Towards Generalizable and Robust Multimodal ML models for Video Emotion Recognition

by

Naresh Kumar Devulapally

May 14, 2024

A Thesis submitted to the
Faculty of the Graduate School of
the University at Buffalo, The State University of New York
in partial fulfillment of the requirements for the
degree of

Master of Science

Department of Computer Science and Engineering

To my family and friends

Acknowledgments

Firstly, I convey my deepest gratitude to my thesis supervisor **Dr. Junsong Yuan** for giving me an opportunity to work in their research team. I would like to thank **Dr. Sreyasee Das Bhattacharjee** for her continued support throughout my thesis. I would like to extend my gratitude to Sidharth Anand, student co-author of the publications made as a part of my thesis, for his support. I would like to thank **Dr. Vishnu Lokhande** for being a part of my MS Thesis defense committee and providing his invaluable feedback on my thesis.

I would like to thank my friends for their immense support. Last, but definitely not the least, I would like to thank my parents and Himavarshini Yarragangu who shaped me into the person I am. I am indebted to their kindness, and unconditional support.

0.1 Disclaimer

It is important to note that throughout this thesis, I have published three works at venues including ACM Multimedia 2023 [1], Big MM 2024 [2], and ICME 2024 [3] as a first author throughout the publications. This thesis encompasses the contributions I have made through these published works while clearly illustrating the motivation, and challenges I tackle throughout the thesis. I request readers to refer to the above-mentioned works for further reading and any similarity found in the above-mentioned works, and this thesis is attributed to my sole contribution as a first-author.

Table of Contents

[1] Field in this thesis refers to the field of Multimodal Machine Learning.

[2] Task in the thesis refers to the task of video emotion recognition utilizing multiple modalities as input

List of Tables

List of Figures

Abstract

Human behavior is inherently multimodal. A modality refers to a way in which a natural phenomenon is perceived or expressed. Examples include text, audio, visual modality. Multimodal Machine Learning (MML) as a field aims to capture, process and infer from information across modalities. However, representation of cross-modal information is a challenging task, due to core characteristics of modalities such as heterogeneity. In addition to Representation, other core challenges in Multimodal Machine Learning systems include Alignment, Reasoning, Transference and Generation. These challenges are further amplified when task specific challenges arise while proposing MML models for tasks such as Video Emotion Recognition, and Visual Question Answering.

In this Thesis, I tackle the core challenges of Representation, Alignment and Reasoning in MML systems and provide a framework to train large MML systems for Video Emotion Recognition. I provide a unified Transformer architecture that utilizes novel Adaptive Fusion and Self-Supervised Machine Learning to extend MML models for Multi-Label Video Emotion Recognition in conversations. Further contributions include an Explainability Module to interpret auxiliary modalities that influence reasoning in MML systems. Finally, I propose a Multimodal Training framework that is robust to missing modalities both during training and inference.

Experiments were conducted on MELD, IEMOCAP, ElderReact, and EmoReact datasets. The proposed Multimodal framework outperforms SoTA methods by 17% (weighted-F1) for cross-dataset performance (Testing on an entire unseen dataset) to demonstrate generalizability of the framework and 5% gain in weighted-F1 with missing-modalities to demonstrate robustness towards missing modality information. I utilize LIME to demonstrate qualitative results for explainability.

Chapter1

Introduction

Developing intelligent computer agents that can understand, reason, incorporate and interpret information from multimodal data sources similar to a human brain has been a grand goal of Artificial Intelligence paving way towards Artificial General Intelligence (See Fig. 1.1). Such Multimodal Machine Learning (MML) systems have significant applications across various fields including Autonomous driving, Robotics, Healthcare and other multidisciplinary fields. These MML systems derive inspiration from the inherent multimodal ability of the human brain. A **modality** refers to a way in which a natural phenomenon is perceived or expressed, examples include text, audio, visual modality [4].

Multimodal Machine Learning as a field, aims to build intelligent machine learning systems, that can effectively represent, and reason with information from multiple sensory modalities. As various modalities arise from a single event in time (e.g., think of the modalities Video, Audio and Text from an online video), there are specific characteristics within modalities [4], such as:

- Modalities are heterogeneous.
- Modalities are connected.
- Interaction between modalities lead to information gain.

Due to the heterogeneity across modalities, combined representation of information across modalities is a core challenge in MML systems.

In addition to representation, there exist other core challenges in MML systems include Alignment, Reasoning, Generation and Transference [4] (See Fig. 1.3).

One primary goal of Machine Learning as a whole is to mimic the behavior of the human brain to interpret and reason from the input information. MML takes this one step further to integrate and reason from multiple modalities, represent the information with the aim to enhance classification (or task-specific performance) in comparison with the uni-modal counterparts.

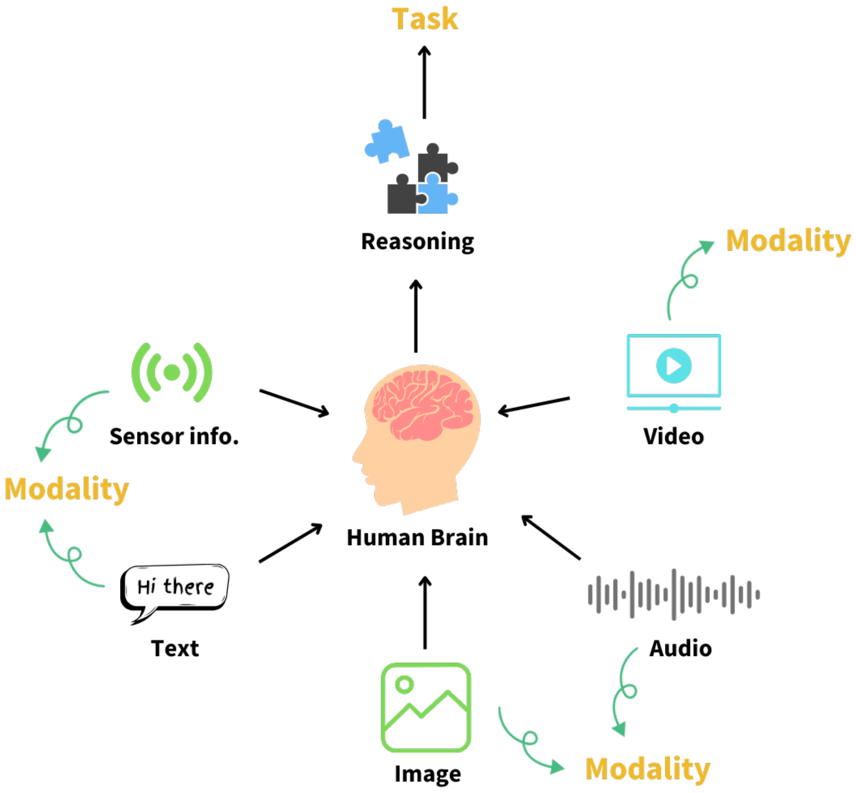


Figure 1.1: Human brain is inherently Multimodal.

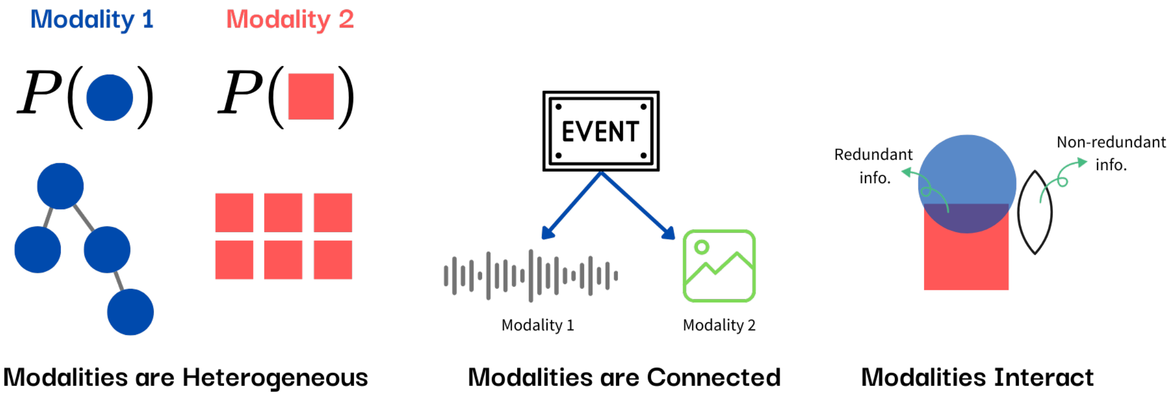


Figure 1.2: Characteristics of Modalities.

1.0.1 Modalities are Heterogeneous

Figure 1.2 illustrates some of the inherent characteristics of modalities. Modalities are heterogeneous at both distribution level and structure. Consider modality 1 to be audio and modality 2 to be images. There is a clear heterogeneity in the way information is represented within modalities 1 and 2, i.e., images are represented in a 2-dimensional structure (and multiple color channels), where are audio is represented as a continuous 1-dimensional sequence.

1.0.2 Modalities are connected

Generally individual modalities arise as a result from a single event in time. Take videos for example, the modalities of video frames, audio and text transcript represent the same event (in the video) individually. This means all the modalities represent the same event in time. Hence, it is essential to note that modalities are connected.

1.0.3 Modalities interact

Consider a simple question “If a blind person get his sight, how would this new information be integrated by the brain to perform various representation and reasoning tasks?”. The above question derives an important insight that is illustrated in the figure 1.2. Interaction between modalities leads to two types of information viz., redundant information (that is present in each of modalities as both the modalities are connected), and non-redundant information that arises as an output of the interaction between modalities which is otherwise not present in each of the modalities individually. This non-redundant information is of great interest to us and is the primary motivation of this thesis. The first phase of the thesis focuses on representing this non-redundant, cross-modal information and then extends the work by proposing a framework for various multi-modal tasks, robust to missing information while training and inference.

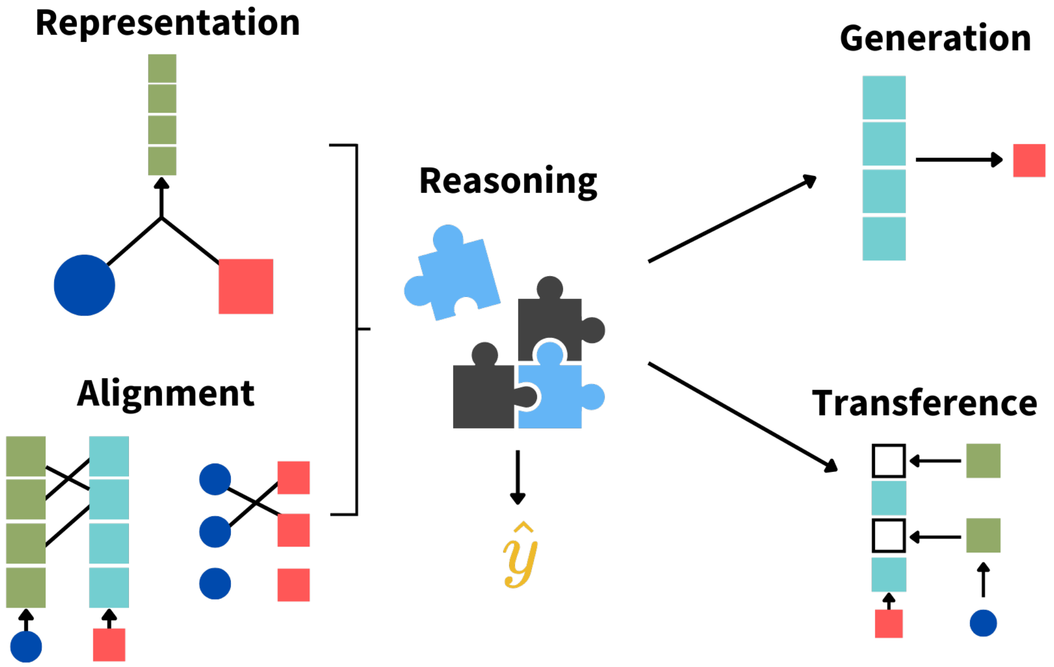


Figure 1.3: Core challenges in MML Systems.

1.1 Multimodal Machine Learning Tasks:

Given the significance of MML Systems, there are numerous tasks that utilize/need MML models. Examples of such tasks include:

- Video Question Answering: Given a video (continuous frames of images) as input, the aim is to train a model that can accept questions in the form of text and provide right answer to the question utilizing the input video as reference.
- Text to Image Generation: Given a text input, the aim is to train a generation model that can generate visual content (images or videos) based on the input text.
- Multimodal ML in healthcare and robotics: The possibilities of Multimodal ML systems in the field of healthcare and robotics are enormous. From crafting robots towards real-time human machine interaction, to utilizing these robots for healthcare and medical applications, multimodal ML has huge potential.

1.1.1 Video Emotion Recognition

In this thesis, I target this multimodal ML task. Given a set of utterances with video, audio, and text modalities as inputs, the aim is to train an multimodal ML model is to predict the most dominant (extended to multi-label scenario in the further phases) emotion state of the speakers in the video. The model should learn to integrate information from multiple modalities to make a decision.

1.2 Problem Statement

Problem Definition: The input utterance u that contains information in the form of raw modalities including video, audio and text, the aim of the model is to predict the most dominant (or set of dominant) emotions of speakers within that utterance.

1.2.1 Motivation

1.2.2 Field[1]-specific motivation

There is a significant need for general-purpose MML systems that can perform various downstream Computer Vision tasks. The motivation arises from the need and potential for MML systems that can represent and integrate information from multiple modalities. Hence, such models can be widely utilized across several fields including Robotics (for human-computer interaction), Healthcare (for better care for elderly population), Education (utilizing multimodal agents to address student queries and concerns) and so on.

1.2.3 Task[2]-Specific motivation

Need for video emotion recognition finds significant interest in the fields of healthcare, robotics, and a few business scenarios (service quality analysis in social environments such as restaurants).

The potential to assist informal caregivers using Emotionally grounded agents is significant. Within industry there exists corporations that aim to build real-time emotion recognition models for various application scenarios.

[1] Field in this thesis refers to the field of Multimodal Machine Learning.

[2] Task in the thesis refers to the task of video emotion recognition utilizing multiple modalities as input

Chapter2

Related works

Towards video emotion recognition majority of the literature is focussed on unimodal machine learning models. These unimodal models are primarily based on text data. This can be attributed to the fact that text data is widely available across the web and considered as a strong indicator of emotion.

Along unimodal ML systems for emotion recognition, there exist several works that extract multimodal information across modalities to perform emotion recognition. Methods including [5, 6, 7] concatenate features across modalities to result in concatenated information across modalities and utilize the information to perform emotion recognition.

Works including [8, 9] use a fixed combination of weighted average feature representation methods to find cross-modal features.

While these works show performance improvements, there is still a significant lack of information fusion across modalities within existing methods including [10, 11].

Chapter3

Proposed Method

3.1 AMuSE

As mentioned in the previous chapters, to tackle the challenges of representation, alignment, and reasoning in MML systems for video emotion recognition, there is a need to develop a framework that can adaptively/dynamically fuse information across modalities. Towards this we propose a two-step solution in the form of Multimodal Attention Network and Adaptive Fusion.

Before diving deep into the details of the proposed method, lets take a look at the overall architecture shown in figure 3.1

The architecture of AMuSE takes in sequence of utterances as raw modalities in the form of video, audio and text as input, learns cross-modal feature representation of the modalities utilizing the Multimodal Attention Network, and then learns to fuse the information across modalities adaptively to learn a common feature representation that across modalities. The cross-modal feature extraction, adaptive fusion are task-agnostic and modality agnostic except for the unimodal feature extraction. Followed by the feature representation from adaptive fusion unit, we add task-specific situation-level processing and dialogue-level processing for the video emotion recognition task at hand, followed by a classification head to predict the most dominant emotion label of all speakers within the utterance.

Next, subsections discuss each component of the proposed method:

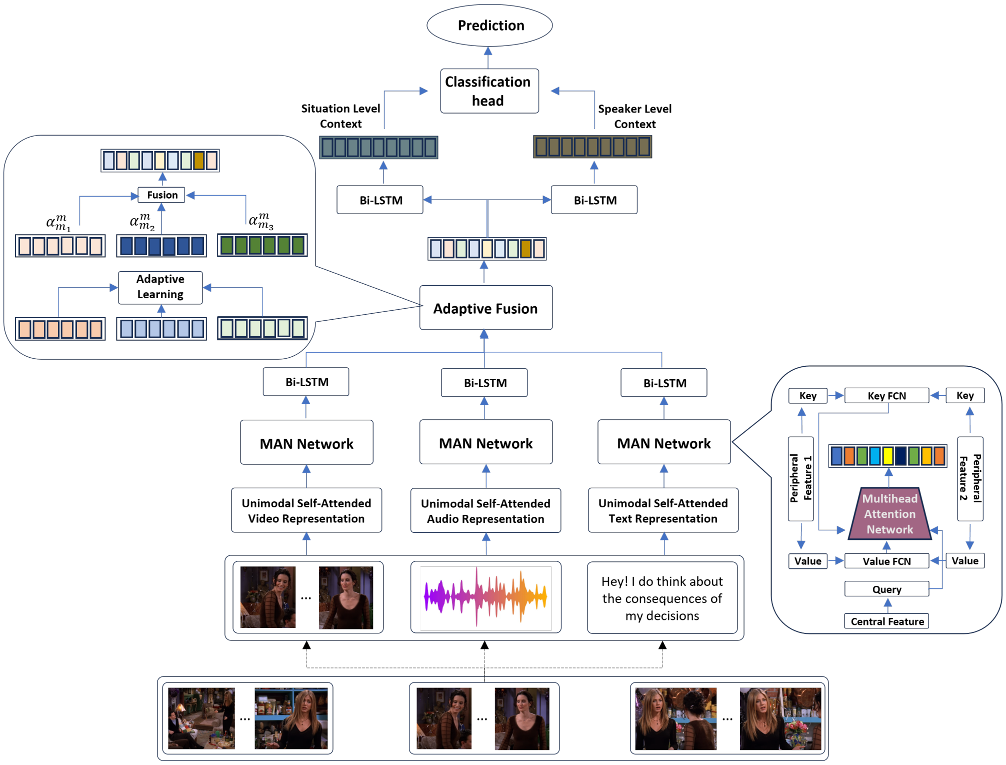


Figure 3.1: Proposed AMuSE model

3.1.1 Representation of unimodal features

Capturing utterance-level spatio-temporal evolution of information is essential for any video emotion recognition pipelines to retrieve crucial features. In this work, I utilize existing, pre-trained feature extractors:

3.1.1.1 Representation of Text modality

Input to the text representation module is the text transcript from a series of utterances u_i. To extract a compact feature representation of the text modality, we utilize the pre-trained MPNet [12] to extract the feature representation of the text modality. Given a sequence of utterances, the MPNet model derives features for each word present in the text.

3.1.1.2 Representation of Video modality

Input to the video feature extractor are the frames present in a video. We utilize FFmpeg to extract k key frames from the video. We then decompose/split the resultant frames into frames with only faces of the speaker, and frames that do not contain the face and contains all other information. Once the frames are extracted from the video, we utilize, Multitask cascaded convolutional networks [13] to extract visual features from the frames.

From the frames that contain face information, we utilize JAA-Net [14] to extract facial action units. We utilize action units as the features from face frames due to the privacy-preserving advantages of Action Units as AUs are local features extracted from regions of human faces and cannot be tracked back to the person's identity. Once features are extracted for face and body frames we employ Bi-LSTM model to extract one final video feature representation.

3.1.1.3 Representation of audio modality

The input to the audio feature extractor is the audio spoken by speakers within the utterance. The datasets are segregated such that per utterance there is only one speaker. To extract the features, we utilize PASST feature extractor that extract audio feature representation. After extracting the audio feature extractors from PASST, similar to utilizing Bi-LSTM for visual features, we utilize Bi-LSTM model to extract final unimodal features from the audio modalities.

3.1.2 Cross-Attended (Multimodal Feature Representation)

We utilize transformer architecture to extract cross-modal feature representation across modalities. We propose a novel Multi-modal Attention Network that extracts cross-modal information across modalities. It is important to note that, the proposed MAN module is modality-agnostic (i.e., once the features are extracted from individual modalities, MAN module learns to extract cross-modal features from the unimodal features extracted). In this work, the unimodal features are extracted such that the final shape of the extracted features is the same.

To achieve this, we extract cross-modal representation for each modality. The MAN network is shown in the figure 3.1. For each modality, the main modality is named as *Central network*, while information from modalities is termed *Peripheral networks*. Then we utilize weighted representation using transformer networks and inject the information from peripheral modalities to the central modality. The MAN module of the proposed method is learned using Self-Supervised Machine Learning. More details regarding the MAN network and the learning procedure can be found at [2].

3.2 Multi-Label Emotion Recognition in conversation

With significant contributions in-terms of dominant emotion classification abilities proposing novel Adaptive Fusion and Multimodal Attention Network, the next phase of this thesis tackles the challenges of multi-label emotion recognition in conversations and builds a complete transformer-based training pipeline to train any large multimodal machine learning model. A significant contribution of this phase of the thesis includes building generalizable framework for multi-label emotion recognition in conversational setting. We test generalizability of the proposed framework in the form of cross-dataset performance testing.

Video emotion recognition datasets are chosen such that each dataset represents speakers from a different age group, demographic background and so on. The motivation to perform multi-label emotion recognition arises from the fact that human emotion expression is inherently multi-label. We, as humans, express multiple emotions at the same time within an utterance in a conversation/dialogue. Capturing these emotion nuances effectively can lead to a generalized framework that could be improved to made robust, reliable for real-time deployment.

Towards this, we propose a complete vision-transformer architecture that takes in raw modalities, learns cross-modal representation of information across modalities, utilized novel calibration loss to perform multi-label emotion recognition in group conversations. More details of our work can be found here [1].

3.3 Robust Multimodal ML systems for video emotion recognition

Robustness in MML systems can be attributed to learning cross-model, aligned representation across modalities during missing information within modalities during training and inference. After our previous contribution towards building a general-purpose framework for multi-label emotion recognition, we extend our contribution towards building robust multimodal machine learning systems for video emotion recognition.

Towards achieving robustness to missing information within MML systems, the proposed framework must be robust to missing entire modalities during training/inference or corrupted information within modalities. We simulate missing information within modalities by randomly dropping representation tensors within a modality, and alternatively drop the entire modality during training/inference. We achieve state of the art performance even with missing information by proposing a novel multimodal architecture utilizing Joint Embedding Learning. More details about the proposed architecture can be found here [3]. This work is accepted at the International Conference on Multimedia and Expo, 2024.

Chapter4

Experimental Setup

4.1 Datasets

Emotion Recognition datasets chosen for this thesis include:

- Multimodal Emotion Lines Dataset (MELD) [15]
- The Interactive Emotional Dyadic Motion Capture (IEMOCAP) [16]
- Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) [17]
- ElderReact [18]
- EmoReact [19]

Each of the datasets contain utterances with majority of the widely known emotion labels for each utterance. ElderReact and EmoReact datasets contain multiple label for each utterance spoken that is relevant to testing the performance of our multi-label emotion recognition model. To test the missing modality robustness of our final model, we simulate missing information within the MELD and IEMOCAP datasets.

4.2 Resources used for training

I thank the Center for Computational Research for providing all the resources to carry out all the experiment for the thesis work. All the experiments were carried of a node containing 2 A100 GPUs.

4.3 Evaluation Metrics

For the first phase of the thesis, the architecture framework is to perform multi-class classification across all the labels present in the dataset. Towards this, we use Accuracy and wF1 as evaluation metrics. Towards further phases in the thesis, the problem statement is changed to perform multi-label classification. Hence, we use wF1 as evaluation metrics for the same.

Chapter5

Results and Discussion

In this chapter, I present the results of all the novel methologies proposed throughout the thesis work.

5.1 AMuSE

The proposed AMuSE method [2] proposes a novel Multimodal Attention Network for cross-modal feature representation, a Adaptive Fusion module to adaptively learn features across modalities. We apply this novel proposed methods for video emotion recognition application and outperform state of the art results of the MELD and IEMOCAP datasets.

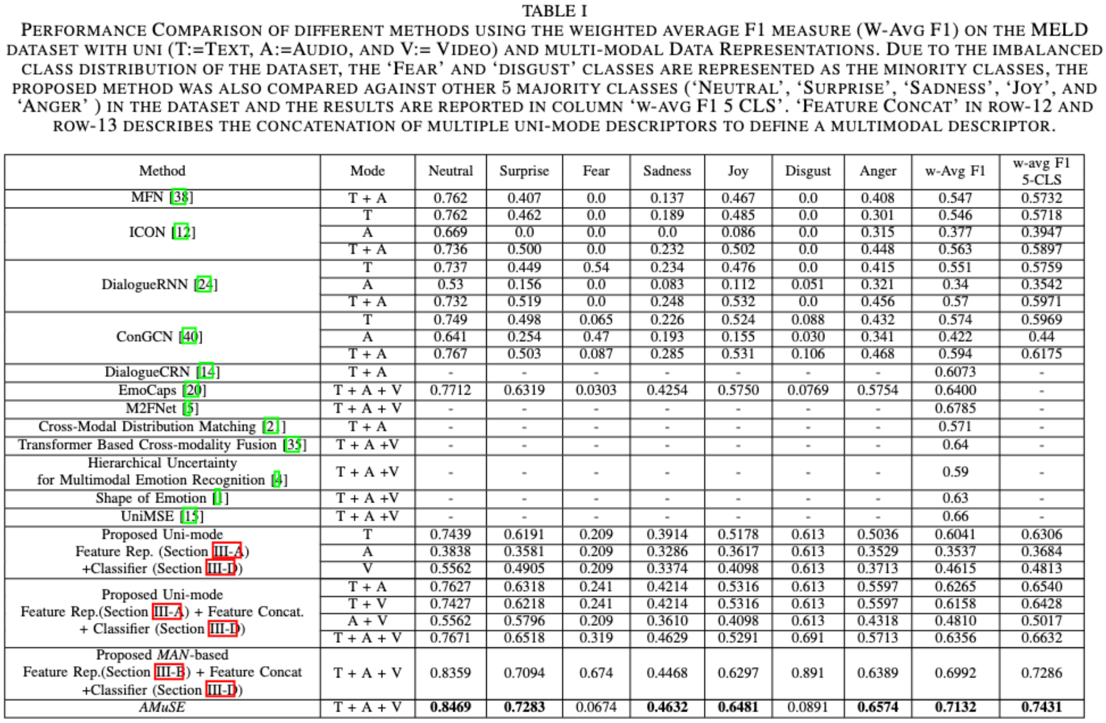


TABLE I

PERFORMANCE COMPARISON OF DIFFERENT METHODS USING THE WEIGHTED AVERAGE F1 MEASURE (W-AVG F1) ON THE MELD DATASET WITH UNI (T:=TEXT, A:=AUDIO, AND V:= VIDEO) AND MULTI-MODAL DATA REPRESENTATIONS. DUE TO THE IMBALANCED CLASS DISTRIBUTION OF THE DATASET, THE 'FEAR' AND 'DISGUST' CLASSES ARE REPRESENTED AS THE MINORITY CLASSES, THE PROPOSED METHOD WAS ALSO COMPARED AGAINST OTHER 5 MAJORITY CLASSES ('NEUTRAL', 'SURPRISE', 'SADNESS', 'JOY', AND 'ANGER') IN THE DATASET AND THE RESULTS ARE REPORTED IN COLUMN 'W-AVG F1 5 CLS'. 'FEATURE CONCAT' IN ROW-12 AND ROW-13 DESCRIBES THE CONCATENATION OF MULTIPLE UNI-MODE DESCRIPTORS TO DEFINE A MULTIMODAL DESCRIPTOR.

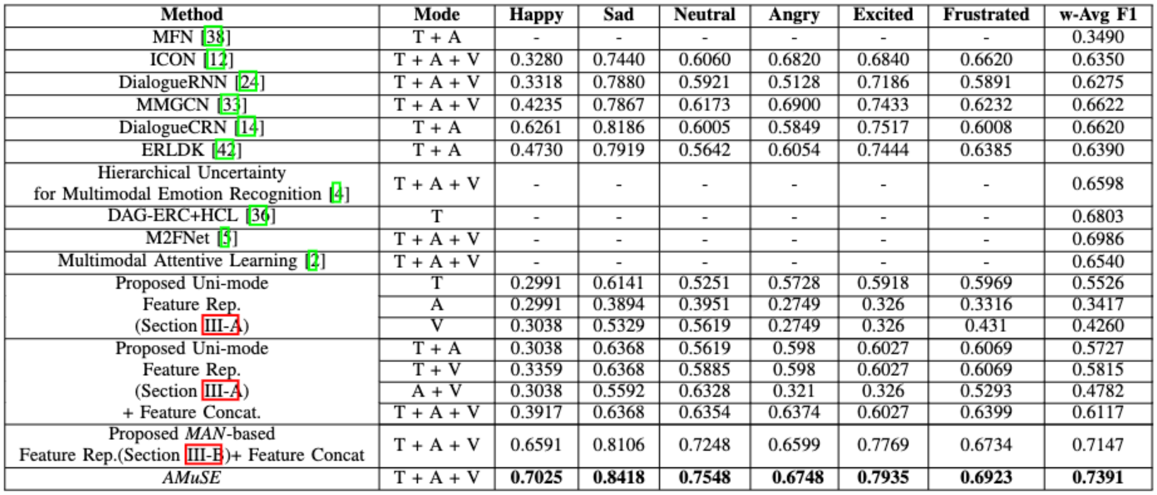
Method	Mode	Neutral	Surprise	Fear	Sadness	Joy	Disgust	Anger	w-Avg F1	w-avg F1 5-CLS
MFN [38]	T + A	0.762	0.407	0.0	0.137	0.467	0.0	0.408	0.547	0.5732
ICON [12]	T	0.762	0.462	0.0	0.189	0.485	0.0	0.301	0.546	0.5718
	A	0.669	0.0	0.0	0.0	0.086	0.0	0.315	0.377	0.3947
	T + A	0.736	0.500	0.0	0.232	0.502	0.0	0.448	0.563	0.5897
DialogueRNN [24]	T	0.737	0.449	0.54	0.234	0.476	0.0	0.415	0.551	0.5759
	A	0.53	0.156	0.0	0.083	0.112	0.051	0.321	0.34	0.3542
	T + A	0.732	0.519	0.0	0.248	0.532	0.0	0.456	0.57	0.5971
ConGCN [40]	T	0.749	0.498	0.065	0.226	0.524	0.088	0.432	0.574	0.5969
	A	0.641	0.254	0.47	0.193	0.155	0.030	0.341	0.422	0.44
	T + A	0.767	0.503	0.087	0.285	0.531	0.106	0.468	0.594	0.6175
DialogueCRN [14]	T + A	-	-	-	-	-	-	-	0.6073	-
EmoCaps [20]	T + A + V	0.7712	0.6319	0.0303	0.4254	0.5750	0.0769	0.5754	0.6400	-
M2FNet [5]	T + A + V	-	-	-	-	-	-	-	0.6785	-
Cross-Modal Distribution Matching [21]	T + A	-	-	-	-	-	-	-	0.571	-
Transformer Based Cross-modality Fusion [35]	T + A +V	-	-	-	-	-	-	-	0.64	-
Hierarchical Uncertainty for Multimodal Emotion Recognition [4]	T + A +V	-	-	-	-	-	-	-	0.59	-
Shape of Emotion [1]	T + A +V	-	-	-	-	-	-	-	0.63	-
UniMSE [15]	T + A +V	-	-	-	-	-	-	-	0.66	-
Proposed Uni-mode Feature Rep. (Section III-A) +Classifier (Section III-D)	T	0.7439	0.6191	0.209	0.3914	0.5178	0.613	0.5036	0.6041	0.6306
	A	0.3838	0.3581	0.209	0.3286	0.3617	0.613	0.3529	0.3537	0.3684
	V	0.5562	0.4905	0.209	0.3374	0.4098	0.613	0.3713	0.4615	0.4813
Proposed Uni-mode Feature Rep.(Section III-A) + Feature Concat. + Classifier (Section III-D)	T + A	0.7627	0.6318	0.241	0.4214	0.5316	0.613	0.5597	0.6265	0.6540
	T + V	0.7427	0.6218	0.241	0.4214	0.5316	0.613	0.5597	0.6158	0.6428
	A + V	0.5562	0.5796	0.209	0.3610	0.4098	0.613	0.4318	0.4810	0.5017
	T + A + V	0.7671	0.6518	0.319	0.4629	0.5291	0.691	0.5713	0.6356	0.6632
Proposed *MAN*-based Feature Rep.(Section III-B) + Feature Concat +Classifier (Section III-D)	T + A + V	0.8359	0.7094	0.674	0.4468	0.6297	0.891	0.6389	0.6992	0.7286
AMuSE	T + A + V	**0.8469**	**0.7283**	0.0674	**0.4632**	**0.6481**	0.0891	**0.6574**	**0.7132**	**0.7431**

The above table is from our published work [2]. As seen in the table, the proposed method significantly outperforms existing SoTA models on the MELD dataset across various emotion labels including Neutral, Surprise, Fear, Surprise, Joy, Disgust, Anger. While the improvement in performance is significant, it is important to note that the performance on the emotion labels including Disgust, Fear, Anger is significantly lower across the board.

This performance represents the difficulty in predicting these labels. Upon further investigation on this work, it was found that many utterance that fall into the bracket of Anger, Fear, and Disgust could indeed be classified to contain multiple emotion labels. Hence, this is another motivation to perform Multi-label emotion recognition in conversation setting.

TABLE II

PERFORMANCE COMPARISON OF DIFFERENCE METHODS USING THE WEIGHTED AVERAGE F1 MEASURE (W-AVG F1) ON THE IEMOCAP DATASET WITH WITH UNI (T:=TEXT, A:=AUDIO, AND V:= VIDEO) AND MULTI-MODAL DATA REPRESENTATIONS. 'FEATURE CONCAT' IN ROW 13 AND ROW 14 DESCRIBE THE CONCATENATION OF MULTIPLE UNI-MODE DESCRIPTORS TO DEFINE A MULTIMODAL DESCRIPTOR.



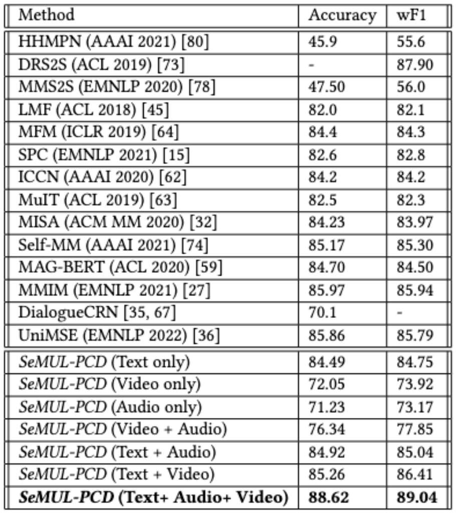
Method	Mode	Happy	Sad	Neutral	Angry	Excited	Frustrated	w-Avg F1
MFN [38]	T + A	-	-	-	-	-	-	0.3490
ICON [12]	T + A + V	0.3280	0.7440	0.6060	0.6820	0.6840	0.6620	0.6350
DialogueRNN [24]	T + A + V	0.3318	0.7880	0.5921	0.5128	0.7186	0.5891	0.6275
MMGCN [33]	T + A + V	0.4235	0.7867	0.6173	0.6900	0.7433	0.6232	0.6622
DialogueCRN [14]	T + A	0.6261	0.8186	0.6005	0.5849	0.7517	0.6008	0.6620
ERLDK [42]	T + A	0.4730	0.7919	0.5642	0.6054	0.7444	0.6385	0.6390
Hierarchical Uncertainty for Multimodal Emotion Recognition [4]	T + A + V	-	-	-	-	-	-	0.6598
DAG-ERC+HCL [36]	T	-	-	-	-	-	-	0.6803
M2FNet [5]	T + A + V	-	-	-	-	-	-	0.6986
Multimodal Attentive Learning [2]	T + A + V	-	-	-	-	-	-	0.6540
Proposed Uni-mode Feature Rep. (Section III-A)	T	0.2991	0.6141	0.5251	0.5728	0.5918	0.5969	0.5526
	A	0.2991	0.3894	0.3951	0.2749	0.326	0.3316	0.3417
	V	0.3038	0.5329	0.5619	0.2749	0.326	0.431	0.4260
Proposed Uni-mode Feature Rep. (Section III-A) + Feature Concat.	T + A	0.3038	0.6368	0.5619	0.598	0.6027	0.6069	0.5727
	T + V	0.3359	0.6368	0.5885	0.598	0.6027	0.6069	0.5815
	A + V	0.3038	0.5592	0.6328	0.321	0.326	0.5293	0.4782
	T + A + V	0.3917	0.6368	0.6354	0.6374	0.6027	0.6399	0.6117
Proposed *MAN*-based Feature Rep.(Section III-B)+ Feature Concat	T + A + V	0.6591	0.8106	0.7248	0.6599	0.7769	0.6734	0.7147
AMuSE	T + A + V	**0.7025**	**0.8418**	**0.7548**	**0.6748**	**0.7935**	**0.6923**	**0.7391**

Similar pattern can be seen with the performance results on the IEMOCAP dataset. The proposed work significantly outperforms existing state of the art models. For ablation studies and other experiments, please refer to our work [2].

5.2 SeMuL-PCD

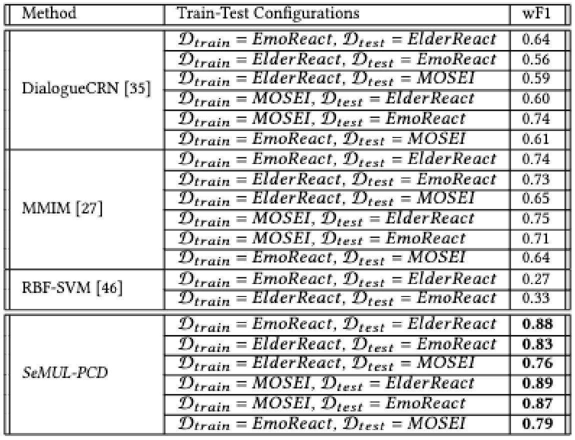
The second phase of this thesis proposes a Vision-Transformer framework for Multi-Label Emotion Recognition. Additionally, we also run experiments to test the generalization ability of the proposed architecture.

Table 1: Comparison of *SeMuL-PCD* with other models using the weighted averaged F1 measure (wF1) and Accuracy scores on the MOSEI dataset. For category specific performances please refer to the Appendix in the supplementary material.



Method	Accuracy	wF1
HHMPN (AAAI 2021) [80]	45.9	55.6
DRS2S (ACL 2019) [73]	-	87.90
MMS2S (EMNLP 2020) [78]	47.50	56.0
LMF (ACL 2018) [45]	82.0	82.1
MFM (ICLR 2019) [64]	84.4	84.3
SPC (EMNLP 2021) [15]	82.6	82.8
ICCN (AAAI 2020) [62]	84.2	84.2
MuIT (ACL 2019) [63]	82.5	82.3
MISA (ACM MM 2020) [32]	84.23	83.97
Self-MM (AAAI 2021) [74]	85.17	85.30
MAG-BERT (ACL 2020) [59]	84.70	84.50
MMIM (EMNLP 2021) [27]	85.97	85.94
DialogueCRN [35, 67]	70.1	-
UniMSE (EMNLP 2022) [36]	85.86	85.79
SeMUL-PCD (Text only)	84.49	84.75
SeMUL-PCD (Video only)	72.05	73.92
SeMUL-PCD (Audio only)	71.23	73.17
SeMUL-PCD (Video + Audio)	76.34	77.85
SeMUL-PCD (Text + Audio)	84.92	85.04
SeMUL-PCD (Text + Video)	85.26	86.41
***SeMUL-PCD* (Text+ Audio+ Video)**	**88.62**	**89.04**

Table 3: The Cross dataset Generalization Performance of *SeMUL-PCD* different Training ($\mathcal{D}_{train}$) and Testing ($\mathcal{D}_{test}$) set pairs. The performance is reported using Weighted average F1 (wF1) as the evaluation metric and it is compared against some recent State of the Art models, for which either the results or the codes were available.



Method	Train-Test Configurations	wF1
DialogueCRN [35]	$\mathcal{D}_{train} = EmoReact, \mathcal{D}_{test} = ElderReact$	0.64
	$\mathcal{D}_{train} = ElderReact, \mathcal{D}_{test} = EmoReact$	0.56
	$\mathcal{D}_{train} = ElderReact, \mathcal{D}_{test} = MOSEI$	0.59
	$\mathcal{D}_{train} = MOSEI, \mathcal{D}_{test} = ElderReact$	0.60
	$\mathcal{D}_{train} = MOSEI, \mathcal{D}_{test} = EmoReact$	0.74
	$\mathcal{D}_{train} = EmoReact, \mathcal{D}_{test} = MOSEI$	0.61
MMIM [27]	$\mathcal{D}_{train} = EmoReact, \mathcal{D}_{test} = ElderReact$	0.74
	$\mathcal{D}_{train} = ElderReact, \mathcal{D}_{test} = EmoReact$	0.73
	$\mathcal{D}_{train} = ElderReact, \mathcal{D}_{test} = MOSEI$	0.65
	$\mathcal{D}_{train} = MOSEI, \mathcal{D}_{test} = ElderReact$	0.75
	$\mathcal{D}_{train} = MOSEI, \mathcal{D}_{test} = EmoReact$	0.71
	$\mathcal{D}_{train} = EmoReact, \mathcal{D}_{test} = MOSEI$	0.64
RBF-SVM [46]	$\mathcal{D}_{train} = EmoReact, \mathcal{D}_{test} = ElderReact$	0.27
	$\mathcal{D}_{train} = ElderReact, \mathcal{D}_{test} = EmoReact$	0.33
SeMUL-PCD	$\mathcal{D}_{train} = EmoReact, \mathcal{D}_{test} = ElderReact$	**0.88**
	$\mathcal{D}_{train} = ElderReact, \mathcal{D}_{test} = EmoReact$	**0.83**
	$\mathcal{D}_{train} = ElderReact, \mathcal{D}_{test} = MOSEI$	**0.76**
	$\mathcal{D}_{train} = MOSEI, \mathcal{D}_{test} = ElderReact$	**0.89**
	$\mathcal{D}_{train} = MOSEI, \mathcal{D}_{test} = EmoReact$	**0.87**
	$\mathcal{D}_{train} = EmoReact, \mathcal{D}_{test} = MOSEI$	**0.79**

As seen in the above table, the performance of the proposed framework outperforms the SoTA model for multi-label emotion recognition for video conversation setting. To test the generalization ability of the proposed model, we utilize cross-dataset performance as a metrics. In Table 3, above it can be seen that the proposed model significantly outperforms SoTA models on cross-dataset evaluation. This demonstrates the generalization ability of the proposed method.

5.3 Missing modality video emotion recognition

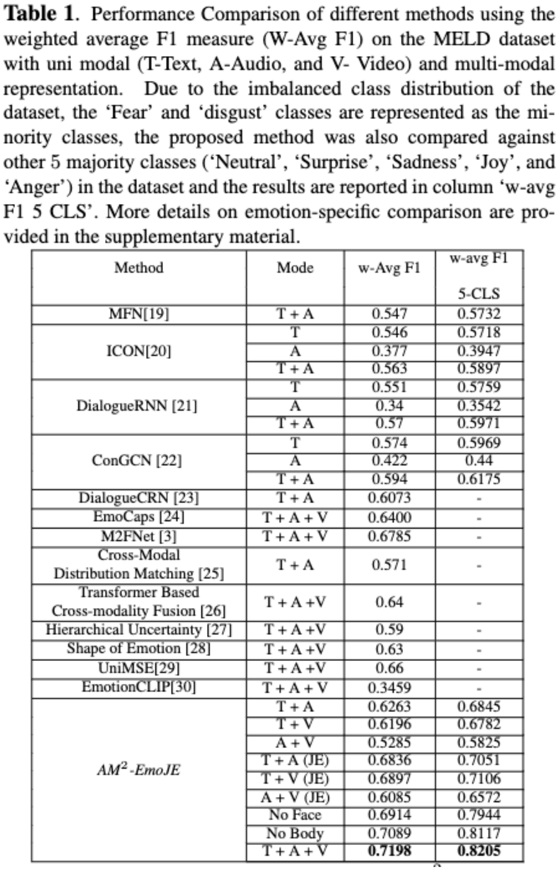


Table 1. Performance Comparison of different methods using the weighted average F1 measure (W-Avg F1) on the MELD dataset with uni modal (T-Text, A-Audio, and V- Video) and multi-modal representation. Due to the imbalanced class distribution of the dataset, the 'Fear' and 'disgust' classes are represented as the minority classes, the proposed method was also compared against other 5 majority classes ('Neutral', 'Surprise', 'Sadness', 'Joy', and 'Anger') in the dataset and the results are reported in column 'w-avg F1 5 CLS'. More details on emotion-specific comparison are provided in the supplementary material.

Method	Mode	w-Avg F1	w-avg F1 5-CLS
MFN[19]	T + A	0.547	0.5732
ICON[20]	T	0.546	0.5718
	A	0.377	0.3947
	T + A	0.563	0.5897
DialogueRNN [21]	T	0.551	0.5759
	A	0.34	0.3542
	T + A	0.57	0.5971
ConGCN [22]	T	0.574	0.5969
	A	0.422	0.44
	T + A	0.594	0.6175
DialogueCRN [23]	T + A	0.6073	-
EmoCaps [24]	T + A + V	0.6400	-
M2FNet [3]	T + A + V	0.6785	-
Cross-Modal Distribution Matching [25]	T + A	0.571	-
Transformer Based Cross-modality Fusion [26]	T + A +V	0.64	-
Hierarchical Uncertainty [27]	T + A +V	0.59	-
Shape of Emotion [28]	T + A +V	0.63	-
UniMSE[29]	T + A +V	0.66	-
EmotionCLIP[30]	T + A + V	0.3459	-
AM^2-EmoJE	T + A	0.6263	0.6845
	T + V	0.6196	0.6782
	A + V	0.5285	0.5825
	T + A (JE)	0.6836	0.7051
	T + V (JE)	0.6897	0.7106
	A + V (JE)	0.6085	0.6572
	No Face	0.6914	0.7944
	No Body	0.7089	0.8117
	T + A + V	**0.7198**	**0.8205**

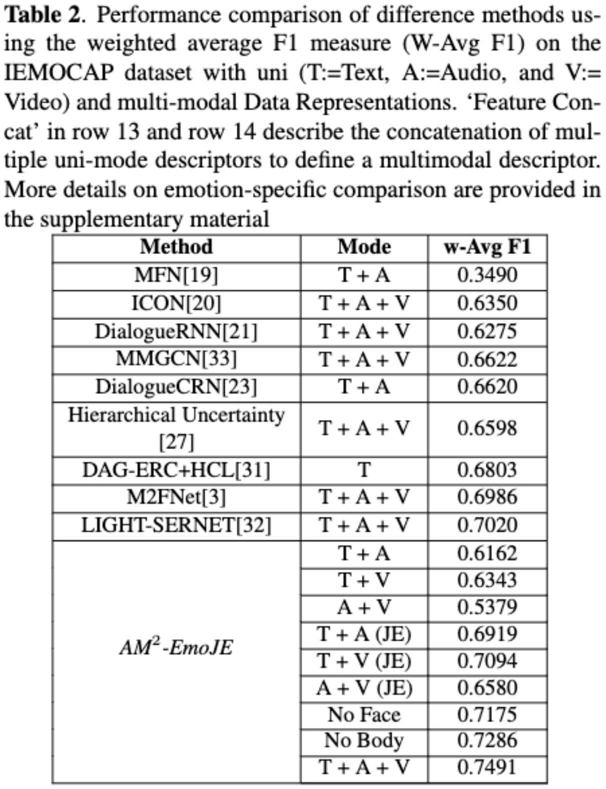


Table 2. Performance comparison of difference methods using the weighted average F1 measure (W-Avg F1) on the IEMOCAP dataset with uni (T:=Text, A:=Audio, and V:= Video) and multi-modal Data Representations. 'Feature Concat' in row 13 and row 14 describe the concatenation of multiple uni-mode descriptors to define a multimodal descriptor. More details on emotion-specific comparison are provided in the supplementary material

Method	**Mode**	**w-Avg F1**
MFN[19]	T + A	0.3490
ICON[20]	T + A + V	0.6350
DialogueRNN[21]	T + A + V	0.6275
MMGCN[33]	T + A + V	0.6622
DialogueCRN[23]	T + A	0.6620
Hierarchical Uncertainty [27]	T + A + V	0.6598
DAG-ERC+HCL[31]	T	0.6803
M2FNet[3]	T + A + V	0.6986
LIGHT-SERNET[32]	T + A + V	0.7020
AM^2-EmoJE	T + A	0.6162
	T + V	0.6343
	A + V	0.5379
	T + A (JE)	0.6919
	T + V (JE)	0.7094
	A + V (JE)	0.6580
	No Face	0.7175
	No Body	0.7286
	T + A + V	0.7491

With respect to missing modalities during training and inference in mutlimodal machine learning systems, we utilize Joint Embedding Learning to align representations across modalities that can compensate for any loss in information during training and inference.

Chapter6

Future Research Directions

Multimodal Machine Learning is an evolving field with a significant need for Foundation models to extract cross-modal information across myriad of modalities. Modality-agnostic, generalizable, robust feature representation is the main focus of this thesis. While the proposed model shows significant impact with state of the art performance, the compute required to train, infer from the proposed model is still pretty huge. This hinders from applying the current method to real-time use cases.

Inference Optimization to utilize the framework for real-time conversational scenarios is an important future direction that can be pursued.

Applying the proposed method to perform multi-task multimodal machine learning tasks to achieve improvements in performance across various downstream Computer Vision tasks. In addition to the multi-task avenue, domain adaptability and generalization of the proposed method across tasks has to be studied.

Chapter7

Conclusion

In this thesis, I study the significance of Multimodal Machine Learning systems across various domains. I study the inherent characteristics across modalities to understand the need for information representation within each modality. Furthermore, several challenges of Representation, Alignment, Reasoning and Transference are studies as a part of this thesis.

Towards Video Emotion Recognition, this thesis tackles the challenges of Representation, Alignment, Reasoning and Transference via proposed methods in phases from proposing novel architecture for cross-modal information representation followed by general-purpose vision-transformer framework for multi-label emotion recognition to proposing a method to incorporate and overcome missing information within multimodal machine learning models.

With numerous experiments across several datasets, I demonstrate superior performance gains of each proposed methods. The significance of this work lies in the fact that a general-purpose MML framework is indeed possible across various downstream Computer Vision tasks and the proposed Adaptive Fusion technique can be used to integrate into upcoming future works to result in superior performance.

Bibliography

[1] Naresh Kumar Devulapally et al. “Multi-label Emotion Analysis in Conversation via Multimodal Knowledge Distillation”. In: *Proceedings of the 31st ACM International Conference on Multimedia*. MM ’23. , Ottawa ON, Canada, Association for Computing Machinery, 2023, 6090–6100. ISBN: 9798400701085. DOI: 10.1145/3581783.3612517. URL: https://doi.org/10.1145/3581783.3612517.

[2] Naresh Kumar Devulapally et al. “AMuSE: Adaptive Multimodal Analysis for Speaker Emotion Recognition in Group Conversations”. In: *2023 IEEE Ninth Multimedia Big Data (BigMM)*. 2023, pp. 40–47. DOI: 10.1109/BigMM59094.2023.00013.

[3] Naresh Kumar Devulapally et al. *AM^2-EmoJE: Adaptive Missing-Modality Emotion Recognition in Conversation via Joint Embedding Learning*. 2024. arXiv: 2402.10921 [cs.AI].

[4] Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. *Foundations and Trends in Multimodal Machine Learning: Principles, Challenges, and Open Questions*. 2023. arXiv: 2209.03430 [cs.LG].

[5] Navonil Majumder et al. *DialogueRNN: An Attentive RNN for Emotion Detection in Conversations*. 2019. arXiv: 1811.00405 [cs.CL].

[6] Dou Hu, Lingwei Wei, and Xiaoyong Huai. “DialogueCRN: Contextual Reasoning Networks for Emotion Recognition in Conversations”. In: *CoRR* abs/2106.01978 (2021). arXiv: 2106.01978. URL: https://arxiv.org/abs/2106.01978.

[7] Soujanya Poria et al. “Convolutional MKL Based Multimodal Emotion Recognition and Sentiment Analysis”. In: *2016 IEEE 16th International Conference on Data Mining (ICDM)*. 2016, pp. 439–448. DOI: 10.1109/ICDM.2016.0055.

[8] Chung-Hsien Wu and Wei-Bin Liang. “Emotion Recognition of Affective Speech Based on Multiple Classifiers Using Acoustic-Prosodic Information and Semantic Labels”. In: *IEEE Transactions on Affective Computing* 2.1 (2011), pp. 10–21. DOI: 10.1109/T-AFFC.2010.16.

[9] Behnaz Nojavanasghari et al. “Deep Multimodal Fusion for Persuasiveness Prediction”. In: *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ICMI ’16. Tokyo, Japan: Association for Computing Machinery, 2016, 284–288. ISBN: 9781450345569. DOI: 10.1145/2993148.2993176. URL: https://doi.org/10.1145/2993148.2993176.

[10] Douwe Kiela et al. “Efficient large-scale multi-modal classification”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32. 1. 2018.

[11] Piao Shi et al. “Learning modality-fused representation based on transformer for emotion analysis”. In: *Journal of Electronic Imaging* 31.6 (2022), p. 063032.

[12] Kaitao Song et al. “MPNet: Masked and Permuted Pre-training for Language Understanding”. In: *CoRR* abs/2004.09297 (2020). arXiv: `2004.09297`. URL: `https://arxiv.org/abs/2004.09297`.

[13] Kaipeng Zhang et al. “Joint face detection and alignment using multitask cascaded convolutional networks”. In: *IEEE Signal Processing Letters* 23.10 (2016), pp. 1499–1503.

[14] Zhiwen Shao et al. “JAA-Net: Joint Facial Action Unit Detection and Face Alignment via Adaptive Attention”. In: *CoRR* abs/2003.08834 (2020). arXiv: `2003.08834`. URL: `https://arxiv.org/abs/2003.08834`.

[15] Soujanya Poria et al. *MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations*. 2019. arXiv: `1810.02508 [cs.CL]`.

[16] Carlos Busso et al. “IEMOCAP: interactive emotional dyadic motion capture database”. In: *Language Resources and Evaluation* 42.4 (2008), p. 335.

[17] Amir Zadeh et al. “Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos”. In: *arXiv preprint arXiv:1606.06259* (2016).

[18] Kaixin Ma et al. “ElderReact: a multimodal dataset for recognizing emotional response in aging adults”. In: *2019 international conference on multimodal interaction*. 2019, pp. 349–357.

[19] Charles E. Hughes Louis-Philippe Morency Behnaz Nojavanasghari Tadas Baltrusaitis. “EmoReact: A Multimodal Approach and Dataset for Recognizing Emotional Responses in Children”. In: *International Conference on Multimodal Interfaces(ICMI)* (2016).