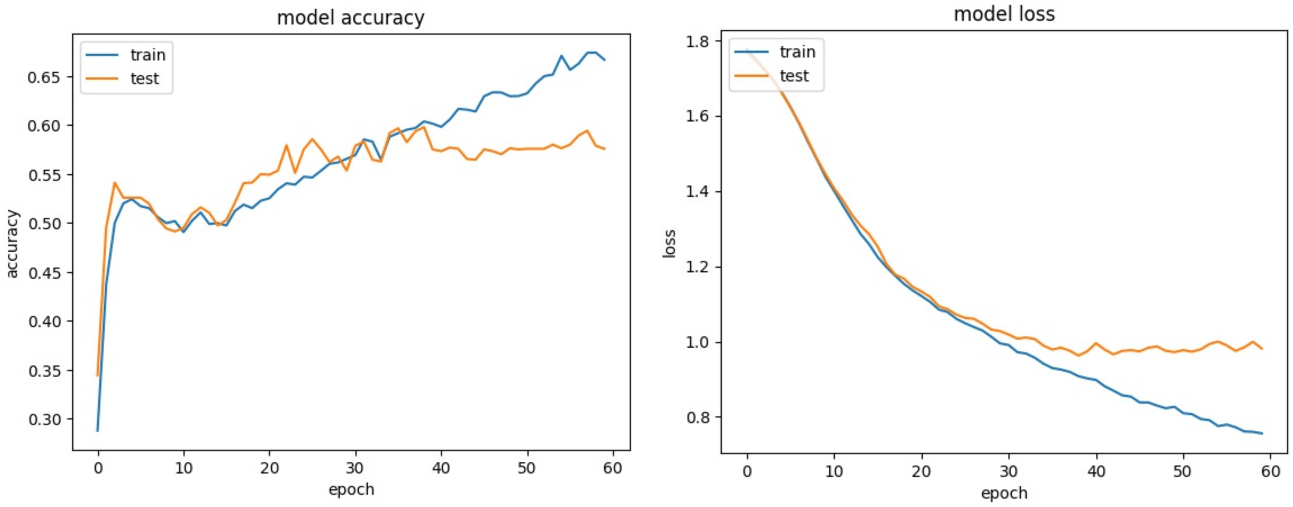


Figure 59: Accuracy and Loss performance during training with joint optimization of text based on transformers and audio in DialogueRNN.