Università degli Studi di Napoli Parthenope

Dipartimento di Scienze e Tecnologie

Corso di laurea in informatica

Tesi di laurea

Adaptive Multimodal Emotion Detection Architecture for Social Robots

Relatori

Prof.ssa Mariacarla Staffa

Candidato

Giuseppe Fiorillo

Matr. 0124002248

Anno Accademico 2022/2023

Contents

List of Figures

List of Tables

A chi ha rinunciato, a chi ogni giorno fatica ad alzarsi, a chi non ha avuto le giuste opportunità dalla vita.

Abstract

Human-robot interaction (HRI) is the study of interactions between humans and robots. In order to improve human-robot interaction, social robots employ emotion recognition. Since human emotions can be represented in a variety of ways (e.g., voice, gesture, and face), several techniques for their recognition has been proposed during the years. However, relying on a single feature could lead to inconsistencies that could influence the expected results. For this reason, multimodal techniques must be integrate for accurate detection.

One important problem about about the multimodal approach is that one modality may be absent or noisy, so it should be evaluated the resilience of the model in unconstrained conditions, in contrast to earlier researches that assume the ideal case of every single modality being consistently present during inference.

In response to this challenge, we have analyzed an emotion recognition architecture based on transformers that demonstrates adaptability and flexibility, using a modality dropout strategy. This architecture seamlessly accommodates multiple sources and modalities of information managing varying levels of data quality and addressing missing data.

To assess the effectiveness of the proposed architecture, various tests to classify emotions are conducted. Each test has been conducted for three main different fusion strategies. The results demonstrate the adaptability

of the approach to varying modalities' quality and presence.

Chapter 1

Introduction

Understanding human emotions is important for machine learning and helps us to comprehend social signals in a variety of applications, such as robotics and human-computer interaction [1], [2]. A wide range of methods and models have been put forth for the recognition of emotions. These include the detection of discrete emotional states (such as "happy", "angry" or "sad") as well as the continuous scale assessment of emotional properties like arousal and valence [3], [4]. Various data formats have been used as input for this task, such as audio [5], images [6], and text [7].

A plethora of techniques are emerging to make the most of this amount of data, which is driving the development of multi-modal approaches [8], [9], [10]. These techniques work in parallel with several kinds of data, such as RGB and skeletal data [11], joint RGB and depth pictures [12], and video data with audio and visual modalities [13]. Social robots use their perception system to gather data necessary for identifying human emotions. A robot's primary means of sensing are its auditory and visual senses; however, alternative methods can be employed. Processing methods for these multi-modal representations vary from straightforward decision-level fusion to sophisticated joint feature learning techniques. While fusion of intermediate features can lead to better performance by jointly learning

representations from various modalities, the simplicity and adaptability of late or early fusion makes them popular in contemporary systems.[8] On the other hand, late fusion mainly takes into account features learnt individually in each modality, whereas early fusion could not be appropriate for significantly distinct data kinds.

Many multi-modal approaches now in use evaluate model performance only in the case when all accepted modalities are present continuously throughout inference. However, in practical implementations, it is likely that one modality will not exist or will exist in poor quality at some points. Therefore, when building multi-modal systems that use real-world data, the model's resilience to such circumstances becomes essential.

There has been a general trend in the field of multi-modal emotion identification to heavily rely on pre-extracted features, especially in the latest transformer-based designs. Instead of creating end-to-end trainable models, these features are then combined with a learned model [14], [15], [16]. Nevertheless, this methodology has constraints on the usefulness of these techniques in actual situations. Prerequisite feature extraction can be difficult, particularly in unrestricted environments, adding another degree of ambiguity to the processing chain as a whole. This problem is most noticeable in approaches that use language information [14], [15], as it is rarely possible to obtain text transcriptions of audio signals in real-world applications and instead requires independent estimation. Consequently, we do not require discrete feature learning and instead concentrate on audiovisual emotion recognition.

The proposed model get past the drawbacks of the multi-modal emotion identification techniques that are currently in use. This is accomplished by developing an end-to-end model that chooses fusion at the intermediate level over previous feature learning. In the presence of noisy or missing

data samples, the model is intended to demonstrate robustness. The key contributions can be outlined as follows:

- Presenting a revolutionary architecture that learns directly from raw videos and eliminates the need for independently learnt features for audiovisual emotion recognition from speech and face films.
- Combining a variety of modality fusion techniques with the introduction of a modality-independent feature-sensitive attention-based intermediate feature fusion method. As far as we know, no one has ever suggested a method like this before.
- In order to improve the model's resilience to missing or noisy data from a single modality, a novel training approach based on modality dropout mechanisms is presented. It is also found that even in the standard case, when both modalities are available, the performance of the recommended strategy is improved.

Chapter 2

Related works

Multimodal emotion recognition typically requires separate assessments of each modality. The outcomes are then combined using a fusion technique. Below are the most popular solutions for identifying emotions from faces and audio, as these are the modalities used in this work. Fusion methods will also be discussed in this section.

2.1 Facial emotion detection

The task at hand involves identifying and categorizing facial image classifications into various moods by utilizing identified facial traits [17]. The movement of the face muscles that create facial expressions is typically responsible for those facial features [17]. Convolutional Neural Networks (CNN), whose inputs are photos or videos, are the foundation of current facial emotion recognition techniques [18]. Certain CNN techniques use tried-and-true image feature extraction techniques to process the facial data [19], which could be adjusted to be more appropriate and produce better outcomes. Additionally, attentional CNN-based deep learning techniques are finding use in the detection of emotions from faces by focusing solely on the most pertinent facial elements within the image [20].

2.2 Audio emotion detection

The voice is a fundamental tool that is frequently employed and crucial for human communication since it may convey "more than simple words." Therefore, because so much rich information can be gleaned, using speech to investigate emotions is important. The resources must be extracted and categorized before voice recognition of emotions can be performed. As a result, resource extraction techniques are frequently used in audio files. One example is the Fourier parameters [21], which are derived from the Discrete Fourier Transform. It is necessary to classify resources using both linear and non-linear classifiers after this data has been extracted.

The Support Vector Machine (SVM) is the most widely used linear classifier; however, nonlinear classifier models, such as the Gaussian Mixture Model (GMM) and Hidden Markov Model (HMM), ensure more effective results for low-level operations. Speech Emotion Recognition (SER) models make greater use of Mel Frequency Cepstral Coefficient (MFCC) and Linear Predictor Coefficients (LPC), and with deep learning's computational capability, these models are better able to automatically detect complex structures and resources. There are various ways to implement SER models, such as using CNN in conjunction with Random Forest to extract speech emotion features from the normalized spectogram [22] or utilizing deep neural networks and extreme machine learning [23], [24] to distinguish between different emotional states. Vocal frequency contrast can also be detected by neural networks trained for emotion recognition.

2.3 Electroencephalography

Brain activity can be measured with the electroencephalography (EEG), a reliable and affordable method. Several sequential procedures must be taken in order to meet the requirements of a brain–computer interface (BCI) in order to detect emotion using EEG signals. These procedures typically involve the following: first, eliminating artifacts from EEG data; second, extracting temporal or spectral information from the time or frequency domain of the EEG signal; and third, creating a multi-class classification approach. The emotion classification strategy's accuracy is significantly improved by feature quality [25].

EEG data provide a unique window into the brain correlates of emotions with high temporal resolution, mobility, and cost-effectiveness, making them a useful and adaptable tool for emotion recognition tasks. As noted by the results from the study in [26] by Staffa et. al, EEG is a valid and interesting method in HRI scenarios.

2.4 Fusion methods

Numerous techniques have been developed to combine different types of data [14], [15], [16] by using different fusion processes. In [27], Chumachenko et al. identify three main approaches of multimodal fusion:

- **early fusion**: the input data of multiple modalities are combined via concatenation;
- **late fusion**: the data of different modalities, or their softmax classification scores, are treated and processed independently, being only combined in the very last layers;

- **intermediate feature fusion**: feature representations of several modalities are jointly learned by performing feature sharing at the network's middle layers.

A significant group of multimodal fusion techniques relies on the use of self-attention. Self-attention is a method of focusing attention that links several points inside a single sequence to create a representation of the sequence [28]. Remember that learnable projection matrices W_q, W_k, and W_v are used to extract input feature representation from queries q, keys k, and values v. An attented representation is then computed using these matrices to formulate self-attention:

$$A_n = softmax\left(\frac{qk^T}{\sqrt{d}}\right)v, \tag{2.1}$$

where d is the dimensionality of a latent space. Vaswani et al. suspected that for large values of d_k, the dot products grow large in magnitude, pushing the softmax function into regions where it has extremely small gradients 4. To counteract this effect, the dot product is scaled by $\frac{1}{\sqrt{d}}$ [28].

A pertinent study is [16], which examines both audio and visual modalities while tackling the emotion recognition task. There, representations learned from each modality are fused with a transformer in a fashion akin to equation 2.1 after each modality is first preprocessed with a distinct transformer block.

2.5 Convolutional neural network

One kind of feed-forward neural network is the convolutional neural network (CNN). The fundamental distinction between CNN and other neural networks is in the implicit assumption made by the CNN architecture that the inputs are image-like, allowing us to encode certain attributes within

the architecture. Convolutions, in particular, encapsulate translation invariance, which states that filters are location-independent. This, in turn, greatly reduces the number of parameters in the forward function, improving its efficiency and making the network less dependent on data size and easier to improve [29].

Unlike conventional neural networks, CNNs feature layers with neurons arranged according to a few dimensions: in the most basic 2D scenario, channels, width, height, and number of filters. Similar to an MLP, a convolution neural network is made up of a series of layers, each of which applies a differentiable function to change the activations or outputs of the layer before it. CNNs use a number of these layers, including the convolution, pooling, and fully connected layers. These layers are comparable to classification layers, dimensionality reduction layers, and feature extractors, in that order. A CNN's entire convolutional layer is formed by stacking these layers.

Specifically, a convolution layer filters the input using a convolutional kernel. These filters are typically several. A filter computes its activation map at each position during a forward pass by slicing over the input volume, multiplying each value pointwise, and summing the results to determine the activation at that point. A convolution is the natural way to build such a sliding filter, and as it is a linear operator, it may be efficiently implemented by writing it as a dot-product.

This implies, intuitively, that during training, the CNN will pick up filters that can identify visual features like edges, orientations, and eventually, whole patterns in a higher layer of the network. There is a whole set of such filters in each such convolution layer, and each of them will yield a unique activation map. The output map, also known as the activation volume of this layer, is created by stacking these activation maps.

Chapter 3

Methods

Three self-attention based modality fusion strategies and the general architecture of the suggested audiovisual emotion identification model are described here. Additionally, a method known as modality dropout is put forth to account for missing data in a particular modality during inference. Generally speaking, the model is comprised of two branches that learn visual and audio features, respectively, and fusion modules that are positioned in the center or at the end of the two branches, depending on the type of feature fusion, as illustrated in 3.1. The 1D Convolutional blocks that are applied in a temporal dimension are mostly utilized in the audio and visual branches.

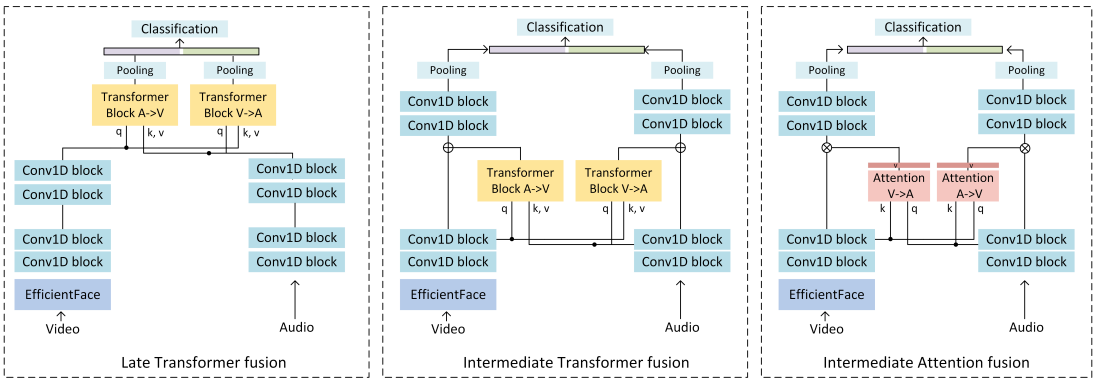

Figure 3.1: Multimodal data fusion appraoches

3.1 Feature extraction

1. *Vision branch*: The vision branch is divided into two sections. The first part involves extracting visual features from individual video frames, while the second part involves learning joint representation for the entire video sequence. In contrast to the vast majority of previous works that separate feature extraction from multimodal fusion and mostly use pre-extracted features, such as facial landmark locations, facial action units, or head pose information [14], [16], we use feature extraction as part of our pipeline and optimize it in tandem with the multimodal fusion module to achieve an end-to-end trainable model capable of learning from raw video. We select EfficientFace [6], one of the recently proposed architectures for facial expression detection, and use it to extract features from individual frames before feeding them into further 1D convolutional blocks. In particular, the average-pooled output of the final convolutional block of EfficientFace is included before the 1D convolutional blocks.

A 2D feature extractor takes an input video sequence consisting of k frames, processes each frame individually, and outputs a single vector descriptor for each frame. These representations undergo additional processing in a temporal dimension, including temporal convolution blocks that will be elaborated upon. We have opted for this method instead of using 3D-convolutions directly, as is typically done in video tasks, since it offers several benefits in the work at hand. The first is the reduced processing cost that 2D convolutional layers bring in comparison to 3D convolutions. One could counter that while temporal relations are less significant in the goal of identifying emotions, 1D convolutional procedures applied in the temporal dimension are adequate to capture this data. The possibility to use 2D feature extractors that have already been pre-trained on bigger

Architecture of the visual branch

	EfficientFace module
Stage1	Reshape
	Conv1D [k=3, d=128, s=1] + BN1D + ReLU
	Conv1D [k=3, d=128, s=1] + BN1D + ReLU
Stage2	Conv1D [k=3, d=256, s=1] + BN1D + ReLU
	Conv1D [k=3, d=256, s=1] + BN1D + ReLU
Predict	Global Average Pooling + Linear

Architecture of the audio branch

Stage1	Conv1D [k=3, d=256] + BN1D + ReLU + MaxPool1d [2x1]
	Conv1D [k=3, d=256] + BN1D + ReLU + MaxPool1d [2x1]
Stage2	Conv1D [k=3, d=512] + BN1D + ReLU + MaxPool1d [k=2]
	Conv1D [k=3, d=128] + BN1D + ReLU + MPool1D [k=2]
Predict	Global Average Pooling + Linear

Table 3.1: Architecture of the visual and audio modules

image-based emotion detection datasets is another important advantage of adopting the suggested approach. This is because pre-trained 2D convolutional models require substantially fewer films, which are needed for pre-training.

The first component, i.e., visual feature extraction, can be separated from the model and any other features can be utilized as input to the second half of the vision branch, even if our main goal is to create an end-to-end pipeline that can learn from raw data. In other words, we assume in the second half of the architecture that a specific feature representation $X_v^{N\times d}$, where N is the temporal dimension and d is the feature dimension, has been retrieved from the input visual data. In this case, X_v can be represented by any kind of feature, such as landmarks or facial action units, which are frequently used for emotion recognition, or by deep features that are

extracted using a trained model. In addition, we employ a sequence of four convolutional blocks to acquire temporal representation knowledge. A 1D convolutional layer with a 3×3 kernel, Batch Normalization, and a ReLU activation make up each convolutional block. Additional information on the vision branch is shown in 3.1, where s stands for stride, d for number of filters in a convolutional layer, and k for kernel size. For multimodal fusion, which will be discussed in more detail, the convolutional blocks are divided into two stages.

2. *Audio branch*: Four blocks of 1D convolutional layers are applied to a feature representation, which can be pre-computed or jointly optimized, in a manner similar to the vision branch. 3.1 defines the specifications for each block, which includes a Convolutional layer, Batch Normalization, ReLU activation, and MaxPooling. Mel-frequency cepstral coefficients are the main features we employ for audio. Other feature representation formats, including spectrograms or chroma features, did not seem to be beneficial in our experience.

3.2 Modality fusion approaches

We discuss the taken into consideration fusion approaches in this part. The two suggested intermediate fusion procedures will be discussed after the late transformer fusion approach, which is comparable to earlier studies that have been documented in the literature, is first detailed.

1. *Late transformer fusion*: In this configuration, a transformer block is utilized to fuse features from two branches. Specifically, we use two transformers to perform the fusion of one modality into the other at the outputs of each branch. These transformer blocks' outputs are concatenated once more before being sent to the last prediction layer. In formal terms, this

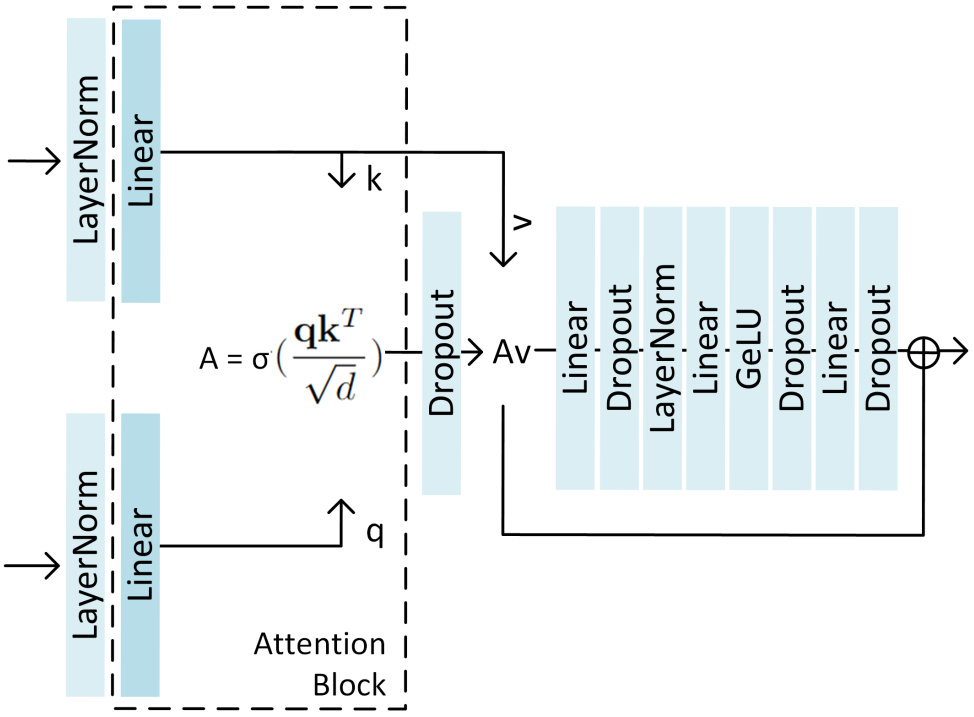

Figure 3.2: Structure of the Transformer block

has the following definition.

After the second feature extraction step, or after the fourth convolutional block, let Φ_a and Φ_v be the feature representations of the audio and vision modalities. In every branch, a transformer block is created that accepts two modalities' representations as inputs. Using the audio branch as an example, the transformer block computes queries from the audio branch features Φ_a, while the vision branch representation Φ_v is fed into it to acquire keys and values. In other words, self-attention is determined as:

$$A = softmax\left(\frac{\Phi_a W_q W_k^T \Phi_v^T}{\sqrt{d}}\right)\Phi_v W_v, \tag{3.1}$$

followed by standard transformer block processing [28]. 3.2 describes the exact architecture of the transformer block.

Concatenated outputs from the two transformer blocks are sent to the last layer for prediction.

2. *Intermediate transformer fusion*: We suggest using transformer blocks

that are comparable to those mentioned before for fusing at intermediate feature levels. In particular, each branch undergoes fusion using a transformer block following the first feature extraction stage, or following two convolutional layers. The fused feature representation is added to the corresponding branch, and the architecture is similar to that shown in 3.1.

Later convolutional layers can learn the features that are jointly important for the task at hand between modalities since data from complimentary modalities is introduced earlier in the design.

3. *Intermediate attention-based fusion*: We also suggest a fusion method that solely relies on the dot-product similarity, which is what the transformer block considers to be attention. This is defined formally as follows. Similar to conventional attention, we construct queries and keys with learned weights given the two feature representations of distinct modalities, Φ_a and Φ_b. Next, the scaled dot-product similarity is computed as

$$A = softmax\left(\frac{\Phi_a W_q W_k^T \Phi_v^T}{\sqrt{d}}\right). \tag{3.2}$$

In order to highlight more significant characteristics or timestamps for each modality and to provide the relevance score of each key in relation to each query—i.e., each representation of modality a in relation to modality b—softmax activation encourages competition in the attention matrix. This makes it possible to aggregate the scores corresponding to all of the modality b attributes for each attribute of modality a, thereby calculating the relative importance of each attribute of modality a. Consequently, we derive an attention vector that may be employed to draw attention to more pertinent properties of the modality a. The attention vector of the vision modality is $v_v = \sum_{i=N_v} A[:, i]$, taking into account the dot-product attention between features of the audio and vision modalities as described

in 3.2.

3.3 Modality dropout

Most multimodal learning techniques that have been documented thus far presume the simultaneous existence of both modalities during the inference process. However, in real-world applications, data of one or more modalities may occasionally be absent or unreliable. Conventional multi-modal tactics usually don't work in these kinds of situations. Here, we seek to address the possibility of missing data and suggest modality dropout as a means of reducing it. As will be demonstrated in more detail, applying this strategy improves performance even in scenarios involving both modalities.

Our suggestion is the modality dropout, which attenuates or arbitrarily obscures one modality's input while it is being trained. We specifically take into account three variations. In the first form, the representation of one modality for a particular sample is maintained while randomly picked data from each sample is replaced with zeros during training. This method mimics missing data and functions as a regularizer in a manner akin to neural networks' Dropout layer. Keep in mind that the third fusion strategy and the lack of bias terms cause the attention block's zero dot similarity matrix, which, after softmax and summation, results in a constant attention vector and, as a result, no information transfer from the zeroed modality.

In the second version, we create a random scaling factor α in the interval $[0, 1]$ [30] for every pair of data samples. We then multiply one of the modalities by α and the other by $1 - \alpha$. This method aims to stop the model from learning from only one modality by attenuating inputs from other modalities at distinct training steps. We also call this strategy "soft" modality dropout. The third variation aims to address the issue of noisy

data, which occurs when one modality's input signal is corrupted. Similar to the previous variant, masking is carried out here, but instead of zero-masking, data is generated at random from a normal distribution with zero mean and unit variance in one of the modalities for every sample.

Chapter 4

Experiments

This section explains the data and experimental technique used to evaluate the suggested approaches' performance. Three fusion variations of the proposed model and two recent multimodal emotion detection techniques, MULT [14] and multimodal transformer [16], are described together with the model's results. Furthermore, [14] examines three modalities, and [16] lacks information on particular architecture hyperparameters, making direct comparison impossible. As a result, we use our feature extraction and solely make comparisons with other publications based on the fusion techniques they outline. In particular, we use a transformer block on top of our two convolutional branches to perform fusion from audio to video or the other way around in order to compare with [16]. We also replace the linear layers in the transformer block with 1D-convolutional ones. In relation to MULT [14], we exclude the transformer blocks in charge of fusion from/to the language modality since we wish to compare with a purely audiovisual model. This results in an architecture that is comparable to our late transformer fusion, except that each branch has one extra single modality transformer block. For a fair comparison, other hyperparameters, including latent space dimensions, are maintained constant throughout the approaches. In the same way as the comparison with [14], we incorporate our

blocks for feature extraction into the model. To create a lightweight model, we always employ a single transformer block with a single head, unless otherwise noted. Naturally, expanding the models' blocks and parameters will improve their performance. For the implementation of transformer blocks, [31] has been used.

We then conduct modality dropout studies in two different scenarios. In the first, we use both soft and hard modality dropout during training to address the issue of missing data from one modality. In this configuration, pairings of complete data, pairs without audio modality, pairs without video modality, and pairs multiplied by random coefficients as previously mentioned make up each batch of data. We present the performance when both modalities are present (shown as 'AV') and when only one modality is present (shown as 'A' or 'V' for the presence of only audio and visual modalities, respectively).

4.1 RAVDESS dataset

The Ryerson Audio-Visual Database of Emotional Speech and Song [32] has been chosen due to availability of raw data as opposed to others. It consists in video recordings of 24 professional actors expressing a range of emotions. The task involves classifying the emotional states into seven classes: calm, happy, sad, angry, fearful, surprise, and disgust (other than neutral). For each actor, there are 60 trials and we crop or zero-pad them to 3.6 seconds, that is the average sequence length.

We extract 10 Mel-frequency cepstral coefficients for further processing in audio processing. In order to obtain visual data, we take 15 uniformly distributed frames out of a 3.6-second video, then use a face detection algorithm to crop the actors' faces [33]. Images are cropped to 224x224 pixels.

We use 15-frame raw videos to train the model. We move the Efficient-Face pre-trained weights from the AffectNet dataset [4], [27]. In order to prevent actor identities from being replicated across sets, we divided the data into training, validation, and test sets. In particular, we employed 16 actors for training, 4 for testing, and 4 for validation. The average result was reported across five folds. Random rotation and random horizontal flip are applied for data augmentation, and the videos are scaled into the $[0, 1]$ scale. All the models are trained for 100 epochs, with SGD, a learning rate of 0.04, momentum of 0.9, weight decay of 1e-3, and a learning rate reduction on a plateau of 10 epochs.

4.2 Results and discussion

The outcomes of the suggested methods on the RAVDESS datasets are displayed in 4.1. Here, the terms 'LT1' and 'LT4' stand for late transformer fusion with one and four heads, respectively, while 'IT' and 'IA' stand for intermediate transformer and attention fusion, respectively. The notations 'TAV' and 'TVA' relate to the fusion approaches discussed in [16] and 'MULT' to [14]. Here is reported the categorical accuracy on the RAVDESS dataset.

As can be shown, late transformer fusion performs best on the RAVDESS dataset when there is no dropout of any kind. It should be noted that, when compared to all other fusion approaches utilizing complete transformer blocks, intermediate attention fusion is the lightest. Nevertheless, there is a marked lack of performance when there is just one modality available. Furthermore, it is evident that using modality dropout significantly enhances performance when one modality's data is inadequate. All fusion techniques profit from this, but intermediate attention fusion gains the

	AV	A	V
LT1	**79.33**	19.83	36.41
LT4	76.42	27.92	30.00
IT1	76.41	21.16	18.33
IT4	78.50	20.33	17.33
IA1	76.00	18.58	22.83
IA4	77.41	20.66	29.83
TAV	77.75	24.25	13.33
TVA	76.00	15.16	42.67
MLT	74.16	22.33	35.42

MODALITY DROPOUT

	AV	A	V
LT1	79.08	59.16	72.66
LT4	79.25	53.00	70.92
IT1	77.33	48.41	73.75
IT4	78.91	44.33	74.92
IA1	**81.58**	58.08	72.83
IA4	79.58	57.16	71.83
TAV	76.58	54.83	13.33
TVA	74.42	44.91	69.58
MLT	78.50	53.58	70.66

MODALITY DROPOUT with NOISE

	AV	A	V
LT1	77.08	53.16	68.50
LT4	**80.33**	54.33	73.00
IT1	76.75	53.75	71.58
IT4	76.08	54.50	71.00
IA1	78.25	58.25	71.66
IA4	78.41	55.75	68.58
TAV	75.83	56.25	12.83
TVA	73.66	41.25	71.41
MLT	77.41	54.16	66.33

Table 4.1: Performance of different fusion methods on RAVDESS

most from it. Additionally, performance is enhanced when both modalities are present, with intermediate attention fusion producing the best results. Out of all the algorithms and dropout levels used on this dataset, this is likewise the best outcome. We still find that the intermediate attention fusion on RAVDESS performs better on the average measure under the noisy condition.

Chapter 5

Conclusions and future works

We suggested an attention-based fusion method and an end-to-end learning model for audiovisual emotion identification. We assessed how robust several approaches to modality fusion were when one of the modalities had noise or wasn't there at all, and we suggested a way to make the model more resilient. Significantly, when both modalities are available, the suggested strategy also enhances performance in the optimal standard environment.

In the future, I plan to implement new features of EEG signal that can improve the performance of the models. Since there aren't any publicly accessible datasets of this kind, I also wish to construct an Audio+Face+EEG dataset.

Acknowledgement

Sono estremamente grato a mia madre e mio padre, le persone più forti che io abbia mai conosciuto, per avermi sempre appoggiato e avermi dato sempre la possibilità di prendere le scelte che più credevo giuste, sbagliate o corrette che fossero. Siete riusciti a trasmettermi molti dei valori e dei principi che rappresentano la mia persona, siete sempre stati il mio punto di riferimento e per questo vi amerò più di ogni altra cosa. Non potrò mai ringraziarvi abbastanza per avermi sostenuto durante questo percorso.

Ringrazio mia sorella Meri per essere stata al mio fianco durante questi anni. Nonostante le infinite discussioni e i molteplici litigi, ci sei sempre stata per me e so che ci saremo sempre l'uno per l'altra. Con te ho condiviso alcuni dei ricordi più belli e sono sicuro che ci sarà occasione per passarne insieme altri mille così.

A Sharpay, la mia cagnolina, che è capace di ravvivare ogni mia giornata e di farmi stare bene con la sua tenerezza anche nei momenti più bui. Sei stata la mia salvezza e l'affetto che ogni giorno mi dai è fondamentale per farmi andare avanti.

Ringrazio con tutto il cuore mia nonna e mio nonno, per tutto l'amore che mi hanno dato e per avermi cresciuto come dei secondi genitori. Anche se non ci siete più, il vostro ricordo è limpido nella mia mente e continuerò a fare tesoro di quanto siete riusciti a trasmettermi.

A Matteo, Daniele e Pierluigi: siete tra gli amici più cari che ho e

sono felice di sapere che su di voi potrò contare sempre e comunque. In voi ho ritrovato una seconda famiglia, sempre pronta a sostenermi e nella quale ognuno potrà sempre fare affidamento sull'altro. La nostra amicizia è qualcosa di davvero speciale e sono sicuro andrà avanti per sempre.

A Emanuele, mio fratello acquisito sin dall'asilo, per essere sempre stata una presenza costante e certa nella mia vita. Sono felice di aver con te condiviso alcuni dei momenti più belli della mia vita, perché persona migliore con cui passarli non potevo trovarla.

A Raffaele per essere stato il mio principale compagno di studi e di progetti durante questi tre anni. Sei stata indubbiamente la presenza più costante nel mio percorso universitario e se ho passato alcuni esami è stato grazie al tempo passato insieme a studiare.

Ringrazio Ylenia, Maria, Giancarlo, Samuele e Gaia per essermi stati vicini nei momenti in cui più ne avevo bisogno. L'università è una bellissima esperienza, non solo per lo studio e per le conoscenze apprese, ma anche per l'ambiente amichevole e molto caloroso che voi avete contribuito a creare. Sono stato molto fortunato nel conoscervi e non nego che senza la vostra presenza il mio percorso universitario sarebbe stato assai più grigio.

Ad Antonio Onorato, Antonio Criscuolo, Simone, Flavio e Attilio. Se sono arrivato fin qui è anche grazie a voi. Le ore passate a studiare, così come le assai più numerose ore passate a perdere tempo e a lamentarsi, sono state una parte fondamentale per questo percorso. Grazie per avermi fatto compagnia e avermi aiutato.

Ringrazio la professoressa Mariacarla Staffa per il supporto e per avermi seguito durante la stesura della tesi; in lei ho trovato una professoressa disponibile e attenta alle problematiche degli studenti. Spero vivamente di poter continuare a lavorare con lei nel futuro del mio percorso accademico.

Bibliography

[1] Z. Liu, M. Wu, W. Cao, L. Chen, J. Xu, R. Zhang, M. Zhou, and J. Mao. A facial expression emotion recognition based human-robot interaction system. *IEEE/CAA Journal of Automatica Sinica*, 4(4):668–676, 2017. 8

[2] J. Chen, Y. Lv, R. Xu, and C. Xu. Automatic social signal analysis: Facial expression recognition using difference convolution neural network. *Journal of Parallel and Distributed Computing*, 131:97–102, 2019. 8

[3] A. Dzedzickis, A. Kaklauskas, and V. Bucinskas. Human emotion recognition: Review of sensors and methods. *Sensors*, 20(3):592, 2020. 8

[4] A. Mollahosseini, B. Hasani, and M. H. Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017. 8, 26

[5] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain. Speech emotion recognition using deep learning techniques: A review. *IEEE Access*, 7:117327–117345, 2019. 8

[6] Z. Zhao, Q. Liu, and F. Zhou. Robust lightweight facial expression recognition network with label distribution training. In *Proceedings*

of the AAAI Conference on Artificial Intelligence, pages 3510–3519, 2021. 8, 17

[7] M. Soleymani, D. Garcia, B. Jou, B. Schuller, S.-F. Chang, and M. Pantic. A survey of multimodal sentiment analysis. *Image and Vision Computing*, 65:3–14, 2017. 8

[8] H. R. V. Joze, A. Shaban, M. L. Iuzzolino, and K. Koishida. Mmtm: Multimodal transfer module for cnn fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13289–13299, 2020. 8, 9

[9] K. Chumachenko, J. Raitoharju, A. Iosifidis, and M. Gabbouj. Speed-up and multi-view extensions to subclass discriminant analysis. *Pattern Recognition*, 111:107660, 2021. 8

[10] Y. Wang, W. Huang, F. Sun, T. Xu, Y. Rong, and J. Huang. Deep multimodal fusion by channel exchanging. In *Advances in Neural Information Processing Systems*, volume 33, 2020. 8

[11] F. Laakom, K. Chumachenko, J. Raitoharju, A. Iosifidis, and M. Gabbouj. Learning to ignore: rethinking attention in cnns. *arXiv preprint arXiv:2111.05684*, 2021. 8

[12] Y. Xiao, F. Codevilla, A. Gurram, O. Urfalioglu, and A. M. López. Multimodal end-to-end autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 2020. 8

[13] F. Xiao, Y. J. Lee, K. Grauman, J. Malik, and C. Feichtenhofer. Audiovisual slowfast networks for video recognition. *arXiv preprint arXiv:2001.08740*, 2020. 8

[14] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, page 6558, 2019. 9, 13, 17, 24, 26

[15] D. Krishna and A. Patil. Multimodal emotion recognition using cross-modal attention and 1d convolutional neural networks. In *Interspeech*, pages 4243–4247, 2020. 9, 13

[16] J. Huang, J. Tao, B. Liu, Z. Lian, and M. Niu. Multimodal transformer fusion for continuous emotion recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3507–3511, 2020. 9, 13, 14, 17, 24, 26

[17] B. Ko. A brief review of facial emotion recognition based on visual information. *Sensors*, 18(2):401, Jan 2018. 11

[18] N. Mehendale. Facial emotion recognition using convolutional neural networks (ferc). *Social Network Applications and Sciences*, 2(3):1–8, Mar 2020. 11

[19] M. Omkar Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 41.1–41.12, 2015. 11

[20] S. Minaee, M. Minaei, and A. Abdolrashidi. Deep-emotion: Facial expression recognition using attentional convolutional network. *Sensors*, 21(9):3046, Apr 2021. 11

[21] K. Wang, N. An, B. N. Li, Y. Zhang, and L. Li. Speech emotion recognition using fourier parameters. *IEEE Transactions on Affective Computing*, 6(1):69–75, Jan 2015. 12

[22] L. Zheng, Q. Li, H. Ban, and S. Liu. Speech emotion recognition based on convolution neural network combined with random forest. In *Proceedings of the Chinese Control and Decision Conference (CCDC)*, pages 4143–4147, Jun 2018. 12

[23] K. Han, D. Yu, and I. Tashev. Speech emotion recognition using deep neural network and extreme learning machine. In *Proceedings of Interspeech*, pages 223–227, Sep 2014. 12

[24] S. Tripathi, S. Tripathi, and H. Beigi. Multi-modal emotion recognition on iemocap dataset using deep learning, 2018. Empty Journal. `arXiv:1804.05788`. 12

[25] S. Gannouni, A. Aledaily, K. Belwafi, and et al. Emotion detection using electroencephalography signals and a zero-time windowing-based epoch estimation and relevant electrode identification. *Scientific Reports*, 11(1):7071, 2021. `doi:10.1038/s41598-021-86345-5`. 13

[26] M Staffa, L D'Errico, S Sansalone, and M Alimardani. Classifying human emotions in hri: applying global optimization model to eeg brain signals. *Frontiers in Neurorobotics*, 17:1191127, Oct 2023. `doi:10.3389/fnbot.2023.1191127`. 13

[27] Kateryna Chumachenko, Alexandros Iosifidis, and Moncef Gabbouj. Self-attention fusion for audiovisual emotion recognition with incomplete data. *arXiv preprint arXiv:2201.11095*, 2022. 13, 26

[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 14, 20

[29] Jonas Teuwen and Nikita Moriakov. Chapter 20 - convolutional neural networks. In S. Kevin Zhou, Daniel Rueckert, and Gabor Fichtinger, editors, *Handbook of Medical Image Computing and Computer Assisted Intervention*, The Elsevier and MICCAI Society Book Series, pages 481–501. Academic Press, 2020. URL: `https://www.sciencedirect.com/science/article/pii/B9780128161760000259`, `doi:10.1016/B978-0-12-816176-0.00025-9`. 15

[30] X. Shen, X. Tian, T. Liu, F. Xu, and D. Tao. Continuous dropout. *IEEE transactions on neural networks and learning systems*, 29(9):3926–3937, 2017. 22

[31] R. Wightman. Pytorch image models, 2019. URL: `https://github.com/rwightman/pytorch-image-models`. 25

[32] S. R. Livingstone and F. A. Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391, 2018. 25

[33] R. Gradilla. Multi-task cascaded convolutional networks (mtcnn) for face detection and facial landmark alignment. *Medium*, 2020. URL: `https://arxiv.org/abs/1404.7486`. 25