Learning from Audio, Vision and Language Modalities for Affect Recognition Tasks

by

Vandana Rajan

Bachelor of Technology in Electronics and Communications Engineering 2013

Master of Technology in Digital Signal Processing 2016

A dissertation submitted to

The School of Electronic Engineering and Computer Science

in partial fulfilment of the requirements of the Degree of

Doctor of Philosophy

in the subject of

Electronic Engineering

Queen Mary University of London
Mile End Road
E1 4NS, London, UK

August, 2022

Declaration

I, Vandana Rajan, confirm that the research included in this thesis is my own work, that is duly acknowledged, and my contributions are indicated. I have also acknowledged previously published materials.

I attest that reasonable care has been exercised to ensure the originality of this work, and, to the best of my knowledge, does not break any UK law, infringe any third party's copyright or other Intellectual Property Right, or contain any confidential material.

I accept that the college has the right to use plagiarism detection software to check the electronic version of the thesis.

I confirm that this thesis has not been previously submitted for the award of a degree to any other university.

The copyright of this thesis rests with the author and no quotation from it or information derived from it may be published without the prior written consent of the author.

Signature: Vandana Rajan

Date: August 01, 2022

Primary supervisor Secondary supervisor Author

Prof. Andrea Cavallaro **Dr. Alessio Brutti** **Vandana Rajan**

Learning from Audio, Vision and Language Modalities for Affect Recognition Tasks

Abstract

The world around us as well as our responses to worldly events are multimodal in nature. For intelligent machines to integrate seamlessly into our world, it is imperative that they can process and derive useful information from multimodal signals. Such capabilities can be provided to machines by employing multimodal learning algorithms that consider both the individual characteristics of unimodal signals as well as the complementariness provided by multimodal signals. Based on the number of modalities available during the training and testing phases, learning algorithms can be of three categories: unimodal trained and unimodal tested, multimodal trained and multimodal tested, and multimodal trained and unimodal tested algorithms. This thesis provides three contributions, one for each category and focuses on three modalities that are important for human-human and human-machine communication, namely, audio (paralinguistic speech), vision (facial expressions) and language (linguistic speech) signals. For several applications, either due to hardware limitations or deployment specifications, unimodal trained and tested systems suffice. Our first contribution, for the unimodal trained and unimodal tested category, is an end-to-end deep neural network that uses raw speech signals as input for a computational paralinguistic task, namely, verbal conflict intensity estimation. Our model, which uses a convolutional-recurrent architecture equipped with attention mechanism to focus on task-relevant instances of the input speech signal, eliminates the need for task-specific meta data or domain knowledge based manual refinement of hand-crafted generic features. The second contribution, for the multimodal trained and multimodal tested category, is a multimodal fusion framework that exploits both cross (inter) and intra-modal interactions for categorical emotion recognition from audio-visual clips. We explore the effectiveness of two types of attention mechanisms, namely, intra- and cross-modal attention by creating two versions of our fusion framework. In many applications, multimodal signals might be available during model training phase, yet we cannot expect the availability of all modality signals during testing phase. Our third contribution addresses this situation wherein we propose a framework for cross-modal learning where paired audio-visual instances are used during training to develop test-time stand-alone unimodal models.

Contents

Publications

Papers

[C2] Vandana Rajan, Alessio Brutti and Andrea Cavallaro. Is Cross-Attention Preferable to Self-Attention for Multi-Modal Emotion Recognition? in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4693–4697, IEEE, 2022.

[C1] Vandana Rajan, Alessio Brutti and Andrea Cavallaro. Robust Latent Representations via Cross-Modal Translation and Alignment. in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4315–4319, IEEE, 2021.

[J1] Vandana Rajan, Alessio Brutti and Andrea Cavallaro. ConflictNET: End-to-end learning for speech-based conflict intensity estimation. *IEEE Signal Processing Letters (SPL)*, vol.26, pp.1668–1672, 2019.

Codes

[P1] https://github.com/smartcameras/SelfCrossAttn: PyTorch implementation of the Self- and Cross-attention models described in [C2]

[P2] https://github.com/smartcameras/ConflictNET: Keras implementation of the ConflictNET, convolutional-recurrent model with attention described in [J1]

[P3] https://github.com/vandana-rajan/1D-Speech-Emotion-Recognition: Keras implementation of an improved version of speech emotion recognition model proposed in [1]

Acknowledgements

First of all, I bow my head humbly in front of the God almighty for all the grace and blessings that have been showered upon me.

Next, I would like to thank my supervisor Professor Andrea Cavallaro for his continuous support and encouragement throughout my Ph.D journey. His experienced guidance has helped me to navigate my research through the pandemic induced difficult times and also to become a more confident version of myself. I would also like to express my gratitude towards my second supervisor, Dr. Alessio Brutti, for his kind and valuable feedbacks, discussions and support during our progress meetings.

It is a pity that most of my Ph.D had to be spent in physical isolation due to the pandemic. I could spend only the first 1.5 years of my Ph.D in our CIS lab (CS 440). However, I really enjoyed that limited time because of all the wonderful people in our lab. I am grateful to all of them - Dr. Alessio, Dr. Xinyuan, Dr. Riccardo, Dr. Ali, Dr. Mohamed, Dr. Mohammad, Dr. Ricardo, Dr. Chau Yi, Dr. Lin, Dr. Changjae, Ashish and Dimitris. I am also thankful to Alina, Elena, Xavier, Dmitrii, Faxian, Tommaso and Long for the happy times during my occassional visits to our post-pandemic lab. Sometimes when you are feeling lost and anxious about the future, it is a great relief to talk to people who have walked the path before you. A big thank you to my Ph.D seniors Dr. Girmaw, Dr. Saumitra and Dr. Amrith for taking their time to share their valuable experiences with me.

I am eternally grateful to my husband for his unconditional love, support and care. I wouldn't have had the courage to even think of doing a Ph.D if it were not for his continuous encouragement. I am thankful to my family for their patience and support. It was not easy for them to send their only child to live half-way across the globe. I am also grateful to my extended family in the UK for all the hand-holdings and care I received during my initial days in this country.

Last, but not the least, I thank all the hardworking people in our research community for their inspirational and motivational works as well as for their insightful comments and reviews on my works.

Chapter 1

Introduction

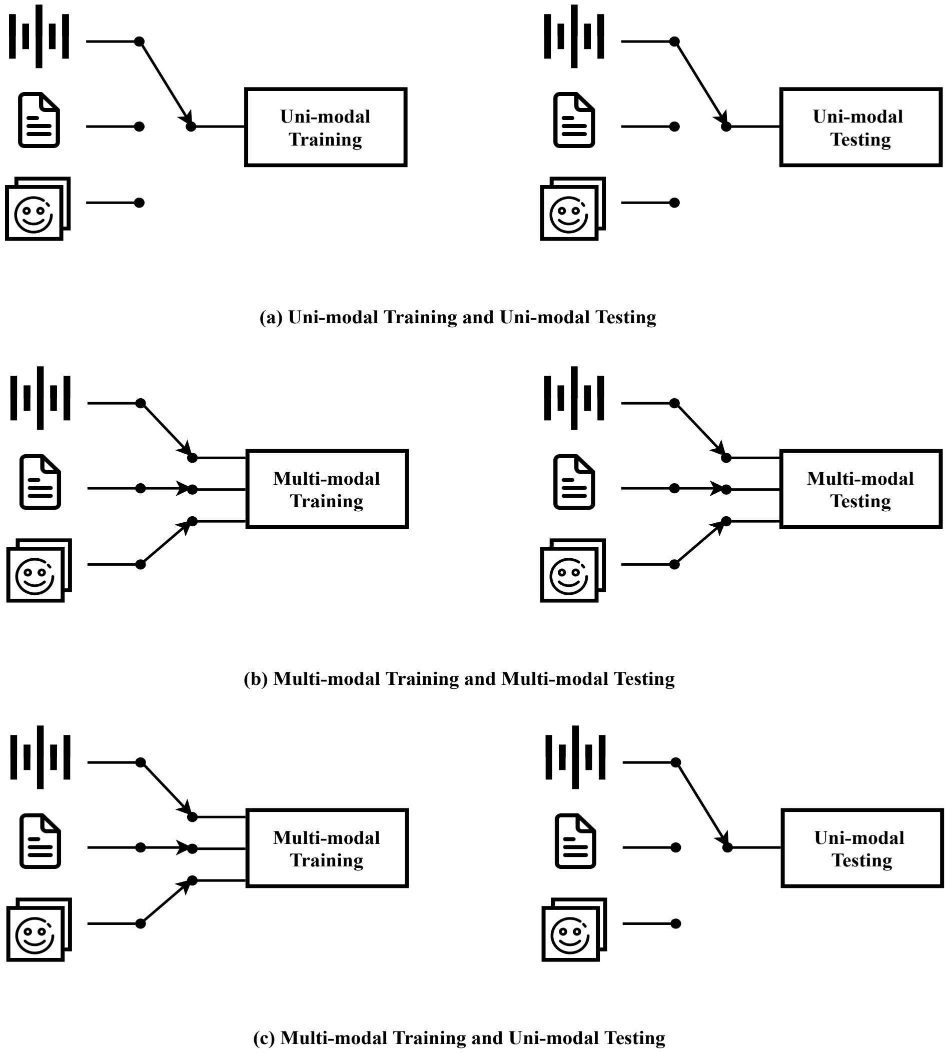
1.1 Motivation

The term ‘modality’ refers to the way in which something exists or is experienced or expressed [2]. Most physical events that we encounter in this world consist of multiple modalities; for example, we can see, hear and touch the rain; objects around us may have their own characteristic shape, sound and smell. We experience this world via biological sensors that capture signals from multiple modalities. Visible aspects of events can be seen using our eyes, the mechanical vibrations that they might produce can be heard, their texture can be felt by touch, their flavour can be tasted and their scent can be smelled. Our brain uses signals provided by one or more of these sensory organs to generate an understanding about the nature of the event/events in question. Our responses to events are also multimodal in nature. We can express ourselves as well as interact with other humans and objects using multiple modalities like linguistic speech, paralinguistic speech, non-verbal audio, facial expressions, gestures, postures, touch etc. In particular, human-human interactions are remarkable examples of events where there exist significant interplays between multimodal signals. This thesis focuses on three modalities that are important for human-human and human-machine interaction scenarios, namely, audio (paralinguistic speech), vision (facial expressions) and language (linguistic speech).

Machine Learning (ML) systems use artificial sensors to automate the functionality of biological sensors. Cameras can be the eyes, microphones can be the ears, tactile sensors can emulate the skin, taste sensors can be the tongue and odor sensors can be the nose. These sensors convert respective attributes of physical events into machine friendly formats. For example, cameras can convert three dimensional world co-ordinates into two dimensional images and microphones can convert mechanical vibrations into electrical audio signals. Outputs from these sensors are thus suitable for being processed by computers. For ML systems to blend seamlessly in our world, they need to be equipped with capabilities to extract relevant information from multi-sensory signals. ML algorithms can be designed to process one or more sensory signals at a time. ML systems that utilise only one modality are called 'unimodal' or 'monomodal' systems and those that process multiple modalities are called 'multimodal' systems [3, 4, 5]. Both unimodal and multimodal systems have their own challenges, advantages and disadvantages and a choice between them is taken after considering various factors like the use case, availability of sensors, computational and memory requirements etc.

Traditional unimodal systems for paralinguistic speech tasks use off-the-shelf classifiers/ regressors on generic hand-crafted acoustic features, which require further manual refinement and time-consuming feature pruning [6, 7, 8, 9]. Such systems require separate training and parameter tuning of the feature extractor and the classifier/regressor. Domain knowledge based hypotheses and metadata may also be needed to extract task-specific features from standard acoustic features [10, 11, 12]. Therefore, an important problem in speech based computational paralinguistics is the design of end-to-end deep learning, which trains models directly from raw input data: since the parameters are trained jointly, the end-to-end model learns task-specific features directly from the input, without requiring any guidance other than the objective function and the training dataset [13, 14].

While unimodal systems are limited by the information content in a single modality, multimodal systems can leverage information from multiple modalities and hence have more potential to provide improved performance compared to their unimodal counterparts [2, 3, 5]. Multi-modal systems should be capable of modelling both intra (within) modality as well as inter (cross) modality interactions. The term intra-modal refers to the interactions between different local temporal positions within a single modality to derive the task-specific global semantics. Intermodal or cross-modal refers to similar interactions across multiple modalities [15]. Multimodal systems should also be capable of handling heterogeneous modalities of varying dimensionalities by designing specific architectural components that can suit the characteristics of their respective modalities [16]. These components should be responsible for modelling the intra-modality inter-



(a) Uni-modal Training and Uni-modal Testing

(b) Multi-modal Training and Multi-modal Testing

(c) Multi-modal Training and Uni-modal Testing

Figure 1.1: The three multimodal paradigms considered in this thesis

actions in the system. The design complexity of multimodal systems stems partly from the fact that the information to explain an event is unevenly spread across the associated modalities and hence the multimodal system should accommodate the following modes of interplay between modalities [17];

- *Equivalence/substitution*: one modality conveys a meaning not borne by the other modalities (it could be conveyed by these other modalities)
- *Redundancy/repetition*: the same meaning is conveyed at the same time via several modalities
- *Complementarity*:
  - Amplification accentuation/moderation: one modality is used to amplify or attenuate

    the meaning provided by another modality

  - Additive: one modality adds congruent information to another modality
  - Illustration/clarification: one modality is used to illustrate/clarify the meaning conveyed by another modality

- *Conflict/contradiction*: the meaning transmitted by one modality is incompatible or contrasting with the ones conveyed by the other modalities
- *Independence*: the meanings conveyed by different modalities are independent and should not be merged.

Thus, an important problem in multimodal learning is to 'fuse' information from multiple modalities by taking these different types of intra- and inter-modal interactions into account [18, 19, 20].

In comparison to multimodal systems, unimodal systems are limited by the information content in a single modality alone. To infuse the advantages of multimodal learning into unimodal systems, a training strategy can be employed where multiple modalities are used during training to improve the test-time performance of unimodal models [21, 22, 23]. This strategy, generally known as, *Co-learning*, is based on the intuition that some modalities can 'aid' or 'help' other modalities in creating better feature representations or stand-alone models [2, 24, 25, 26]. The helper modality/modalities are usually used only during the training and not the testing phase.

This thesis focuses on the three paradigms of multimodal learning, namely, unimodal learning (training and testing with a single modality), multimodal fusion (training and testing with multiple modalities) and co-learning (multimodal training and unimodal testing), as shown in Figure 1.1. We first propose a unimodal end-to-end modelling framework that can process raw data without the need for meta information or domain knowledge based hand-crafted features. We then propose a multimodal fusion framework that can exploit intra- and cross-modal interactions using two types of attention mechanisms. Finally, we propose a multimodal co-learning framework that transfers knowledge from multiple modalities into unimodal models to improve the unimodal test-time stand-alone performance.

1.2 Problem Definition

Let $\mathbf{X}$ represent an audio-visual clip containing either human-human verbal communication interactions or monologues. $\mathbf{X}$ can be split into $\mathbf{X}_a$ and $\mathbf{X}_v$ representing audio (speech) and vision (RGB image sequence) channels respectively. Language modality represented by $\mathbf{X}_l$ can be extracted from $\mathbf{X}_a$ by means of manual or automatic transcription. Since all signals are derived from the same audio-visual clip, the temporal lengths of all signals are the same (in terms of seconds, minutes or hours). However, the number of samples within each modality can be different because of the difference in the sampling frequencies of their corresponding sensors.

Since speech is the most natural means of human-human communication, our first research goal is to ascertain if it is possible to design an end-to-end deep learning model for computational paralinguistics. Here the objective is to develop an end-to-end trainable deep learning based regression model that uses one dimensional raw speech signal $\mathbf{X}_a$ as input and maps it to a decision space of continuous labels $\mathbf{Y}_c$. The model should not rely on explicit meta information like the number of speakers, speech overlaps or onset-offset instances, instead it should learn task-relevant features directly from the input data.

Our second research goal is to achieve multimodal fusion using audio, vision and language modalities for affect recognition tasks. Here the objective is to develop a fusion model that can learn both intra (within) and cross (across) modality interactions for an important multimodal para-linguistic task, namely, categorical emotion classification. The model should take combinations of audio, vision and language modality features as input and map them to a discrete ($\mathbf{Y}_d$) or continuous label space ($\mathbf{Y}_c$).

Our third and final research goal is to achieve multimodal co-learning that can improve the unimodal stand-alone performance by employing multiple modalities during training phase. The objective here is to develop a co-learning model that can learn better unimodal feature representations by using paired multimodal signals during the training phase alone. Both $\mathbf{X}_a$ and $\mathbf{X}_v$ can be used during training to develop better unimodal features when either $\mathbf{X}_a$ or $\mathbf{X}_v$ alone is available during testing. Let the performance of a unimodal trained and unimodal tested model be denoted by an evaluation metric as $\mathbf{A}_1$ and the performance of the multimodal trained and unimodal tested model in terms of the same metric be denoted as $\mathbf{A}_2$. Then $\mathbf{A}_2$ should be greater than $\mathbf{A}_1$, where higher the value better is the metric.

1.3 Research Questions

In order to address the above mentioned research goals, the following research questions(RQs) have been explored in this thesis.

- **RQ1:** *Is it possible to design end-to-end models for speech based computational paralinguistic tasks? and if yes, can the predictions of such 'black-box' models be interpreted using any existing explainable-AI [27] methods?*

  Exploration of the existing research literature shows that majority of the end-to-end neural networks for speech modality are designed for the task of speech recognition [28, 29, 30]. Although a few end-to-end works exist for emotion recognition [14], such models for other computational paralinguistics tasks are scarce. Also, these works do not use any existing explainable-AI methods to interpret the predictions of their end-to-end models. We aim to address these research gaps by designing the first speech based end-to-end model for a complex paralinguistic task, namely, verbal conflict intensity estimation, and adapting the explainable-AI method called LIME [31] to explain its predictions.

- **RQ2:** *Is cross-modal attention preferable to intra-modal attention in multi-modal fusion models for emotion recognition? Also, how robust are the performances of these models in missing modality conditions and what are the ways to alleviate missing-modality performance deterioration?*

  A plethora of research literature exists on multi-modal fusion for affective computing, that uses audio, vision and language modalities as input. Lately, attention [32] based multi-modal models have been shown to be effective in focusing on emotionally salient regions across signals from multiple modalities [33, 20, 34]. Recently, cross-modal attention [19], that computes relevance score for each time-step in one modality by utilising another modality, has been shown effective for multi-modal fusion. Since then, multiple works on multi-modal affective computing models incorporate cross-modal attention [35, 34, 36, 37], self-attention [38, 39] or a combination of both [40, 20, 41, 42] in their architectures. However, the existing literature is unclear on whether cross-modal attention is indeed better than using only self-attention mechanism in a model architecture. Such a comparison could be useful for making an informed choice between the two for future research works. Hence the reasoning behind our second research question.

- **RQ3:** *Is it possible to design a framework that utilises multi-modal signals during training phase to develop a model that is intended to have a single modality input during test time for affective computing applications?*

  The research questions we have so far, dealt with cases where the inherent assumption is that the same modalities are available during both the training and testing phases. A natural extension would be the case where all modalities available during training are not expected or required during the testing phase. Hence, the reasoning beind our third research question. Such a framework is known as 'co-learning' [2, 43] or 'learning with privileged information' [25, 44]. Existing literature contains multiple works on multi-modal co-learning where the involved modalities are different types of image modalities, like RGB, depth and optical-flow images [21, 45, 26]. These works cater to image modalities that have strong correspondence with each other, for example, a discontinuity in the depth image can be directly mapped to an edge in the RGB image. However, when it comes to heterogeneous modalities like audio, vision and language signals derived from affective videos, such correspondences become less obvious and there is a dearth of co-learning research in multi-modal affective computing. We aim to explore this research gap.

1.4 Contributions

Given audio, vision and language signals corresponding to human-human or human-machine interaction scenarios, our aim is to develop models that fit into the three multimodal paradigms, namely, unimodal learning, multimodal fusion and co-learning. We propose in this thesis, three frameworks, one for each of these paradigms. The main contributions of this thesis are as follows:

- **Contribution 1:** An end-to-end trainable deep learning model for verbal conflict intensity estimation from raw speech signals. The model uses a convolutional-recurrent architecture with attention for finding task-relevant features directly from the input. The network training solely relies on the raw speech signals and their corresponding labels in the training set. The need for task specific meta data like the number of speakers and speech interruptions, as well as domain knowledge based hypothesis is eliminated. An extensive ablation study confirms the choice of individual components in the architecture. Furthermore, a subjective as well as an interpretability analysis of the model points out what specific instances from the input signal have been picked by the network to create predictions. [J1]

Novelty/Improvements:

(a) The first end-to-end model for verbal conflict intensity estimation, *ConflictNET* that takes raw speech signal as input. This answers the first part of our RQ1.

(b) *ConflictNET* outperforms in terms of the evaluation metric (Pearson Correlation Coefficient) all but one method [7] (see table 3.1). *ConflictNET* achieves almost the same performance as [7], a model that uses speech overlap feature set and feature pruning based conflict specific subset of standard acoustic features. This could indicate that our end-to-end architecture has automatically learned task-specific information from the raw speech input.

(c) We were able to adapt an existing explainable-AI method, LIME [31], to explain the predictions of an end-to-end paralinguistic model, thus pinpointing the salient portions of input speech which are relevant for the model's prediction. This answers the second part of our RQ1.

- **Contribution 2:** A multimodal fusion framework that uses audio, vision and language modalities as input for multi-class emotion classification task. Modality-specific and common components are used to model intra-modal and cross-modal interactions respectively. Two types of attention mechanisms, namely, self- and cross-attentions are used to develop two versions of the multi-modal framework. [C2]

  Novelty/Improvements:

  (a) The first work to perform an extensive comparison between the two most commonly used attention mechanisms in multi-modal affective computing literature, thus answering the first part of our RQ2.

  (b) Our proposed fusion models achieve state-of-the-art results for categorical tri-modal emotion recognition on one of the most widely used datasets, IEMOCAP [46] (see table 4.1). In terms of weighted accuracy metric, our tri-modal self and cross-attention models outperform the state-of-the-art model AMH [33] by 4.0 and 3.1 percentage points (pp) respectively. Similar numbers in terms of unweighted accuracy metric are 2.5 and 1.1 pp respectively.

  (c) A thorough study of the performance deterioration of our multi-modal models during missing modality scenarios indicate that both self and cross-attention based models are susceptible to this issue. Nevertheless Moddrop [47] and KNN based missing modality imputation method [48] could be used to reduce the extend of performance degradation. Infact, we show that Moddrop training could even bring performance improvement for full modality situations (see table 4.5). This answers the second part of our RQ2.

- **Contribution 3:** A co-learning framework that uses audio and vision modalities during training so as to improve either audio or vision alone during the testing phase. This framework is developed based on the observation that not all modalities provide equal performance on the same task which can be attributed to the variations in the task-specific information that they contain. Our co-learning framework can be used to improve the performance of a weakly performing modality by using a stronger modality during training. The framework consists of two core components, translation from weaker to stronger modality (cross-modal translation) and correlation based latent space alignment. Modality-specific unimodal encoders are used to map their respective features into a common latent space and a correlation based loss is applied over this space to align the weaker modality components with those of the stronger modality. Based on the intuition that cross-modal translation can create intermediate features that are representative of both modalities, a decoder is used to translate the weaker modality into the stronger modality representations. Once the multi-loss based training is over, all model components corresponding to the stronger modality can be discarded. [C1]

  Novelty/Improvements:

  (a) Two versions of co-learning framework are developed - one for non-sequential data and another for sequential data. Our models are able to use both stronger and weaker modalities during training to improve the test-time performance of weaker modality. Evaluation of our models on two affective computing tasks answers RQ3.

  (b) Our non-sequential model is able to either outperform or be on par with the state-of-the-art method Emobed [23] for visual-to-audio and audio-to-visual emotional knowledge transfer. Out of all the methods compared (including the best uni-modal models), our method occupies first or second position in terms of performance on the evaluation metric (Concordance correlation coefficient) for continuous emotion recognition task using RECOLA [49] dataset (see table 5.3).

(c) Our sequential model is able to further improve upon the results for continuous emotion recognition task. This shows that incorporating contextual information from neighbouring time-steps is beneficial. Our sequential model is thus able to improve upon the state-of-the-art method Emobed [23] and achieve the best results in terms of performance on the evaluation metric (Concordance correlation coefficient) for continuous emotion recognition task using RECOLA [49] dataset (see table 5.7). Our model is also able to improve upon the state-of-the-art results of HMTL [50] for the binary sentiment classification task using the CMU-MOSI [51] dataset (see table 5.6).

1.5 Thesis Structure

This thesis is structured as follows:

- Chapter 1: We introduce and formulate the three paradigms of multimodal learning using modalities that most commonly occur in human-human and human-machine interaction scenarios, namely, speech (paralinguistic), vision (facial expressions) and language (linguistic speech) and we list the contributions made.

- Chapter 2: We provide a review of the relevant background literature for multimodal fusion and co-learning. We also discuss methods specific to the multiple paralinguistic problems used in this thesis, namely, verbal conflict intensity estimation/detection, multi-class multimodal emotion classification, multimodal continuous emotion recognition and multimodal binary sentiment classification. We also provide details about the datasets and the metrics used to validate our approaches.

- Chapter 3: We present an end-to-end trainable model design, called *ConflictNET*, for verbal conflict intensity estimation from raw speech signals. We describe in detail the architectural components, the loss function and other training details. We present the results, ablation study and comparison with state-of-the-art methods in terms of three metrics. We also present details about the subjective and interpretability based analysis of our model. Finally we present the observations and conclusions derived from the experiments.

- Chapter 4: We present a multimodal fusion framework that uses audio, vision and language signals for seven class emotion classification task. We design two models, one based on intra- and another based on cross-modal attention and present a thorough comparison of the uni-, bi- and tri-modal modality combinations of both models. We present details of our model architectures, loss function and other training details. We also present our results in terms of two metrics for 7-class emotion classification. We then present details about the test-time missing modality handling analysis of the two models. Finally, we present our observations and conclusions.

- Chapter 5: We present a co-learning framework, called Stronger-Enhancing-Weaker, for improving the unimodal performance of a weaker modality model by using a stronger modality during training phase. We present details about two versions of our framework, one each for non-sequential and sequential data and discuss the choice of architectural components used. We also detail about each individual losses in the multi-loss function used for the training. We present our results on the tasks of continuous emotion recognition and binary sentiment classification. We also provide details of the ablation study that quantifies the effect of individual components in the model. Finally, we present our observations and conclusions.
- Chapter 6: We summarise the methods and achievements in the thesis and provide a discussion on future works.

Chapter 2

Background

2.1 Introduction

In this chapter, we review the literature on deep learning based multimodal learning techniques for heterogeneous modalities. The purpose of our literature review is to understand existing research on multimodal machine learning that uses various image modalities, audio as well as language modalities for different human-machine interaction applications. We approach this review from the point-of-view of machine learning model designing, focusing on the particular aspect of how to effectively combine task specific information originating from multimodal signals. Since the specific focus of this thesis is on multimodal affective computing applications, we provide a brief description on how the subject of affective computing evolved over the years and what are some popular tasks contained within its broad umbrella. Specifically, we describe in detail two particular focus areas in multimodal learning, namely, multimodal fusion and multimodal co-learning in sections 2.2-2.3. Since attention mechanism has been used extensively in the unimodal and multimodal contributions in this thesis, we detail this mechanism in section 2.4. We explain the four multimodal learning application scenarios considered in this thesis in section 2.5. The details of the datasets used, the feature representations and evaluation metrics are discussed in section 2.6.

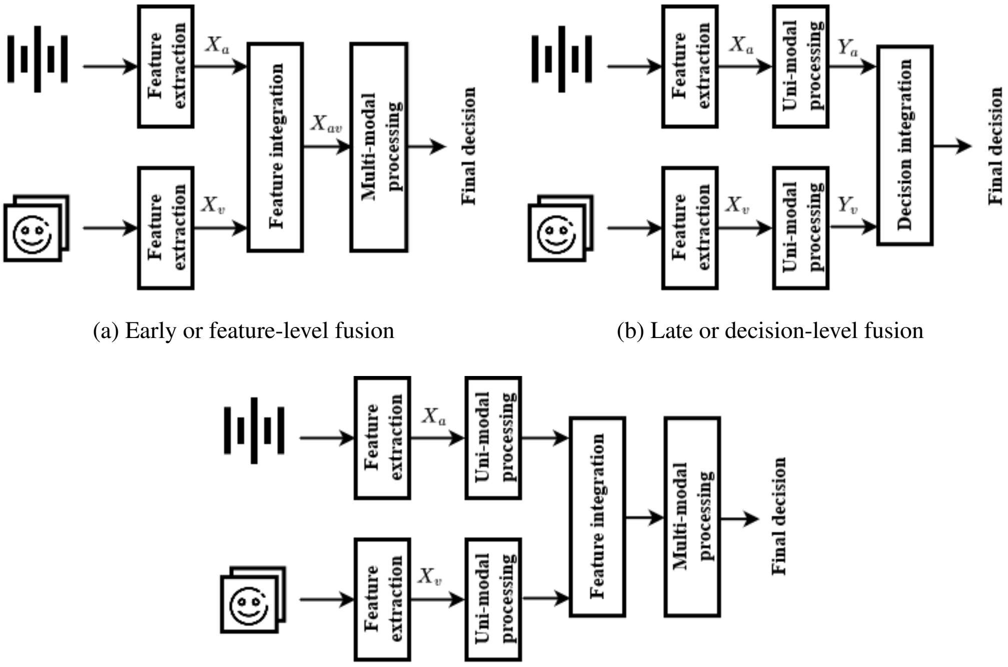
2.2 Multimodal Fusion

The multimodal learning approach that combines information from multiple modalities for a classification or regression task is popularly known as 'multimodal fusion'. Multimodal fusion has been an active research area for the past few decades. To the best of our knowledge, the earliest work on multimodal fusion using neural networks, was published in the IEEE Communications Magazine in 1989 [52]. This work used a weighted combination of speech and vision modality representations for audio-visual speech recognition task. With technological advancements in computing, as deep neural networks started to become popular in 2010s, deep learning based multimodal fusion models were also introduced in 2011 [5]. An auto-encoder model was used to create a fusion framework using audio and vision modalities for multimodal speech recognition task. Since then, there has been a plethora of research in deep learning based multimodal fusion methods. Earliest works categorize multimodal fusion approaches into three categories - *early* or *feature-level* fusion, *late* or *decision-level* fusion and *model-level* or *intermediate* fusion (see Figure. 2.1).

2.2.1 Early, late and model-level fusion

Early fusion methods concatenate the multiple modalities into a unified representation prior to proceeding through the learning/feature extraction process [53, 54, 55, 56, 57]. Despite the simplicity in formulation, an obvious downside of the method is that since the early fusion techniques avoid explicit modeling of the different modalities, they fail to model both the fluctuations in the relative reliability and the asynchrony problems between the distinct (e.g., audio and visual) streams [58]. Because of the simple concatenation at the input, the processing model lacks the ability of capturing the complex correlations across modalities when the data sources are significantly varied from each other in terms of sampling rate, data dimensionality and unit of measurement [3, 59].

On the other hand, late fusion methods use multiple unimodal models and combine their decisions or predictions by a voting scheme or averaging [57, 60], a bilinear product [61] or a simple pooling operator [62]. Thus, late fusion methods are inherently capable of handling missing modality scenarios. Another reason for the popularity of late fusion is that the architecture of each unimodal stream is carefully designed over years to achieve state-of-the-art performance for each modality. This enables the unimodal streams of a multimodal model to be initialized



(a) Early or feature-level fusion

(b) Late or decision-level fusion

(c) Intermediate or model-level fusion

Figure 2.1: Early, late and model-level fusion

with weights that have been pre-trained with a large number of unimodal training samples [16]. However, late fusion does not provide scope for interaction between modalities until at the last stage of prediction. This limits the ability of the methods to exploit inter-modal correlations for deriving high level semantic concepts related to the prediction task. Moreover, as different models are used to obtain the unimodal decisions, the learning process for them becomes tedious and time-consuming [59].

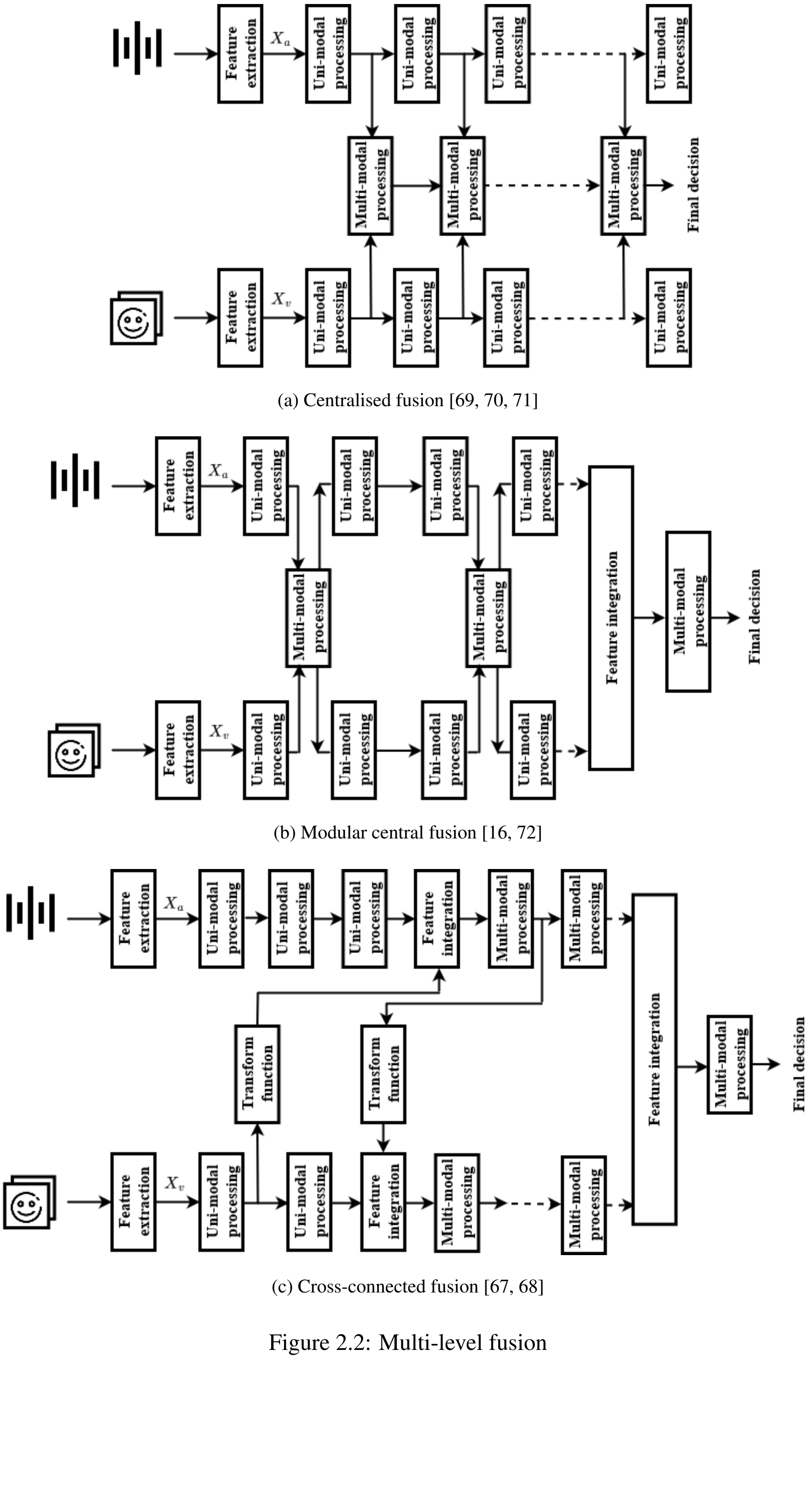
The model-level or intermediate fusion strategy first processes unimodal components using modality-specific parts of the fusion model. The processed unimodal components are then combined via concatenation [63, 64, 65] or weighted addition [66] before being further processed using a common feature extractor. While model-level fusion can improve the inference of shared semantics, they are susceptible to failures due to missing modality. Also, intermediate level features of different modalities have different or unaligned spatial dimensions making the intermediate fusing more challenging [67, 68].

2.2.2 Multi-level fusion

Even though the early, late and model-level fusion paradigm provides a generic framework to decide the level at which fusion is done in a multimodal model, determining the exact layer or depth at which the fusion would provide optimal performance is not straightforward. An interesting approach to solve this problem is to fuse unimodal features at multiple levels instead of only one (see Figure 2.2). This is in line with a common interpretation of deep neural models considering that features learned at different layers of the network carry varying levels of semantic meanings. Thus features from different layers at different modalities can give different insights from the input data.

CentralNet [69] uses two modality specific unimodal networks and connects them using an additional central network dedicated to the projection of the features coming from different modalities into the same common space. The central network combines features issued from different modalities, by taking, as input of each one of its layers, a weighted sum of the layers of the corresponding unimodal networks and of its own previous layers. A global loss allows to back propagate some global constraints on each modality, coordinating their representations. The approach is multitask since it simultaneously tries to satisfy per modality losses as well as the global loss defined on the joint space. A contemporary work, Dense Multimodal Fusion (DMF) [70] also combines unimodal features from multiple levels using a central network. The difference from CentralNet is that DMF does not use multi-task training and uses only a single loss for optimising the entire network using the output at the central layer. The resulting model is claimed to have faster convergence, lower training loss, and better performance.

Instead of using a central network branch, multiple MultiModal Transfer Modules (MMTM) can be used to combine unimodal features of different spatial dimensions at different levels of feature hierarchy [16]. MMTM uses a modified form of Squeeze and Excitation [73] operation and each MMTM module contains 2 units (1) multimodal squeeze unit that receives the features from all modalities at a given level of representation across the branches, generating a global joint representation of these features, and 2) an excitation unit that uses this joint representation to adaptively emphasize on more important features and suppress less important ones in all modalities. While MMTM is specific to CNN architectures, Multimodal Split Attention Fusion (MSAF) [72] extends this idea for non-convolutional architectures as well. Another line of work infuses cross-modal features at different levels into the unimodal branches to periodically allow for information exchange between them [67, 68]. In XCNN [67], the constituent unimodal networks are individually designed to learn the output function on their own subset of the input data, after which cross-connections between them are introduced after each pooling operation to pe-



(a) Centralised fusion [69, 70, 71]

(b) Modular central fusion [16, 72]

(c) Cross-connected fusion [67, 68]

Figure 2.2: Multi-level fusion

riodically allow for information exchange between them. This method is shown to be beneficial especially when the amount of training data is less. While XCNN is limited by the use of unimodal networks that are compatible (CNNs), XFlow [68] removes this restriction by proposing generalised cross-connections which transfer information between streams that process incompatible data.

Yet another line of research extends the previous idea that considering features extracted from all the hidden layers of independent modalities could potentially increase performance with respect to only using a single combination of late (or early) features. However, unlike CentralNet, DMF, XCNN and XFlow, where the locations of fusion are either pre-defined or are determined empirically, we can automatically find optimal locations for fusion by formulating the problem as a multimodal neural architecture search problem. A sequential search algorithm called, sequential model-based optimization (SMBO) [74] scheme, which has previously been applied to the related problem of neural architecture search or AutoML [75], can be adapted for multimodal fusion architecture search [71]. In this way, the number of possible fusion layers and the type of fusion operation (concatenation or weighted fusion) are search parameters. This design enables the search space to contain a large number of possible fusion architectures, including the networks defined in CentralNet and DMF.

2.2.3 Location agnostic fusion

Orthogonal to the literature on finding optimal locations for fusion in a multimodal model, another line of research focuses on the methodology of fusion, i.e; the nature of combining information from multiple modalities.

A fusion mechanism inspired by the flow control in recurrent architectures like GRU or LSTM is Gated Multimodal Unit (GMU) [76]. GMU can be used as an internal unit in any neural network architecture and it learns an input dependent gate activation pattern that determines how each modality contributes to the output of hidden units. Multimodal channel exchanging fusion is a parameter-free fusion framework that dynamically exchanges channels between subnetworks of different modalities [77]. The scaling factor of batch normalisation is used as the importance measurement of each corresponding channel. The method replaces the channels associated with close-to-zero scaling factors of each modality with the mean of other modalities. Such message exchanging is parameter-free and self-adaptive as it is dynamically controlled by the scaling factors that are determined by the training itself. Parameters except batch-norm layers of all sub-networks are shared with each other. By using private batch-norms, we can determine the channel importance of each modality. By sharing convolutional filters, the corresponding channels among different modalities are embedded with the same mapping, thus more capable of modelling the modality common information. This design further compacts the multimodal architecture to be as small as the unimodal one.

Bi-linear pooling [78], originally proposed for fine-grained visual recognition, is a method to fuse feature maps from different networks processing the same image. The intuition here is that different networks can capture different types of information from the input image and combining the feature maps would result in a fine-grained representation of the input image. Bi-linear pooling has been extended to multimodal fusion by combining feature representations corresponding to different modalities. To reduce the computational complexity, matrix factorization based compact bi-linear fusion [79] can be used for multimodal models. Tensor fusion layer [80] explicitly models the unimodal, bi-modal and tri-modal interactions using a 3-fold Cartesian product from modality embeddings. Unimodal vectors are first obtained using modality specific networks and outer-product between the vectors is taken to obtain the 3D cube of all possible combination of unimodal embeddings. Since Tensor Fusion is mathematically formed by an outer product, it has no learnable parameters and it is empirically observed that although the output tensor is high dimensional, chances of overfitting are low.

Multiplicative fusion [81] explicitly models the fact that on any particular sample not all modalities may be equally useful. The method first makes decisions on each modality independently and then the multimodal combination is done in a differentiable and multiplicative fashion. This multiplicative combination suppresses the cost associated with the noisy or weak modalities and encourages the discovery of truly important patterns from informative modalities. Attention [82] mechanism can be used for multimodal fusion by dynamically adjusting the relative importance of time-steps in input sequences of multiple modalities [83]. One of the benefits of attention based fusion is that modalities that are most helpful for the task can dynamically receive stronger weights. Also, the network can detect interference (noise) and other sources of uncertainty in each modality and dynamically down-weigh the modalities that are less certain.

2.2.4 Factorised fusion

A different approach to multimodal fusion focuses on learning better feature representations instead of novelty in fusion methodology or model architecture.

MISA [18] learns two distinct representations for each modality - *modality-invariant* and *modality specific*. The modality-invariant representations are aimed to reduce modality gaps. Modality gaps refer to the differences across modality representations with respect to dimensionality and sampling rates, which make it computationally difficult to find task relevant common information across modalities. Modality-invariant mappings help to capture underlying commonalities and correlations by aligning the projections of multiple modalities onto a shared sub-space. Modality-specific representations are, on the other hand, private to each modality. By explicitly factorising representations of each modality into modality-invariant and specific components, MISA relieves the extra burden on the multimodal model to implicitly bridge modality gaps and learn common features. Multimodal input can also be factorized into *multimodal discriminative* and *modality-specific generative* factors [84]. The discriminative factors are shared across all modalities and contain intra-modal and cross-modal features required for discriminative tasks. The generative factors are unique for each modality and contain information for generating the data which allows the model to infer missing modalities at test-time and deal with the presence of noisy modalities.

2.2.5 Sequential fusion

Multimodal fusion models for heterogeneous sequential data should be able to exploit the contextual information in sequences across modalities.

Sequence-to-sequence (Seq2Seq) [82] models, originally designed for language translation tasks, can be adapted for learning fused representations of sequential multimodal data. Unsupervised seq2seq modality translation using recurrent encoder-decoder architecture can be used to create multimodal representations useful for sentiment analysis tasks [85]. An attention mechanism is used at the encoder output to provide varying weights to different time-steps in the encoded sequence. A hierarchical version of this model can be used to create multimodal representations when the number of involved modalities is more than two.

Memory Fusion Network [86] is a deep network for multimodal sequential learning. MFN uses a system of LSTMs in which modality specific interactions are learned by assigning an LSTM function to each modality. Cross-modal interactions are learned using a special attention mechanism called Delta Memory Attention Network (DMAN) and summarised through time using a multimodal gated memory mechanism. DMAN identifies the multimodal interactions by associating a relevance score to the memory dimensions of each LSTM. The gated memory mechanism updates its contents based on the outputs of the DMAN and its previously stored contents, acting as a dynamic memory module for learning crucial multimodal interactions throughout the sequential data. Graph Memory Fusion Network (GMFN) [87], an extended version of MFN, replaces the original delta-memory attention network with a dynamic fusion graph.

Multi-Attention Recurrent Network (MARN) [88] is a model for human communication comprehension that can discover interactions between modalities through time using multi-attention blocks and store them in the hybrid memory of a recurrent component. Recurrent Attended Variation Embedding Network (RAVEN) [89] models human language by shifting word representations based on the accompanying nonverbal behaviors such as facial expressions and vocal patterns. Recurrent Multistage Fusion Network (RMFN) [90] decomposes fusion into three stages: a 'highlight' stage for identifying and highlighting a subset of multimodal signals, a 'fuse' stage for conducting local fusion of highlighted features and integrating representations of the previous stage, and a 'summarize' stage for drawing final prediction.

Fine-tuning Attention Fusion (FAF) [91] preserves the original unimodal attentions and provides a fine-tuning attention for the final prediction. It utilizes word-level alignment to model time-dependent interactions among modalities. A multi-hop attention can be used to alternatively finding relevant time-steps in one modality by conditioning on the other modalities [33]. The sequential unimodal features are first processed using individual recurrent layers (GRUs) and a context vector using last-step hidden representations of other modalities is used to obtain attention scores for each modality. This process is done iteratively across modalities for the task of multi-class emotion recognition.

A summary of all the fusion categories described so far, along with their specific characteristics, pros and cons, is given in Table 2.1.

2.3 Multimodal Co-learning

Multimodal co-learning, in comparison to multimodal fusion, is a less explored research area. Co-learning involves cross-modal knowledge transfer where multiple modalities are used during the training phase to help a target modality to function independently of other modalities during the testing phase.

Multimodal Training Unimodal Testing (MTUT) [21] is a method for supervised cross-modal knowledge transfer using RGB, depth and optical-flow modalities where the transferred knowl-

Table 2.1: Summary of multimodal fusion methods

Type	Characteristics	Pros	Cons	Examples
Early Fusion	[1] feature-level fusion [2] no explicit unimodal modelling	[1] simple & easy to implement [2] needs only one model	[1] fails to model asynchorny between modalities [2] cannot model incompatible dimensions [3] cannot capture complex correlations across modalities	[53] [54] [55] [56] [57]
Late Fusion	[1] decision-level fusion [2] explicit unimodal modelling	[1] can handle missing modality cases [2] can use pre-trained unimodal streams [3] can handle modalities with varying dimensions	[1] limited inter-modal interactions [2] time-consuming & separate learning of unimodal models	[57] [60] [61] [62]
Model-level Fusion	[1] intermediate-level fusion [2] has both modality-specific & common architectural components	[1] better inference of shared semantics [2] models both intra- & inter-modal interactions	[1] susceptible to failures due to missing modality [2] incompatible dimensions of modality-specific intermediate features	[63] [64] [65]
Centralised Fusion	[1] extension of model-level fusion [2] central branch connecting unimodal branches at multiple levels	[1] more fine-grained fusion [2] fusion of multimodal semantics at multiple levels	[1] increased architectural complexity [2] difficult to extend for more than two modalities	[69] [70]
Modular Central Fusion	[1] two inputs-two outputs modules [2] dedicated fusion modules connecting selected levels of unimodal branches	[1] eliminates cumbersome architecture of central branch	[1] data dependent location of fusion modules [2] difficult to extend for more than two modalities	[16] [72]
Cross-connected Fusion	[1] cross-connections across unimodal branches at multiple levels [2] connections are irrespective of feature hierarchy	[1] easier dataflow between unimodal branches [2] simpler implementation compared to centralised methods	[1] limited to two modalities [2] data & unimodal architecture dependent location of cross-connections	[67] [68]
Location Agnostic Fusion	[1] focus on ‘how’ to fuse rather than ‘where’ to fuse	[1] generic across modalities & model architectures [2] can be extended to more than 2 modalities	[1] needs careful selection based on task & data properties	[76] [77] [78] [80]
Factorized Fusion	[1] focuses on factorization of modality features	[1] leads to interpretable features [2] architecture agnostic	[1] increase in number of components & factors in loss function with increase in modalities	[18] [84]
Sequence Fusion	[1] designed to use contextual information within & across multimodal sequences [2] uses recurrence & attention mechanisms	[1] can model temporal dynamics across modalities [2] suitable for more than 2 modalities	[1] expects availability of entire sequences of modalities without missingness [2] can be complex architecture than non-sequential models	[85] [86] [33]

edge works as an extra supervision in addition to the class labels. The unimodal 3D CNN networks share knowledge by aligning the semantics of the deep representations. This is done by selecting an in-depth layer in the network and enforcing them to share a common correlation by minimising the distance between their correlation matrices during the training phase. A regularisation parameter is also used to avoid the weaker networks from transferring knowledge to the stronger networks. Multimodal co-learning is closely related to the concept of *Learning Under Privileged Information* (LUPI) [25]. LUPI is based on the principle of knowledge transfer from an 'intelligent' teacher to a student. During training stage, an intelligent teacher provides the student with information that contains, along with classification of each example, additional privileged information (for example, explanation) of this example. When the additional information is from another modality, LUPI becomes similar to supervised multimodal co-learning.

2.3.1 Co-learning for image modalities

Modality 'hallucination' technique considers depth modality as a side information during training to create an RGB only model for testing [44]. Multi-layer CNNs are used as unimodal architectures. In a multi-step training process, firstly, RGB and depth modality streams are independently trained. Then, a hallucination network, which takes RGB images as input, is initialised with the learned depth network weights. Finally, the three streams are trained jointly. The entire network is trained using a composite loss function. Based on the intuition that the deeper layers of depth and hallucination streams should have similar activations, the hallucination network, in addition to classification loss, is also trained using a Euclidean loss between the activations. Finally during the testing phase, the two stream multimodal model composed of the RGB network and the hallucination network, both using RGB data as input, is deployed. The Euclidean loss for increasing the similarity between depth and hallucinated feature maps, is part of the total loss along with more than ten classification and localization losses, thus making its effectiveness dependent on hyperparameter tuning to balance the different values, as the model is trained jointly in one step by optimizing the aforementioned composite loss.

As an improvement, the hallucination learning can be encouraged by design, by using cross-stream multiplicative connections from depth to RGB network [45]. After the first step of pre-training unimodal streams independently, they are trained jointly with the multiplicative cross-connection from depth to RGB stream as well as fine tuning on the late fusion model. In the next step, the depth stream is frozen and hallucination network is initialised with depth stream's weights and a teacher-student training loss combined with Euclidean loss between activation maps is used to train the hallucination network. Finally the RGB stream and hallucination network are fine-tuned in the late fusion model. In order to further reduce the dependency on balancing multiple losses and hyper-parameter tuning, an adversarial learning strategy is used to further improve the method.

Adversarial Discriminative Modality Distillation [26] also uses a two-step algorithm to learn representations from RGB and depth modalities while relying only on RGB during test. In the first step, RGB and depth modality networks are trained individually as two standard, separate supervised learning models. In the next step, a 'hallucination' network, initialised using the depth stream weights, is trained using an adversarial training strategy to generate depth features using RGB video as input. Finally during the testing phase, the two stream multimodal model composed of the RGB network and the hallucination network, both using RGB data as input, is deployed. The adversarial strategy uses the hallucination network as a 'generator' to produce the corresponding depth modality features and a 'discriminator' network which not only generates a 'true' or 'fake' label for the generator output but also has the auxiliary task of classifying feature vectors with their correct class.

It is to be specifically noted that these methods have all been used for image based modalities like RGB, depth and optical-flow which have strong correlations and correspondences with each other compared to heterogeneous modalities like audio, visual and text. For example, a discontinuity in the depth image can be directly mapped to an edge in the RGB image, while such a one-to-one correspondence between speech and facial expressions or gestures cannot be guaranteed or are less explicit.

2.3.2 Co-learning for heterogeneous multi-modal data

Cross-modal knowledge transfer using acoustic, visual and/or textual modalities exploit the complementary information from these modalities during training to develop a unimodal [22, 23, 94] or bi-modal [50] system. While the majority of these systems are based on supervised learning, there are some works based on un-supervised and self-supervised knowledge transfer.

A joint audio-visual training and cross-modal triplet loss [97, 98] based fully supervised framework, called EmoBed, can be used for multimodal training in face/speech emotion recognition task [23]. Two layer GRUs are used as unimodal streams for audio and visual modalities. These are followed by a shared network made up of two GRU layers. The shared network can

Table 2.2: Summary of cross-modal knowledge transfer methods used in multimodal training and unimodal testing models.

Ref.	Arch.	Modalities						Method	Seq.	Task
		Vis.				Aud.	Lan.			
		$\mathcal{I}$	$\mathcal{D}$	$\mathcal{F}$	$\mathcal{V}$					
[21]	I3D (3D CNN)	●	●	●	○	○	○	correlation alignment	●	gesture recognition
[44]	AlexNet/VGG	●	●	○	○	○	○	weight initialisation, activation matching	●	object classification
[45]	Resnet-50	●	●	○	○	○	○	cross-connections, weight initialisation adversarial training	●	action recognition,
[26]	Resnet-50	●	●	○	○	○	○	weight initialisation, adversarial training	●	action recognition, object classification
[22]	LSTM-attn.	○	○	○	●	●	●	contrastive loss	●	emotion recognition
[92]	CNN	○	○	○	●	●	○	T-S knowledge distillation	●	emotion recognition
[93]	LSTM enc.-dec.	○	○	○	●	●	●	cross-modal translation	●	sentiment analysis
[94]	GRU	○	○	○	●	●	○	alternate shared layers training,	○	emotion recognition
[23]	GRU	○	○	○	●	●	○	alternate shared layers training, cross-modal triplet training	○	emotion recognition
[50]	GRU-attn.	○	○	○	●	●	●	cross-modal translation, adversarial training	●	sentiment analysis, emotion recognition
[95]	MLP	○	○	○	●	●	○	cross-modal translation, correlation alignment	○	sentiment analysis
[24]	LSTM	○	○	○	●	●	●	mutual learning	●	sentiment analysis
[96]	U-Net	○	○	○	●	●	○	cross-modal translation	●	emotion recognition

KEY - Ref.: reference, Arch.: architecture, Vis.: vision, Aud.: audio, Lan.: language, $\mathcal{I}$: RGB image; $\mathcal{D}$: depth image; $\mathcal{F}$: optical flow image; $\mathcal{V}$: RGB video; enc.: encoder, dec.: decoder, attn.: attention, trans.: transformer, MLP: multi-layer perceptron, T-S: teacher-student, Seq.: sequential

be trained by alternatively providing it with embeddings from one modality at a time. A hyper-parameter in the loss function can be used to control the significance of each modality for the shared network. Also, a cross-modal triplet loss is applied to the unimodal embeddings in the shared space. The combined loss based training is shown to be useful for developing a test-time stand-alone model made of audio or vision modality. Even though GRUs are used to obtain unimodal embeddings, EmoBed uses a single time-step GRU and does not take contextual information into account. Also, an inherent caveat in this system is that a weaker modality can degrade the performance of a stronger modality.

Heterogeneous Multimodal Transfer Learning (HMTL) [50] uses a cross-modal decoder and discriminator for fully supervised knowledge transfer from text modality to audio/vision modalities for sentiment classification. A cross-modal decoder and discriminator is used for the knowledge transfer. Unimodal networks are made of GRU, dense and self-attention layers followed by dense layers for classification. The cross-modal decoder takes the pre-classifier embeddings of audio/vision modality as input and maps them to the corresponding text modality embeddings. A GAN style adversarial training strategy is used with the audio/vision modality network as generator and a discriminator with input from generator as well as text modality embeddings. The generator tries to output embeddings that are as close as possible to the text embeddings and discriminator tries to label them as true or fake. One drawback of the method is the use of discriminator based adversarial training which demands additional parameters along with added complexities such as oscillations in loss values during training [99].

While the works mentioned so far transfer knowledge from one modality to another, multimodal models can also be used as the source of knowledge. A multimodal acoustic-lexical model can be used as a source or teacher for supervised knowledge transfer to an audio-only model using contrastive loss [22]. The models use words to obtain semantic audio features. The multimodal model, made of bi-directional LSTM, GRU and attention layers, is first trained on audio and text features obtained from word aligned acoustic-lexical data. An audio-only network is trained using a combination of contrastive loss between the multimodal embeddings and unimodal acoustic embeddings and a KL divergence based teacher-student loss.

An un-supervised knowledge transfer method, called deep canonically correlated cross-modal autoencoder (DCC-CAE) [95], uses a combination of correlation based feature alignment and cross-modal translation to develop unimodal audio or vision based sentiment classification model. The encoders are made up of multi-layer perceptrons and do not take contextual information in utterances into account. Un-supervised knowledge transfer from vision to audio modality can also be done using a teacher-student modelling framework [92]. Squeeze and Excitation architecture [73] is used for the vision modality, which serves as the teacher model. It is pre-trained on the VGG-Face2 dataset [100] for speaker identity verification and then fine-tuned on FER-Plus dataset [101] for face emotion recognition by matching with the distribution of annotated labels. The student model, which is tasked with performing emotion recognition from speech, is based on the VGG-M architecture [102]. Respective modalities from an audio-visual dataset, (VoxCeleb [103]) which has not been labelled for emotion recognition, is then given as input to the teacher and student networks. The trained teacher outputs labels which are then used to train the student network.

An unsupervised cross-modal translation based framework, called Seq2SeqSentiment [85], uses an LSTM based encoder-decoder model to create intermediate features that are representative of both modalities. However, the absence of supervision during translation can create representations that might not be discriminative for the task at hand. Hence, an extension of this work, called Multimodal Cyclic Translation Network (MCTN) [93], tries to solve this issue by feeding the intermediate features to a classifier. MCTN can be used to transfer knowledge from auxiliary modalities (audio and vision) to text modality for sentiment analysis tasks. A hierarchical version of the model is used when there are more than two modalities involved. Once the training phase is over, the encoder and classifier of the model is separated and used with text input alone during inference phase to obtain improved text representations for the downstream tasks. Because of the supervised learning setting, MCTN is able to provide improved performance in comparison to its unsupervised predecessor.

A self-supervised audio-visual training system can be used to obtain improved audio-only representations at test-time for multiple downstream tasks like multi-class emotion recognition, continuous emotion recognition and speech recognition [96]. A U-Net [104] architecture fed with a single face frame is used to output video of talking face by infusing the U-Net decoder with the output of an audio encoder. The system is trained using a combination of video reconstruction loss and cross-entropy loss from audio self-supervision. Thus, the audio encoder, with help from visual modality, is driven to produce useful speech features that correlate with mouth and facial movements.

It has also been shown that a multimodal model trained using all modalities when tested with only the strongest modality can perform better than a unimodal model trained and tested on the stronger modality [24]. (Here the notion of strength is based on the individual performance of different modalities on the same downstream task.) This is because individual modalities, with the help of model parameters, are able to distil information from other modalities and perform better on unimodal tasks.

A summary of all co-learning methods described so far is given in Table 2.2.

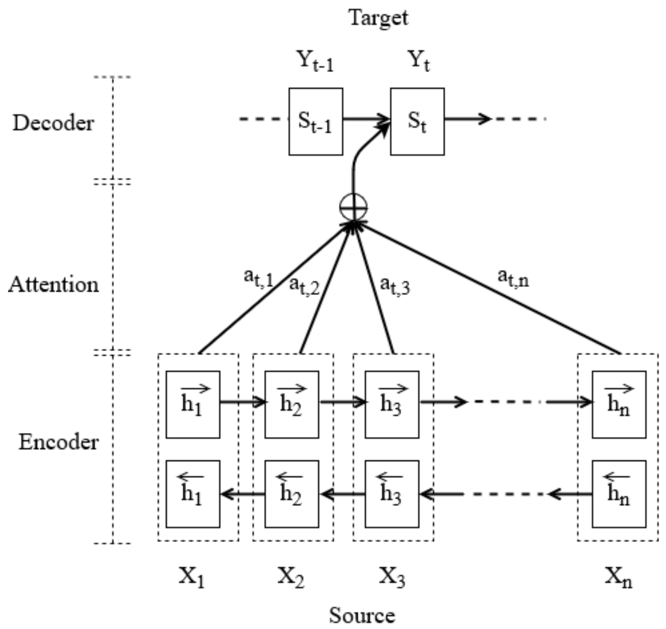


Figure 2.3: Bahdanau attention [105] in an encoder-decoder model. Figure adapted from [105].

2.4 Attention in deep neural networks

Attention mechanism is an important component in many unimodal, multimodal and cross-modal learning models in the literature. In this section, we describe in detail the motivation and methodology of attention mechanisms.

The concept of 'attention' was introduced in the Natural Language Processing (NLP) literature for machine translation tasks [105]. Prior to the introduction of attention, translation from a source language to a target language used sequence-to-sequence (seq2seq) models [106] with an encoder-decoder mechanism. The encoder and decoder are usually made of recurrent networks like LSTM or GRU. The encoder generates a fixed length context vector, which is a compressed summary representation of the whole source language sequence. The decoder is initialized with this context vector to generate the transformed output. A severe limitation of the use of fixed length context vector is its incapability to remember long sentences. The attention mechanism was introduced to solve this problem.

Conceptually, attention in deep networks emulate the human visual attention mechanism that allows us to focus on a certain region with "high resolution" while perceiving the surrounding image in "low resolution", and then adjust the focal point or do the inference accordingly [32]. Rather than building a fixed length context vector out of the encoder's last hidden state, attention creates shortcuts between the context vector and the entire source input. The weights of these shortcut connections are customizable for each output element. The alignment between the source and target is learned and controlled by the context vector. Since the context vector now has access to the entire source sequence via attention, it will not forget long sentences. Fig. 2.3 shows the Bahdanau attention [105] incorporated into an encoder-decoder model. The vectors [X_1, X_2, X_3, ..., X_n] and [Y_1, Y_2, Y_3, ..., Y_m] represent the source and target sentences of lengths n and m respectively. The encoder is a bi-directional RNN with forward and backward states represented by $\overrightarrow{\mathrm{h}_i}$ and $\overleftarrow{\mathrm{h}_i}$ respectively. A simple concatenation of the two represents the encoder state given by h_i = [$\overrightarrow{\mathrm{h}_i}$; $\overleftarrow{\mathrm{h}_i}$]. The decoder hidden state at time-step t is given by

$$S_t = f(S_{t-1}, Y_{t-1}, C_t), \tag{2.1}$$

where the context vector C_t is given by

$$\begin{cases} C_t = \sum_{i=1}^{n} (\mathrm{a}_{t,i} \cdot \mathrm{h}_i), \\ \mathrm{a}_{t,i} = \frac{exp(score(S_{t-1}, \mathrm{h}_i))}{\sum_{j=1}^{n} exp(score(S_{t-1}, \mathrm{h}_j))}, \\ score(S_t, \mathrm{h}_i) = v_a^T \cdot tanh(W_a[S_t; \mathrm{h}_i]) \end{cases} \tag{2.2}$$

The attention mechanism assigns a weight $\mathrm{a}_{t,i}$ to the pair of input at position i and the output at position t based on how well they match. The set of weights $\{\mathrm{a}_{t,i}\}$ define how much of each source hidden state should be considered for each output. These weights are parameterized by a feed-forward network with a single hidden layer and this network is jointly trained with other parts of the model as given by the '*score*' function in eq. 2.2. Both v_a and W_a are weight matrices to be learned.

With the success of Bahdanau attention in machine translation, the concept of attention got extended into other fields like computer vision [107, 108] and audio processing [109, 110, 111]. Over time, different types of attention mechanisms have also emerged where the functions used to compute the alignment scores are different [112, 113]. For example, [112] proposed two types of scoring functions as follows;

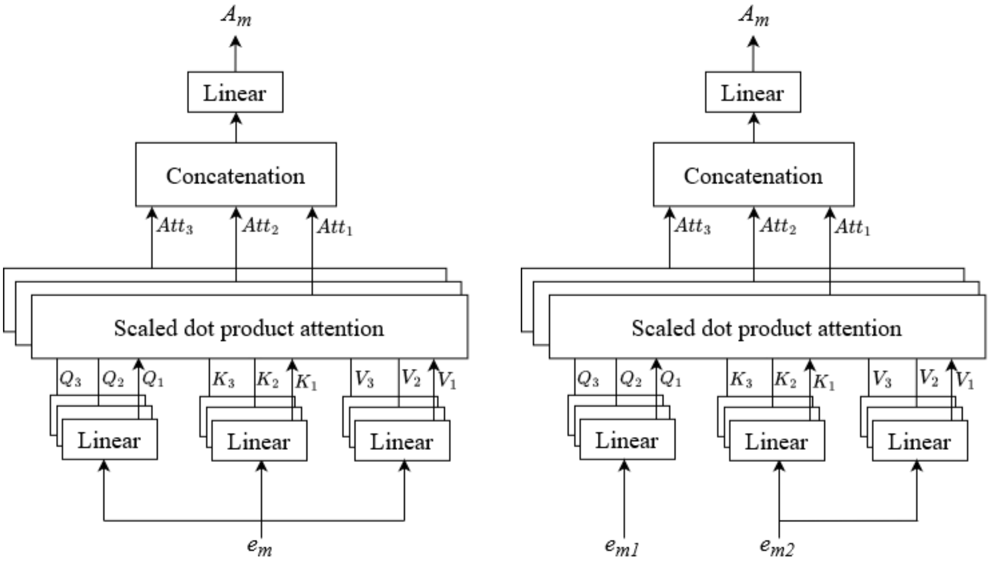
$$\begin{cases} score(S_t, \mathrm{h}_i) = S_t^T \cdot \mathrm{h}_i, & \text{called dot-product attention and} \\ score(S_t, \mathrm{h}_i) = S_t^T \cdot W_a \cdot \mathrm{h}_i, & \text{called general attention} \end{cases} \tag{2.3}$$

where W_a is learnable. Another categorisation of attention can be based on the scope of the input, namely, global (or soft attention) and local (or hard attention). The original attention mechanism [82] is global since it looks at the entire source input, while local attention focuses only on parts of the given input [107, 112]. When the target sequence is replaced with the same input source sequence, the form of attention is called 'self-attention' or 'intra-attention'. In multimodal learning, when attention mechanism is used to find the similarity or alignment between the same sequence belonging to a single modality, it is called 'self-' or 'intra-modal' attention and if it is between two different modality sequences, it is called 'cross-' or 'inter-modal' attention.

2.4.1 Multi-head attention

The Transformer model [113] introduced in 2017 elevated the importance of attention mechanism by proposing an architecture for machine translation tasks that completely relied on a form of dot product attention by eliminating the recurrent layers in seq2seq models. The major component of a transformer model is 'multi-head self-attention' (MHA). Rather than only computing the attention once, the MHA mechanism runs through the scaled dot-product attention multiple times in parallel and hence the name 'multi-head'. Each MHA module performs multiple scaled dot-product attention on three inputs, named as, Query (Q), Key (K) and Value (V). The terminology for Query/Key/Value can be considered analogous to retrieval systems. For example, when we search for videos on any video hosting site on the internet, the search engine will map our query (text in the search bar) against a set of keys (video title, description, etc.) associated with candidate videos in their database, then present the best matched videos (values). In the case of self or intra-modal attention, the Q, K and V are all derived from the same input modality sequence.

Fig. 2.4 shows the intra-modal MHA mechanism with 3 heads in detail. In intra-modal case, the input sequence from a single modality, e_m, is projected into multiple heads or sub-spaces via linear layers. Each sub-space contains a set of Q, K and V values.



(a) Intra-modal attention

(b) Cross-modal attention

Figure 2.4: Multi-Head Attention (MHA) [113] with 3 heads (H=3) for (a) intra-modal and (b) cross-modal cases. Note that the intra-modal model is fed with only one input whereas the cross-modal model is fed with two inputs.

$$\begin{cases} Q_h = W_h^Q e_m, \\ K_h = W_h^K e_m, \\ V_h = W_h^V e_m, \end{cases} \tag{2.4}$$

where h denotes head index, $m \in \{a, v, l\}$ denotes the modality and W refers to learnable weights of the linear layers. On each set of Q,K,V values, a scaled dot-product attention operation is performed in parallel. For a sub-space h, the attention operation is given as,

$$Att_h(Q_h, K_h, V_h) = Softmax(\frac{Q_h {K_h}^T}{\sqrt{d_k}})V_h, \tag{2.5}$$

where Att_h and d_k refer to the attention operation in head h and feature dimensionality, respectively. The outputs of all attentions are concatenated and passed through a linear layer to obtain the final output of an MHA module.

In the cross-modal case, a source modality (e_{m2}) is used to generate K and V, whereas a target modality (e_{m1}) is fed as Q. The intuition behind such an approach is to discover cross-modal interactions by adapting the source modality to the target modality [19]. As an example, let us take the case of audio as target modality and vision as the source modality. Due to different sampling frequencies of both modalities, the sequence lengths of audio and vision representing the same temporal event can be different. Thus, let the audio and vision features with common feature dimensionality d'' be represented by $e_a \in \mathbb{R}^{t'_a \times d''}$ and $e_v \in \mathbb{R}^{t'_v \times d''}$ respectively. The e_a is transformed to Q and e_v to K and V using eq. 2.4. The cross-modal MHA module then maps vision to audio modality and outputs vision features adapted to audio $A_m \in \mathbb{R}^{t'_a \times d''}$. Thus, cross-modal attention re-inforces a source modality by calculating its attention weights using a different target modality. Note that the sequence length of the cross-attention weighted output is the same as the target modality audio. With 3 modalities, we can have 6 different combinations of source-target modalities.

2.5 Baseline methods

Out of all the multi-modal fusion and co-learning methods discussed so far, we provide details about the specific methods that are baselines for our thesis contributions in Table 2.3.

For multi-modal fusion based emotion classification, the state-of-the-art results are obtained by methods that use attention mechanism along with a recurrent network like GRU or LSTM. Nevertheless, none of these methods provide a comparison between the 2 major types of attention mechanisms being used, namely, self and cross-attention, thus leading us to our research question RQ2. In case of the co-learning paradigm, first of all, the number of co-learning methods for affective computing tasks is lesser compared to tasks involving different types of image modalities. Secondly, out of the few methods available, we find that, albeit they achieve co-learning from stronger to weaker modality, each are limited by disadvantages that we intent to avoid or improve upon. Thus via RQ3, we intend to improve upon the results of compared methods, along with eliminating some of the drawbacks associated with them.

2.6 Application Scenarios

In this section, we look at the application scenarios considered in this thesis. Since one important and practical application of unimodal and multimodal representation learning involving heterogeneous multimodal data is human-machine communication, we focus on affective computing applications. In particular, we look at four specific tasks involving vocal prosodic information from speech signals, text representing the spoken words and visual representations from dynamic

Table 2.3: Summary of multimodal baseline methods to this thesis

	Method	Details	Pros	Cons
fusion	MDRE [65]	late fusion of audio and text modality recurrent encoder outputs	[1] simple bi-modal fusion	[1] uses only two of three modalities [2] subpar performance compared to attention based models
	AMH [33]	audio, vision and text modality recurrent encoders output multi-hop attention	[1] tri-modal fusion [2] achieved SOTA on 7-class emotion classification	[1] needs to wait for last time-step output of 2 modalities to compute attention score for the 3rd
	Cross-attn [19]	audio, vision and text cross-modality transformer encoders	[1] tri-modal fusion [2] uses both self and cross-attention	[1] unclear on effectiveness of cross-attention compared to self
	MCSAN [114]	audio and text modality encoders with self and cross-attention fusion	[1] achieved SOTA on bi-modal 7-class emotion classification [2] uses both self and cross-attention	[1] unclear on effectiveness of cross-attention compared to self
co-learning	Seq2SeqSent. [85]	un-supervised encoder-decoder based cross-modal translation	[1] method to generate un-supervised multi-modal representations	[1] subpar performance compared to uni-modal baselines [2] un-supervised representations need not be task-specific
	HMTL [50]	cross-modal decoder & discriminator for co-learning from stronger to weaker modality	[1] successful co-learning from text to audio & vision modalities	[1] complex GAN based mechanism which adds complexities such as loss value oscillations
	Emobed [23]	Joint audio-visual training with cross-modal triplet loss	[1] simple architecture with GRU & dense layers	[1] Method claims to perform co-learning between any two modalities, results indicate failure for weaker to stronger cases

facial images of the speaker, namely, unimodal verbal conflict intensity estimation, multimodal multi-class emotion classification, multimodal continuous emotion recognition and multimodal binary sentiment classification.

The first steps towards automatic processing of emotions in speech occurred by mid-1990s. With Picard's book on affective computing published in 1997 [115] and the International Speech Communication Association (ISCA) Workshop on emotion and speech in 2000 [116], the research community began to recognize the importance and challenges associated with estimating the 'paralinguistics' of speech. The term paralinguistics refers to 'alongside' linguistics and focuses on 'how' you say rather than 'what' you say. 'Computational paralinguistics', which refers to the automatic inference of paralinguistic cues using a computer, was not popularly recognized as a discipline on its own merit by the research community until after 2000. Computational paralinguistics deals with 'traits' and 'states'; traits being long-term events and states short-term. The long-term events include age, gender, personality etc and short-term events include emotions, mood, inter-personal stances etc. The first INTERSPEECH computational paralinguistics challenge [117] was introduced in 2013 and consisted of 4 sub-challenges, namely, the detection of social signals, conflict, emotion and autism from mono or conversational speech. Evidently, it is not only speech that communicate emotion, affect, personality etc, but facial expressions, gestures and body movements/postures as well [3, 118]. In fact, Darwin (1872) [119], who initiated the evolutionary theory of emotions in the late 19th century, considered the face to be the '*chief seat of expression and the source of the voice*', an opinion which is shared by many in the research community and evidenced by the abundance of literature on face based analysis for various affect recognition tasks [120, 121]. Till mid-1990s, recognition of emotions via face modality and speech modality were considered as separate research paradigms and it was in 1997 that the first paper on multimodal emotion recognition was published [122]. Since then there has been a growing interest in using multiple modalities for affect recognition tasks.

In this thesis, we use a sub-set of problems under the umbrella of computational paralinguistics for validation of our proposed speech-based and multimodal models. The details and background of each of these tasks are given as follows:

2.6.1 Unimodal verbal conflict intensity estimation from speech

Verbal conflict is an interaction process between parties who pursue incompatible goals [123]: each party perceives that their interests are being opposed or negatively affected by another party

[124]. While goals and interests are not directly observable, they influence human behaviour through gestures, facial expressions and speech [125]. Inter-personal conflicts are found to not only negatively affect the lives of involved parties to a significant extent [126] but also cause long-term negative effects on the rapport between them. Conflicts can span from minor disagreements to physical assaults and can become one of the most concerning causes of stress [127]. Thus, detection and monitoring of such inter-personal conflicts is a desired ability for socially intelligent technologies that are expected to understand and seamlessly integrate human interactions [128]. In particular, the automatic estimation of conflict from speech signals has several important applications, such as monitoring conflicts during meetings and in call centers to help employees handle difficult interactions and thereby reduce stress and anxiety.

The problem of verbal conflict intensity estimation from speech has been popularised by the 2013 INTERSPEECH computational paralinguistics challenge [117]. It can be formulated as a detection or estimation problem. *Conflict detection* aims to identify if a given temporal interval of speech contains an instance of verbal conflict [8, 11, 12]. *Conflict intensity estimation* is a regression task that aims to determine a continuous level of conflict intensity [7][129], which is more informative than the binary class label generated by conflict detection methods [7]. Most of the prior methods relied on the baseline features provided in the INTERSPEECH challenge, which are 6,373 acoustic features extracted using OpenSMILE [130]. Relevance of these baseline features can be determined by repeated classification using random feature subset selection [9], canonical correlation analysis based discriminative projection [8], greedy forward-backward feature selection [6] or ensemble Nyström method on manually partitioned feature subsets [129]. A major drawback of these methods is that they require extra techniques to filter out redundant features and identify conflict-specific features. For example, [9] performs 300,000 iterations to identify 349 conflict specific features out of the 6,373 baseline features.

A Support Vector Machine (SVM) classifier can be used for conflict detection using predicted speech overlap ratio [12] or speech overlap based features [11]. Speech overlap predictions generated by a bi-directional LSTM can also be used for conflict detection using a DNN classifier [10]. Utterance-level features, obtained by combining frame-level DNN predicted speech overlap posteriors along with a subset of the baseline features, can be used for conflict intensity estimation using Support Vector Regressors (SVR) [7]. These methods require the availability of metadata, like the number of speakers and speech overlap duration. To our knowledge, there is

Table 2.4: Summary of features, refinement methods and classifiers/regressors.

Ref.	Input	Feature Refinement Method	Class/Reg
[9]	IS13	relevance adjustment by rep. class.	KNN
[8]	IS13	canonical correlation analysis	SVM
[6]	IS13	forward-backward pass	SVR
[129]	IS13	manual feature partitioning + ensemble Nyström	ensemble SPLSR
[12]	IS13	speech overlap ratio using SVR	SVM
[10]	conv. & pros.	speech overlap ratio using BLSTM	DNN
[7]	IS13 & over.	forward-backward pass	SVR
[11]	IS10 & IS13	overlap detection using SVR + backward selection	SVM
[131]	FPF & LLD	LSTM based encoder-decoder network	
[132]	raw speech	End-to-End Convolutional Neural Network	

KEY - IS13: INTERSPEECH 2013 Conflict sub-challenge baseline features; IS10: INTERSPEECH 2010 paralinguistics challenge baseline features; rep. class.: repeated classification; conv.: conversational features; pros.: prosodic features; over.: overlap features; Class/Reg: Classifier/Regressor; KNN: K Nearest Neighbour; SVM: Support Vector Machine; SVR: Support Vector Regressor; LSTM: Long Short Term Memory; BLSTM: Bi-directional LSTM; SPLSR: Sparse Partial Least Squares Regression; FPF: Facial Point Features; LLD: Low Level Descriptors; CRNN: Convolutional Recurrent Neural Network

only one multimodal conflict estimation method and it uses a concatenation of audio and visual features as input to an LSTM-based encoder-decoder architecture with attention. This method focuses on visual features (facial gestures) and uses 65 audio Low-Level Descriptors (LLD) features, sampled at 25 Hz [131]. The key methods are summarised in Table 2.4.

2.6.2 Multi-class multimodal emotion classification

Since speech is the most natural means of communication between humans, researchers are motivated to use it as an efficient medium for human-machine interaction. While the initial focus of researchers on human-machine interaction via speech signals was on the speech recognition task that started in the late 1950s, later they realised that for having a more *natural* interaction, machines have to understand the emotional state of the speaker. Thus began the interest in the computational paralinguistics task of Speech Emotion Recognition (SER), which is defined as extracting the emotional state of a speaker from his or her speech. Applications of SER include interfacing with robots, audio surveillance, web-based E-learning, commercial applications, clinical studies, entertainment, banking, call centers, computer games, etc [133, 134, 135].

An important aspect of SER is the need to decide a set of emotion categories to be classified by an SER system. The emotions that are most distinct and commonly occurring in our lives are called *archetypal* emotions and include the categories of *angry*, *happy/joy*, *sad*, *fearful*, *surprised*, *disgust* and *neutral* [136, 137]. However, different researchers create different sets of categories depending upon the use cases and the relative relevance of different emotions in their use cases. For example, SER can be formulated as a binary classification problem to recognize negative/non-negative [138, 139], angry/neutral [140] or fearful/neutral [141] emotions for call center monitoring applications. In most general applications of SER, however, researchers mostly focus on multi-class emotion classification, where the number of classes can vary from four [65, 110, 142, 143, 144] to seven [33, 145] with additional classes being *frustration*, *excitement* and *other* to accommodate for all other emotions not included.

To design a successful SER system, we need to take into account the following [146];

- the choice of an appropriate database that captures different emotions
- the type of features to be extracted from speech and
- the design of a reliable machine learning algorithm

A classical SER system consists of two stages [133]:

- a feature extraction unit that extracts the task relevant features from the input speech data
- a classifier that maps the output of feature extraction unit to the decision space

Over the years, various types of classifiers have been used for SER like HMM [147, 148], GMM [149, 150], SVM [151], artificial neural networks (ANN) [1, 152], KNN [153] and many others, with each having its own advantages and limitations. An ensemble of classifiers [154] can also be created to derive the merits of different types of classifiers.

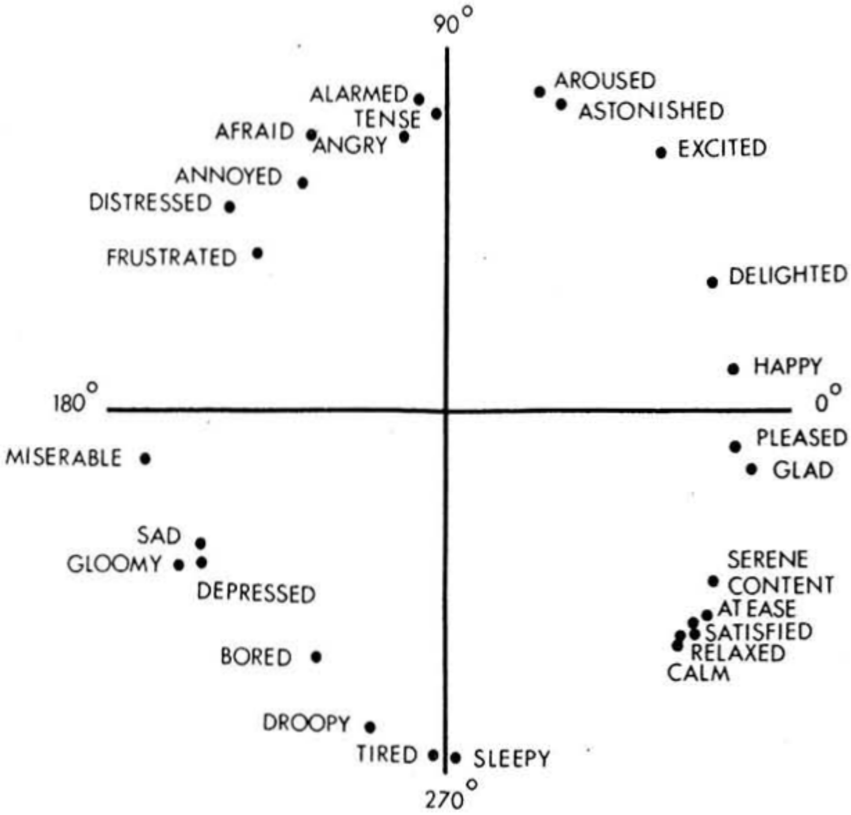
Apart from speech, other modalities such as visual modality (facial expressions, gestures, postures, gait) [155, 156], language modality (text transcripts) [157, 158] and physiological modality (brain or muscle electrical activity, temperature, skin conductance, cardiac function) [159, 160] have also been explored for emotion recognition. Due to the requirement for contact based or invasive sensors, we do not consider physiological modality within the scope of this thesis. We focus on a combination of speech (paralinguistics), visual (facial expressions) and text (speech transcripts) modalities for the multimodal fusion based multi-class emotion classification problem. As mentioned in section 2.2, there are various creative multimodal fusion techniques that can be applied for emotion classification.

2.6.3 Multimodal continuous emotion recognition

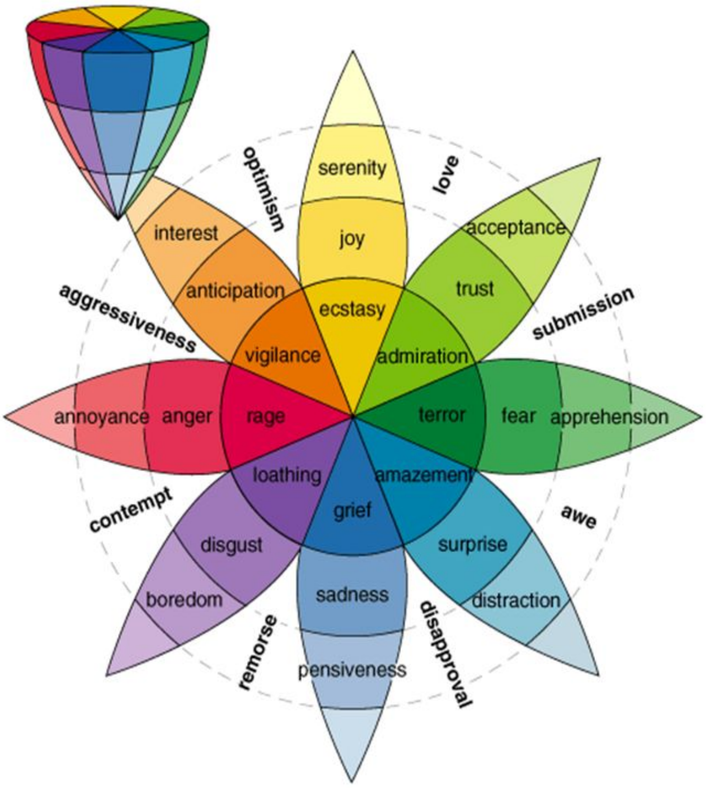
Apart from discrete emotion classification, another popular line of research is emotion recognition when the label space is continuous. This is based on the studies in psychology that represents emotions as coordinates in a multi-dimensional space [161]. The circumplex model of emotion posits that different emotional states are processed and represented as points in an emotional space, along the dimensions of valence and arousal [162]. There exist different types of circumplex models for emotion mappings. Russell's circumplex model uses the dimensions of arousal and valence to plot 28 affective labels [163], while Whissell considers emotions as a continuous 2D space whose dimensions are evaluation (or valence) and activation (or arousal), where the evaluation dimension measures how a human feels, from positive to negative and the activation dimension measures whether humans are more or less likely to take some action under the emotional state, from active to passive [164]. Plutchik's wheel of emotions (Figure. 2.5.b) is another 2D model of emotions, which consists of 8 basic emotions and 8 advanced emotions each composed of 2 basic ones, the vertical dimension represents intensity and the radial dimension represents degrees of similarity among emotions [165]. Besides the 2D approaches, a commonly used 3D emotion representation framework is the (valence, arousal, dominance) set, which is known in the literature by different names, including (evaluation, activation, power) and (pleasure, arousal, dominance) [166].

2.6.4 Multimodal binary sentiment classification

The concepts of emotion recognition and sentiment recognition sit under the broad umbrella of affect recognition and it is important to understand the differences between them. Even though both emotions and sentiments refer to "experiences that result from the combined influences of the biological, the cognitive, and the social" [167], sentiments can be differentiated from emotions by the duration in which they are experienced [168]. Emotions are brief episodes of brain, autonomic, and behavioral changes [169], sentiments have been found to form and be held for a longer period. Furthermore, sentiments are formed and directed toward an object, whereas emotions are not always targeted towards an object. In this context, an object refers to a person,



(a) Russel's model



(b) Plutchik's model

Figure 2.5: Circumplex models of emotions [163, 165]

a thing, a condition, a place or an event at which a mental state is directed [170].

In affective computing literature, compared to emotion recognition, sentiment recognition is a more coarse grained task since it is usually considered as a binary (positive or negative) or ternary (positive, negative or neutral) classification task [3]. While earlier works on sentiment classification focused on text modality alone, later works added visual and acoustic modalities and formulated the task as a multimodal problem [171, 172]. To the best of our knowledge, the very first work on multimodal learning for sentiment analysis task that used three modalities (audio, text and visual) came only in 2011 [172]. Even though, by this time there was a significant amount of work done on audio-visual emotion recognition, it is interesting to note that text modality and audio-visual modalities were scarcely considered for emotion recognition and sentiment analysis tasks respectively.

In this thesis, we use the verbal conflict intensity estimation task to verify our first research paradigm, namely, unimodal training and unimodal testing. For our next research paradigm, multimodal training and multimodal testing, we use multi-class multimodal emotion classification task. For the third and final research paradigm of multimodal training and unimodal testing, we use multimodal continuous emotion recognition and multimodal binary sentiment classification tasks.

2.7 Datasets, Features and Metrics

In this sub-section, we explain about the datasets used for various tasks mentioned previously. Specifically, we use *SSPNet Conflict Corpus* [128] for verbal conflict intensity estimation, *IEMOCAP* [46] for multi-class multimodal emotion classification, *RECOLA* [173] for multimodal continuous emotion recognition and *CMU-MOSI* [174] for multimodal binary sentiment classification.

2.7.1 SSPNet Conflict Corpus

SSPNet Conflict Corpus [128] is a subset of *Canal 9* [175], an audio-visual database of political debates televised in Switzerland during 2005. The Canal9 debates were segmented into uniform, non-overlapping windows of 30 seconds and only the segments portraying at least two persons were retained. Compared to shorter windows or analysis units, 30 seconds long segments are less ambiguous and, therefore, the annotations are more likely to converge. The result is a collection

Table 2.5: Number of 30 seconds duration audio clips in the Train-Val-Test Split [117] for the SSPNet Conflict Corpus [128]

	Train	**Val**	**Test**	**Total**
Low (conflict<0)	471	127	226	824
High (conflict≥0)	322	113	171	606
Total	793	240	397	1430

of 1430 clips - the SSPNet Conflict Corpus - showing 138 subjects for a total length of 11 hours and 55 minutes. Each clip is rated by 10 different non-French speaking assessors and the conflict intensity value assigned to each clip is the average of individual scores [176]. These values are in the range [-10,10], from no conflict (-10) to high level of conflict (+10), thus making the dataset suitable for regression tasks. Audio signals are sampled at 48KHz, resulting in 1,440,000 samples per clip. This dataset was adopted in the *conflict sub-challenge* of the INTERSPEECH 2013 Computational Paralinguistics Challenge [117] using only the audio signal. The training-validation-testing data split as defined in the challenge is shown in Table 2.5. All clips with the female moderator (speaker #50) were assigned to the training set. The development set consists of all broadcasts moderated by a male (speaker #153), and the test set comprises the rest (male moderators).

2.7.2 Interactive Emotional Dyadic Motion Capture (IEMOCAP)

Interactive Emotional Dyadic Motion Capture (IEMOCAP) [46] is a multimodal dataset which contains approximately 12 hours of audio-visual dyadic emotional interactions in acted and spontaneous settings. The dataset, recorded with 5 male and 5 female speakers, includes the ground-truth text transcripts. The labelling of each utterance was determined by majority voting from 3 annotators. Emotional labels present in the dataset are anger, happiness, excitement, sadness, frustration, fear, surprise, disgust, other and neutral. Fleiss' Kappa (k) statistic [177] was used to measure agreement between annotators and was found to be $k = 0.48$, indicating moderate agreement. There is lack of consensus amongst researchers on the use of IEMOCAP dataset. Some use it for 4 class classification [19] by merging different classes (*happy* and *excited*, *angry* and *frustrated*), while others [20, 65, 33, 142] perform 7-class classification. Class sizes smaller than 100 utterances (*fear*, *disgust*, *other*) are usually eliminated [33]. The final dataset contains 7,487 utterances in total (1,103 *angry*, 1,041 *excited*, 595 *happy*, 1,084 *sad*, 1,849 *frustrated*, 107 *surprise* and 1,708 *neutral*).

For IEMOCAP audio modality, 40D MFCC features (frame size is set to 25 ms at a rate of 10 ms with the Hamming window) are extracted and concatenated with their first and second order derivatives to obtain the final acoustic feature dimension of 120. These features are then standardised by removing the mean and scaling to unit variance. For vision data, cropped face images of speakers are fed into a ResNet-101 [178] to obtain 2048D features at a frame rate of 3 Hz. For text modality, each word in an utterance is represented by a 300D GloVe [179] embedding. Note that the modalities are sampled at different rates and the maximum sequence length of audio, vision and text modalities is set to 1,000, 32 and 128 respectively.

2.7.3 Remote Collaborative and Affective Interactions (RECOLA)

Remote Collaborative and Affective Interactions (RECOLA) [173] is an audiovisual dimensional emotion recognition dataset and has been used in multiple Audio Visual Emotion Challenges (AVEC) over the years [49, 180]. It contains audiovisual recordings of spontaneous and natural interactions from 27 French-speaking participants in order to investigate socio-affective behaviours in the context of remote collaborative tasks. Moreover, time and value continuous dimensional emotion annotations (in the range [-1,1]) in terms of arousal and valence are given with a constant frame rate of 40 ms for the first five minutes of each recording, by averaging all six annotators and meanwhile taking the interevaluator agreement into consideration. The interevaluator agreement, measured on the basis of Cronbach's α [181], show good ($\alpha > 0.8$) and acceptable ($\alpha > 0.7$) agreement for arousal and valence annotations respectively. The dataset is further equally divided into three disjoint parts, by balancing the gender, age, and mother tongue of the participants. Therefore, each part consists of nine unique recordings, resulting in 67.5 k segments in total for each part (training, development, or test). Data from first 9 speakers comprise the training set, the next 9 speakers comprise the development set. Only the training and development sets are made publicly available and hence we use these sets primarily for obtaining our results. The last 9 speakers' data is the test-set or the held out evaluation set. Only the features are publicly available and the annotations are privately held by the dataset creators. We sent our final model's predictions on this set and obtain the results from the evaluation done by the dataset creators.

The audio and vision features are provided by the AVEC 2016 and 2018 baselines [49, 180]. These are 88-D extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [182] features extracted using openSMILE, LGBP-TOP based 168-D video-appearance features and 49 facial landmarks based 632-D video-geometric features. The arithmetic mean and the standard derivation of both audio and vision features were computed over the sequential handcrafted features of each frame using a sliding window of 8 s with a step size of 40 ms.

2.7.4 CMU Multimodal Corpus of Sentiment Intensity (CMU-MOSI)

CMU Multimodal Corpus of Sentiment Intensity (CMU-MOSI) [51], is a collection of 2199 opinion video clips from 93 YouTube movie review videos in English language. Each video inherently contains three modalities: language in the form of spoken text, vision via perceived gestures and facial expressions, and acoustics through intonations and prosody. The videos are limited to setups where the speaker's attention is exclusively towards the camera and to have manual and properly punctuated transcriptions provided by the uploader. There are 89 distinct speakers (41 females, 48 males). The training, validation and test sets have 52 (1151), 10 (296) and 31 (752) videos (utterances), respectively. The videos are segmented into utterances with each utterance's sentiment label scored between +3 (strong positive) to -3 (strong negative) by 5 annotators. The inter annotator agreement was 0.77 in terms of Krippendorf's Alpha [183]. The average of these five annotations is taken as the sentiment polarity to create two classes (positive and negative) [184, 185, 186].

For CMU-MOSI, the audio, vision and language features are provided by the creators of the dataset [51]. A CNN is used for textual feature extraction, which takes utterances represented as a matrix of Google word2vec [187] vectors. The CNN has two convolutional layers: the first layer has two kernels of size 3 and 4, with 50 feature maps each and the second layer has a kernel of size 2 with 100 feature maps. The convolution layers are interleaved with max-pooling layers of window 2×2. This is followed by a fully connected layer of size 500 and softmax output. ReLU is used as the activation function. The activation values of the fully-connected layer are taken as the features of utterances for text modality. Audio features are extracted with 30 Hz frame-rate and a sliding window of 100 ms using openSMILE [130] toolkit. The features extracted consist of several low-level descriptors, e.g., voice intensity, pitch, and their statistics, e.g., mean, root quadratic mean. A 3D CNN [188] is applied on video clips to obtain visual features. The features from last convolution layer are passed through max-pooling operation to remove irrelevant features. This is followed by a fully-connected layer of size 100. The dimensions of textual, visual and acoustic features thus obtained are 100, 100 and 73 respectively.

2.7.5 Metrics and loss functions

For regression tasks that use continuous labels, correlation based metrics and loss functions can be used. An example is Pearson Correlation Coefficient (PCC) based loss/metric function. PCC is given as

$$PCC = \frac{1}{N\sigma\hat{\sigma}} \sum_{i=1}^{N} (y_i - \mu)(\hat{y}_i - \hat{\mu}), \tag{2.6}$$

where N is the number of labels; y_i and $\hat{y}_i$ are true and predicted labels, respectively; and (μ,σ) and $(\hat{\mu},\hat{\sigma})$ are their corresponding mean and standard deviation pairs.

Then the loss function using PCC can be formulated as

$$L = 1 - PCC, \tag{2.7}$$

Similar to PCC, another correlation based metric popularly used for regression tasks in AVEC challenges [49, 180] is Concordance Correlation Coefficient (CCC), given as,

$$CCC = \frac{2\sigma_{xy}^2}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}, \tag{2.8}$$

where x and y are the true and the predicted labels, respectively, and μ_x, μ_y, σ_x, σ_y and σ_{xy} refer to their means, variances and covariance, respectively.

Apart from correlation based metrics, Mean Absolute Error (MAE) or Mean Squared Error (MSE) are also widely used for regression tasks. For classification tasks, categorical cross-entropy loss function is used.

$$L_{cross-entropy}(y,\hat{y}) = -\sum_{i=1}^{C} y_i log(\hat{y}_i) \tag{2.9}$$

where y and $\hat{y}$ are the true and predicted posterior class probabilities, C is the total number of classes.

The evaluation metrics for classification tasks are Un-Weighted Accuracy (UWA) and Weighted

Accuracy (WA), given as,

$$
\begin{aligned}
UWA &= \frac{1}{N}\sum_{i=1}^{N} 1(\hat{y}_i = y_i), \\
WA &= \frac{1}{C}\sum_{c=1}^{C}(\frac{1}{N_c}\sum_{i=1}^{N} 1(\hat{y}_i = y_i = c)),
\end{aligned}
\tag{2.10}
$$

where N is the total number of samples; y_i and $\hat{y}_i$ are true and predicted labels, $1(\mathrm{x})$ is the indicator function, C is the total number of classes and N_c is the total number of samples belonging to class c.

2.8 Summary

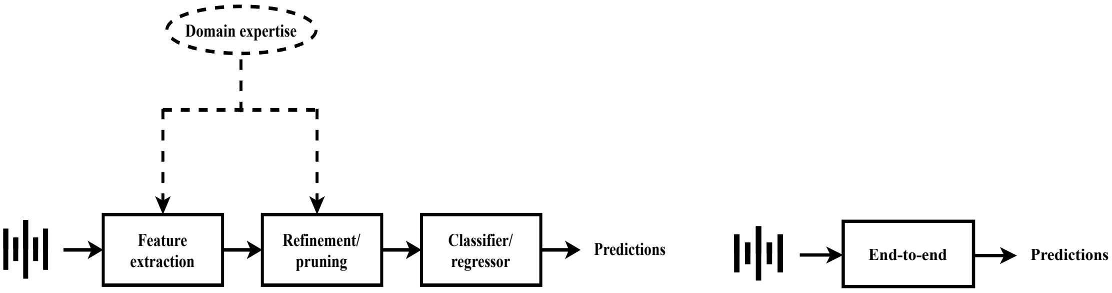
In this chapter, we provide the background literature relevant for this thesis. Specifically, we provide a holistic overview of multimodal fusion techniques involving audio, vision and language modalities. We categorize these methods according to their characteristics and provide a comparison based on their advantages and disadvantages. We also provide a detailed discussion on multimodal co-learning techniques applied to audio, vision and language modalities. We then describe the four tasks involving the three modalities, namely, unimodal verbal conflict intensity estimation, multi-class multimodal emotion classification, multimodal continuous emotion recognition and multimodal binary sentiment classification. Furthermore, we describe the datasets, the features as well as the loss functions and metrics used for validation in this thesis.

Chapter 3

Single Modality Modelling for Computational Paralinguistics

3.1 Introduction

In this chapter, we discuss details about the design of an end-to-end DNN for a speech based computational paralinguistic task, namely, verbal conflict intensity estimation. Traditional speech-based conflict detection and conflict intensity estimation methods use off-the-shelf classifiers or regressors on generic hand-crafted acoustic features, which require further manual refinement and time consuming feature pruning [6, 7, 8, 9]. Task-specific hypotheses and metadata, like the number of speakers and the ratio of their speech overlaps, may also be needed to extract conflict-specific features from standard acoustic features [10, 11, 12]. Another drawback of these methods is the need for separate training and parameter tuning of the feature extractor and the classifier/regressor. An alternative approach is end-to-end learning, which trains models directly from raw input data: since the parameters are trained jointly, the end-to-end model learns task-specific features directly from the input signal, without requiring any guidance other than the objective function and the training dataset (see Figure. 3.1). While hand-crafted features may facilitate interpretation of specific characteristics of the speech signal that are used as predictors for the task at hand, it is worthwhile to explore if an end-to-end learning framework can be used for a complex paralinguistic task such as verbal conflict intensity estimation by automatically learning relevant acoustic features for this task. We aimed to explore this gap in literature and this resulted in the design of an end-to-end deep neural network for the task of verbal conflict



(a) Classical model

(b) End-to-end model

Figure 3.1: Traditional v/s end-to-end speech based detection/estimation models

intensity estimation.

3.2 End-to-End Model Design

According to our RQ1, the aim was to design an end-to-end network, *ConflictNET*, after considering the following aspects about the input data:

(a) Features relevant for the task have to be automatically extracted from the raw speech waveform.

(b) The nature of temporal evolution of speech should be taken into account.

(c) Instances of verbal conflict are unevenly spread across the entire duration of speech signal.

We use the following rationale to design our end-to-end network.

Since the input is a 1D temporal raw signal, in order to extract features (as required by item (a) from above), we could use a series of temporal convolutional layers. Each 1D convolutional layer is composed of multiple learnable filters of specific (chosen) dimensionality. A progressive increase in the number of filters as well as decrease in filter size after the first convolutional layer could be employed due to the fact that, with increased depth, the network learns more detailed features. In order to take into the account the fact that speech is a temporally evolving signal across which task-relevant cues are spread (items (b) and (c) from above), we need to exploit the sequential nature by means of a recurrent network like LSTM or GRU. Thus, the output of the last convolution layer should be provided as input to one or more recurrent layers for further processing. Next, while hearing the audio samples, we can find that only some temporal instances in each signal contain task-related cues with varying degrees of relevance. Some temporal instances have zero contribution towards the task while some others might contribute more. In order for the

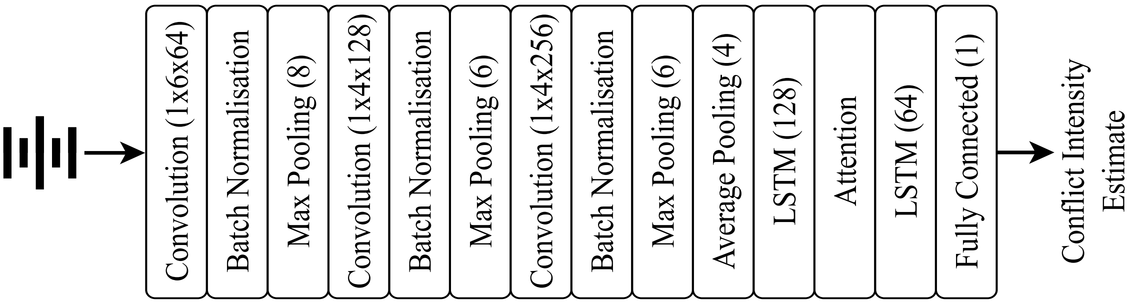


Figure 3.2: The proposed ConflictNET architecture for conflict intensity estimation.

network to provide a task relevance weightage to temporal instances, we could use an attention layer. Finally, to provide a mapping the final feature space to the label space, we could use a fully connected layer.

A more detailed explanation of the architecture is as follows:

3.2.1 Convolutional-recurrent network with attention

ConflictNET contains six types of layers arranged as a single stream and combines feature extraction and regression in a unified framework. Features from the speech signal are extracted by 1D *convolutional* layers with learnable filters. There are 3 1D strided convolutional layers, with 64, 128 and 256 filters respectively. Each convolutional layer uses ReLU activation. 1D filters of successive convolutional layers, each with stride 1, are of sizes 6, 4 and 4 respectively. Changes in the parameters of network layers during training modify the distribution of the input to their subsequent layers, a phenomenon known as internal covariate shift [189]. To reduce the effect of this phenomenon and thereby accelerate the training, we perform *batch-normalization* after each convolutional layer. Successive *max-pooling* layers downsample the convolution outputs and reduce the number of network parameters. The pooling size is determined by considering the rate of overlap, R, between convolution filter size, F, and pooling size, P [14]:

$$R = \frac{F-1}{F+P-1}. \tag{3.1}$$

We keep $R < 0.4$ and use a stride size equal to pool size in all the max pooling layers. Even though the common choice to model temporal sequential data like speech is to use a Recurrent Neural Network (RNN), vanilla RNNs are hard to train due to the vanishing gradient problem [190], which can be attenuated using *Long-Short-Term-Memory* (LSTM). Thus, we use two Tanh activated LSTM layers, with 128 and 64 units respectively, to capture the interdependencies between features across time. Although, theoretically, there is no limitation on the number of time-steps an LSTM can process, our experiments showed that restricting the number of time steps to fewer than 250 improves performance. Thus, we use a *temporal average pooling* layer of pool size 4 to reduce the number of input time-steps to the first LSTM layer. Since, not all portions of an input speech signal will contribute equally towards the conflict intensity estimate of the entire signal, to enable the network to focus on portions of the signal that are more relevant for conflict intensity estimation, we use an *attention mechanism* between the LSTM layers. The LSTM layer with 128 units provides a sequence output rather than a single value to the attention layer, which assigns different weights to hidden states across different time-steps. We use a global additive self-attention mechanism [108], which considers the whole context to calculate relevance:

$$\begin{cases} g(t,t^{'}) = \tanh(W_g h_t + W_{g^{'}} h_{t^{'}} + b_g), \\ e(t,t^{'}) = \tanh(W_a g(t,t^{'}) + b_a), \\ a(t) = \mathrm{softmax}(e(t)), \\ l_t = \sum_{t'} a(t,t^{'}) h_t^{'}, \end{cases} \tag{3.2}$$

where W_g and $W_{g^{'}}$ are weight matrices corresponding to hidden states h_t and $h_{t^{'}}$ respectively; W_a is the weight matrix corresponding to their non-linear combination; b_g and b_a are the bias vectors; $a(t,t^{'})$ captures the similarity between h_t and $h_{t^{'}}$; l_t represents the attention focused hidden state representation, which is then given as input to the second LSTM at time-step t. A final *fully connected* layer, with a linearly activated single output neuron connected to the final time-step of last LSTM layer, provides the continuous conflict intensity value. The resulting *ConflictNET* model has 420,418 trainable and 896 non-trainable parameters.

3.2.2 Pearson correlation-based loss function

In line with the previous works in the literature [6, 8], we use Pearson Correlation Coefficient (PCC) as the performance evaluation metric and the loss function was designed to maximise this metric as given by:

$$L = 1 - PCC = 1 - \frac{1}{N\sigma\hat{\sigma}} \sum_{i=1}^{N} (y_i - \mu)(\hat{y}_i - \hat{\mu}), \tag{3.3}$$

where N is the number of labels; y_i and $\hat{y}_i$ are true and predicted labels, respectively; and (μ,σ) and $(\hat{\mu},\hat{\sigma})$ are their corresponding mean and standard deviation.

3.2.3 Training

The model was developed, trained and tested using Keras with Tensorflow backend [191]. The model was trained using the training set and the validation set was used to identify the epoch for early stopping and model saving callbacks. We used the Adam optimizer [192] with a learning rate of 0.01 and decay of 0.6 for training the network with mini-batches of size 32. The model was selected based on the highest PCC value on the validation set. We follow the same training-validation-testing data split as defined in the challenge. Note that the challenge considered a binary classification task, obtained by classifying the conflict level into high ($\geq$0) or low (<0). Also, we convert the target labels from the range [-10,10] to [-1,1] for compatibility with the activations of the neural network. Since the input to the network is raw speech waveform, due to memory considerations, we downsample the speech signals to 8KHz. Thus a 30s duration input signal will have 240000 samples. The downsampling operation loses information above 4KHz which is perceptually significant but assumed to contain little information relevant to conflict recognition. To normalize the energy level across the entire input signal S, we perform root-mean-square normalization as follows:

$$s = \frac{S}{\sqrt{\frac{\sum_{i=1}^{M} |S_i|^2}{M}}}, \tag{3.4}$$

where S_i is the i^{th} sample, M is the total number of samples of the input signal and s is the normalized signal.

3.3 Results and Ablations

In our experiments, we focus on improving the PCC metric, in line with the related previous works. We also measure the UAR and WAR after binarising the predicted continuous labels into high and low conflict levels. We map the predicted output values to the same range as the training labels before calculating UAR and WAR, which helps to improve these evaluation measures without changes in the PCC value. The results we report are average values obtained after training and testing the model for 10 times.

We compared the performance of *ConflictNET* with previous works in literature as well as with a baseline convolutional recurrent network, which we call *ParaNET*. *ParaNET* is composed of the three sets of convolution, max pooling and batch normalization layers as well as two LSTM layers and a fully connected layer from the *ConflictNET*. The following ablation study observations were made using the SSPNet corpus [128].

The performance of our baseline model *ParaNET* is much better than the expected measure by chance values (PCC = -0.008 ± 0.023, UAR = 50%) given in [117]. An average pooling operation at the input of the first LSTM layer improves the performance on all the 3 evaluation measures, which can be attributed to the better performance of the LSTM obtained by reducing the number of input time-steps. An attention layer added to *ParaNET* improves its performance by a noticeable margin of 0.162, 9.8% and 8.1% in PCC, UAR and WAR, respectively. This supports our intuition that weighted combinations of hidden states across multiple time-steps can result in performance improvement of the LSTM layers. Further, adding both average pooling and attention layers to *ParaNET* improves the PCC value to 0.853 ± 0.003. We also experimented by using a Global Average Pooling layer that took a temporal average over the entire output sequence of the second LSTM layer before feeding it to the fully connected layer. However, adding this layer resulted in a slight decrease of 0.002 in PCC and a slight improvement of 0.2% and 0.5% in UAR and WAR values, respectively. It is worthwhile to note that the standard deviation of UAR and WAR values (0.43% and 0.51%, respectively) are higher than that of PCC. This is not surprising since we optimized our network in terms of PCC alone.

The comparison[1] in Table 3.1 shows that the performance of *ParaNET*+AP is similar to that of the end-to-end solution in [132]. Our best performing model *ConflictNET* outperforms in terms of PCC all but one method ([7]). *ConflictNET* achieves almost the same performance as [7] , a model with DNN based speech overlap feature set and feature pruning based conflict specific subset of standard acoustic features. This suggests that our end-to-end architecture has automatically learned task-specific information from the raw speech input.

[1] As the results of [131] on the SC2 are not available, this method in not included in the comparison.

Table 3.1: Performance comparisons on the SSPNet Conflict Corpus test set. Note that the range of PCC is [-1,1], and that of UAR and WAR are in percentage.

Ref	Method	PCC	UAR	WAR
[117]	INTERSPEECH'13 baseline	.826*	80.8*	-
[9]	Random subset feature selection	.826	81.6	82.1
[8]	Random discriminative projection	-	84.6*	-
[10]	Deep hierarchical neural networks	.838*	84.3*	-
[6]	Greedy forward-backward	.842*	85.6*	-
[129]	Ensemble Nyström method	.849*	-	-
[12]	Detection using speaker overlap	-	83.1	-
[11]	Speech interruption detection	-	85.3	-
[7]	DNN-based feature extraction	.856	84.7	-
[132]	End-to-End Convolutional NN	.779	79.8	-
	ParaNET	.675	72.4	75.3
	ParaNET + AP	.781	79.9	81.3
	ParaNET + Attn	.837	82.2	83.4
	ParaNET + AP + Attn + GAP	.850	84.5	84.8
	ConflictNET: *ParaNET* + AP + Attn	.853	84.3	84.3

KEY - '*' results reported by training on both training and validation sets; '-' values not reported; Ref: Reference; PCC: Pearson Correlation Coefficient; WAR: Weighted Average Recall; UAR: Unweighted Average Recall; NN: Neural Network; DNN: Deep Neural Network; AP: Average Pooling; Attn: Attention; GAP: Global Average Pooling

3.4 ConflictNET: Model Analysis

Since *ConflictNET* is an end-to-end architecture, it is not straightforward to understand what cues from the input signal are being used by the network to generate predictions. To understand what instances of input speech are being used to predict a conflict intensity estimate, we performed the following steps. (a) Manual analysis by creating a ground truth versus predicted labels graph and (b) Interpretation using Local Interpretable Model-agnostic Explanation (LIME) algorithm [31].

3.4.1 Manual analysis

We created a Graphical User Interface (GUI) in Python [193], where each speech sample is represented as a circle and upon clicking on a circle, we can listen to the sample. The speech samples in SSPNet corpus [128] were given as input to *ConflictNET* and the predictions were obtained. Since the predictions are in the range [-1 1], they were scaled up by a factor of 100 for ease of visualisation of the circles. Figure 3.3 shows the GUI graph. The X axis represents ground-truth/actual labels ($y \in \mathbb{R}^{N\times 1}$) and Y axis represents predicted labels ($\hat{y} \in \mathbb{R}^{N\times 1}$). The quadrants shown in yellow indicate samples where the polarity of actual and predicted labels are opposite. The straight line through origin is drawn as reference to understand which samples have predicted values that are exactly equal to their actual labels. The two other straight lines are shown to indicate a margin of error $\varepsilon = 0.5$ (50 after scaling). We categorised the entire samples into 4 cases for analysis purposes.

- Samples for which actual and predicted values have same polarity and are within a margin of error, i.e., $(y > 0 \;\&\; \hat{y} > 0)$ or $(y < 0 \;\&\; \hat{y} < 0) \;\&\; |y - \hat{y}| \leq \varepsilon$
- Samples for which actual value has positive polarity and predicted value has negative polarity, i.e., $y > 0 \;\&\; \hat{y} < 0$
- Samples for which actual value has negative polarity and predicted value has positive polarity, i.e., $y < 0 \;\&\; \hat{y} > 0$
- Samples for which actual and predicted values have same polarity but differ by more than the margin of error, i.e., $(y > 0 \;\&\; \hat{y} > 0)$ or $(y < 0 \;\&\; \hat{y} < 0) \;\&\; |y - \hat{y}| > \varepsilon$

We took an error margin of 0.5. The rationale behind this choice is as follows. Even though we formulated conflict analysis as a regression task, it can also be considered as a classification task, where conflict level > 0 means presence of conflict and conflict level < 0 indicates absence of conflict. Thus for a regression output of range [-1,1], [-1,0) means absence of conflict and [0,1] means presence of conflict. We wanted to refine it further with [-1,-0.5) as very low, [-0.5,0) as low, [0,0.5) as high and [0.5,1] as very high levels of conflict. In this sense, an error margin of 0.5 means that the predicted label belongs to a different sub-class than the actual label.

3.4.2 Local Interpretable Model Agnostic Explanation (LIME) for ConflictNET

LIME is an algorithm proposed in [31] for model agnostic interpretation of machine learning models. Model agnostic means that this algorithm can be applied to interpret any classifier or regressor regardless of the architecture. LIME provides instance-based explanations to the predictions of a model. This means that, given an instance of input data (one text/image/audio sample), LIME can identify which portions of that particular instance are relevant for the model's prediction. For example, for an e-mail classification system, LIME can provide a list of words contained in the e-mail as an explanation for its classification to some category. SoundLIME (SLIME) [194] has shown that LIME can be applied to machine learning models whose input

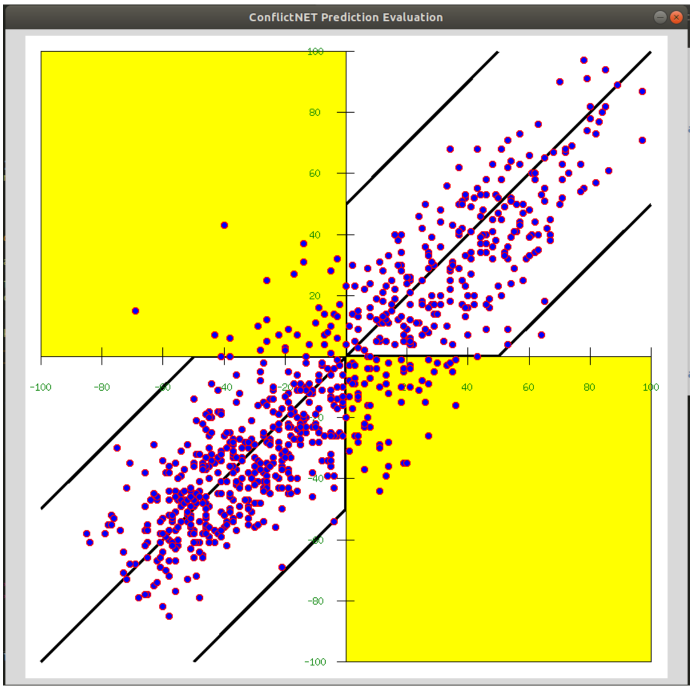


Figure 3.3: GUI for *ConflictNET* analysis

can be audio waveforms or spectrogram representations. SLIME pinpoints the time or time-frequency regions that contribute most to a decision. We have adapted SLIME for interpreting *ConflictNET*.

To use LIME, we need to first define what is an *interpretable data representation* for our specific case. For example, [31] states that a possible interpretable representation for text classifier is a binary vector indicating the presence/absence of a word, even if the classifier uses more complex and incomprehensible features like word embeddings. In our case, the input is raw audio waveform. Thus, similar to [194], we consider temporal segments (also called super-samples) as analogous to words in a text classifier. The interpretable data representation would then be a binary vector indicating the presence/absence of a temporal segment. Thus, an input audio signal, $\mathbf{x}_i$, can be uniformly split into several super-samples $\mathbf{T}_j$. Thus $\mathbf{x}_i = [\mathbf{T}_0,\mathbf{T}_1,....,\mathbf{T}_n]$, with n number of super-samples and $\mathbf{x}'_i = [1,1,1,....,1]$ is considered as the interpretable data representation for $\mathbf{x}_i$. Let us call the space of interpretable data representation as *sparse binary space*.

LIME provides local interpretation, meaning that it explains the local behaviour of the model in the vicinity of the instance being predicted. To do this, LIME defines the neighbourhood/vicinity using a synthetic dataset of perturbed data. If $\mathbf{x}_i$ is the input sample for which an explanation is required, LIME first creates a set of perturbed instances $\mathbf{Z}$ by randomly keeping/removing super-samples from $\mathbf{x}_i$. The corresponding sparse binary space representation for $\mathbf{Z}$ is the set $\mathbf{Z}'$. For example, if the number of super-samples is 4, then $\mathbf{z}'_k$=[1,0,0,1] is a possible perturbed instance , where 1/0 indicates the presence/absence of that particular super-sample. The neighborhood is defined by using a distance metric, that calculates the distance between the instance of interest $\mathbf{x}_i$ and each of the perturbed instances in the sparse binary space.

LIME defines an *interpretable explanation* as a model $\mathbf{g}$, that belongs to a class of interpretable models (linear models, decision trees etc). Following [194], we have chosen $\mathbf{g}$ to be a linear model. This means that LIME will find a linear model $\mathbf{g}$ that approximates the classifier/regressor for the specific case of input $\mathbf{x}_i$ using the vicinity defined by the perturbed instances. In case of a linear regression model, the weights of the model indicate the relevance of the corresponding features and their polarity indicate whether the features have a positive/negative influence on the prediction.

Each speech input to *ConflictNET* is of 30s duration and 8KHz sampling frequency. Thus, each speech file has 240,000 samples and there exists only one label for each 30s long speech file. To create *interpretable data representation*, we split each 30s long file into 10 super-samples, each of 3s duration and having 24,000 samples. Thus an input speech waveform, $\mathbf{x}_i$, was split into [$\mathbf{T}^0$,$\mathbf{T}^1$,.......,$\mathbf{T}^9$]. The intention here was to use LIME to understand which of these 10 super-samples are relevant for $\mathbf{y}_i$, the model's prediction of $\mathbf{x}_i$. We can choose the number of super-samples in the explanation of LIME, as less than or equal to the total number of super-samples. If it is less than 10, say 6, then LIME will explain only the most prominent 6 super-samples. We chose the number of super-samples to be equal to 10, to understand the amount of contribution from each of them.

The next step was to create the synthetic dataset $\mathbf{Z}$ of perturbed samples to define the neighborhood of local prediction. We chose the number of perturbed samples as 2000. The set $\mathbf{Z}'$ contains the corresponding binary representations. Note that the first element of $\mathbf{Z}'$ contains all ones, indicating the original input speech sample for which we require a LIME explanation. To apply *ConflictNET* to the perturbed samples, we projected these samples from the sparse space back to the input space, where absence of a super-sample is indicated by a sequence of zeros. Then, *ConflictNET* was applied on each element of $\mathbf{Z}$ and their predicted labels were obtained. Next, we need to determine the 'importance' of each perturbed sample $\mathbf{z}_i \in \mathbf{Z}$ by finding its distance from the original speech sample $\mathbf{x}_i$ in the sparse binary space. Samples that are closer to $\mathbf{x}_i$ has more 'importance' or 'weight' compared to samples that are farther from $\mathbf{x}_i$. The function to obtain these weights is given by,

$$\rho(\mathbf{x}'_i, \mathbf{z}'_k) = \exp(-D(\mathbf{x}'_i, \mathbf{z}'_k)^2/\sigma^2), \tag{3.5}$$

where we chose ‘cosine’ as the distance metric D and the width of the kernel σ as 25 (default value used in [194]).

Using the sparse binary space representations $\mathbf{Z}'$, predicted labels and the importance weights, a linear model was obtained using ridge regression (*sklearn.linearmodel.Ridge*) with the locally weighted square loss [31, 194] as given below.

$$L(f,g,\rho) = \sum_{z_k \in Z, z'_k \in Z'} \rho(\mathbf{x}'_i, \mathbf{z}'_k)(f(z_k) - g(z'_k))^2, \tag{3.6}$$

where ρ is the importance weight function from eq.3.5, g is the linear model and f is the original model (*ConflictNET* in our case). After the model is fit, super-samples were sorted in order of the coefficient magnitudes. Higher magnitude indicates that the particular super-sample has more relevance on the prediction. The polarity refers to the correlation between the super-sample and the prediction value. The LIME code [194] requires us to provide labels denoting the classification or regression value in probability. For example, if we have 2 classes cats and dogs, and if the classifier gives a probability of 0.7 that it is a dog, then the label to be given to LIME is [0.3,0.7]. In case of *ConflictNET*, the output is a regression value between -1 and 1. So to create probability based labels, we did the following mapping.

$$\begin{aligned} P_p &= (1+\hat{y})/2.0 \\ P_n &= (1-\hat{y})/2.0, \quad \forall \quad \hat{y} \in [-1,1] \end{aligned} \tag{3.7}$$

where P_p and P_n denote probabilities associated with positive and negative classes.

We performed LIME analysis on random samples taken from the 4 quadrants of the GUI (see Figure 3.3).

1. $(y > 0$ & $\hat{y} > 0)$ or $(y < 0$ & $\hat{y} < 0)$ & $|y - \hat{y}| \leq \varepsilon$: In this category, actual and predicted labels have same polarity and their magnitude difference is less than 0.5. The network is focusing on instances where speakers are interrupting each other with raised voices. For example, in case of sample ‘*07−01−31_1710_1740.wav*’, the actual label is 0.43 and

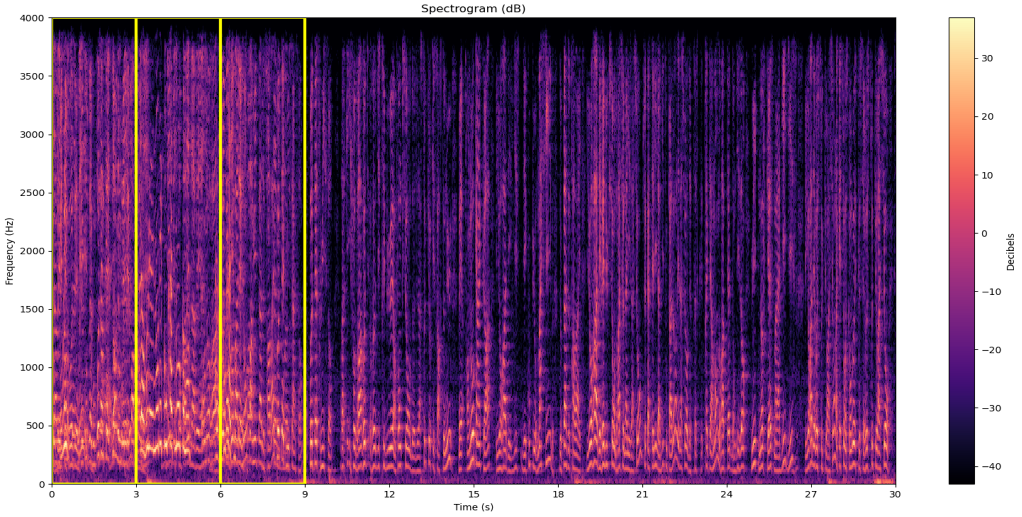


Figure 3.4: Log magnitude spectrogram for sample '07−01−31_1710_1740.wav'. Time-frequency representation of the three most relevant super-samples according to LIME [31] are highlighted in yellow. Notice the distortions in the harmonics in these super-samples compared to the rest of the spectrogram.

predicted label is 0.68. The order of super-samples given by LIME is [0,1,2,9,6,4,5,7,3,8] (see Figure. 3.4). Upon hearing the super-samples, 0, 1 and 2 contain speech interruptions with high energy. All other samples are more or less monologues and LIME gives negative weights to these super-samples, indicating that they have 'negative' contribution towards the predicted conflict level. In order to make sure that the network is not biased towards the placement of conflict instances in the speech waveform, other speech files that have conflict instances at the middle as well as end were also analysed. An example is, sample '*06−12−13_2160_2190.wav*', whose actual label is 0.31 and predicted label is 0.44. This speech file has verbal interruptions in the last 10 seconds. The order of super-samples given by LIME is [7,6,8,0,5,2,3,1,4,9]. Here, 7,6 and 8 actually contain verbal interruptions during a heated discussion between three people.

2. $y > 0$ & $\hat{y} < 0$: It is observed that for samples belonging to this category, more than 70% duration contain monologues. This means that the amount of time that has verbal interruptions is less. It is also observed that for many samples in this category, the network's prediction seems to be better than the actual label for a person who does not understand French language. However, LIME analysis has pointed out some discrepancies in the network's prediction. For example, for sample '*06-11-08_1980_2010.wav*', the

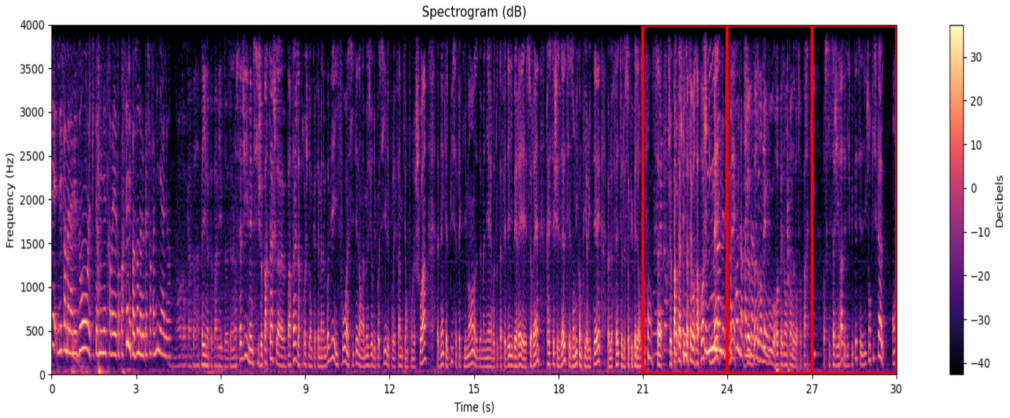


Figure 3.5: Log magnitude spectrogram for sample '06-11-08_1980_2010.wav'. Time-frequency representation of the three super-samples with negative LIME [31] coefficients are highlighted in red.

actual label is 0.27 and predicted label is -0.26 (see Figure 3.5). The order of super-samples given by LIME is [0,1,2,5,6,4,7,8,9,3], where 7, 8 and 9 are given negative polarity, which means that they are negatively correlated with the prediction. However, on hearing the samples, only 7 and 8 contain verbal interruptions. 9 contains a single speaker's voice. In fact, super-sample 9 sounds similar to 5 and it is not clear why the network thinks that 9 contributes towards conflict and 5 contributes towards non-conflict prediction. Similarly, for samples '*06-12-20_690_720.wav*', '*08-01-30_2280_2310.wav*', '*08-01-15_1740_1770.wav*', '*07-02-14_900_930.wav*' and '*08-01-15_1170_1200.wav*', LIME analysis shows that the network thinks a few super-samples that actually contain verbal interruptions as contributing towards non-conflict prediction and a few super-samples that contain only monologues as contributing towards conflict prediction.

3. $y < 0$ & $\hat{y} > 0$: Most speech samples in this category contain laughter instances. Some instances contain strong background music or mic tapping noise (someone is touching on the wearable or hand-held mic) as well. The network is giving more importance to super-samples containing laughter and mic taps. For example, for sample '*07-02-28_1260_1290.wav*', the actual label is -0.69 and predicted label is 0.15. The order of super-samples, as given by LIME, is [0,5,4,6,8,9,2,3,7,1], where 0 is given a weight of +0.5 and all others have negative weights. Upon hearing the super-samples, it is observed that 0 contains laughter and all the rest contains monologues with little or no cross-talks. This means that super-sample 0 is the reason for the network's prediction of this sample as containing conflict.

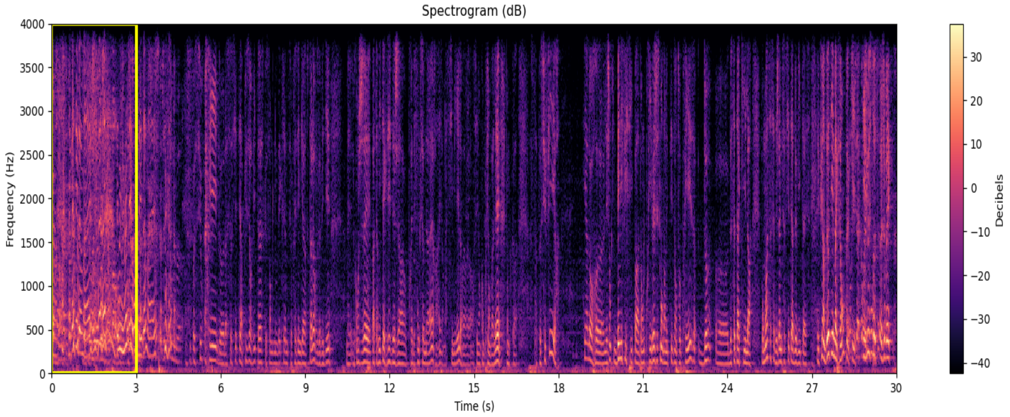


Figure 3.6: Log magnitude spectrogram for sample '*07-02-28_1260_1290.wav*'. Time-frequency representation of the super-sample containing laughter instance is highlighted in yellow.

Similarly, for sample '*08-01-30_810_840.wav*', actual value is -0.4 and predicted value is 0.43. The order of super-samples, given by LIME, was [8,1,6,7,4,3,9,2,5,0], in which 8,1,6 and 7 contain laughter by a cross-talker.

4. Actual and predicted values have same polarity but exceeds a margin of error ($|y - \hat{y}| > \varepsilon$): The margin of error is selected as 0.5. This category contains fewer samples compared to the other categories. It has been observed that for samples whose actual and predicted values are both positive, the 2 speakers sound very similar to each other. For example, for sample '*08-01-30_720_750.wav*', whose actual and predicted values are 0.64 and 0.07 respectively, 1 male speaker speaks for roughly half the time and other male speaker speaks for the other half. The transition interruption is of very short duration as well. For all the samples, whose actual and predicted values are negative and their difference exceeds 0.5, the network's prediction is lower than the actual value. For example for samples '*07-05-02_1530_1560.wav*' and '*06-09-27_1230_1260.wav*', the actual values are -0.04 and -0.21 respectively and the corresponding predicted values are -0.55 and -0.73 respectively. This means that the network perceives these samples as containing lower amount of conflict than the annotators.

From these observations, we can conclude that the network is focusing on energy variations in the input speech. Verbal conflicts are associated with raised voices and hence high energy segments of conversations, which means that *ConflictNet* is using the right cues to identify instances of verbal conflict. However, the network is prone to errors that can occur with other high energy segments like laughter instances, music or other high energy noise instances. Also, since the network is language independent, it cannot understand a passive aggressive verbal conflict where the people are not exactly shouting at each other.

3.5 Conclusion

Verbal conflicts can occur during inter-personal arguments and can be the cause or effect of human aggressive behaviour, thus making the estimation of verbal conflict level an important research topic under the umbrella of affective computing. Previous works focused on using hand-crafted generic acoustic features along with off-the-shelf classifier/regressor models, which can lead to time-consuming feature pruning and/or manual refinement. Some works also use extra meta data like the number of speakers and speech overlap information. One of our research questions was whether or not it is possible to design a network that can automatically learn conflict specific features given raw speech signals as input. We develop a convolution-recurrent neural network model, equipped with attention mechanism, that directly maps the given raw speech signal to a continuous label space indicating the verbal conflict intensity. Thus, we conclude that it is possible to design an end-to-end model that can predict continuous conflict intensity estimation values from raw speech signals. Performance evaluation of the model on the SSPNet Conflict Corpus [128] showed that it is competitive with respect to the state-of-the-art methods. Since our end-to-end model is a ‘black-box’ from an interpretability point of view, we adapt a popular explainable AI method called LIME [31] to provide sample based localised explanations of our model. This process showed that our network uses energy variations in the input speech as cues for detecting conflicts.

Chapter 4

Attention Based Multimodal Fusion

4.1 Introduction

In this chapter, we discuss details of the design of multimodal fusion models using audio (speech), vision and language modalities as input. Multimodal fusion models fuse complementary information from multiple modalities to outperform their unimodal counterparts. However, a successful model that fuses modalities requires components that can effectively aggregate task-relevant information from each modality. Recently, cross-modal attention [19, 20, 195], that uses one modality to compute attention scores for another modality, is being viewed as an effective mechanism for multimodal fusion. However, the current literature is unclear on the gain that such a mechanism brings compared to the corresponding intra-attention mechanism, that relies only on one modality to compute attention scores for itself. We aim to fill this research gap and quantify the performance differences between the two types of attention mechanisms. To this end, we design two models, one based on cross-attention and another based on intra-attention. In addition to attention mechanism, each model uses convolutional layers for local feature extraction and recurrent layers for global sequential modelling. We validated the effectiveness of our models on the task of 7-class emotion classification using the audio, vision and language modalities from IEMOCAP [46] dataset. We also analyse the behaviour of our trimodal models when one or more modalities are missing during the test-time and use two strategies, namely, Moddrop [47] and KNN based imputation [196] to combat the performance drop.

4.2 Intra- and Cross-Modal Models

We want to verify the hypothesis that multimodal recognition models benefit from cross-attention mechanism [19, 20, 195] and hence we contrast this mechanism with the corresponding models using intra-modal (self) attention mechanism. To this end, we design two multimodal emotion recognition models, each employing one of the two attention mechanisms. To enable a direct comparison between the two types of attention mechanisms, we use only the attention module and not the transformer [113] encoder module. In addition to the attention mechanisms, our models also contain convolutional and recurrent layers for effective modelling of temporal sequential data.

Our proposed cross- and intra-attention models (see Figures. 4.1 & 4.2) first process individual modalities using modality-specific encoders. Each modality specific encoder is provided with their corresponding features obtained from audio-visual utterance clips. The encoded features are then fed into intra- or cross-modal Multi-Head-Attention (MHA) [113] modules, respectively. A global representation of the entire utterance clip is then generated as temporal average at the outputs of each MHA module. The resulting features are then concatenated and their mean and standard deviation are obtained using a statistical pooling layer. The concatenation of mean and standard deviation vectors is then fed to fully connected layers. The class predictions are then obtained through a softmax operation. A detailed explanation is given as follows:

Let $X_a \in \mathbb{R}^{t_a \times d_a}$ be the audio features corresponding to an utterance clip, where t_a is the sequence length and d_a is the feature dimension. The audio encoder consists of a 1 dimensional convolution layer followed by a bi-directional GRU. The convolution layer, which refines the input feature sequence by finding task-relevant patterns, operates as follows:

$$X'_a(t') = b(t') + \sum_{k=0}^{t_a-1} (W(t',k) * X_a(k)), \tag{4.1}$$

where $X'_a \in \mathbb{R}^{t'_a \times d'_a}$ is the output with length t'_a and dimension d'_a, $t' \in [0, t'_a - 1]$, $*$ is the convolution operator, W are the weights and b are the biases associated with the layer. Thus, the convolution layer modifies the sequence length as well as the feature dimension. The bi-GRU layer models contextual inter-dependence of the features across time. For each element in the sequence, the bi-GRU layer computes the following functions:

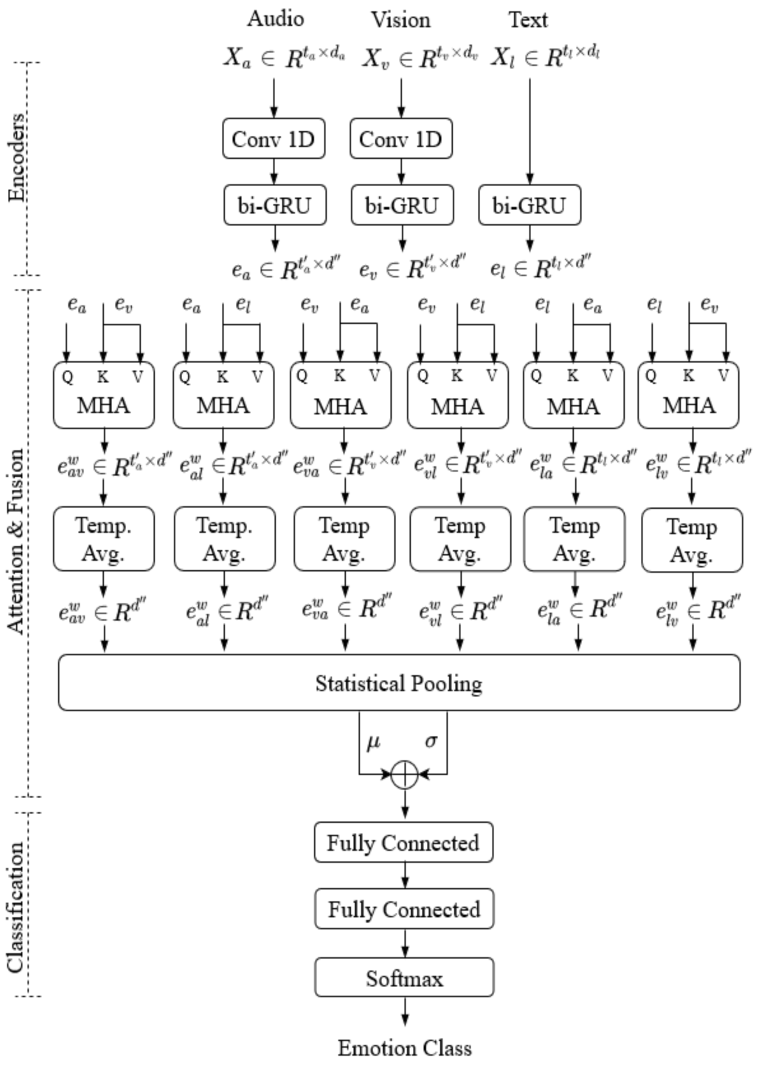


Figure 4.1: The architecture of our proposed Cross-attention model. The input features $\bigoplus$ represents concatenation operation. KEY - MHA: Multi-Head Attention, Temp.: Temporal, Avg.: Averaging, a: audio, v: visual, l: language/text, μ: mean, σ: standard deviation

$$\begin{cases} r_t = \sigma(W_{ir}X'_a(t) + b_{ir} + W_{hr}h_{t-1} + b_{hr}), \\ z_t = \sigma(W_{iz}X'_a(t) + b_{iz} + W_{hz}h_{t-1} + b_{hz}), \\ n_t = \phi_h(W_{in}X'_a(t) + b_{in} + r_t \odot (W_{hn}h_{t-1} + b_{hn})), \\ h_t = (1 - z_t) \odot n_t + z_t \odot (h_{t-1}), \end{cases} \tag{4.2}$$

where h_t and h_{t-1} are the hidden states at times t and $t-1$, $X'_a(t)$ is the input at time t. r_t, z_t and n_t are the reset, update and new gates, W and b are the corresponding weights and biases, σ and ϕ_h are the sigmoid and hyperbolic tangent functions and $\odot$ is the Hadamard product. At the

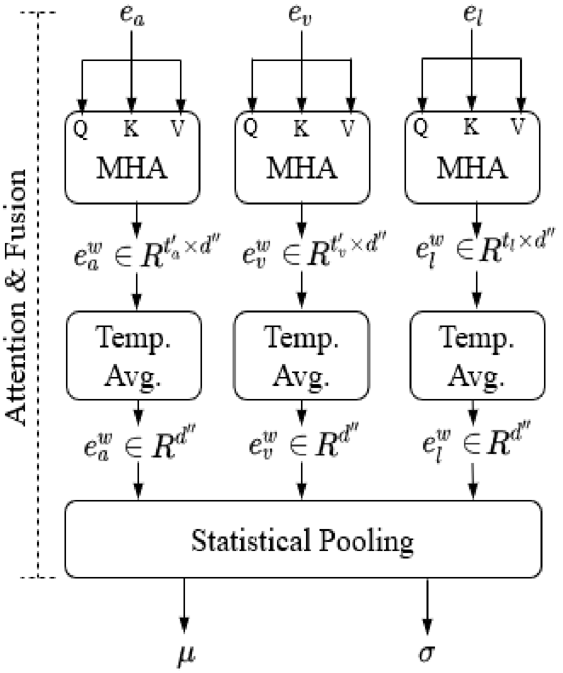


Figure 4.2: Attention and fusion module in our proposed intra-attention model. Rest of the model is same as the cross-attention model. Note that the number of required MHA and Temp.Avg. modules is half that of the cross-attention model.

output of bi-GRU, the forward and backward hidden states for each time-step are concatenated and the refined audio features can be represented as $e_a \in \mathbb{R}^{t'_a \times d''}$, where d'' is twice the number of hidden neurons in the GRU.

Similar to audio, the video encoder consists of one 1D convolution layer followed by a bi-GRU layer. If $X_v \in \mathbb{R}^{t_v \times d_v}$ represents the video features corresponding to an utterance, then at the output of video encoder, the features are refined to $e_v \in \mathbb{R}^{t'_v \times d''}$. For the text modality, the encoder consists of only one bi-GRU layer. The input and output of text encoder can be represented by $X_l \in \mathbb{R}^{t_l \times d_l}$ and $e_l \in \mathbb{R}^{t_l \times d''}$ respectively.

The encoder outputs of all the three modalities are then fed into intra- or cross-modal MHA modules (see Figure.2.4 for details on MHA module). For cross-attention model, since there are 6 pairs of modalities, statistical pooling is done across the concatenation of the temporal averages of 6 cross-modal sequences, whereas for the intra-attention model, it is done across the concatenation of the temporal averages of the intra-attended sequences of all the 3 modalities. The classifier for both models is:

$$\hat{y} = Softmax(f_{\theta_2}(f_{\theta_1}([\mu, \sigma]))), \tag{4.3}$$

where μ is mean, σ is standard deviation, f_{θ_1} and f_{θ_2} denote the 2 fully connected layers with parameters θ_1 and θ_2 respectively and $\hat{y}$ denotes the vector of class predictions.

4.2.1 Architecture, loss function and training

The models are implemented using PyTorch [48]. The bimodal and unimodal versions of the tri-modal models are created by removing components corresponding to the unused modality/modalities. We use Adam [197] optimiser with an initial learning rate of 0.001. The learning rate is reduced by a factor 0.1 when the validation loss has stopped decreasing for 10 consecutive epochs. Training is stopped when UnWeighted Accuracy (UWA) does not improve in the validation set for 10 consecutive epochs and the model with best validation UWA is used for testing. The batch size is 32 and all models are trained using the categorical cross-entropy loss. The audio and vision encoders contain one 1 dimensional convolution layer each. The kernel size and stride length are both set to 1. The number of input and output channels for audio convolution layer are 1,000 and 500 respectively while for vision they are 32 and 25 respectively. The number of bi-GRU layers for all the 3 modalities is 1. The number of hidden neurons in each bi-GRU layer is 60. The number of attention heads in all MHA modules is 6 and a dropout rate of 0.1 is applied to reduce overfitting. The number of neurons in the first and second fully connected output layers are 60 (same as number of bi-GRU neurons) and 7 (number of output classes) respectively. All parameters were chosen based on the performance on validation set. Specifically, a grid search over hyper-parameters was done for the intra and cross tri-modal models. In order to make sure that the tuning is not biased towards one model, we made sure that the same hyper-parameter combination was applied to both models after fixing the seed value. The best hyper-parameters were selected by the best UWA result of either models on the validation set.

4.3 Results and Ablations

We use the IEMOCAP [46] dataset and evaluate our models on 7 class emotion classification. Table 4.1 shows the results of comparing the intra- and cross-modal models on 7-class unimodal, bimodal and trimodal emotion recognition tasks. We report the mean and standard deviation

Table 4.1: Results of a 7-class emotion classification task presented as mean ± standard deviation. AMH refers to AMH [33] for trimodal models and to MHA [142] for bimodal models.

	Weighted Accuracy			
Modality	MDRE [65]	AMH [33]	Cross	Intra
T	-	-	-	.474 ± .030
V	-	-	-	.454 ± .019
A	-	-	-	.365 ± .018
T+V	.524 ± .021	.526 ± .024	**.567** ± **.022**	.563 ± .022
T+A	.418 ± .077	.491 ± .028	.501 ± .026	**.518** ± **.031***
V+A	.376 ± .024	.371 ± .042	.481 ± .024	**.483** ± **.026**
T+V+A	.490 ± .056	.547 ± .025	.578 ± .024	**.587** ± **.022***
	Unweighted Accuracy			
Modality	MDRE [65]	AMH [33]	Cross	intra
T	-	-	-	.535 ± .016
V	-	-	-	.513 ± .018
A	-	-	-	.452 ± .017
T+V	.579 ± .015	.580 ± .019	**.617** ± **.015**	.614 ± .020
T+A	.498 ± .059	.543 ± .026	.562 ± .017	**.574** ± **.018***
V+A	.477 ± .025	.471 ± .047	.566 ± .022	**.567** ± **.026**
T+V+A	.564 ± .043	.617 ± .016	.636 ± .017	**.642** ± **.019**

KEY - A: audio; V: vision; T: text; intra: intra-attention model; Cross: cross-attention model. The best results in each row are in bold font. The symbol * refers to the only three results with statistically significant difference between the intra and cross models.

obtained across 50 runs (5 folds × 10 repetitions) for each model. We also applied two-tailed t-test with the null hypothesis that the accuracy values of both intra and cross-attention models have identical average (expected) values. Comparison of the unimodal performances shows that the text outperforms the vision and audio modalities. This result is consistent with previous works [19, 65]. Since unimodal performance evaluation is not possible with the cross-modal model, we report results with the unimodal version of the intra-attention model. Among bimodal models, the combination of vision and text modalities gives the best performance for both models. These results are also consistent with previous works [65, 142]. Overall, both models provide comparable performances for bi- and trimodal cases. Intra-attention significantly outperforms cross-attention (P value $< .05$) only for T+A (text and audio) and the Weighted Accuracy (WA) of T+V+A (text, vision, and audio).

We compare with methods that use the same set of features and dataset partition. The trimodal models are compared with AMH [33], the current state-of-the-art model, which uses a combination of unimodal GRU layers and an iterative attention mechanism[1]. Note that the intra-attention model exceeds the performance of AMH by 4.0 and 2.5 percentage points (pp) over mean in terms of WA and UWA, respectively. Similar figures for the cross-attention model are 3.1 pp and 1.9 pp. We also compare with MDRE [65], which uses recurrent layers to model unimodal signals followed by aggregation and classification using fully connected layers. The better performance of the intra and cross-attention models, as well as AMH, compared to MDRE can be attributed to the effectiveness of the attention mechanism. For bimodal models, we compare with the bimodal version of AMH called MHA [142] and MDRE. Again, both models outperform MHA and MDRE in all the 3 bimodal cases. Note that we obtain bimodal results by ablating the trimodal models and not by fine-tuning for individual bimodal cases. This means that only the original tri-modal models' architectures were optimized for best performance using tri-modal training. For the bi-modal results, we simply removed the components corresponding to the unavailable modality from the tri-modal model and trained the remaining layers using the available two modalities. A much better optimization for the bi-modal models could be done by doing architecture search and hyper-parameter optimization for the individual bi-modal cases. This might have resulted in improved results compared to simply training the bi-modal models ablated from the tri-modal models. However, since the objective of our whole experiments is to compare the self- and cross-attention performance on tri-modal models that use audio, vision and text modalities, fine-tuning for individual bi-modal cases might be out of scope and not relevant for the problem.

Furthermore, AMH, MHA and MDRE use prosody features in addition to MFCC features for audio, whereas we use only MFCC features. The state-of-the-art result for text+audio case is obtained by [20] (0.560 WA and 0.612 UWA) which is significantly higher than the bimodal T+A (text and audio) results. We hypothesize two reasons for this: (1) unlike [20], the bimodal models are not fine-tuned for the bimodal cases; (2) [20] uses transformer encoders that contain additional parameters that might help in learning more complex inter-modal relationships, whereas we use only the multi-head attention mechanism. Nevertheless, both models improve the state-of-the-art trimodal results of AMH.

[1] We use the revised results of AMH, MHA and MDRE from `https://github.com/david-yoon/attentive-modality-hopping-for-SER`. We note that the WA and UWA values were swapped by the authors and we rectify this error in Table 4.1.

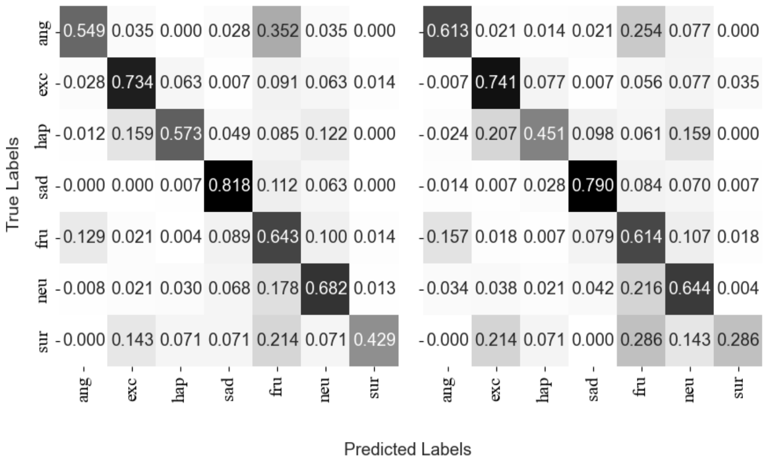


Figure 4.3: Confusion matrices of intra- (left) and cross-attention models (right) for trimodal 7-class classification using a random fold. The emotions classes are abbreviated with their first 3 letters.

Table 4.2: Weighted accuracy (WA) and Unweighted accuracy (UWA) for 7-class emotion classification using additional trimodal model configurations. Intra and Cross model results are also shown for comparison.

Model	**WA**	**UWA**
Cross-noSP	.570 ± .021	.634 ± .015
Cross	.578 ± .024	.636 ± .012
Intra-noSP	.584 ± .021	.638 ± .019
Intra	**.587 ± .022**	**.642 ± .019**
Cross+Intra	.585 ± .028	.642 ± .020

KEY - SP: statistical pooling; Cross-noSP and Intra-noSP: cross and intra-attention models without SP; Cross+Intra: combination model that concatenates mean and standard deviation vectors from intra and cross-attention models.

Figure. 4.3 shows the confusion matrices for the intra- and cross-attention models. For both models we can observe that the classes *angry* and *frustrated* are more often confused with each other, and the class *happy* gets confused with *excited* (these 2 classes are inherently similar). The poor performance of both models on the class *surprise* can be attributed to the fact that this has the smallest sample size in the dataset. These observations are consistent with the previous literature [33].

In addition to the two described model configurations, we also experimented with different variations of the trimodal models. We removed the statistical pooling layer from both models to assess its significance. The outputs from all temporal averaging modules (see Figures. 4.1 & 4.2) were concatenated and passed to the classifier module. These models are shown as ‘Cross-noSP’ and ‘Intra-noSP’ in Table 4.2. We can make two observations. Firstly, the intra-attention model outperforms the cross-attention model (P value $<$ .05 for WA) even after ablating statistical pooling. Secondly, the performance of both models decreases without the statistical pooling layer. We also assessed the performance of a combined model created by merging the intra and cross-attention models (Cross+Intra). The statistical pooling output from both models were concatenated and fed to a common classifier module. Surprisingly, we can see that the performance is similar to that of the intra-attention model. This might indicate that the cross-attention model does not contribute any additional, relevant information compared to that of the intra-attention model.

We also conducted further ablation on the cross-attention model to assess the role of individual modalities as sources and targets in the MHA mechanism. Table 4.3 shows the results. For bimodal combinations with a single source and a single target modality, using vision modality to generate attention scores for text modality (T2V) provides the best results. Intuitively, this combination is using the second best performing modality to produce attention scores for the best performing modality. The next highest values in this group are obtained when vision is used as source and text as target. In this case, the best performing modality is used to find the attention scores for the second best performing modality. On the other hand, the lowest performance is obtained when audio is used as target and vision as source modality and audio is used to find attention scores for vision modality. Next, we kept the target modality fixed and used the other two modalities as sources. In this group, we find that keeping vision as target modality gives the best performance (A2V+T2V). In the case when we keep the source fixed, text as source modality gives the best performance (T2V+T2A). Audio and vision modalities are used to find the attention scores for text. This might point to the intuition that text is the most informative modality and that audio and vision modalities are auxiliary modalities that help the text by finding relevant time-steps in its sequence.

4.4 Missing Modality Behaviour Analysis

One common assumption associated with developing multimodal models is that all the modalities available during training are also available during the testing or deployment or inference phase. In practice, however, this cannot be guaranteed. Missing modality situations are far too common and hence it is important to assess the behaviour of multimodal models during missing modality scenarios. We check the behaviour of our trimodal models during scenarios where one or two

Table 4.3: Weighted accuracy (WA) and Unweighted accuracy (UWA) for 7-class emotion classification by ablating the trimodal configuration of the cross-attention model. Modality combination is shown in the format S2T where S and T represent source and target respectively.

Model	**WA**	**UWA**
T2V only	**.550 ± .025**	**.597 ± .018**
T2A only	.504 ± .023	.561 ± .016
V2T only	.520 ± .026	.597 ± .022
V2A only	.448 ± .020	.532 ± .017
A2T only	.456 ± .026	.533 ± .020
A2V only	.471 ± .027	.558 ± .023
V2T + A2T	.534 ± .027	.613 ± .020
V2A + T2A	.557 ± .023	.611 ± .021
A2V + T2V	**.569 ± .018**	**.622 ± .015**
T2V + T2A	**.540 ± .024**	**.590 ± .020**
V2A + V2T	.504 ± .027	.580 ± .019
A2T + A2V	.511 ± .030	.589 ± .021

modalities are missing.

Table 4.4 shows the results when both intra- and cross-attention trimodal models are trained using 3 modalities but when tested using one or two modalities. For comparison purposes, the original unimodal and bimodal trained and tested model counterparts are also shown. Note that for computational reasons, the missing modality situation is simulated by representing that modality as a vector of zeros. When comparing the intra- models, we can see that the performance has dropped significantly for all the unimodal and bimodal cases. Similar observations apply for the bimodal cross-attention models as well. This means that our trimodal models which were trained with all the three modalities consistently under-perform during missing modality situations when compared to their unimodal and bimodal trained and tested counterparts.

In order to improve the robustness of our multimodal models, we exploit two ideas:

- Apply the Moddrop [47] technique where modalities are dropped out during training.
- Impute missing modality intermediate features using a K Nearest Neighbour (KNN) algorithm based technique [48].

The fundamental difference between both techniques is that while Moddrop has to be applied during the training time, the KNN based method is to be applied after the training. The latter will suit situations where we do not have a provision to train the model but need to accommodate for missing modality scenario at test-time.

Table 4.4: Weighted accuracy (WA) and Unweighted accuracy (UWA) for 7-class emotion classification by ablating the trimodal intra- and cross-attention models. orig refers to results using the original unimodal and bimodal models that were trained and tested using same modalities. ●and ○indicate the presence and absence of any modality respectively.

T	V	A	Intra		Intra-orig	
			WA	UWA	WA	UWA
●	○	○	.306 ± .063	.351 ± .049	.474 ± .030	.535 ± .016
○	●	○	.233 ± .033	.330 ± .040	.454 ± .019	.513 ± .018
○	○	●	.255 ± .034	.350 ± .035	.365 ± .018	.452 ± .017
●	●	○	.446 ± .055	.500 ± .052	.563 ± .022	.614 ± .020
●	○	●	.468 ± .038	.515 ± .032	.518 ± .031	.574 ± .018
○	●	●	.413 ± .027	.520 ± .023	.483 ± .026	.567 ± .026

T	V	A	Cross		Cross-orig	
			WA	UWA	WA	UWA
●	○	○	.237 ± .062	.285 ± .048	N.A	N.A
○	●	○	.220 ± .033	.312 ± .040	N.A	N.A
○	○	●	.229 ± .036	.315 ± .040	N.A	N.A
●	●	○	.361 ± .059	.423 ± .060	.567 ± .022	.617 ± .015
●	○	●	.446 ± .057	.495 ± .039	.501 ± .026	.562 ± .017
○	●	●	.403 ± .023	.502 ± .022	.481 ± .024	.566 ± .022

4.4.1 Moddrop training for missing modality

The Moddrop technique where modalities are dropped out during training makes the model aware of missing or dropped modality situations. In Moddrop training, our training set contains the following combinations for each sample; (T, V, A), (T, V, 0), (T, 0, A) and (0, V, A) where 0 represents a vector of the same feature dimensionality as the missing modality. Thus, Moddrop can be considered as a data augmentation scheme as well.

Figure. 4.4 shows the results of applying Moddrop training on the trimodal version of intra-modal attention model. The original models which are trained and tested on same modality combinations are shown as Unimodal-Training-and-Unimodal-Testing(UTUT)/Bimodal-Training-and-Bimodal-Testing(BTBT). The Multimodal-Training-and-Unimodal-Testing(MTUT)/Multimodal-Training-and-Bimodal-Testing(MTBT) shows the trimodal trained but unimodal or bimodal tested versions. It can be seen that Moddrop is able to improve the performance over the MTUT/MTBT versions. Similar experiments on the cross-attention model is shown in Figure. 4.5. We can observe that Moddrop based training strategy is effective for the cross-modal model as well. Another interesting result, as shown in Table 4.5, is that the trimodal trained and trimodal tested versions of both intra- and cross-attention models also show performance improvement with the

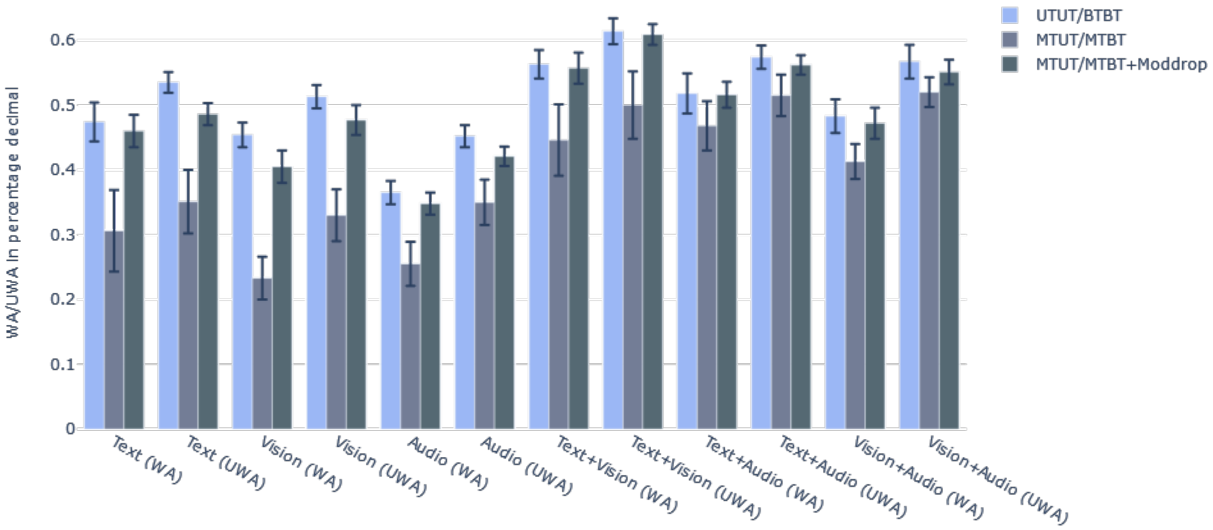


Figure 4.4: Results in mean and standard deviation format for 50 runs (5 folds x 10 seeds) using intra-attention model for original, missing modality and Moddrop trained scenarios. KEY: UTUT - Unimodal Training and Unimodal Testing, BTBT - Bimodal Training and Bimodal Testing, MTUT - Multimodal Training and Unimodal Testing, MTBT - Multimodal Training and Bimodal Testing

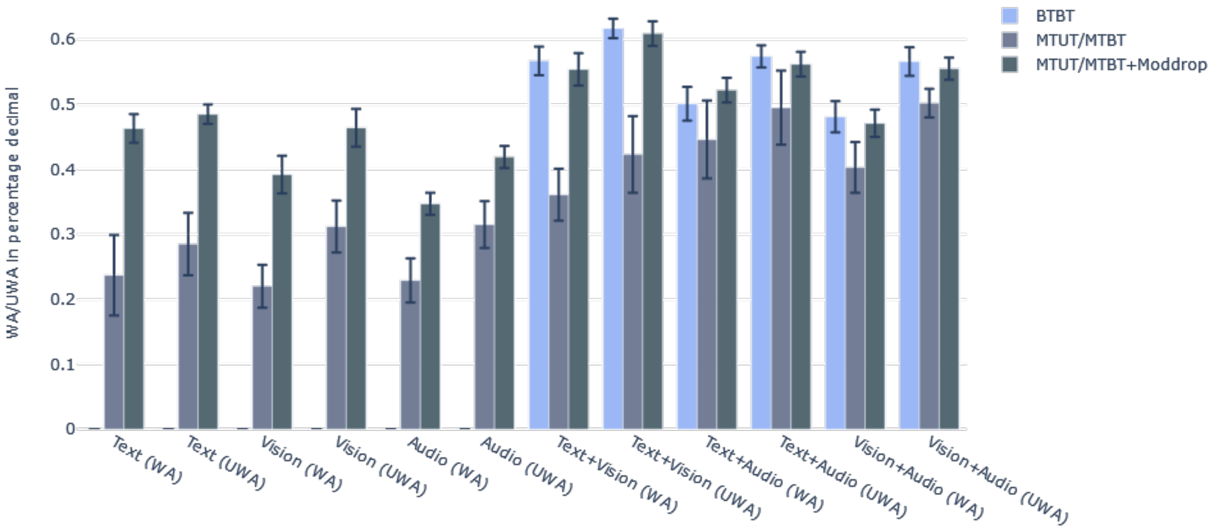


Figure 4.5: Results in mean and standard deviation format for 50 runs (5 folds x 10 seeds) using cross-attention model for original, missing modality and mod-drop trained scenarios. KEY: BTBT - Bimodal Training and Bimodal Testing, MTUT - Multimodal Training and Unimodal Testing, MTBT - Multimodal Training and Bimodal Testing. Note that cross-attention model does not have unimodal training and unimodal testing versions.

Table 4.5: Weighted accuracy (WA) and Unweighted accuracy (UWA) for 7-class emotion classification using the intra- and cross-attention trimodal models trained using Moddrop.

Model	**WA**	**UWA**
Cross	.578 ± .024	.636 ± .012
Cross+Moddrop	**.599 ± .019**	**.654 ± .014**
Intra	.587 ± .022	.642 ± .019
Intra+Moddrop	.593 ± .022	.648 ± .014

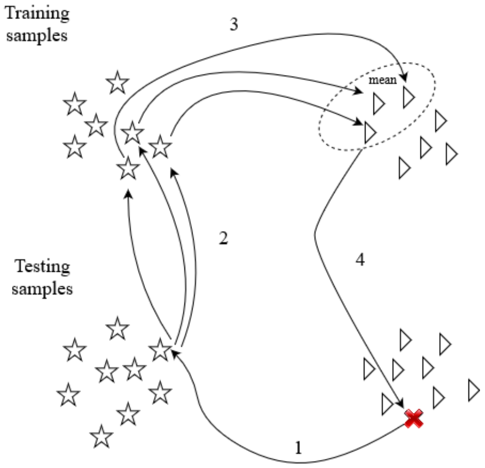


Figure 4.6: KNN based missing modality imputation strategy. Only two modalities are shown for simplicity. Stars and triangles represent samples from two different modalities.

Moddrop training. Albeit both models show performance gains with Moddrop, the gain obtained for cross-attention model is higher compared to the intra-attention model. Thus, Moddrop training strategy is not only useful for missing modality situations, but also for full modality situations as well.

4.4.2 KNN imputation for missing modality

Next, we use the KNN based test-time missing modality imputation strategy. The method is shown in Figure 4.6. Given paired modality samples in the training set, if for any given sample in the test set one modality is missing (as shown by the red cross in Figure 4.6), then it is imputed by using a 4 step strategy. First, the corresponding test sample in the other modality is taken. Then its K nearest neighbours in the training set is chosen based on some distance metric like Euclidean or Cosine. Then the corresponding samples in the other modality is found, their mean is taken. This mean value is used to replace the missing sample in the test-set. In practise, since the input modality features are high dimensional, we use this imputation strategy on intermediate features of lower dimensionality. The feature imputation was performed at the input of statistical

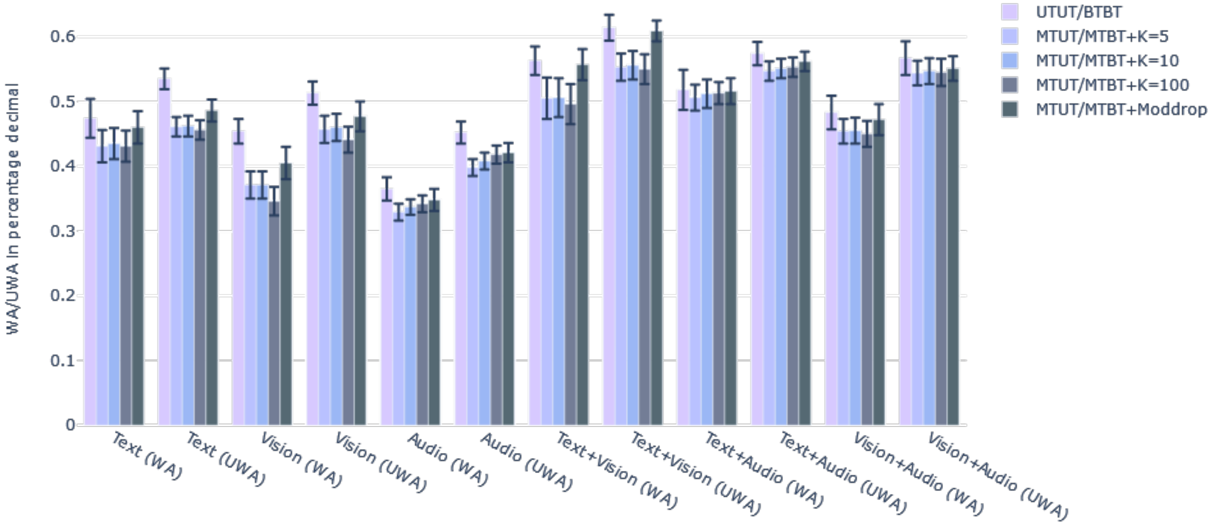


Figure 4.7: Results in mean and standard deviation format for 50 runs (5 folds x 10 seeds) using intra-attention model for KNN based missing modality imputation with K=5,10 and 100. Moddrop results are shown for comparison. KEY: UTUT - Unimodal Training and Unimodal Testing, BTBT - Bimodal Training and Bimodal Testing, MTUT - Multimodal Training and Unimodal Testing, MTBT - Multimodal Training and Bimodal Testing. Note that the distance computation metric used is Euclidean.

pooling layer (see Figure 4.2). Figure 4.7 shows the results obtained by different values of K for the intra-attention model. We also show the original UTUT/BTBT and Moddrop based results for comparison. We can see that the performance is inferior to the Moddrop strategy. Also, we do not observe any consistent performance differences across different K values. Even though for all K values, the performance is better than with no imputation, a major drawback of this method is the increased space complexity since we need to store the entire training set in memory.

4.5 Conclusion

Even though many contemporary works in multimodal fusion literature use intra-modal and cross-modal attention mechanisms, it is unclear whether there is a clear advantage in using one over the other. To understand this, in this chapter, we introduced two fusion models, one based on intra-modal attention and another based on cross-modal attention for the task of multi-class, multimodal emotion recognition using audio, vision and language modalities. Based on our unimodal, bimodal and trimodal experiments on the IEMOCAP dataset, we conclude that there is no consistent statistically significant performance differences across both models in terms of weighted and un-weighted accuracy measures. Ablation studies on both models validate the choice of statistical pooling layer in the model architecture. Furthermore, a combination model created by merging the intra- and cross-attention models did not out-perform the intra-attention model indicating that the cross-attention model might not contribute any useful information compared to the intra-attention model. Also, both our tri-modal models improve upon the state-of-the-art models in seven class emotion recognition task. Albeit both models under-perform in case of missing modality situations, which is a commonly reported issue with multimodal models, it can be concluded from our experiments that modality-dropping based training strategy as well as a KNN based post-training intermediate feature imputation strategy can be used to combat the performance drop. We also note that the modality-dropping based training strategy is not only beneficial for test-time missing modality cases but also improves the performance of test-time full modality cases.

Chapter 5

Co-Learning for Improving Feature Descriptiveness

5.1 Introduction

Most multimodal machine learning algorithms assume that all modalities that are present during training phase would be available during testing phase as well. Multimodal fusion models aggregate the unevenly distributed, complementary information across the available modalities to outperform unimodal models. However, several applications or use cases require only unimodal models. For example, emotion recognition models for emergency telephone lines [198] or customer-support call centers [132] can rely on speech data only. Vision based drowsiness detection models [199, 200] use only face images to detect the level of lethargic state of a driver. These unimodal models are limited by the characteristics of their respective modalities and their performance often falls short of their multimodal counterparts. The research question in this case is whether it is possible to utilise all available modalities during training to create a stand-alone unimodal model that can provide an improved test-time performance compared to the case when only one modality is available for training. This challenge is addressed by a learning paradigm called Multimodal Co-Learning (MCL) [2, 43]. MCL algorithms work by cross-modal knowledge transfer, i.e; transferring knowledge from one modality to another. In this chapter, we explain our cross-modal knowledge transfer method, called Stronger Enhancing Weaker (SEW), that can be used to improve the test-time performance of a weaker modality by using a stronger modality during training phase alone. Here the notion of strength for a modality is based on its individual unimodal performance. We present two versions of the SEW method, one for non-sequential data and the other for sequential data. We explain the model design, architectural details and the loss functions used for training. We validate the method on continuous emotion recognition using the RECOLA [173] dataset and binary sentiment classification using CMU-MOSI [174] dataset.

5.2 Cross-Modal Knowledge Transfer Modelling

The term ‘knowledge’ in Cross-Modal Knowledge Transfer (CMKT) refers to the task-specific or useful information available in the features of a modality. In other words, it refers to the feature descriptiveness of a modality. As the information to explain any multimodal event is unevenly spread across all the modalities involved, some modalities might be more informative compared to the others [21, 23, 24, 50]. For example, in multimodal sentiment analysis research, it is often shown that text transcripts are more informative of sentiment compared to vision and acoustic modalities [50, 184]. Similarly, for continuous emotion recognition task, information about arousal or the level of activation is more evident from speech while valence or the level of pleasantness information is conveyed better by the vision (facial expressions) modality [23, 49, 201, 202, 203].

We create a novel framework for CMKT from a stronger modality to a weaker modality. Our CMKT framework consists of two main components, namely, cross-modal translation and latent-feature alignment. The intuition behind using cross-modal translation is that translating from one modality to another creates intermediate representations that capture joint information between both modalities [85, 204]. In our framework, we translate from the weaker to the stronger modality by using an encoder-decoder model. An explicit alignment between the intermediate and the stronger modality latent features further encourages the framework to discover components of the weaker modality that are maximally correlated with the stronger modality.

5.2.1 Modality ranking

For any given task and dataset, we first determine the strength of individual modalities by employing unimodal classifiers or regressors. This way, we can rank modalities according to their performance. The best and the least performing modalities can be called the strongest and the weakest respectively. Specifically, let a, v and t represent acoustic, visual and textual modalities, respectively. The sequence of features for modality $i \in \{a, v, t\}$ are given by

$M_i = [M_{i_1}, M_{i_2}, ..., M_{i_N}] \in \mathbb{R}^{N \times d_i}$, where d_i is the feature dimensionality and N is the sequence length. Let the corresponding labels, which are common for all the modalities, be represented by $Y = [Y_1, Y_2, ..., Y_N] \in \mathbb{R}^{N \times 1}$.

Let Γ^c (Γ^r) denote the unimodal classifier (regressor) with parameters θ_i for modality i. Let the performance score for each modality, e_i, be given by the evaluation metric $\mathcal{E}$ as

$$e_i = \mathcal{E}\left(\Gamma^c(M_i; \theta_i), Y\right). \tag{5.1}$$

Using this performance score, we can rank the modalities for a specific classification or regression task: a modality s is said to be stronger than modality w if $e_s > e_w$ (the opposite if $\mathcal{E}$ measures errors). Then our objective is to improve the task performance of feature M_w (the weaker modality) using the stronger modality during training.

It should be noted here that our objective is to improve the weaker modality using the stronger and not vice-versa. This is because we hypothesize the following. The reason for a modality to be stronger is because of its increased task-specific feature discriminativeness compared to the weaker modality. Hence, a weaker modality model can take up the role of a ‘student’ that learns to improve its ‘discriminative feature learning’ by using the stronger modality model as a ‘teacher’. However, the opposite case where the weaker modality model becomes a ‘teacher’ could result in negative knowledge transfer [21] and can cause performance deterioration of the stronger modality model. Thus, even though theoretically it could be plausible to design a system to control the effect of negative knowledge transfer, we consider such a case as a potential follow-up of our current objective. Another reason for choosing the stronger to enchance weaker modality model is that it is practically more useful since based on the margin of difference between their uni-modal performances, the former could improve the latter much more than vice-versa.

5.2.2 Auto-encoder based model for non-sequential data

For CMKT on non-sequential data, we use an auto-encoder based architecture. The framework, called Stronger Enhancing Weaker (SEW), employs a supervised neural network model that uses paired modality data during training. The key concepts of our framework are cross-modality translation and latent-feature alignment. These concepts are implemented using four main modules: a cross-modal translator, an intra-modal auto-encoder, a feature alignment module and a task-specific regressor or classifier (see Figure. 5.1). These modules are described as follows.

The cross-modal translator contains an encoder, W_E and a decoder, S_{D1}. The translator takes the features of the weaker modality, M_W, as input and produces the features of the stronger modality, $\hat{M}_{SW}$, as output. The encoder of the cross-modal translator, creates intermediate representations, m_{sw}, that capture joint information across modalities. This is achieved by using a translation loss, $\mathcal{L}_{tr}$, between the true, M_S, and the predicted, $\hat{M}_{SW}$, features of the stronger modality:

$$\mathcal{L}_1 = \mathcal{L}_{tr}(M_{\mathrm{S}}, \hat{M}_{\mathrm{SW}}). \tag{5.2}$$

W_E is encouraged to discover components of the weaker modality that are inclined towards the stronger modality by increasing the alignment between m_{sw} and the representations of the stronger modality. For this purpose, we project the stronger modality features into the same latent space as m_{sw}. We use an intra-modal auto-encoder to create stronger modality representations, m_{ss}, of the same dimensionality as that of the inter-modal translator representations, m_{sw}. To this end, we employ an auto-encoding loss, $\mathcal{L}_{ae}$, between the true, M_S, and the predicted, $\hat{M}_S$, features:

$$\mathcal{L}_2 = \mathcal{L}_{ae}(M_{\mathrm{S}}, \hat{M}_{\mathrm{S}}). \tag{5.3}$$

For modality reconstructions, we use Mean-Square-Error (MSE) as $\mathcal{L}_{tr}$ and $\mathcal{L}_{ae}$ [93].

A feature alignment loss, $\mathcal{L}_{al}$, ensures that the intermediate representations of the cross-modal translator are maximally aligned to the stronger modality representations:

$$\mathcal{L}_3 = \mathcal{L}_{al}(m_{\mathrm{ss}}, m_{\mathrm{sw}}). \tag{5.4}$$

Following [114, 95], we use Canonical Correlation Analysis (CCA) for feature alignment, such that $\mathcal{L}_{al}$ = -CCA. CCA for deep neural networks, also known as Deep CCA or DCCA, is a method to learn complex nonlinear transformations of data from two different modalities, such that the resulting representations are highly linearly correlated [205]. For a training set of size p, $M_s \in \mathbb{R}^{p \times d_s}$ and $M_w \in \mathbb{R}^{p \times d_w}$ are the input matrices corresponding to the stronger and the

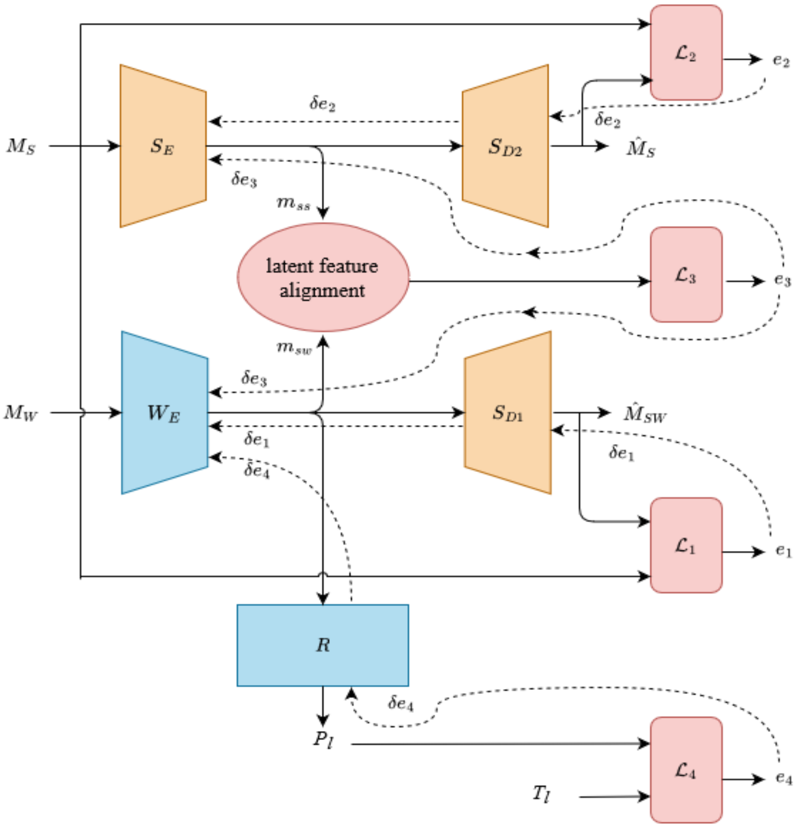


Figure 5.1: The proposed SEW training framework. (M_S,M_W) denotes a pair of stronger and weaker modality instances, S_E and S_{D2} represent intra-modal autoencoder, W_E and S_{D1} represent inter-modal translator, $\hat{M}_S$ and $\hat{M}_{SW}$ are the reconstructions of stronger modality from the encodings of stronger and weaker modalities respectively, R denotes the regressor/classifier connected to the inter-modal encoder, T_l and P_l stands for true and predicted labels, respectively, m_{ss} and m_{sw} represent the two latent representations, $\mathcal{L}_1$-$\mathcal{L}_4$ represent the 4 components of the total loss and e_1-e_4 are their respective error values. Dotted arrows represent the back-propagation of component error gradients. Only the blocks in cyan are retained during the deployment/inference phase.

weaker modalities, respectively. $m_{ss} \in \mathbb{R}^{p\times d}$ and $m_{sw} \in \mathbb{R}^{p\times d}$ are the representations obtained by nonlinear transformations introduced by the layers in the encoders S_E and W_E, respectively. Note that S_E and W_E bring the individual modalities with dimensions d_s and d_w into a common latent dimension d. If θ_{es} and θ_{ew} denote the vectors of all parameters of S_E and W_E, respectively, then the goal of DCCA is to jointly learn parameters for both the views such that correlation, (ρ), between m_{ss} and m_{sw} is as high as possible, i.e.,

$$\begin{aligned}(\theta_{es}^{\star}, \theta_{ew}^{\star}) &= \underset{\theta_{es},\theta_{ew}}{\arg\max}\, \rho(m_{ss}, m_{sw}) \\ &= \underset{\theta_{es},\theta_{ew}}{\arg\max}\, \rho(S_E(M_S;\theta_{es}), W_E(M_W;\theta_{ew})).\end{aligned} \tag{5.5}$$

If $\bar{m}_{ss}$ and $\bar{m}_{sw}$ are the mean-centred versions of m_{ss} and m_{sw}, respectively, then the total correlation of the top-K components of m_{ss} and m_{sw} is the sum of the top-K singular values of the matrix, $T = {\Sigma_s}^{-1/2}\Sigma_{sw}{\Sigma_w}^{-1/2}$, in which the self (Σ_s, Σ_w) and cross covariance (Σ_{sw}) matrices are given by

$$\Sigma_{sw} = \frac{1}{p-1}\bar{m}_{ss}\bar{m}_{sw}^T. \tag{5.6}$$

$$\Sigma_s = \frac{1}{p-1}\bar{m}_{ss}\bar{m}_{ss}^T + r_1 I. \tag{5.7}$$

$$\Sigma_w = \frac{1}{p-1}\bar{m}_{sw}\bar{m}_{sw}^T + r_2 I. \tag{5.8}$$

where $r_1 > 0$ and $r_2 > 0$ are regularisation constants. We use the gradient of correlation obtained on the training data to determine $(\theta_{es}^\star, \theta_{ew}^\star)$.

Finally, the task-specific regressor or classification module, which takes the cross-modal translator representations as input, ensures the discriminative ability of the resulting latent space. We use a prediction loss, $\mathcal{L}_{pr}$, that operates on the true, T_l, and predicted task labels, P_l, as:

$$\mathcal{L}_4 = \mathcal{L}_{pr}(T_l, P_l). \tag{5.9}$$

The total training loss, $\mathcal{L}$ combines the four components:

$$\mathcal{L} = \alpha\mathcal{L}_1 + \beta\mathcal{L}_2 + \gamma\mathcal{L}_3 + \mathcal{L}_4, \tag{5.10}$$

where α, β and γ are hyper-parameters. After training, all the components except the encoder, W_E, and the regressor, R, are removed and the stronger modality is not required at the testing (deployment) phase. The encoders, decoders as well as the classifiers of SEW are implemented using multi-layer perceptrons.

5.2.3 Recurrence based model for sequential data

We extend the SEW framework for sequential data by using recurrent networks and transformer encoders [113] for the encoders and recurrent networks for the decoders. We name the new framework as SeqSEW. SeqSEW is an improved version of SEW with the following changes.

- Unlike SEW, that uses an auto-encoder (encoder-decoder) model for creating the stronger modality representations, SeqSEW uses an encoder-classifier model that can create task-specific discriminative representations for the stronger modality. While the auto-encoder is trained to minimise the reconstruction loss and thereby recreate the input features, it is only encouraging the network to discover compressed representations of the input. Such representations need not be optimal in terms of task-specific feature discriminativeness. This motivates us to replace the auto-encoder model with an encoder-classifier model that is trained to map the stronger modality features to task-specific labels.

- Unlike the single step training in SEW, we split the training process into two steps in SeqSEW. In the first step, we train the encoder-classifier model for the stronger modality. The parameters of this model are then kept unchanged (or frozen) for the rest of the training process. In the second step, the encoder, decoder and classifier for the weaker modality are trained. This is based on the intuition that the stronger modality representations need not be changed, rather they have to be used only as a ‘guide’ or ‘reference’ for improving the discriminativeness of the weaker modality representations.

A detailed explanation of the SeqSEW model is as follows. We first create a model for the stronger modality, which acts as a ‘source’ of the knowledge to be transferred to another modality. This source model consists of an encoder and a classifier. The purpose of the encoder is to effectively model the information contained in the input stronger modality features and to map them into a latent space of desired dimensionality. The classifier ensures that the latent representations thus obtained are discriminative for the specific task.

In sequential data, every component in a sequence can have a dependence or correlation with neighbouring components in the same sequence [184, 186]. To model this inter-dependence across time, we use bi-GRU as the first layer of our encoder to transform the input feature sequence into ‘context aware’ representations, which are then input to a dense layer. The dense (fully connected) operation is shared across the time-steps, for projecting ‘context aware’ features onto a fixed dimension d. Even though the bi-GRU layers can capture contextual information via the hidden states, they cannot provide varying focus on which hidden states carry more valuable task-specific information. This can be accomplished by the use of self-attention mechanism that provides varying levels of attention weights to the time-steps in the same sequence.

For this purpose, we exploit the stacked self-attention mechanism using transformer encoder layers [113]. The use of Multi Head Attention (MHA), which contains multiple self-attention operations, allows the transformer to capture richer interpretations of the input sequence. Note that the transformer maintains same feature dimensionality (d) at the output of both MHA module as well as the dense layers to facilitate residual addition.

Additionally, for the transformer encoder layers to be aware of the temporal order of the input sequence, positional information in the form of sinusoidal position embeddings is added to the input of the first transformer encoder layer [113]. Similar to other multimodal machine learning works that utilise transformers [19, 206, 207], our intuition for using positional encoding for the features of all modalities (audio, vision and text) is that, similar to word embeddings, for audio/vision sequences it could be beneficial for the succeeding blocks of MHA modules to infer the order of their arrangement.

As output of the multi-layer transformer encoder, we obtain attention-weighted and context aware stronger modality representations of a desired dimensionality. In order to ensure that these representations are discriminative for the specific task, they are fed into a classifier made up of two dense layers. The first dense layer has ReLU activation whereas the second layer has Sigmoid or Tanh activation for classification or regression task respectively. The upper part of Figure.5.2 shows the encoder and classifier of source model as S_E and C_S. S_E takes a sequence of stronger modality features $M_S \in \mathbb{R}^{N \times d_s}$ as input and converts them into attention-weighted, context aware latent representations $m_{ss} \in \mathbb{R}^{N \times d}$. Note that the encoder has brought the feature dimension from d_s to d. The classifier C_S then takes m_{ss} as input and provides the predicted labels $\hat{Y} \in \mathbb{R}^{N \times 1}$.

We train the entire source model in an end-to-end fashion to map the input stronger modality features into the task-specific label space. Following [23, 50], we use Mean-Square-Error or Binary-Cross-Entropy for training depending upon whether the task is regression or binary classification respectively. If $\hat{Y}_n$ denotes the predicted label for the n^{th} time-step at the classifier output and Y_n represents the corresponding true label, then the prediction loss for regression and classification are given by,

$$\mathcal{L}_p = \frac{1}{N} \sum_{n=1}^{N} (Y_n - \hat{Y}_n)^2 \tag{5.11}$$

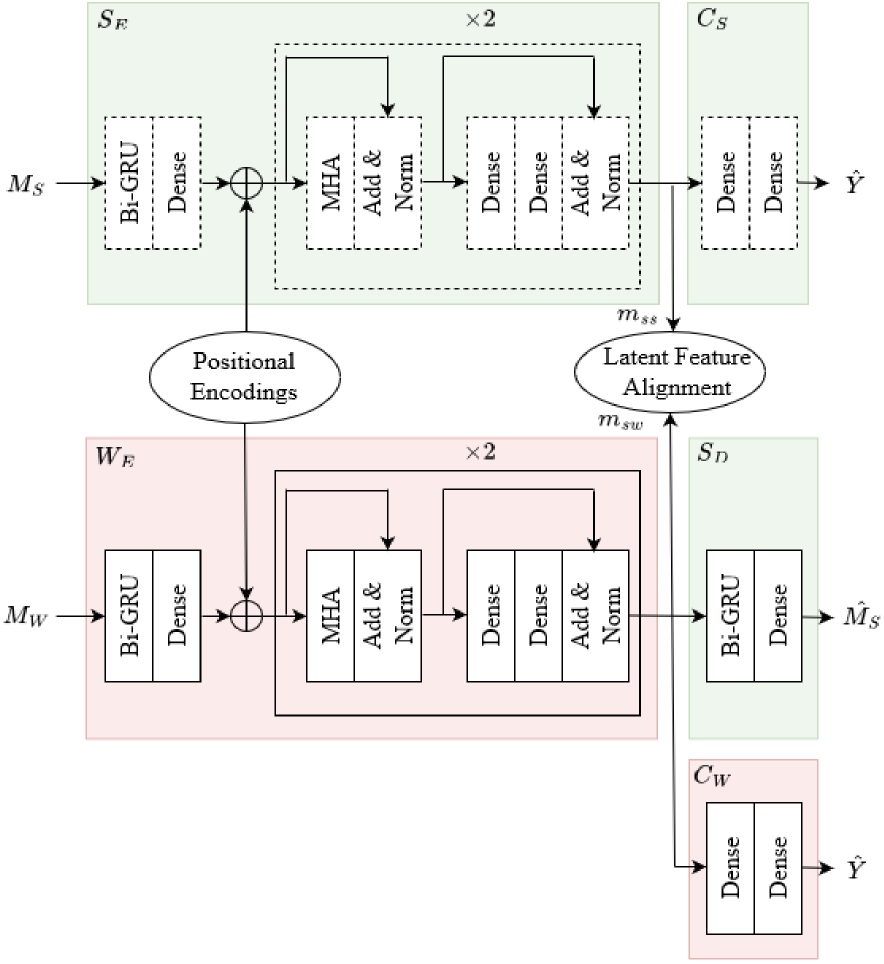


Figure 5.2: SeqSEW model. M_S and M_W are stronger and weaker modality inputs, $\hat{Y}$ denotes predicted labels, m_{ss} and m_{sw} are the two intermediate representations, $\hat{M}_S$ denotes the reconstructed stronger modality features. ×2 shows that two transformer encoder layers are used. Dashed lines indicate layers whose parameters are fixed. Only the blocks in red background (W_E and C_W) have to be retained after the training. KEY: MHA-multi-head attention, Add: addition, Norm: layer normalisation, Bi: bi-directional

and

$$\mathcal{L}_p = \frac{1}{N}\sum_{n=1}^{N}(Y_n.\log(\hat{Y}_n) + (1-Y_n).\log(1-\hat{Y}_n)) \tag{5.12}$$

respectively, where N represents the total number of segments in the sequence. Once the training is over, the source model parameters are fixed and only the encoded representations m_{ss} are retrieved for the next steps of knowledge transfer.

Similar to the source model, the weaker modality model also uses an encoder (W_E) and a classifier (C_W) to obtain attention-weighted, context aware and task-specific discriminative features. The encoder maps the weaker modality features into latent representations of same dimensionality d as the encoded representations from the source model. The encoder and classifier architecture are the same as the source model. Additionally, a decoder (S_D) is used to map the output of encoder to the stronger modality features. Thus, the encoder output is given as input to both the decoder as well as the classifier. To facilitate the sequential aspect of cross-modal translation, the decoder is made up of bi-GRU layers, which are followed by a dense layer. The dense operation is shared across the time-steps, for projecting the representations from each time step onto the same dimension as the stronger modality features.

Similar to auto-encoder based SEW, for latent feature alignment, SeqSEW uses Deep Canonical Correlation Analysis (DCCA) [205]. The encoder, decoder and classifier are jointly trained by optimising three objective functions: a translation loss between the decoder output and the stronger modality features, alignment loss between the encoder output and the stronger modality representations obtained from the source model and a task-specific prediction loss between the true and predicted labels. Following [85], we use Mean-Absolute-Error as cross-modal translation loss. If M_{s_n} and $\hat{M}_{s_n}$ denote the stronger modality features and the decoder output for segment n respectively, then, translation loss $\mathcal{L}_t$ is given by,

$$\mathcal{L}_t = \frac{1}{N}\sum_{n=1}^{N} |M_{s_n} - \hat{M}_{s_n}|. \tag{5.13}$$

For alignment loss $\mathcal{L}_a$, we use the negative of correlation obtained using DCCA. Thus,

$$\mathcal{L}_a = -corr(m_{ss}, m_{sw}). \tag{5.14}$$

Similar to the source model, we use Mean-Square-Error or Binary-Cross-Entropy as prediction loss $\mathcal{L}_p$ depending upon whether the task is regression or classification respectively. Thus, $\mathcal{L}_p$ is same as in eq. 5.11 and eq. 5.12. Hence the total training loss is given by,

$$\mathcal{L} = \mathcal{L}_p + \alpha\mathcal{L}_a + \beta\mathcal{L}_t, \tag{5.15}$$

where α, β are scalar weighting hyper-parameters. Once the training is over, only the encoder and classifier of the weaker modality model are to be retained for the inference or deployment phase.

Table 5.1: Unimodal results on the RECOLA [173] dataset. The best unimodal results in arousal and valence are highlighted in bold.

	Audio	Video-geo	Video-app
Arousal	**0.761**	0.482	0.492
Valence	0.543	**0.643**	0.489

KEY - CCC: Concordance Correlation Coefficient, geo: geometric, app: appearance.

5.3 Non Sequential Model: Validation

In this section, we evaluate our proposed auto-encoder model based SEW on the task of continuous emotion recognition using audio and vision modalities from the RECOLA [173] dataset. Note that the dataset contains recordings of 27 different speakers equally split into train (9 speakers), val/dev (9 speakers) and test (9 speakers). The train and val sets are publicly available and the test set labels are kept private by the dataset owners. Our experimental results on the non sequential model are reported on the publicly available val set, while our results on the sequential model are reported on both the publicly available val set and the held out test set (evaluation done by the owners of the dataset upon request). Note that since RECOLA [173] dataset has annotations for every time-step in the sequence, it can be considered and processed as a non-sequential data by processing each time-step in isolation. SEW framework that uses an MLP based auto-encoder does not take contextual information into account and uses RECOLA [173] dataset in a non-sequential manner.

5.3.1 Architecture

We use a regressor (see R in Figure. 5.1) similar to [23] and it consists of 4 single time-step GRU-RNN layers, each made up of 120 neurons, followed by a linear layer and trained using the MSE loss. Because the proposed method combines multimodal data with different characteristics, it was necessary to design various architectural parameters according to the characteristics of the given modalities rather than solving the problem using a generic model. Specifically, the encoder and decoder for each modality differ in terms of the number of linear layers and the number of neurons in each layer. Since the provided video-appearance features were already refined using PCA, we did not reduce the dimensionality further and used a single linear layer of size 168 for both its encoder and decoder. Thus, for all modality combinations that contain video-appearance features, the size of the latent dimension was 168. For all the rest, it was 128. The encoder

Table 5.2: Ablation results for SEW using RECOLA [173] in terms of CCC and binary classification accuracy.

	Arousal		Valence		
	vid-geo(+aud)	vid-app(+aud)	aud(+vid-geo)	vid-app(+aud)	vid-app(+vid-geo)
SEW	**0.565**	**0.544**	0.552	**0.554**	**0.549**
-S_{D2}	0.532	0.519	0.486	0.539	0.540
-CCA	0.512	0.508	0.496	0.532	0.546
-S_{D1}	0.514	0.523	**0.556**	0.514	0.505
-(CCA & S_{D1})	0.484	0.497	0.545	0.497	0.491
unimodal-weaker	0.482	0.492	0.543	0.489	0.489
unimodal-stronger	0.761	0.761	0.643	0.543	0.643

KEY - vid: video, aud: audio, geo: geometric, app: appearance.

and decoder for video-geometric features use linear layers of size [512, 256, 128] and [256, 512, 632], respectively with Tanh activation between layers. For audio features, these were [108, 128] and [108, 88]. Note that 632 and 88 were chosen to match the dimensionality of the video-geometric and audio features, respectively. All the models were developed, trained and tested using PyTorch. We used the SGD optimiser with learning rate 0.001, momentum 0.7 and weight decay regularisation. The batch size was 32. The number of CCA components, K, was 10 in all the experiments. The contribution of each loss component was found to be equally important (i.e, $\alpha = \beta = \gamma = 1$) via hyperparameter searching across values [1e-3,1e-2,0.1,1,10].

5.3.2 Results and ablations

In order to identify the stronger and weaker modalities, we first assess the unimodal performances of audio, video-geometric and video-appearance features for arousal and valence. The unimodal results obtained are shown in Table 5.1. For arousal, the performance of audio surpasses both video-geometric and video-appearance features. For valence, the video-geometric features outperform audio and video-appearance features. We have 5 cases for cross-modal knowledge transfer from stronger to weaker modalities, namely video-geo(+audio) and video-app(+audio) for arousal and audio(+video-geo), video-app(+audio) and video-app(+video-geo) for valence, where the modality in parenthesis indicates the stronger modality. Note that we have not considered the case of video-geo(+video-app) for arousal because of the insignificant difference between the unimodal scores.

Table 5.2 reports the results using the full SEW framework as well as after ablating individual components. The bottom two rows provides the unimodal results for the weaker and stronger modalities respectively for ease of comparison with the SEW results. We can see that the SEW-(CCA&S_{D1}) results are close to the unimodal results of the weaker modality. This is as expected since SEW-(CCA&S_{D1}) contains only the $\mathrm{W_E}$ and regressor with no interaction with the stronger modality. In all the 5 cases, SEW was able to improve the results from the unimodal and SEW-(CCA&S_{D1}) models both in terms of CCC and binary accuracy. For arousal video-geo(+audio) and video-app(+audio), removing the CCA based alignment causes a drop of 0.053 and 0.036, respectively in CCC. The corresponding number for valence audio(+video-geo) is 0.056. These observations support the significance of the CCA based feature alignment in the SEW framework. For valence video-app(+audio) and video-app(+video-geo), removing the decoder of the inter-modal translator causes a drop of 0.040 and 0.044, respectively in CCC, which indicates the effectiveness of the weaker-to-stronger modality translation.

5.3.3 Performance comparison with literature

In Table 5.3, we compare the best unimodal results of SEW with the 4 most relevant unimodal models [49, 208, 209, 210] and a cross-modal training method [23] in terms of CCC. [49] provides the baseline results on the RECOLA [173] dataset for the AVEC 2016 challenge. The unimodal baseline used an SVM based classifier on the individual features. SEW significantly outperforms the baseline unimodal results for all the weaker modalities considered. Our method is also able to improve the unimodal results for all the cases from [210], which uses difficulty awareness based training and [208, 209] which use multi-task learning. SEW outperforms EmoBed [23] for arousal video-geo(+audio) and valence audio(+video-geo) by a margin of 0.038 and 0.031, respectively, in CCC. For arousal video-app(+audio), the performance of SEW and EmoBed [23] are very close (0.544 and 0.549, respectively). However, for the valence video-app features, EmoBed [23] outperforms SEW. The top and bottom rows of Table 5.2 show that SEW improves the unimodal performance of the weaker modalities. Specifically, to the best of our knowledge, the best results to date on the task of unimodal arousal estimation using video-geometric features and on the task of unimodal valence estimation using audio features have been achieved by SEW.

Table 5.3: Performance comparison of SEW using RECOLA [173] with other methods in terms of CCC.

	Arousal		Valence	
	video-geo	video-app	audio	video-app
SVR + offset [49]	0.379	0.483	0.455	0.474
MTL (RE) [208]	0.502	0.512	0.519	0.529
MTL (PU) [209]	0.508	0.502	0.506	0.468
DDAT (RE) [210]	*0.544*	0.539	0.508	0.528
DDAT (PU) [210]	0.513	0.518	0.498	0.514
EmoBed [23]	0.527	**0.549**	*0.521*	**0.564**
SEW	**0.565**	*0.544*	**0.552**	*0.554*

KEY - geo: geometric, app: appearance. Best and second best results are shown in bold and italics, respectively.

5.4 Sequential Model: Validation

In this section, we evaluate our proposed recurrence based model SeqSEW on the task of continuous emotion recognition using audio and vision modalities and sentiment classification using audio, vision and language modalities. We use RECOLA [173] dataset for continuous emotion recognition and CMU-MOSI [51] for binary sentiment classification.

Next, we use RECOLA [173] in a sequential manner by considering a sequence of time-steps. In SeqSEW, for processing each time-step, we are using contextual information provided by its neighbours. Similar to [211], to ensure that sequences are long enough to capture the contextual information and to increase the number of training samples, we split the 9 recordings in RECOLA [173] training set by applying a sliding window of 3s with hop-size 1s. Thus, for each recording in training set, 299 clips are used, each including $N = 75$ time-steps. For the development set, we use non-overlapping 3s sequences. Similarly, CMU-MOSI [51] has labels for each time-step (also called utterance) in a sequence. Thus, instead of processing each time-step in isolation, SeqSEW uses the contextual information provided by neighbouring time-steps in the sequence.

5.4.1 Architecture

The network architecture for all SeqSEW models is kept generic except for a few modality-specific and dataset-specific differences. The front-end of the encoder consists of 2 bi-GRU layers, except for the audio models in RECOLA [173] and the text model in CMU-MOSI [51], where only a single layer bi-GRU is used. The number of neurons in each bi-GRU layer is equal to the input feature dimensionality. Dropout rates ($\{0.2,0.3,0.4,0.5,0.6,0.7\}$) are optimized for each model depending on the performance on validation set. The number of neurons in the dense layer succeeding the bi-GRU layers is 100 which is the dimension of the latent representations. This is because the transformer encoder layers do not change the feature dimensionality of their inputs [113]. Two transformer encoder layers are then employed with 2 self-attention operations in each transformer encoder layer. The first and second dense layers after the MHA modules in each transformer encoder layer have 400 and 100 neurons, respectively. Dropout regularisation is applied to the residual connections, first dense layer after MHA module as well as the attention weights in the transformer encoder to prevent overfitting.

For the decoder, we use a single layer bi-GRU for the RECOLA [173] valence audio model, while for all other models, we use 2 layers of bi-GRU. For RECOLA [173], the number of neurons in the decoder bi-GRU layers is kept 250 for arousal models and 500 for valence audio models. This is because for the latter, the decoder has to map the 100 dimensional latent features to higher dimensional video features (632 and 168). Finally, the number of neurons in the decoder dense layer is kept equal to the feature dimensionality of the stronger modality it is mapping into. The classifier module contains a ReLU activated input dense layer with 300 neurons and an output dense layer with a single neuron activated using Sigmoid or Tanh for classification or regression respectively. A dropout of rate of 0.3 is applied after its first dense layer to prevent overfitting. We use a batch size of 32 and the Adam optimizer [197] for training all the unimodal baselines and the SeqSEW models. We use a learning rate of 1e-4 and 1e-5 for experiments on CMU-MOSI [51] and RECOLA [173], respectively. We keep $r_1 = r_2 = 0.001$ in eq. 5.7-5.8 which is within the recommended range of [1e-8,10] [205]. The α and β values were empirically determined from the range [0.001, 1] using a grid search and values that gave the best validation set performance were retained.

5.4.2 Results and ablations

For sentiment classification, we assess the unimodal performance of each modality with a classifier only model (the 2 dense layer model C_S in Figure.5.2) and an encoder with classifier model (S_E - C_S in Figure.5.2). Results are reported in Table. 5.4. In accordance with previous works which found linguistic features to be more discriminative [50, 184, 186], we confirmed that the unimodal performance of textual features surpasses that of acoustic or visual. Thus, we consider two cases for stronger-to-weaker cross-modal knowledge transfer, namely, textual to acoustic and

Table 5.4: Unimodal baseline results on CMU-MOSI [51] for binary sentiment classification task.

	Method	*Acc.*	$\overline{F}_1$
Textual	Classifier	74.9	75.2
	Bi-GRU + Classifier	76.7	77.2
	Bi-GRU + TE + Classifier	80.3	80.1
Acoustic	Classifier	54.7	55.2
	Bi-GRU + Classifier	56.6	56.3
	Bi-GRU + TE + Classifier	61.0	60.0
Visual	Classifier	51.5	52.2
	Bi-GRU + Classifier	59.6	59.3
	Bi-GRU + TE + Classifier	60.8	59.7

KEY - *Acc.*: binary classification accuracy, $\overline{F}_1$: weighted F_1 score, TE: Transformer Encoder

textual to visual.

Similar to sentiment classification, for emotion regression, we assess the unimodal performance of acoustic, visual-geometric and visual-appearance features with a classifier only model (the 2 dense layer model C_S in Figure.5.2) and an encoder with classifier model (S_E - C_S in Figure. 5.2). Results are reported in Table. 5.5. For arousal, acoustic modality is stronger than visual modality whereas for valence, visual modality performs better than acoustic modality. These results are consistent with previous studies which found that acoustic and visual features are more discriminative for arousal and valence respectively [23, 49]. Hence, for stronger-to-weaker cross-modal knowledge transfer, we consider acoustic to visual-appearance and acoustic to visual-geometric for arousal and visual-appearance to acoustic and visual-geometric to acoustic for valence, respectively.

The results obtained using the two unimodal baseline models in Tables. 5.4 & 5.5 for both sentiment classification and emotion regression clearly indicate the performance improvement obtained by the addition of the encoder block to the classifier module. This validates our hypothesis that incorporating contextual information using the recurrence and attention mechanisms from the bi-GRU and transformer encoder can help in better understanding the underlying affective behaviour.

For the sentiment classification task, Table. 5.6 reports the results obtained using SeqSEW

Table 5.5: Unimodal baseline results in terms of Concordance Correlation Coefficient (CCC) on RECOLA [173] for continuous emotion regression task. Note that the range of CCC is [-1,1].

	Method	Acoustic eGeMAPS	Visual appearance	Visual geometric
Arousal	Classifier	0.769	0.517	0.470
	Bi-GRU + Classifier	0.773	0.530	0.488
	Bi-GRU + TE + Classifier	0.786	0.541	0.536
Valence	Classifier	0.490	0.496	0.560
	Bi-GRU + Classifier	0.521	0.542	0.594
	Bi-GRU + TE + Classifier	0.525	0.570	0.601

KEY - TE: Transformer Encoder

when the weaker acoustic and visual modalities are improved using textual modality during training. Compared to the best unimodal baseline results, SeqSEW improved the performance of both acoustic and visual models in terms of both accuracy (by 2.6 percentage points (pp) and 3.8pp for visual and acoustic respectively) and $\overline{F}_1$ (by 4.2pp and 4.3pp for visual and acoustic respectively) metrics. The ablation results are obtained by removing the cross-modal decoder (- Decoder) or by removing the latent feature alignment mechanism (- LFA). It can be seen that both the decoder as well as LFA contribute towards the SeqSEW knowledge transfer process, with the contribution of LFA component being slightly higher than the decoder component.

For the emotion regression task, Table. 5.7 shows the results obtained on both arousal and valence estimation. For arousal estimation, SeqSEW improves the performance of the unimodal baselines for both types of visual modality models. Specifically, an improvement of 0.033 and 0.050 are obtained for visual-appearance and visual-geometric models respectively, which are 6.1% and 9.3% of improvement over their best unimodal baselines. The ablation results indicate that both decoder as well as LFA contribute towards the SeqSEW method, with the contribution of decoder being slightly higher than the LFA component.

For valence estimation, Table. 5.7 shows that the performance of the weaker acoustic modality model could be improved using a visual-appearance modality model or a visual-geometric modality model. Specifically, we obtain an improvement of 0.035 and 0.030 when performing knowledge transfer from video-appearance and video-geometric models respectively, which are improvements of 6.7% and 5.7% over the unimodal acoustic baseline model (which in turn out-

Table 5.6: Results obtained using our proposed SeqSEW on CMU-MOSI [51] in terms of Binary Accuracy (*Acc.*) and Weighted F_1 ($\overline{F_1}$).* indicates results obtained using our evaluation on the publicly available codes.

Ref.	Method	Visual		Acoustic	
		Acc.	$\overline{F_1}$	*Acc.*	$\overline{F_1}$
[185]	CAT-LSTM-Uni	-	55.5	-	60.1
[185]	CAT-LSTM-Uni*	55.0	55.6	62.1	60.2
[186]	MU-SA	63.7	-	62.1	-
[186]	MU-SA*	62.8	61.9	59.7	58.4
[85]	Seq2SeqSent.-Uni	-	48.0	-	56.0
[50]	HMTL-Uni	62.1	61.3	58.2	58.2
-	our unimodal (Table. 5.4)	60.8	59.7	61.0	60.0
[85]	Seq2SeqSent.(+Textual)	-	58.0	-	56.0
[50]	HMTL(+Textual)	**64.8**	61.7	62.6	60.8
	SeqSEW(+Textual)	63.4	**63.9**	**64.8**	**64.3**
	- LFA	62.8	63.3	63.6	63.2
	- Decoder	63.0	63.4	64.1	63.5

performs other unimodal methods in the literature). Our ablation study shows that even though addition of both decoder and LFA components to the unimodal baseline model improve performance, the absence of the decoder provides similar results to the whole non-ablated system. We hypothesise that this might be due to the presence of zero frames (features corresponding to frames where the face detector failed). The decoder might be mapping multiple acoustic features to the same visual features (zeros) thus decreasing the discriminative ability of the intermediate features. Nevertheless, the results are comparable to those obtained with the full SeqSEW model.

Lastly, we used our best performing models to obtain predictions on a held-out test set and sent the results to the RECOLA [173] database administrators for evaluation. Results are shown in Tables. 5.8 & 5.9. Unimodal results on the held-out test set also shows that acoustic modality is the strongest for arousal estimation whereas it is the weakest for valence estimation. Application of SeqSEW has improved the performance of both visual features for arousal estimation. Similarly, visual features have been able to improve the test-time performance of acoustic model for the valence estimation task as well.

Table 5.7: Results obtained using our proposed SeqSEW on RECOLA [173] for Arousal and Valence predictions in terms of Concordance Correlation Coefficient (CCC).

Ref.	Method	Arousal		Valence
		Visual appearance	Visual geometric	Acoustic eGeMAPS
[49]	SVR + offset	0.379	0.483	0.455
[208]	MTL (RE)	0.502	0.512	0.519
[209]	MTL (PU)	0.508	0.502	0.506
[210]	DDAT (RE)	0.544	0.539	0.508
[210]	DDAT (PU)	0.513	0.518	0.498
-	our unimodal (Table. 5.5)	0.541	0.536	0.525
[23]	EmoBed(+Acoustic)	0.527	0.549	-
[23]	EmoBed(+Visual-app.)	-	-	0.514
[23]	EmoBed(+Visual-geo.)	-	-	0.521
Ours	SEW(+Acoustic)	0.565	0.544	-
Ours	SEW(+Visual-geo.)	-	-	0.552
	SeqSEW(+Acoustic)	**0.574**	**0.586**	-
	- LFA	0.571	0.580	-
	- Decoder	0.561	0.564	-
	SeqSEW(+Visual-app.)	-	-	0.560
	- LFA	-	-	0.538
	- Decoder	-	-	**0.563**
	SeqSEW(+Visual-geo.)	-	-	0.555
	- LFA	-	-	0.531
	- Decoder	-	-	**0.556**

Table 5.8: Results (sent by RECOLA [173] database administrators) obtained using our whole unimodal model, Bi-GRU + TE + Classifier, on RECOLA [173] held-out test set for Arousal and Valence predictions in terms of Concordance Correlation Coefficient (CCC).

	Acoustic eGeMAPS	Visual appearance	Visual geometric
Arousal	0.647	0.388	0.418
Valence	0.379	0.460	0.516

5.4.3 Performance comparison with literature

For both datasets, we considered only methods that reported results on the same set of features and the same dataset partition as we used.

For the sentiment classification task, from Table. 5.6 we observe that our unimodal baseline model, comprising the encoder and classifier, performs comparable to or better than four unimodal models from the literature. This makes the improvement provided by SeqSEW significant as the unimodal models are very competitive. Specifically, we compare the performance of SeqSEW on CMU-MOSI [51] using four unimodal models (CAT-LSTM-Uni [185], MU-SA [186], Seq2SeqSentiment-Uni [85], HMTL-Uni [50]) and two cross-modal knowledge transfer frameworks (Seq2SeqSentiment [85] and HMTL [50]). Since [185] and [186] did not report *Acc.* and $\overline{F}_1$ scores respectively, we used their publicly available codes to assess the performance. Considering the cross-modal knowledge transfer frameworks HMTL [50] and Seq2SeqSentiment(+Textual) [85], except for unimodal visual model accuracy, our models achieve better results, thus validating the effectiveness of knowledge transfer from the richer textual modality via the proposed methodology.

For arousal estimation experiments on RECOLA [173], from Table. 5.7, we can see that, with respect to other unimodal methods from the literature, our unimodal model provides better or comparable performance for both types of video features. This might be attributed to the fact that, unlike the compared models, our model takes contextual information into account. We compare with five unimodal models (SVR+offset [49], MTL(RE) [208], MTL(PU) [209], DDAT(RE) [210] and DDAT(PU) [210]) and two cross-modal knowledge transfer frameworks (EmoBed [23], SEW). Comparing our improved visual modality models with their counterparts obtained using other cross-modal knowledge transfer frameworks (EmoBed [23], SEW), we observe that our models perform better than both EmoBed [23] and SEW, thus validating the effectiveness of our knowledge transfer method. For valence estimation experiments

Table 5.9: Results (sent by RECOLA [173] database administrators) obtained using our proposed SeqSEW on RECOLA [173] held-out test set for Arousal and Valence predictions in terms of Concordance Correlation Coefficient (CCC).

Ref.	Method	Arousal		Valence
		Visual appearance	Visual geometric	Acoustic eGeMAPS
[49]	SVR + offset	0.343	0.272	0.375
[208]	MTL (RE)	0.425	0.324	0.331
[209]	MTL (PU)	0.406	0.327	0.416
[210]	DDAT (RE)	0.437	0.400	0.422
[210]	DDAT (PU)	0.438	0.397	0.407
-	our unimodal (Table. 5.8)	0.388	0.418	0.379
[23]	EmoBed(+Acoustic)	**0.475**	0.417	-
[23]	EmoBed(+Visual-app.)	-	-	0.434
[23]	EmoBed(+Visual-geo.)	-	-	0.439
	SeqSEW(+Acoustic)	0.434	**0.438**	-
	SeqSEW(+Visual-app.)	-	-	**0.465**
	SeqSEW(+Visual-geo.)	-	-	0.417

on RECOLA [173], (see Table. 5.7), comparison with the corresponding models from SEW and EmoBed [23], shows that our SeqSEW models achieve the best results in terms of CCC. Evaluation of our SeqSEW framework (Table. 5.9) on the held-out test set shows competitive performance with respect to the compared methods. Specifically, the SeqSEW enhanced video-geometric arousal model and acoustic valence model outperforms EmoBed [23] and achieves state-of-the-art results. Nevertheless, it should be noted that the results on the held-out test set is poorer compared to the development set, an observation which is consistent with the previous literature [23]. This could be attributed to the distribution mismatch between the development and test partitions.

5.5 Conclusion

An important paradigm in multimodal learning research is co-learning where multiple modalities can be present during training while only one modality is available during testing phase. Our research question was whether it is possible to design a framework that utilises multi-modal signals during training phase to develop a model that is intended to have a single modality input during test time for affective computing applications. To this end, we developed a framework called Stronger Enhancing Weaker (SEW) that uses a combination of cross-modal translation from weaker to stronger modality and correlation based latent feature alignment. The intuition behind using cross-modal translation is that translating from one modality to another creates intermediate representations that capture joint information between both modalities. Also, a correlation based alignment between the intermediate and the stronger modality latent features further encourages the framework to discover components of the weaker modality that are maximally correlated with the stronger modality. We created two versions of this framework, one for non-sequential data (SEW) and another for sequential data (SeqSEW). Our experiments on the SEW model achieve state-of-the-art results for test-time weaker modality models using the audio and vision modalities from the RECOLA dataset for continuous emotion recognition task. For SeqSEW, we modify the architecture of our model using components for sequential data processing, like recurrent layers and transformer encoders, and verify the resulting model on continuous emotion recognition using RECOLA and binary sentiment classification using CMU-MOSI. Comparison of our results with the state-of-the-art uni-modal models as well as two cross-modal knowledge transfer methods indicate that our models improve upon the state-of-the-art results on test-time weaker modality models.

Chapter 6

Conclusion and Future Works

In this chapter, we first provide a summary of our contributions in this thesis. Then, we provide directions for potential future works based on the limitations of existing literature.

6.1 Summary of Contributions

Our research so far has focused on developing unimodal and multimodal deep learning models for computational paralinguistics tasks like verbal conflict intensity estimation, emotion recognition and sentiment analysis.

- The first research contribution during this Ph.D. is an end-to-end convolutional-recurrent neural network called ConflictNET that provides an estimate of verbal conflict from raw speech waveforms of conversations between two or more people [3]. Such a system can have several applications like security and surveillance, providing feedback to call centre employees, helping journalists to navigate through long videos of political debates and identify instances and topics of strong disagreements etc. Apart from the performance evaluation of ConflictNET using multiple metrics (Pearson Correlation, Weighted and Unweighted Average Recall), a LIME [4] based explainability analysis was also done to 'open-up' the end-to-end model and understand what instances in the input speech the network was focusing on. We found that the network relied on relevant cues like speech overlaps and raised voice instances. Also, the network is prone to mistake other high energy segments like laughter, music or microphone tapping instances as conflicts.

- Next, we developed novel multimodal fusion models for an emotion classification task [5]. The involved modalities were speech (paralinguistics), vision (face images) and text (speech transcript). We used a convolutional-recurrent model with multi-head-attention [6] mechanism. Two versions of the model were created, one using intra-modality attention and another using inter (or cross) modality attention. While the former uses the same modality sequence to find task-relevant instances, the latter uses one modality to find task-relevant instances in the other modality's sequence. Extensive comparison between the models were done using uni-, bi- and tri-modal modality combinations in terms of weighted and unweighted accuracy metrics. Amongst the unimodal models, text modality performed the best indicating that explicit emotional cues can be obtained from the spoken words. Although both our intra- and inter-modal models improved upon the state-of-the-art results on IEMOCAP [7] dataset, our experiments did not indicate a clear edge for one attention mechanism over the other.

- Our final contribution is in multimodal co-learning, where multiple modalities can be present during the training phase but only one modality is available during the testing phase. Co-learning sits in between unimodal and multimodal fusion paradigms by leveraging the best of both worlds. Multimodal co-learning models aim to provide better performance compared to their unimodal counterparts. The unimodal performance of different modalities on the same task can vary and we can rank modalities according to their performance. The best performing modalities are called stronger and low performing ones are called weaker. We developed a novel co-learning framework called Stronger Enhancing Weaker (SEW) that uses a stronger modality as a 'helper' during training to improve the stand-alone test-time performance of a weaker modality [8]. SEW uses a combination of cross-modal translation and latent space alignment. This is based on the intuition that translation from one modality to another can create intermediate representations that are representative of both modalities. A latent space alignment of the weaker modality features with respect to the stronger modality features further enhances the feature discriminativeness of the weaker modality. SEW was validated on the task of audio-visual continuous emotion recognition and found to be effective in improving the unimodal performance of weaker modalities.

6.2 Future Works

Based on our experience so far, we identify some future directions to pursue in line with the research in this thesis.

- To the best of our knowledge, no work has explored the incorporation of text modality along with audio for the task of verbal conflict intensity estimation. The usage of explicit spoken words can aid the network in cases of passive aggressive statements and politer disagreements. It would be interesting to extend ConflictNET into a multimodal end-to-end network.

- Most of the multimodal co-learning methods proposed so far including ours (SEW & Seq-SEW) are limited to two modalities. However, real-world data can contain more than two modalities with rich information related to the task. It would be useful to devise strategies that can accommodate all the available modalities during the training phase instead of only two. An exception is MCTN [93] that uses a hierarchical strategy with $N-1$ number of training steps for N modalities present. The drawback in this case is that each step in the training phase is again limited to two modalities and the network is never exposed to discover the joint information present in all modalities together. More research is needed in this direction to address the limitations.

- One very interesting direction for future research on uni-modal as well as multimodal learning is based on explainable-AI.

  - Even though end-to-end models like ConflictNET eliminates the need for domain knowledge, they are difficult to interpret and it is unclear what features from the input data have been picked up by the model. Such a lack of clarity can be critical when analyzing from a privacy or fairness point of view. Even though we were able to gain some insights using LIME [31], it is not a foolproof method of analysis since the explanations are dependent on the choice of hyper-parameters (sample weights, number of samples used etc) for LIME. Exploration of Explainable AI algorithms suitable for speech based end-to-end models could be useful for the research community.
  - Another way towards model behaviour understanding would be to create inherently interpretable models. These models contain components that facilitate the interpretation of features. An example would be SincNet [212], a Convolutional Neural

Network (CNN) that encourages the first layer to discover more meaningful filters by exploiting parametrized sinc functions. The feature maps thus obtained are more human readable, clearly indicating the portions of spectrum that are learned for the task. It would be an interesting future direction to modify our unimodal and multimodal architectures using interpretable components like SincNet.

- One of the least addressed problems in multimodal learning is to develop model agnostic explanation methods for multimodal models. So far, most of the research community is focused on developing novel fusion mechanisms and they compare the fusion model performance with unimodal model variants as well as state-of-the-art fusion models. But the improvement obtained using the new fusion model can be attributed to several factors like better hyper-parameter selection, random seeds, the number of models compared. Ideally, a multimodal model should model unimodal combinations (UC) as well as multimodal interactions (MI). But it is not clear if the performance improvement is indeed due to MI and that the model is not just an ensemble of UCs. So, the question is '*Can we create a model-agnostic method that can explain whether a multimodal model utilises MI in addition to UC as well?*'. If we can create such a method, then we can analyse the state-of-the-art multimodal fusion models from this perspective and can provide an inference on what type of fusion mechanisms are effective in cross-modal interaction modelling. A work which addresses this problem is EMAP [9]. However, there are some limitations of this method. EMAP has been validated on VQA datasets. However, for emotion datasets, where modalities are highly correlated with each other, their empirical approximation of partial dependence of modalities might be invalid. Also, EMAP is validated for bi-modal datasets, and it doesn't scale well for tri-modal cases (the number of factors to be considered doubles). A proposed direction can be to explore statistical methods on functional decomposition of models to see how we can improve over EMAP.

Bibliography

[1] J. Zhao, X. Mao, and L. Chen, Speech Emotion Recognition using Deep 1D & 2D CNN LSTM Networks, *Biomedical Signal Processing and Control*, vol. 47, pp. 312–323, 2019.

[2] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, Multimodal Machine Learning: A Survey and Taxonomy, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[3] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, A Review of Affective Computing: From Unimodal Analysis to Multimodal Fusion, *Information Fusion*, vol. 37, pp. 98–125, 2017.

[4] Louis-Philippe Morency, Tadas Baltrusaitis, Tutorial on Multimodal Machine Learning. `https://www.cs.cmu.edu/~morency/MMML-Tutorial-ACL2017.pdf`. Online; accessed 01 September 2020.

[5] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Ng, Multimodal Deep Learning, in *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 689–696, ACM, 2011.

[6] G. Gosztolya, Conflict Intensity Estimation from Speech using Greedy Forward-Backward Feature Selection, in *Proceedings of the INTERSPEECH*, pp. 1339–1343, ISCA, 2015.

[7] G. Gosztolya and L. Tóth, DNN-based Feature Extraction for Conflict Intensity Estimation from Speech, *IEEE Signal Processing Letters*, vol. 24, no. 12, pp. 1837–1841, 2017.

[8] H. Kaya, T. Özkaptan, A. A. Salah, and F. Gürgen, Random Discriminative Projection based Feature Selection with Application to Conflict Recognition, *IEEE Signal Processing Letters*, vol. 22, no. 6, pp. 671–675, 2014.

[9] O. Räsänen and J. Pohjalainen, Random Subset Feature Selection in Automatic Recognition of Developmental Disorders, Affective States, and Level of Conflict from Speech., in *Proceedings of the INTERSPEECH*, pp. 210–214, ISCA, 2013.

[10] R. Brueckner and B. Schuller, Be at odds? Deep and Hierarchical Neural Networks for Classification and Regression of Conflict in Speech, in *Conflict and Multimodal Communication*, pp. 403–429, Springer, 2015.

[11] M.-J. Caraty and C. Montacié, Detecting Speech Interruptions for Automatic Conflict Detection, in *Conflict and Multimodal Communication*, pp. 377–401, Springer, 2015.

[12] F. Grezes, J. Richards, and A. Rosenberg, Let Me Finish: Automatic Conflict Detection using Speaker Overlap, in *Proceedings of the INTERSPEECH*, pp. 200–204, ISCA, 2013.

[13] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, Adieu Features? End-to-end Speech Emotion Recognition using a Deep Convolutional Recurrent Network, in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5200–5204, IEEE, 2016.

[14] P. Tzirakis, J. Zhang, and B. W. Schuller, End-to-end Speech Emotion Recognition using Deep Neural Networks, in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5089–5093, IEEE, 2018.

[15] J. Mao, S. Qiu, W. Wei, and H. He, Cross-modal guiding and reweighting network for multi-modal rsvp-based target detection, *Neural Networks*, 2023.

[16] H. R. V. Joze, A. Shaban, M. L. Iuzzolino, and K. Koishida, MMTM: Multimodal Transfer Module for CNN Fusion, in *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13289–13299, IEEE/CVF, 2020.

[17] J.-C. Martin, L. Devillers, A. Raouzaiou, G. Caridakis, Z. Ruttkay, C. Pelachaud, M. Mancini, R. Niewiadomski, H. Pirker, B. Krenn, *et al.*, Coordinating the Generation of Signs in Multiple Modalities in an Affective Agent, in *Emotion-Oriented Systems*, pp. 349–367, Springer, 2011.

[18] D. Hazarika, R. Zimmermann, and S. Poria, MISA: Modality-Invariant and Specific Representations for Multimodal Sentiment Analysis, in *Proceedings of the ACM International Conference on Multimedia*, pp. 1122–1131, 2020.

[19] 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*

Annual Meeting of the Association for Computational Linguistics (ACL), pp. 6558–6569, 2019.

[20] L. Sun, B. Liu, J. Tao, and Z. Lian, Multimodal Cross and Self-Attention Network for Speech Emotion Recognition, in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4275–4279, 2021.

[21] M. Abavisani, H. R. V. Joze, and V. M. Patel, Improving the Performance of Unimodal Dynamic Hand-gesture Recognition with Multimodal Training, in *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1165–1174, IEEE/CVF, 2019.

[22] G. Aguilar, V. Rozgic, W. Wang, and C. Wang, Multimodal and Multi-view Models for Emotion Recognition, in *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 991–1002, 2019.

[23] J. Han, Z. Zhang, Z. Ren, and B. Schuller, EmoBed: Strengthening Monomodal Emotion Recognition via Training with Crossmodal Emotion Embeddings, *IEEE Transactions on Affective Computing*, vol. 12, no. 3, pp. 553–564, 2019.

[24] A. Zadeh, P. P. Liang, and L.-P. Morency, Foundations of Multimodal Co-learning, *Information Fusion*, vol. 64, pp. 188–193, 2020.

[25] V. Vapnik, R. Izmailov, *et al.*, Learning Using Privileged Information: Similarity Control and Knowledge Transfer., *Journal of Machine Learning Research*, vol. 16, no. 1, pp. 2023–2049, 2015.

[26] N. C. Garcia, P. Morerio, and V. Murino, Learning with Privileged Information via Adversarial Discriminative Modality Distillation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 10, pp. 2581–2593, 2019.

[27] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, Explainable ai: A review of machine learning interpretability methods, *Entropy*, vol. 23, no. 1, p. 18, 2020.

[28] A. Graves and N. Jaitly, Towards end-to-end speech recognition with recurrent neural networks, in *International conference on machine learning*, pp. 1764–1772, PMLR, 2014.

[29] Y. Zhang, M. Pezeshki, P. Brakel, S. Zhang, C. Laurent, Y. Bengio, and A. Courville, Towards end-to-end speech recognition with deep convolutional neural networks, *Interspeech 2016*, pp. 410–414, 2016.

[30] Y. He, T. N. Sainath, R. Prabhavalkar, I. McGraw, R. Alvarez, D. Zhao, D. Rybach, A. Kannan, Y. Wu, R. Pang, *et al.*, Streaming end-to-end speech recognition for mobile devices, in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6381–6385, IEEE, 2019.

[31] M. T. Ribeiro, S. Singh, and C. Guestrin, Why Should I Trust You? Explaining the Predictions of any Classifier, in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144, 2016.

[32] Lilian Weng, Attention? Attention!. https://lilianweng.github.io/posts/2018-06-24-attention/. Online; accessed 01 June 2022.

[33] S. Yoon, S. Dey, H. Lee, and K. Jung, Attentive Modality Hopping Mechanism for Speech Emotion Recognition, in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3362–3366, IEEE, 2020.

[34] M. Ren, X. Huang, X. Shi, and W. Nie, Interactive multimodal attention network for emotion recognition in conversation, *IEEE Signal Processing Letters*, vol. 28, pp. 1046–1050, 2021.

[35] W. Chen, X. Xing, X. Xu, J. Yang, and J. Pang, Key-sparse transformer for multimodal speech emotion recognition, in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6897–6901, IEEE, 2022.

[36] Y. Wang, J. Wu, P. Heracleous, S. Wada, R. Kimura, and S. Kurihara, Implicit knowledge injectable cross attention audiovisual model for group emotion recognition, in *Proceedings of the 2020 International Conference on Multimodal Interaction*, pp. 827–834, 2020.

[37] N.-H. Ho, H.-J. Yang, S.-H. Kim, and G. Lee, Multimodal approach of speech emotion recognition using multi-level multi-head fusion attention-based recurrent neural network, *IEEE Access*, vol. 8, pp. 61672–61686, 2020.

[38] S. Dutta and S. Ganapathy, Multimodal transformer with learnable frontend and self attention for emotion recognition, in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6917–6921, IEEE, 2022.

[39] E. Ghaleb, J. Niehues, and S. Asteriadis, Multimodal attention-mechanism for temporal emotion recognition, in *2020 IEEE International Conference on Image Processing (ICIP)*, pp. 251–255, IEEE, 2020.

[40] V. John and Y. Kawanishi, Audio and video-based emotion recognition using multimodal transformers, in *2022 26th International Conference on Pattern Recognition (ICPR)*, pp. 2582–2588, IEEE, 2022.

[41] D. Priyasad, T. Fernando, S. Denman, S. Sridharan, and C. Fookes, Attention driven fusion for multi-modal emotion recognition, in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3227–3231, IEEE, 2020.

[42] J. Yoon, C. Kang, S. Kim, and J. Han, D-vlog: Multimodal vlog dataset for depression detection, in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, pp. 12226–12234, 2022.

[43] A. Rahate, R. Walambe, S. Ramanna, and K. Kotecha, Multimodal Co-learning: Challenges, Applications with Datasets, Recent Advances and Future Directions, *Information Fusion*, vol. 81, pp. 203–239, 2022.

[44] J. Hoffman, S. Gupta, and T. Darrell, Learning with Side Information through Modality Hallucination, in *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 826–834, IEEE/CVF, 2016.

[45] N. C. Garcia, P. Morerio, and V. Murino, Modality Distillation with Multiple Stream Networks for Action Recognition, in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 103–118, Springer, 2018.

[46] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, IEMOCAP: Interactive Emotional Dyadic Motion Capture Database, *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[47] N. Neverova, C. Wolf, G. Taylor, and F. Nebout, ModDrop: Adaptive Multi-Modal Gesture Recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 8, pp. 1692–1706, 2015.

[48] S. Moon, S. Kim, and H. Wang, Multimodal Transfer Deep Learning with Applications in Audio-Visual Recognition, in *Proceedings of the Multi Modal Machine Learning Workshop at Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

[49] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, AVEC 2016: Depression, Mood, and Emotion Recognition Workshop and Challenge, in *Proceedings of the International Workshop on Audio/Visual Emotion Challenge*, pp. 3–10, 2016.

[50] S. Seo, S. Na, and J. Kim, HMTL: Heterogeneous Modality Transfer Learning for Audio-Visual Sentiment Analysis, *IEEE Access*, vol. 8, pp. 140426–140437, 2020.

[51] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, Multimodal Sentiment Intensity Analysis in Videos: Facial Gestures and Verbal Messages, *IEEE Intelligent Systems*, vol. 31, no. 6, pp. 82–88, 2016.

[52] B. Yuhas, M. Goldstein, and T. Sejnowski, Integration of Acoustic and Visual Speech Signals using Neural Networks, *IEEE Communications Magazine*, vol. 27, no. 11, pp. 65–71, 1989.

[53] A. Pompili, T. Rolland, and A. Abad, The INESC-ID Multi-Modal System for the ADReSS 2020 Challenge, *Proceedings of the INTERSPEECH*, pp. 2202–2206, 2020.

[54] J.-W. Hwang, R.-H. Park, and H.-M. Park, Efficient Audio-Visual Speech Enhancement Using Deep U-Net With Early Fusion of Audio and Video Information and RNN Attention Blocks, *IEEE Access*, vol. 9, pp. 137584–137598, 2021.

[55] Y. Ito, T. Ogawa, and M. Haseyama, Personalized Video Preference Estimation based on Early Fusion using Multiple Users Viewing Behavior, in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3006–3010, IEEE, 2017.

[56] Y. Sun, J. F. Gemmeke, B. Cranen, L. ten Bosch, and L. Boves, Early Fusion of Sparse Classification and GMM for Noise Robust ASR, in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, pp. 1495–1499, IEEE, 2011.

[57] L. Pepino, P. Riera, L. Ferrer, and A. Gravano, Fusion Approaches for Emotion Recognition from Speech using Acoustic and Text-based Features, in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6484–6488, IEEE, 2020.

[58] A. Jaimes and N. Sebe, Multimodal Human-Computer Interaction: A Survey, *Computer Vision and Image Understanding*, vol. 108, no. 1-2, pp. 116–134, 2007.

[59] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, Multimodal Fusion for Multimedia Analysis: A Survey, *Multimedia Systems*, vol. 16, no. 6, pp. 345–379, 2010.

[60] Q. Jin, C. Li, S. Chen, and H. Wu, Speech Emotion Recognition with Acoustic and Lexical Features, in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4749–4753, IEEE, 2015.

[61] H. Ben-Younes, R. Cadene, M. Cord, and N. Thome, MUTAN: Multimodal Tucker Fusion for Visual Question Answering, in *Proceedings of the International Conference on Computer Vision (ICCV)*, pp. 2612–2620, IEEE, 2017.

[62] P. Natarajan, S. Wu, S. Vitaladevuni, X. Zhuang, S. Tsakalidis, U. Park, R. Prasad, and P. Natarajan, Multimodal Feature Fusion for Robust Event Detection in Web Videos, in *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1298–1305, IEEE/CVF, 2012.

[63] P. Tzirakis, J. Chen, S. Zafeiriou, and B. Schuller, End-to-end Multimodal Affect Recognition in Real-World Environments, *Information Fusion*, vol. 68, pp. 46–53, 2021.

[64] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, End-to-end Multimodal Emotion Recognition using Deep Neural Networks, *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1301–1309, 2017.

[65] S. Yoon, S. Byun, and K. Jung, Multimodal Speech Emotion Recognition using Audio and Text, in *Proceedings of the Spoken Language Technology Workshop*, pp. 112–118, IEEE, 2018.

[66] G. Shen, R. Lai, R. Chen, Y. Zhang, K. Zhang, Q. Han, and H. Song, WISE: Word-Level Interaction-Based Multimodal Fusion for Speech Emotion Recognition., in *Proceedings of the INTERSPEECH*, pp. 369–373, ISCA, 2020.

[67] P. Veličković, D. Wang, N. D. Lane, and P. Liò, X-CNN: Cross-modal Convolutional Neural Networks for Sparse Datasets, in *Proceedings of the Symposium Series on Computational Intelligence (SSCI)*, pp. 1–8, IEEE, 2016.

[68] C. Cangea, P. Veličković, and P. Lio, XFlow: Cross-modal Deep Neural Networks for Audiovisual Classification, *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 9, pp. 3711–3720, 2019.

[69] V. Vielzeuf, A. Lechervy, S. Pateux, and F. Jurie, CentralNet: A Multilayer Approach for Multimodal Fusion, in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pp. 575–589, Springer, 2018.

[70] D. Hu, C. Wang, F. Nie, and X. Li, Dense Multimodal Fusion for Hierarchically Joint Representation, in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3941–3945, IEEE, 2019.

[71] J.-M. Pérez-Rúa, V. Vielzeuf, S. Pateux, M. Baccouche, and F. Jurie, MFAS: Multimodal Fusion Architecture Search, in *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6966–6975, IEEE/CVF, 2019.

[72] L. Su, C. Hu, G. Li, and D. Cao, Msaf: Multimodal split attention fusion, *arXiv preprint arXiv:2012.07175*, 2020.

[73] J. Hu, L. Shen, and G. Sun, Squeeze-and-Excitation Networks, in *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7132–7141, IEEE/CVF, 2018.

[74] F. Hutter, H. H. Hoos, and K. Leyton-Brown, Sequential Model-based Optimization for General Algorithm Configuration, in *Proceedings of the International Conference on Learning and Intelligent Optimization*, pp. 507–523, Springer, 2011.

[75] C. Liu, B. Zoph, M. Neumann, J. Shlens, W. Hua, L.-J. Li, L. Fei-Fei, A. Yuille, J. Huang, and K. Murphy, Progressive Neural Architecture Search, in *Proceedings of the European conference on computer vision (ECCV)*, pp. 19–34, Springer, 2018.

[76] J. Arevalo, T. Solorio, M. Montes-y Gómez, and F. A. González, Gated Multimodal Units for Information Fusion, in *Proceedings of the International Conference on Learning Representations (ICLR), Workshop Track*, 2017.

[77] Y. Wang, W. Huang, F. Sun, T. Xu, Y. Rong, and J. Huang, Deep Multimodal Fusion by Channel Exchanging, in *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 4835–4845, 2020.

[78] T.-Y. Lin, A. RoyChowdhury, and S. Maji, Bilinear CNN Models for Fine-grained Visual Recognition, in *Proceedings of the International Conference on Computer Vision (ICCV)*, pp. 1449–1457, IEEE, 2015.

[79] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell, Compact Bilinear Pooling, in *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 317–326, IEEE/CVF, 2016.

[80] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, Tensor Fusion Network for Multimodal Sentiment Analysis, in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1103–1114, 2017.

[81] K. Liu, Y. Li, N. Xu, and P. Natarajan, Learn to Combine Modalities in Multimodal Deep Learning, *arXiv preprint arXiv:1805.11730*, 2018.

[82] D. Bahdanau, K. Cho, and Y. Bengio, Neural machine translation by jointly learning to align and translate, in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

[83] C. Hori, T. Hori, T.-Y. Lee, Z. Zhang, B. Harsham, J. R. Hershey, T. K. Marks, and K. Sumi, Attention-based Multimodal Fusion for Video Description, in *Proceedings of the International Conference on Computer Vision (ICCV)*, pp. 4193–4202, IEEE, 2017.

[84] Y.-H. H. Tsai, P. P. Liang, A. Zadeh, L.-P. Morency, and R. Salakhutdinov, Learning Factorized Multimodal Representations, in *Proceedings of the International Conference on Learning Representations*, 2018.

[85] H. Pham, T. Manzini, P. P. Liang, and B. Póczos, Seq2seq2sentiment: Multimodal Sequence to Sequence Models for Sentiment Analysis, in *Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML)*, pp. 53–63, 2018.

[86] A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria, and L.-P. Morency, Memory Fusion Network for Multi-View Sequential Learning, in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, pp. 5634–5641, 2018.

[87] A. Zadeh and P. Pu, Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph, in *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 2236–2246, 2018.

[88] A. Zadeh, P. P. Liang, S. Poria, P. Vij, E. Cambria, and L.-P. Morency, Multi-Attention Recurrent Network for Human Communication Comprehension, in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, pp. 5642–5649, 2018.

[89] Y. Wang, Y. Shen, Z. Liu, P. P. Liang, A. Zadeh, and L.-P. Morency, Words Can Shift: Dynamically Adjusting Word Representations using Nonverbal Behaviors, in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 7216–7223, 2019.

[90] P. P. Liang, Z. Liu, A. B. Zadeh, and L.-P. Morency, Multimodal Language Analysis with Recurrent Multistage Fusion, in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 150–161, ACL, 2018.

[91] Y. Gu, K. Yang, S. Fu, S. Chen, X. Li, and I. Marsic, Multimodal Affective Analysis using Hierarchical Attention Strategy with Word-Level Alignment, in *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 2225–2235, 2018.

[92] S. Albanie, A. Nagrani, A. Vedaldi, and A. Zisserman, Emotion Recognition in Speech using Cross-Modal Transfer in the Wild, in *Proceedings of the ACM International Conference on Multimedia*, pp. 292–301, 2018.

[93] H. Pham, P. P. Liang, T. Manzini, L.-P. Morency, and B. Póczos, Found in Translation: Learning Robust Joint Representations by Cyclic Translations between Modalities, in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 6892–6899, 2019.

[94] J. Han, Z. Zhang, Z. Ren, and B. Schuller, Implicit Fusion by Joint Audiovisual Training for Emotion Recognition in Mono Modality, in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5861–5865, IEEE, 2019.

[95] S. H. Dumpala, I. Sheikh, R. Chakraborty, and S. K. Kopparapu, Audio-Visual Fusion for Sentiment Classification using Cross-Modal Autoencoder, in *Proceedings of the Vigil workshop at Advances in Neural Information Processing Systems (NeurIPS)*, pp. 1–4, 2018.

[96] A. Shukla, S. Petridis, and M. Pantic, Does Visual Self-Supervision Improve Learning of Speech Representations for Emotion Recognition, *IEEE Transactions on Affective Computing*, pp. 1–1, 2021.

[97] G. Chechik, V. Sharma, U. Shalit, and S. Bengio, Large scale online learning of image similarity through ranking., *Journal of Machine Learning Research*, vol. 11, no. 3, 2010.

[98] F. Schroff, D. Kalenichenko, and J. Philbin, Facenet: A unified embedding for face recognition and clustering, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015.

[99] I. Goodfellow, NeurIPS 2016 Tutorial: Generative Adversarial Networks, *arXiv preprint arXiv:1701.00160*, 2016.

[100] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, VGGFace2: A Dataset for Recognising Faces Across Pose and Age, in *Proceedings of the International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pp. 67–74, IEEE, 2018.

[101] E. Barsoum, C. Zhang, C. C. Ferrer, and Z. Zhang, Training Deep Networks for Facial Expression Recognition with Crowd-Sourced Label Distribution, in *Proceedings of the ACM International Conference on Multimodal Interaction*, pp. 279–283, 2016.

[102] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, Return of the Devil in the Details: Delving Deep into Convolutional Nets, in *Proceedings of the British Machine Vision Conference (BMVC)*, BMVA Press, 2014.

[103] A. Nagrani, J. S. Chung, and A. Zisserman, VoxCeleb: A Large-Scale Speaker Identification Dataset, *Proceedings of the INTERSPEECH*, pp. 2616–2620, 2017.

[104] O. Ronneberger, P. Fischer, and T. Brox, U-net: Convolutional Networks for Biomedical Image Segmentation, in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241, Springer, 2015.

[105] D. Bahdanau, K. H. Cho, and Y. Bengio, Neural Machine Translation by Jointly Learning to Align and Translate, in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

[106] I. Sutskever, O. Vinyals, and Q. V. Le, Sequence to Sequence Learning with Neural Networks, *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, vol. 27, pp. 3104–3112, 2014.

[107] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, Show, Attend and Tell: Neural Image Caption Generation with Visual Attention, in *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 2048–2057, PMLR, 2015.

[108] G. Zheng, S. Mukherjee, X. L. Dong, and F. Li, OpenTag: Open Attribute Value Extraction from Product Profiles, in *Proceedings of the International Conference on Knowledge Discovery & Data Mining*, pp. 1049–1058, 2018.

[109] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, Attention-based Models for Speech Recognition, *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 28, pp. 577–585, 2015.

[110] S. Mirsamadi, E. Barsoum, and C. Zhang, Automatic Speech Emotion Recognition using Recurrent Neural Networks with Local Attention, in *Proceedings of the International conference on acoustics, speech and signal processing (ICASSP)*, pp. 2227–2231, IEEE, 2017.

[111] A. Pankajakshan, H. L. Bear, V. Subramanian, and E. Benetos, Memory Controlled Sequential Self Attention for Sound Recognition, *Proceedings of the INTERSPEECH*, pp. 831–835, 2020.

[112] M.-T. Luong, H. Pham, and C. D. Manning, Effective Approaches to Attention-based Neural Machine Translation, in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1412–1421, 2015.

[113] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, Attention is All You Need, *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, pp. 5998–6008, 2017.

[114] Z. Sun, P. K. Sarma, W. Sethares, and E. P. Bucy, Multi-Modal Sentiment Analysis using Deep Canonical Correlation Analysis, in *Proceedings of the INTERSPEECH*, pp. 1323–1327, ISCA, 2019.

[115] R. Picard, *Affective Computing*. Inteligencia artificial, Cambridge, Mass., 1997.

[116] R. Cowie, E. Douglas-Cowie, and M. Schröder, Speech and Emotion, ISCA Tutorial and Research Workshop (ITRW), 09 2000.

[117] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, *et al.*, The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social signals, Conflict, Emotion, Autism, in *Proceedings of the INTERSPEECH*, pp. 148–152, ISCA, 2013.

[118] B. Schuller and A. Batliner, *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*. John Wiley & Sons, 2013.

[119] C. Darwin, The expression of the emotions in man and animals, in *The expression of the emotions in man and animals*, University of Chicago press, 2015.

[120] C. A. Corneanu, M. O. Simón, J. F. Cohn, and S. E. Guerrero, Survey on RGB, 3D, Thermal, and Multimodal Approaches for Facial Expression Recognition: History, Trends, and Affect-Related Applications, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 8, pp. 1548–1568, 2016.

[121] D. Canedo and A. J. Neves, Facial Expression Recognition using Computer Vision: A Systematic Review, *Applied Sciences*, vol. 9, no. 21, p. 4678, 2019.

[122] L. De Silva, T. Miyasato, and R. Nakatsu, Facial Emotion Recognition using Multi-modal Information, in *Proceedings of the International Conference on Information, Communications and Signal Processing*, vol. 01, pp. 397–401, 1997.

[123] J. A. Wall Jr and R. R. Callister, Conflict and its Management, *Journal of management*, vol. 21, no. 3, pp. 515–558, 1995.

[124] C. M. Judd, Cognitive Effects of Attitude Conflict Resolution, *Journal of Conflict Resolution*, vol. 22, no. 3, pp. 483–498, 1978.

[125] V. W. Cooper, Participant and Observer Attribution of Affect in Interpersonal Conflict: An Examination of Noncontent Verbal Behavior, *Journal of Nonverbal Behavior*, vol. 10, no. 2, pp. 134–144, 1986.

[126] R. F. Baumeister, A. Stillwell, and S. R. Wotman, Victim and Perpetrator Accounts of Interpersonal Conflict: Autobiographical Narratives about Anger., *Journal of Personality and Social Psychology*, vol. 59, no. 5, p. 994, 1990.

[127] P. E. Spector and S. M. Jex, Development of Four Self-Report Measures of Job Stressors and Strain: Interpersonal Conflict at Work Scale, Organizational Constraints Scale, Quantitative Workload Inventory, and Physical Symptoms Inventory., *Journal of Occupational Health Psychology*, vol. 3, no. 4, p. 356, 1998.

[128] S. Kim, F. Valente, M. Filippone, and A. Vinciarelli, Predicting Continuous Conflict Perception with Bayesian Gaussian Processes, *IEEE Transactions on Affective Computing*, vol. 5, no. 2, pp. 187–200, 2014.

[129] D.-Y. Huang, H. Li, and M. Dong, Ensemble Nyström Method for Predicting Conflict Level from Speech, in *Proceedings of Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pp. 1–5, 2014.

[130] F. Eyben, M. Wöllmer, and B. Schuller, OpenSMILE: the Munich Versatile and Fast Open-Source Audio Feature Extractor, in *Proceedings of the ACM International Conference on Multimedia*, pp. 1459–1462, 2010.

[131] R. Vereecken, S. Petridis, Y. Panagakis, and M. Pantic, Online Attention for Interpretable Conflict Estimation in Political Debates, in *Proceedings of the International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pp. 389–393, IEEE, 2018.

[132] C. Segura, D. Balcells, M. Umbert, J. Arias, and J. Luque, Automatic Speech Feature Learning for Continuous Prediction of Customer Satisfaction in Contact Center Phone Calls, in *Proceedings of the International Conference on Advances in Speech and Language Technologies for Iberian Languages*, pp. 255–265, 2016.

[133] M. El Ayadi, M. S. Kamel, and F. Karray, Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases, *Pattern recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[134] B. W. Schuller, Speech Emotion Recognition: Two Decades in a Nutshell, Benchmarks, and Ongoing Trends, *Communications of the ACM*, vol. 61, no. 5, pp. 90–99, 2018.

[135] M. B. Akçay and K. Oğuz, Speech Emotion Recognition: Emotional Models, Databases, Features, Preprocessing Methods, Supporting Modalities, and Classifiers, *Speech Communication*, vol. 116, pp. 56–76, 2020.

[136] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, Emotion Recognition in Human-Computer Interaction, *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[137] P. Ekman, An Argument for Basic Emotions, *Cognition & emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[138] C. M. Lee, S. S. Narayanan, and R. Pieraccini, Combining Acoustic and Language Information for Emotion Recognition, in *Proceedings of the INTERSPEECH*, ISCA, 2002.

[139] M. Kotti, F. Paternò, and C. Kotropoulos, Speaker-Independent Negative Emotion Recognition, in *Proceedings of the International Workshop on Cognitive Information Processing*, pp. 417–422, 2010.

[140] W. Kim and J. H. Hansen, Angry Emotion Detection from Real-Life Conversational Speech by Leveraging Content Structure, in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5166–5169, IEEE, 2010.

[141] S.-a. Yoon, G. Son, and S. Kwon, Fear Emotion Classification in Speech by Acoustic and Behavioral Cues, *Multimedia Tools and Applications*, vol. 78, no. 2, pp. 2345–2366, 2019.

[142] S. Yoon, S. Byun, S. Dey, and K. Jung, Speech Emotion Recognition using Multi-hop Attention Mechanism, in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2822–2826, IEEE, 2019.

[143] Z. Aldeneh and E. M. Provost, Using Regional Saliency for Speech Emotion Recognition, in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2741–2745, IEEE, 2017.

[144] P. Li, Y. Song, I. McLoughlin, W. Guo, and L. Dai, An Attention Pooling Based Representation Learning Method for Speech Emotion Recognition, *Proceedings of the INTERSPEECH*, pp. 3087–3091, 2018.

[145] M. Lech, M. Stolar, C. Best, and R. Bolia, Real-time Speech Emotion Recognition using A Pre-trained Image Classification Network: Effects of Bandwidth Reduction and Companding, *Frontiers in Computer Science*, vol. 2, p. 14, 2020.

[146] M. A. Martin, S. Shah, G. S. Charith, N. Singh, and D. Bhulakshmi, Automatic Speech Emotion Recognition using Machine Learning, *International Journal of Advanced Research in Computer Science*, vol. 12, pp. 101–106, 2021.

[147] B. Schuller, G. Rigoll, and M. Lang, Hidden Markov Model-based Speech Emotion Recognition, in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, pp. II–1, IEEE, 2003.

[148] T. L. Nwe, S. W. Foo, and L. C. De Silva, Speech Emotion Recognition using Hidden Markov Models, *Speech communication*, vol. 41, no. 4, pp. 603–623, 2003.

[149] S. Yun and C. D. Yoo, Loss-Scaled Large-Margin Gaussian Mixture Models for Speech Emotion Classification, *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 585–598, 2011.

[150] D. Ververidis and C. Kotropoulos, Emotional Speech Classification using Gaussian Mixture Models, in *Proceedings of the International Symposium on Circuits and Systems*, vol. 03, pp. 2871–2874, IEEE, 2005.

[151] H. Hu, M.-X. Xu, and W. Wu, Gmm Supervector based SVM with Spectral Features for Speech Emotion Recognition, in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4, pp. IV–413, IEEE, 2007.

[152] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, Learning Salient Features for Speech Emotion Recognition using Convolutional Neural Networks, *IEEE Transactions on Multimedia*, vol. 16, no. 8, pp. 2203–2213, 2014.

[153] R. B. Lanjewar, S. Mathurkar, and N. Patel, Implementation and Comparison of Speech Emotion Recognition System using Gaussian Mixture Model (gmm) and K-Nearest Neighbor (K-NN) Techniques, *Procedia computer science*, vol. 49, pp. 50–57, 2015.

[154] W. Zehra, A. R. Javed, Z. Jalil, H. U. Khan, and T. R. Gadekallu, Cross Corpus Multilingual Speech Emotion Recognition using Ensemble Learning, *Complex & Intelligent Systems*, vol. 7, no. 4, pp. 1845–1854, 2021.

[155] S. Li and W. Deng, Deep Facial Expression Recognition: A Survey, *IEEE Transactions on Affective Computing*, pp. 1–1, 2020.

[156] F. Noroozi, C. A. Corneanu, D. Kamińska, T. Sapiński, S. Escalera, and G. Anbarjafari, Survey on Emotional Body Gesture Recognition, *IEEE Transactions on Affective Computing*, vol. 12, no. 2, pp. 505–523, 2018.

[157] N. Alswaidan and M. E. B. Menai, A Survey of State-of-the-art Approaches for Emotion Recognition in Text, *Knowledge and Information Systems*, vol. 62, no. 8, pp. 2937–2987, 2020.

[158] F. Calefato, F. Lanubile, and N. Novielli, EmoTxt: A Toolkit for Emotion Recognition from Text, in *Proceedings of the International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pp. 79–80, IEEE, 2017.

[159] L. Shu, J. Xie, M. Yang, Z. Li, Z. Li, D. Liao, X. Xu, and X. Yang, A Review of Emotion Recognition using Physiological Signals, *Sensors*, vol. 18, no. 7, p. 2074, 2018.

[160] K. Gouizi, F. Bereksi Reguig, and C. Maaoui, Emotion Recognition from Physiological Signals, *Journal of Medical Engineering & Technology*, vol. 35, no. 6-7, pp. 300–307, 2011.

[161] C. E. Osgood, The Nature and Measurement of Meaning, *Psychological Bulletin*, vol. 49, no. 3, p. 197, 1952.

[162] C. Gerfen, J. Bolam, A. Parent, C. Wilson, A. Reiner, D. Oorschot, D. Plenz, J. Wickens, J. Goldberg, J. Tepper, *et al.*, Handbook of Behavioral Neuroscience, 2010.

[163] J. A. Russell, Affective Space is Bipolar, *Journal of Personality and Social Psychology*, vol. 37, no. 3, p. 345, 1979.

[164] C. M. Whissell, The Dictionary of Affect in Language, in *The Measurement of Emotions*, pp. 113–131, Elsevier, 1989.

[165] R. Plutchik, Emotion: A Psychoevolutionary Synthesis, 1980.

[166] A. Mehrabian, Pleasure-arousal-dominance: A General Framework for Describing and Measuring Individual Differences in Temperament, *Current Psychology*, vol. 14, no. 4, pp. 261–292, 1996.

[167] J. E. Stets, Emotions and Sentiments, in *Handbook of Social Psychology*, pp. 309–335, Springer, 2006.

[168] M. Munezero, C. S. Montero, E. Sutinen, and J. Pajunen, Are They Different? Affect, Feeling, Emotion, Sentiment, and Opinion Detection in Text, *IEEE Transactions on Affective Computing*, vol. 5, no. 2, pp. 101–111, 2014.

[169] R. J. Davidson, K. R. Sherer, and H. H. Goldsmith, *Handbook of Affective Sciences*. Oxford University Press, 2009.

[170] J. A. Russell and L. F. Barrett, Core Affect, Prototypical Emotional Episodes, and Other Things Called Emotion: Dissecting the Elephant., *Journal of Personality and Social Psychology*, vol. 76, no. 5, p. 805, 1999.

[171] A. Yadav and D. K. Vishwakarma, Sentiment Analysis using Deep Learning Architectures: A Review, *Artificial Intelligence Review*, vol. 53, no. 6, pp. 4335–4385, 2020.

[172] L.-P. Morency, R. Mihalcea, and P. Doshi, Towards Multimodal Sentiment Analysis: Harvesting Opinions from the Web, in *Proceedings of the International Conference on Multimodal Interfaces*, pp. 169–176, ACM, 2011.

[173] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, Introducing the RECOLA Multimodal Corpus of Remote Collaborative and Affective Interactions, in *Proceedings of the International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pp. 1–8, IEEE, 2013.

[174] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, Multimodal Sentiment Intensity Analysis in Videos: Facial Gestures and Verbal Messages, *IEEE Intelligent Systems*, vol. 31, no. 6, pp. 82–88, 2016.

[175] A. Vinciarelli, A. Dielmann, S. Favre, and H. Salamin, Canal9: A Database of Political Debates for Analysis of Social Interactions, in *Proceedings of the International Conference on Affective Computing and Intelligent Interaction and Workshops*, pp. 1–4, IEEE, 2009.

[176] A. Vinciarelli, S. Kim, F. Valente, and H. Salamin, Collecting Data for Socially Intelligent Surveillance and Monitoring Approaches: The Case of Conflict in Competitive Conversations, in *Proceedings of the International Symposium on Communications, Control and Signal Processing (ISCCSP)*, pp. 1–4, IEEE, 2012.

[177] J. L. Fleiss, B. Levin, and M. C. Paik, *Statistical methods for rates and proportions*. john wiley & sons, 2013.

[178] K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition, in *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, IEEE/CVF, 2016.

[179] J. Pennington, R. Socher, and C. D. Manning, GloVe: Global Vectors for Word Representation, in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, ACL, 2014.

[180] F. Ringeval, B. Schuller, M. Valstar, R. Cowie, H. Kaya, M. Schmitt, S. Amiriparian, N. Cummins, D. Lalanne, A. Michaud, *et al.*, AVEC 2018 Workshop and Challenge: Bipolar Disorder and Cross-Cultural Affect Recognition, in *Proceedings of the International Workshop on Audio/Visual Emotion Challenge*, pp. 3–13, 2018.

[181] L. J. Cronbach, Coefficient alpha and the internal structure of tests, *psychometrika*, vol. 16, no. 3, pp. 297–334, 1951.

[182] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, The Geneva Minimalistic Acoustic Parameter Set (gemaps) for Voice Research and Affective Computing, *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[183] K. Krippendorff, Estimating the reliability, systematic error and random error of interval data, *Educational and Psychological Measurement*, vol. 30, no. 1, pp. 61–70, 1970.

[184] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, Context-Dependent Sentiment Analysis in User-Generated Videos, in *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 873–883, 2017.

[185] S. Poria, E. Cambria, D. Hazarika, N. Mazumder, A. Zadeh, and L.-P. Morency, Multi-level Multiple Attentions for Contextual Multimodal Sentiment Analysis, in *Proceedings of the International Conference on Data Mining (ICDM)*, pp. 1033–1038, IEEE, 2017.

[186] D. Ghosal, M. S. Akhtar, D. Chauhan, S. Poria, A. Ekbal, and P. Bhattacharyya, Contextual Inter-modal Attention for Multi-modal Sentiment Analysis, in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3454–3466, 2018.

[187] T. Mikolov, K. Chen, G. Corrado, and J. Dean, Efficient Estimation of Word Representations in Vector Space, in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2013.

[188] S. Ji, W. Xu, M. Yang, and K. Yu, 3d Convolutional Neural Networks for Human Action Recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221–231, 2012.

[189] S. Ioffe and C. Szegedy, Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift, in *Proceedings of International Conference on Machine Learning (ICML)*, pp. 448–456, 2015.

[190] S. Hochreiter, The Vanishing Gradient Problem during Learning Recurrent Neural Nets and Problem Solutions, *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 6, no. 02, pp. 107–116, 1998.

[191] F. Chollet *et al.*, Keras. `https://keras.io`, 2015. (last accessed on 26 July 2022).

[192] D. P. Kingma and J. Ba, Adam: A Method for Stochastic Optimization, in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

[193] G. Van Rossum and F. L. Drake Jr, *Python Reference Manual*. Centrum voor Wiskunde en Informatica Amsterdam, 1995.

[194] S. Mishra, B. L. Sturm, and S. Dixon, Local Interpretable Model-Agnostic Explanations for Music Content Analysis, in *Proceedings of the International Society for Music Information Retrieval Conference*, pp. 537–543, 2017.

[195] J. Huang, J. Tao, B. Liu, Z. Lian, and M. Niu, Multimodal Transformer Fusion for Continuous Emotion Recognition, in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3507–3511, IEEE, 2020.

[196] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, Automatic Differentiation in PyTorch, in *Proceedings of the Autodiff Workshop at Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[197] D. P. Kingma and J. Ba, Adam: A Method for Stochastic Optimization, in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

[198] J. Lefter, L. Rothkrantz, D. van Leeuwen, and P. Wiggers, Automatic Stress Detection in Emergency (telephone) Calls, *International Journal of Intelligent Defence Support Systems*, vol. 4, pp. 148–168, 2011.

[199] R. Jabbar, K. Al-Khalifa, M. Kharbeche, W. Alhajyaseen, M. Jafari, and S. Jiang, Real-time Driver Drowsiness Detection for Android Application using Deep Neural Networks Techniques, *Procedia computer science*, vol. 130, pp. 400–407, 2018.

[200] B. Reddy, Y.-H. Kim, S. Yun, C. Seo, and J. Jang, Real-time Driver Drowsiness Detection for Embedded System using Model Compression of Deep Neural Networks, in *Proceedings of the Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 121–128, IEEE/CVF, 2017.

[201] A. Triantafyllopoulos, H. Sagha, F. Eyben, and B. Schuller, audEERINGs approach to the One-Minute-Gradual emotion challenge, *arXiv preprint arXiv:1805.01222*, 2018.

[202] D. Kollias and S. P. Zafeiriou, Exploiting Multi-CNN Features in CNN-RNN based Dimensional Emotion Recognition on the OMG in-the-wild Dataset, *IEEE Transactions on Affective Computing*, pp. 1–1, 2020.

[203] Y. Yin, B. Huang, Y. Wu, and M. Soleymani, Speaker-Invariant Adversarial Domain Adaptation for Emotion Recognition, in *Proceedings of the International Conference on Multimodal Interaction*, p. 481–490, ACM, 2020.

[204] Z. Wang, Z. Wan, and X. Wan, Transmodality: An End2end Fusion Method with Transformer for Multimodal Sentiment Analysis, in *Proceedings of The Web Conference*, pp. 2514–2520, 2020.

[205] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, Deep Canonical Correlation Analysis, in *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 1247–1255, ACM, 2013.

[206] W. Boes and H. Van hamme, Audiovisual transformer architectures for large-scale classification and synchronization of weakly labeled audio events, in *Proceedings of the 27th ACM International Conference on Multimedia*, pp. 1961–1969, 2019.

[207] A. Khare, S. Parthasarathy, and S. Sundaram, Self-supervised learning with cross-modal transformers for emotion recognition, in *2021 IEEE Spoken Language Technology Workshop (SLT)*, pp. 381–388, IEEE, 2021.

[208] J. Han, Z. Zhang, F. Ringeval, and B. Schuller, Reconstruction-error-based Learning for Continuous Emotion Recognition in Speech, in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2367–2371, IEEE, 2017.

[209] J. Han, Z. Zhang, M. Schmitt, M. Pantic, and B. Schuller, From Hard to Soft: Towards more Human-like Emotion Recognition by Modelling the Perception Uncertainty, in *Proceedings of the ACM International Conference on Multimedia*, pp. 890–897, 2017.

[210] Z. Zhang, J. Han, E. Coutinho, and B. Schuller, Dynamic Difficulty Awareness Training for Continuous Emotion Prediction, *IEEE Transactions on Multimedia*, vol. 21, no. 5, pp. 1289–1301, 2019.

[211] H. Chen, D. Jiang, and H. Sahli, Transformer Encoder with Multi-modal Multi-head Attention for Continuous Affect Recognition, *IEEE Transactions on Multimedia*, pp. 1–1, 2020.

[212] M. Ravanelli and Y. Bengio, Interpretable Convolutional Filters with SincNet, in *Proceedings of the Interpretability and Robustness for Audio, Speech and Language Workshop at Advances in Neural Information Processing Systems (NeurIPS)*, 2018.