A Self-Supervised Study of Multimodal Interactions in Opinion Videos

Jiaxin Shi

CMU-CS-22-147

August 2022

Computer Science Department
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Louis-Philippe Morency, Chair
Robert E Frederking

Submitted in partial fulfillment of the requirements for the degree of Master of Science.

Copyright © 2022 **Jiaxin Shi**

Keywords: Sentiment Analysis, Multimodal Interactions, Self-Supervised Learning

Abstract

Our experience of the world is inherently multimodal. Analyzing human multimodal language is an increasingly popular area of research that often focuses on sentimental analysis and emotion recognition, where three main modalities are present: language, acoustic, and vision. The advancements in deep learning rely heavily on the abundance of data available for the model to learn rich patterns. Due to the heavy labor required to annotate large-scale data, it is beneficial to explore what we could achieve from self-supervised learning methods. In this work, we propose a self-supervised task to study the cross-modal interactions present in the multimodal language datasets (with language, acoustic and visual modalities). We study bimodal interactions between two source modalities through our proposed self-supervised task by generating the third modality, the target modality, given the two source modalities. In other words, we quantify the information overlap between the source and target modalities while studying which multimodal interactions are used for this self-supervised task. A secondary advantage of our proposed self-supervised task is that it can also be used in downstream tasks where one of the modalities is missing. Our approach builds on the intuition that observed modalities may be able to generalize information about the missing modality. For example, people may be able to imagine the voice of a speaker when watching muted videos. In summary, this thesis is a self-supervised study on multimodal interactions in opinionated videos. Our work investigates how much information overlap exists between different modalities, quantifies the amount of cross-modal interactions, and evaluates how much information can be learned from a missing modality given other available modalities.

Acknowledgments

I would like to thank my advisor Professor Louis-Philippe Morency for his invaluable guidance and support for my previous projects and this thesis. Thank you for encouraging me to move forward with my research and become a better researcher.

I would like to thank Professor Robert E Frederking for joining my thesis committee and providing valuable feedback for my project.

I would also like to thank Dr. Amir Zadeh and Torsten Wörtwein for supporting me over the course of this project in any possible way.

Last bust not the least, I would like to thank my parents, my friends for their enormous support during the pursuit of my master degree at Carnegie Mellon University.

Contents

List of Figures

List of Tables

Chapter 1

Motivation and Introduction

1.1 Motivation

As human beings, emotions are indispensable parts of our lives: emotions aid decision-making, learning, communication, and situation awareness in human-centric environments [11]. Moreover, with the rapid development of technology, vast amounts of data are uploaded in the format of videos rather than plain texts. For example, many bloggers and consumers use cameras to record their product reviews or opinions on news, movies, and books; then, they upload the videos on social media platforms such as YouTube and Instagram. These videos tend to be full of opinions, as the speaker in the video usually compares the subject they are talking about to other similar or related matters.

With such a strong trend of using videos as the format to express opinions and emotions, researchers inevitably started attempting to build AI systems that can recognize and interpret emotions. A modality refers to the way in which something happens or is experienced [4]. Three modalities are present in videos: language, visual and audio streams. Since videos contain behavioral cues that are important for identifying sentiments of the opinion holder [11, 13], multimodal machine learning has gained popularity in sentiment analysis for its ability to interpret and reason about multimodal behaviors.

Modeling interactions across modalities is particularly crucial for multimodal emotion recognition tasks. In this paper, we divide multimodal interactions into two categories: additive and non-additive interactions. A set of modalities are said to have additive interactions if these modalities encode similar information, but combining them together will amplify or diminish the information shared among them. On the other hand, we need to model non-additive interactions when the information across modalities cannot be linearly combined. In this case, we need multimodal models to connect the complementary information across modalities. For example, if a person says, "oh my gosh, that is so ridiculous," with a smile, this person could be surprised and therefore having a strong positive emotion. However, if the person says the same sentence with a frown, then this person probably holds a negative emotion.

Despite the importance of modeling cross-model interactions, it is reported in [7] that most multimodal models are simply additive models that are equivalent to ensembles of unimodal classifiers. These models only learn unimodal contribution(UC) from each modality, meaning that

the models fail to utilize any non-additive interactions for the task. Moreover, it will be helpful to know how much information overlap exists among modalities. Information overlap refers to the shared redundant information among modalities. Two modalities have information overlap if we can transform information in one modality to the other one through modeling the unimodal contribution and multimodal interactions. Knowing how much information overlap exists between modalities allows us to know how much information we can retrieve if one modality is missing.

A significant challenge of multimodal sentiment analysis is that we lack labeled datasets as it requires much labor work to annotate the videos. The advancements in deep learning rely heavily on the abundance of data available for the model to learn rich patterns [8]. This is also the case for multimodal machine learning. There is a need for large-scale multimodal datasets, but there is a limited amount of labeled multimodal data available [3]. Moreover, the supervised strategy may introduce a biased system as it excludes the vast amount of unlabeled, unstructured video data [1]. Using self-supervised methods allows the model to be trained on large-scale datasets without requiring expensive annotations.

As humans, we form representations of the world by drawing connections among the modalities, mostly in a self-supervised way. Therefore, we believe it is meaningful to model multimodal interactions in unlabeled data. Although there has been works [1, 2] to learn multimodal representations using self-supervised strategies, these methods do not explicitly model the multimodal interactions in these unlabeled videos. To our knowledge, no previous work uses self-supervised learning to study multimodal interactions in opinionated videos. We would like to use a self-supervised task to quantify the information overlap, the amount of additive and non-additive interactions in unlabeled videos and evaluate if the representation learned from modeling interactions contains useful information for downstream tasks.

1.2 Thesis Statement

In this work, we propose a self-supervised task to study the cross-modal interactions in the multimodal language datasets (with language, acoustic and visual modalities). Through our proposed self-supervised task, we can know how much information overlap exists between different modalities, quantifies the amount of additive and non-additive interactions, and evaluates how much information useful for downstream tasks can be learned.

1.3 Contribution

Our first contribution is the Bimodal Information Transformation(BIT) model that uses two source modalities to predict the third modality(target modality) in a self-supervised task. In addition, we propose a model variant called BIT-MRO that allows us to quantify the amount of unimodal and bimodal contributions from source modalities to predict the target modality. We also present metrics to quantify (a)the information overlap, (b)unimodal contribution and bimodal non-additive interactions from the source modalities, and (c) the amount of useful information learned for emotion prediction.

1.4 Thesis Outline

This thesis is outlined as follows. In Chapter 2, we discuss the background of multimodal sentiment analysis and the related work of our project. In Chapter 3, we first state the problem formulation and then present our approach and evaluation metrics. Next, we will discuss the experiments in detail in Chapter 4: we will use metrics to quantify the information overlap, the amount of non-additive interactions and evaluate the learned representation in downstream tasks. Finally, in Chapter 5, we give out conclusions and discuss potential future works in self-supervised tasks for studying multimodal interactions.

Chapter 2

Related Work

This section discusses previous research relevant to our approach, including multimodal sentiment analysis, multimodal interaction detection, decomposition of multimodal interactions, and self-supervised method of learning multimodal representations of videos.

2.1 Multimodal Sentiment Analysis

Multimodal sentiment analysis is a vibrant area of research. The core challenge is modeling different types of multimodal interactions to fuse information from different modalities. Deep neural networks have been making consistent progress in modeling these complex interactions across modalities.

For example, the Seq2Seq-based model [10] with the attention mechanism that translates between modalities learns robust joint representations. This work is a supervised method that shares a very similar insight with ours: we can implicitly learn a joint representation through translation from a source to a target modality. The Tensor Fusion Network [23] uses a three-fold Cartesian product that explicitly learns unimodal, bimodal, and trimodal dynamics. Memory Fusion Network [24] uses a system of LSTMs to encode unimodal interactions. It uses an attention mechanism and a memory storage unit to encode cross-modal interactions. The multimodal transformer proposed in [17] learns pairwise cross-modal attention to model interactions between unaligned multimodal sequences at different time steps.

2.2 Detection of Cross-Modal Interactions

Despite the success of these deep neural networks, it is hard to interpret and reason the decision made by these black-box models. To better understand models and improve performances and make their decisions more interpretable, it is critical to understand the complex inter- and intra-modal dynamics.

M^2Lens [20] is an interactive system that provides explanations of intra- and inter-modal interactions at the global, subset, and local levels. It summarizes three interaction types(dominance, complement, and conflict) on the model predictions. GLIDER [18] detects and encodes global

feature interactions in black-box recommender systems. It utilizes gradient-based Neural Interaction Detection and LIME [15].

There are also post-hoc explanation techniques for black-box multimodal models, such as using guided backpropagation and occlusion to interpret VQA models [6]. They often identify important features that influence the model decision the most. However, these post-hoc methods often only give local explanations in one modality.

2.3 Decomposing Multimodal Interactions

To increase model explainability, diagnostic tools are developed to distinguish different types of multimodal interactions learned in multimodal frameworks.

EMAP [7] uses function projection to diagnose if cross-modal interactions improve model performance. It finds the closest approximation of a bimodal model $f(A, B)$ such that the approximation has the form $g(A) + h(B)$. The approximation learns only so-called additive interactions between the two modalities A and B. However, it is only able to disentangle unimodal contributions and does not give any explanation of multimodal interactions.

DIME [9] is a LIME-based multimodal explanation tool that provides information on both unimodal interactions and bimodal interactions. Moreover, it provides visualizations of each type of interaction on text and image modalities.

2.4 Self-Supervised Multimodal Learning

With the vast amount of unlabeled multimodal data available and so limited resources of annotated videos, researchers have started to self-supervised methods to learn multimodal representations from large-scale unlabeled videos.

Self-Supervised Multimodal Versatile Networks(MMV) [2] learn representations from multimodal videos in a self-supervised manner. For each modality, a backbone network learns the representation that respects the specificity of that modality. A modality embedding graph is constructed to store the embeddings of all modalities. Moreover, the learned embeddings are evaluated on various downstream tasks.

VATT [1] proposes a transformer-based architecture for learning representations from unlabeled videos. Similar to MMV, it learns modality-specific backbone networks and a modality-agnostic network for learning the shared information among all modalities. The difference between VATT and MMV is that it uses the Transformer [19] as the modality-specific and modality-agnostic backbone model.

Chapter 3

Self-Supervised Study of Multimodal Interactions

This chapter provides background and settings for our project. Then we describe the formulation, model, and evaluation metrics of our self-supervised study of multimodal interactions.

3.1 Background

With the advent of the Internet and the widespread use of electronic devices with cameras, individuals can broadly express their opinions in videos instead of text. As a result, multimodal sentimental analysis is an increasingly popular area of research that focuses on generalizing text-based sentiment analysis to opinionated videos. In this paper, we focus on three communicative modalities present: language (spoken words), visual (gestures), and acoustic (voice) [23].

Figure 3.1 shows three types of multimodal interactions: unimodal, bimodal, and trimodal interactions, and illustrates how these interactions can affect how we predict emotion. Sometimes we can predict the emotion using only unimodal interactions. As shown in figure 3.1, the utterance ”The book is amazing” is sufficient for predicting the speaker’s attitude. However, in other cases, we may need more context before confidently making a prediction. For example, the sentence ”The book is sick” could contain an ambiguous attitude that is hard to determine if it is positive without any other information. If the speaker, at the same time, is smiling, then it will be perceived as a favorable opinion. On the contrary, if the speaker frowns when saying the sentence, the opinion will be negatively perceived. However, the speaker’s attitude may remain unclear if the speaker says, ”The book is sick,” using a loud voice. In this case, we still cannot determine if the speaker’s opinion of the book is favorable. These examples above are illustrations of unimodal and bimodal interactions. Figure 3.1 also shows that we can obtain more information, such as the sentiment degree that describes how strong the opinions are. For example, a person speaking ”the book is fair” with a smile and a loud voice presents a much stronger positivity than speaking the same sentence with a smile but a low voice.

As illustrated above, we can see that modeling different types of interactions is crucial in sentiment analysis. However, most studies on multimodal interactions are done in supervised settings. Therefore we would like to know: can we model these interactions in a self-supervised

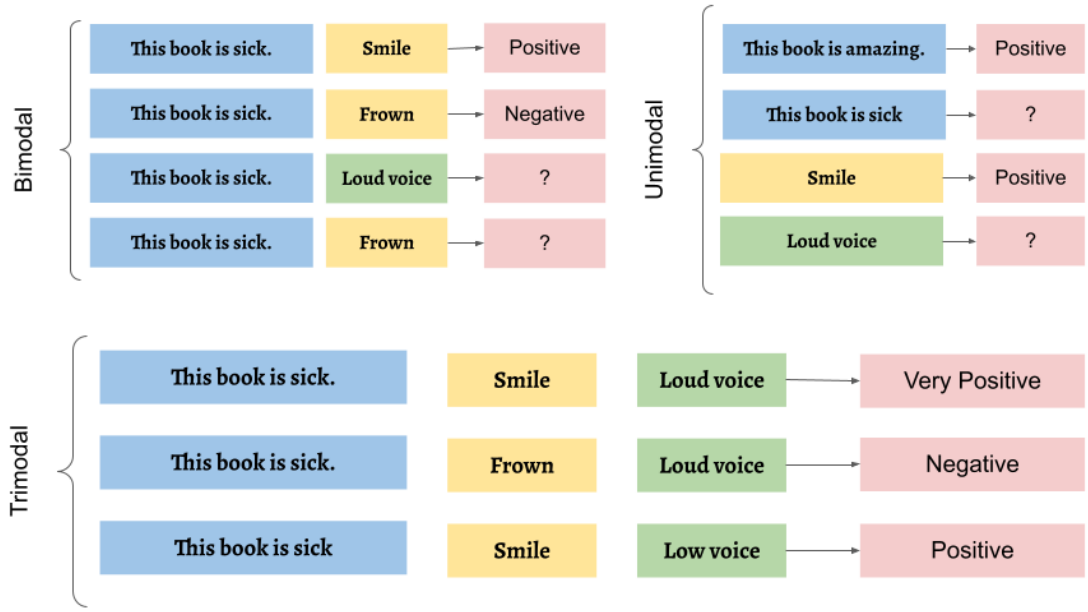

Figure 3.1: Different types of multimodal interactions in sentiment analysis.

setting? More specifically, how much about multimodal interactions could we learn with unannotated multimodal data?

This project is one of the first attempts to understand multimodal interactions in opinionated videos without the aid of any human annotation. Our approach builds on the intuition that observed modalities might be able to generalize information about the missing modality. For example, people may be able to imagine the voice of a speaker when watching muted videos. Given two modalities and two video clips, there should be multimodal interactions that can predict the similarity of the third missing modality of the two video clips. In other words, if we are given two video clips with no sound, can we predict how similar the sound of the two videos is? In the following part of the paper, the two present modalities will be called source modalities, and the missing modality will be called the target modality. A formal definition of the task will be stated in section 3.2.

With this main task of predicting the similarities of the target modality given two source modalities of two clips, we are also interested in knowing:

- How much information overlap exists between the source and target modalities?
- When we try to transform information from source modalities to the target modality, how many additive and non-additive interactions are present?
- What can the learned representation achieve in downstream tasks?

3.1.1 Information Overlap

Information overlap in multimodal data refers to having shared, redundant information among two or three modalities. Our self-supervised method tries to transform information from two source modalities to the target modality. The information transformation involves modeling both additive and non-additive interactions. The unimodal contribution from each source modality can be learned through additive interactions, and the non-additive interactions allow the model to translate more complex information that requires reasoning over both source modalities.

Quantifying information overlap allows us to know the consistency across modalities. A video is consistent across modalities if all three modalities contain similar emotional cues. For example, a speaker with strong emotion will speak with a louder voice and have exaggerated facial expressions. Likewise, a speaker with a negative attitude will show redundant cues in the speaker's voice and facial expression. However, when the modalities are less consistent, the information overlap will be much smaller. For example, if the video is recorded in a very dark environment, it is hard to transform information from the visual modality. Therefore, measuring information overlap allows us to detect noise in videos and gives us an idea of how consistent the emotional cues are across the modalities.

Modeling Additive and Non-Additive Interactions

To quantify source modalities' unimodal and bimodal contributions, we will measure additive and non-additive interactions as a proxy metric for information overlap. Since it was found that we can use a simple additive model such as MLP to predict emotions if there is redundant information across the modalities [12], we would conduct experiments to verify if the model performance on our self-supervised task is also mainly attributed to unimodal contributions.

We say f is an additive model if it is an ensemble of two unimodal models(f_{M_1} and f_{M_2}) of modality M_1 and M_2, respectively.

$$f(m_1, m_2) = f_{M_1}(m_1) + f_{M_2}(m_2) \tag{3.1}$$

As defined in 3.1, we use an additive model where two modalities are involved in learning the unimodal contributions of each modality.

It is also important to model non-additive interactions. For example, some sentences can have different meanings accompanying different cues from other modalities. A video of a person saying "this movie is sick" with a smile will have very different acoustic features from a video with a person saying the same sentence but with a frown. In this case, we need to interactively transform information from both source modalities to connect the information.

To quantify additive and non-additive interactions, we will use cosine similarity to evaluate how well the additive and the non-additive model transforms information from source modalities to the target modality, respectively. To measure information overlap, we will evaluate how well the overall model(consisting of additive and non-additive models) transforms information.

3.1.2 Evaluate the usefulness of the Representation in Downstream Tasks

We are also interested in knowing if modeling interactions from source to target modalities allows the representation to encode some useful information for downstream tasks such as predicting

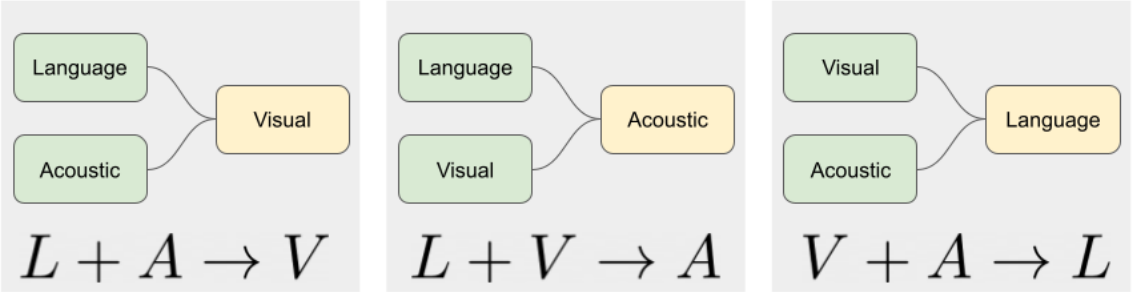

Figure 3.2: Three problem setups: (1) $L + A \rightarrow V$: predict visual from language and acoustic modalities; (2) $L + V \rightarrow A$: predict acoustic from language and visual modalities. (3) $V + A \rightarrow L$: predict language from visual and acoustic modalities.

emotion, and if yes, to what extent. We will use our learned representations of the target modality as the unimodal input to a unimodal model that predicts emotion. The mean absolute error(MAE) of the model prediction and the correlation between the prediction and true labels will serve as metrics for evaluating how much information relevant to predicting emotions can be transferred from the source to target modalities through modeling additive and non-additive interactions.

3.2 Problem Setup

In this project, we focus on modeling multimodal interactions in multimodal language datasets consisting of three modalities: language, acoustic and visual modalities.

We propose a self-supervised task to predict the target modality from two source modalities through modeling multimodal interactions. More specifically, we would like to know how much information is translatable from source to target modalities.

As shown in figure 3.2, two modalities in the green boxes are source modalities, and the target modality is in the yellow box. We would like to measure how much information overlap between the source and the target modalities. Moreover, we would like to know if a certain combination of the source and target modalities has more information overlap than the other combination. Since we have three modalities, we have three problem setups:

1. using language and acoustic to predict visual modality
2. using language and visual to predict acoustic modality
3. using visual and acoustic to predict language modality

3.3 Our Method: BIT(-MRO)

This section describes our approach to the self-supervised study of multimodal interactions. We propose the Bimodal Information Translation(BIT) model, which is designed to model the unimodal additive interactions and the bimodal cross-modal interactions to translate information from two source modalities to the target modality. The vanilla version of our BIT model does not learn the additive and non-additive interactions separately; therefore, we propose BIT-MRO to learn and decompose the learned embedding into unimodal and bimodal interactions. Hugely inspired by Multimodal Residual Optimization(MRO) [21], the BIT-MRO uses a different loss function from BIT. This loss function prioritizes learning the additive interactions before learning non-additive ones. This method allows us to quantify how much unimodal and bimodal interactions are useful in learning a meaningful representation of the target modality. We then evaluate the effectiveness of the learned representation with different metrics and downstream tasks.

3.3.1 Problem Formulation

In this section, we give a formal mathematical formulation of our problem. We will introduce the model architecture and the loss functions of the BIT(-MRO) model.

We consider models $f(\cdot)$ such that takes a pair of input $(d_1, d_2) = ((m_1^{d_1}, m_2^{d_1}), (m_1^{d_2}, m_2^{d_2}))$, each input consists of 2 modalities, denoted by m_1 and m_2, the superscript describes which data point the modality belongs to. The output of the model $f(d_1, d_2) = \hat{y} \in \mathcal{R}^{\dim(M_3)}$, where $\hat{y}$ is the learned representation of the third modality, given d_1 and d_2, each having two modalities M_1, M_2. The output $\hat{y}$ is of the same size of the embedding of M_3, denoted by $\dim(M_3)$. Intuitively, we would doubt if d_1 and d_2 contains enough information to reconstruct a good representation of the third modality M_3. Therefore we design the model to predict the differences between the third modalities. More specifically, $\hat{y}$ should approximate $m_3^{d_1} - m_3^{d_2}$.

3.3.2 Model Architecture

Bimodal Information Translation(BIT)

We first introduce the BIT model, which acts as a baseline model of our BIT-MRO model. As shown in figure 3.3, it consists of two branches: a unimodal branch that takes in one modality and a bimodal branch that takes inputs from both modalities.

Therefore, the model makes predictions from both predictions made by modeling the additive and non-additive interactions:

$$\hat{y} = \hat{y}_{\text{uni}} + \hat{y}_{\text{bi}}, \tag{3.2}$$

where $\hat{y}_{\text{uni}}$, $\hat{y}_{\text{bi}}$ are the prediction made from the unimodal(additive) and the bimodal(non-additive) branch, respectively. More specifically,

$$\hat{y}_{\text{uni}} = f_{M_1}(m_1) + f_{M_2}(m_2) \tag{3.3}$$

and

$$\hat{y}_{\text{bi}} = f_{(M_1,M_2)}(m_1, m_2) \tag{3.4}$$

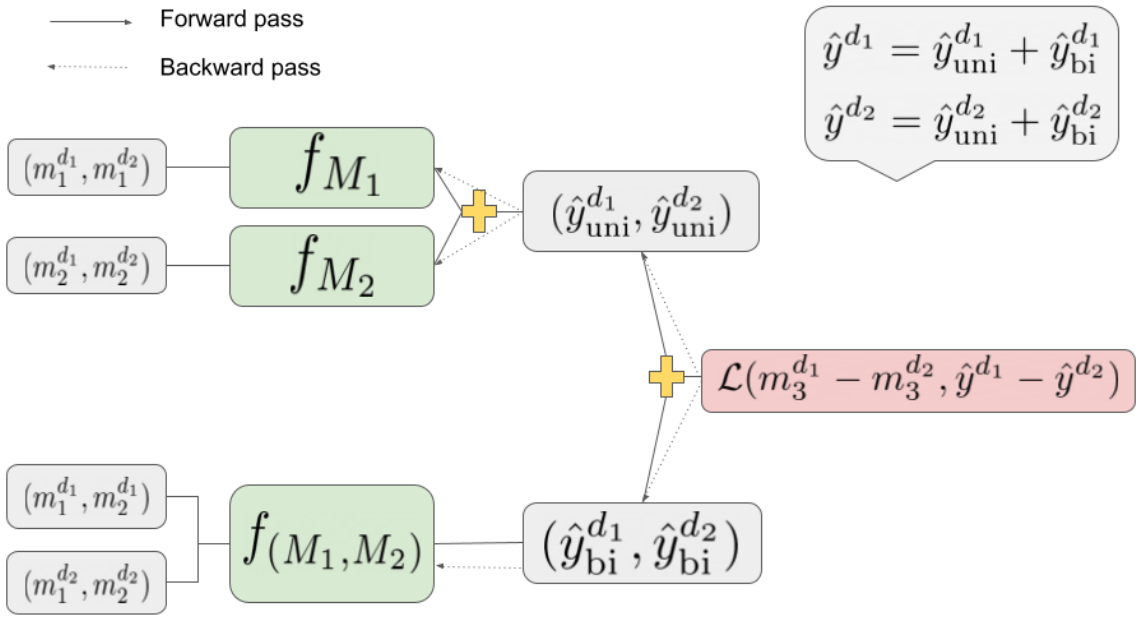

Figure 3.3: Overview of the BIT model architecture

where $f_{M_1}(\cdot)$ and $f_{M_2}(\cdot)$ are neural networks that only uses one modality as an input, and $f_{(M_1,M_2)}(\cdot,\cdot)$ are neural networks that takes both modalities as input. These models are independent of each other and do not share any parameters.

Given $d_1 = (m_1^{d_1}, m_2^{d_1})$ and $d_2 = (m_1^{d_2}, m_2^{d_2})$, the unimodal branch will compute $\hat{y}^{d_1}_{\text{uni}}$ and $\hat{y}^{d_2}_{\text{uni}}$ to approximate the additive part of $m_3^{d_1}$ and $m_3^{d_2}$. Likewise, the bimodal branch will output $\hat{y}^{d_1}_{\text{bi}}$ and $\hat{y}^{d_2}_{\text{bi}}$. Combining results from both branches, $\hat{y}^{d_1}_{\text{uni}} + \hat{y}^{d_1}_{\text{bi}}$ will approximate $m_3^{d_1}$.

To maximize the similarity between the true and the predicted embeddings, we optimize the BIT model by minimizing the negative cosine similarity as follows:

$$\mathcal{L}(m_3^{d_1} - m_3^{d_2}, \hat{y}^{d_1} - \hat{y}^{d_2}) = -\text{CosineSimilarity}(m_3^{d_1} - m_3^{d_2}, \hat{y}^{d_1} - \hat{y}^{d_2}) \tag{3.5}$$

This model structure aims to use two branches to learn additive and non-additive interactions. Ideally, we would like the bimodal branch to only approximate non-additive interactions. However, the two branches are optimized simultaneously, and the hollow structure of the bimodal branch may not be able to capture complex non-additive interactions; therefore, we cannot guarantee that the bimodal branch only learns non-additive interactions and vice versa.

Bimodal Information Translation with MRO Loss(BIT-MRO)

Instead of using a single loss function to optimize both branches simultaneously, we use MRO loss [21] to approximate $m_3^{d_1} - m_3^{d_2}$ in a way such that the model first learns unimodal contributions and then uses bimodal contributions to correct the mistakes made by purely using unimodal

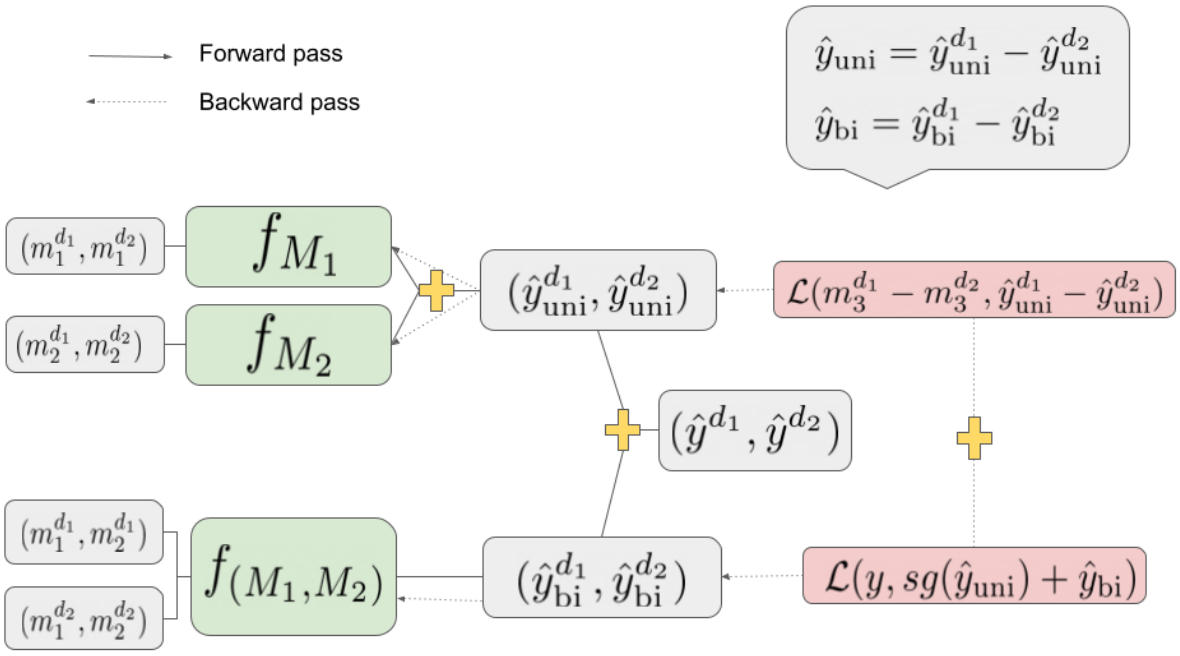

Figure 3.4: Overview of the BIT-MRO model architecture: f_{M_1}, f_{M_2} learns unimodal features from M_1 and M_2 respectively, and $f_{(M_1,M_2)}$ learns bimodal contributions that can't be learned from the unimodal models.

contributions. As mentioned in [21], the high-level intuition of MRO is that "(simpler) unimodal interactions should be learned before learning (more complex) bimodal and trimodal interactions.

The loss function of BIT-MRO is defined as:

$$L(y, \hat{y}) = -\text{CosineSimilarity}(y, \hat{y}_{\text{uni}}) - \text{CosineSimilarity}(y, sg(\hat{y}_{\text{uni}}) + \hat{y}_{\text{bi}}) \tag{3.6}$$

where $\hat{y}_{\text{uni}} = \hat{y}_{\text{uni}}^{d_1} + \hat{y}_{\text{uni}}^{d_2}$, $\hat{y}_{\text{bi}} = \hat{y}_{\text{bi}}^{d_1} + \hat{y}_{\text{bi}}^{d_2}$. sg refers to stop-gradient [14], as we do not back-propagate again through the unimodal branches when learning bimodal contributions. Figure 3.4 shows that we first update model parameters of the unimodal branch to calculate $L(y, \hat{y}_{\text{uni}})$ using only the unimodal branch. The second term in the loss function uses predictions from both unimodal and bimodal branches, but we do not back-propagate through the unimodal branch and only update the parameters in the bimodal branch.

The MRO loss allows us to know what types of compositionality that f uses over the two input modalities to predict the difference between the third modalities of the two data inputs so that we can quantify the additive and non-additive interactions. This stop gradient part forces the unimodal branch to learn additive contributions first, then learns bimodal contributions to correct the mistakes made in $\hat{y}_{\text{uni}}$. As a result, the unimodal branch should only learn additive interactions, and the bimodal branch should learn things that cannot be learned by the unimodal branch, which is non-additive interactions.

3.4 Experiments and Evaluation

We design the following three experiments to evaluate our model BIT(-MRO) and to answer the three questions we asked in section 3.1:

1. Sanity check: we will run the model on synthetic datasets with a controlled amount of additive and non-additive interactions to verify that our model properly quantifies additive and non-additive interactions. We also run EMAP [7] to ensure the unimodal and bimodal branches of BIT-MRO are only learning additive and non-additive interactions, respectively. Moreover, we conduct experiments on the synthetic dataset with different levels of information overlap to confirm that cosine similarity is a proxy metric for measuring the amount of information overlap.
2. Quantifying information overlap: we use cosine similarity to quantify both unimodal and bimodal contributions and use it as a proxy metric for quantifying information overlap between the source and the target modalities in MOSI dataset [22]. Moreover, we will compare if additive or non-additive interactions contribute more to the self-supervised information transformation.
3. Quantifying useful feature learned for emotion prediction: we use the learned representation of the target modality as the unimodal input to predict emotion. We would like to use this task as a metric to quantify how much information learned through our self-supervised task can be useful for predicting emotion.

Chapter 4

Experimental Methodology

In this chapter, we will discuss our experimental methods in detail. First, we will introduce the synthetic dataset and the real multimodal dataset we used for the experiments. Two experiments are conducted to evaluate if BIT(-MRO) can learn the additive and non-additive interactions separately. Then, we will evaluate how we use the amount of additive and non-additive interactions as a proxy metric for information overlap. To evaluate what else is learned besides information overlap, we use the representation output by BIT(-MRO) to perform a downstream task and compare MAE and correlation for both BIT and BIT-MRO to investigate if disentangling additive and non-additive interactions influences how the representation performs on the downstream task.

4.1 Dataset

4.1.1 Synthetic Dataset

We generate a set of synthetic multimodal data that we have control over the amount of additive and non-additive interactions to check if our model learns the two types of interactions separately. Moreover, we generate a set of synthetic multimodal data with different levels of information overlap between the source and target modalities to verify that using cosine similarity to measure the amount of unimodal and bimodal contributions can quantify the amount of additive and non-additive interactions when we transform information from source modalities to the target modality. Therefore, using cosine similarity to quantify interactions can serve as a proxy metric for information overlap. (Note: the synthetic dataset is generated fully randomly and does not contain information about any real data.)

The synthetic dataset contains three modalities, denoted by (l, v, a), with l and a being the source modalities and v being the target modality according to the following process:

1. Sample N vectors of l and a from a multivariate normal distribution with zero mean and unit variance.
2. To avoid finite sampling biases, we will multiply the sampled l and a with the following scalars: $(1, 1), (-1, 1), (1, -1), (-1, -1)$. These $4N$ samples will become the l and a modality of our synthetic dataset.
3. We would like to control the amount of additive and non-additive information overlap;

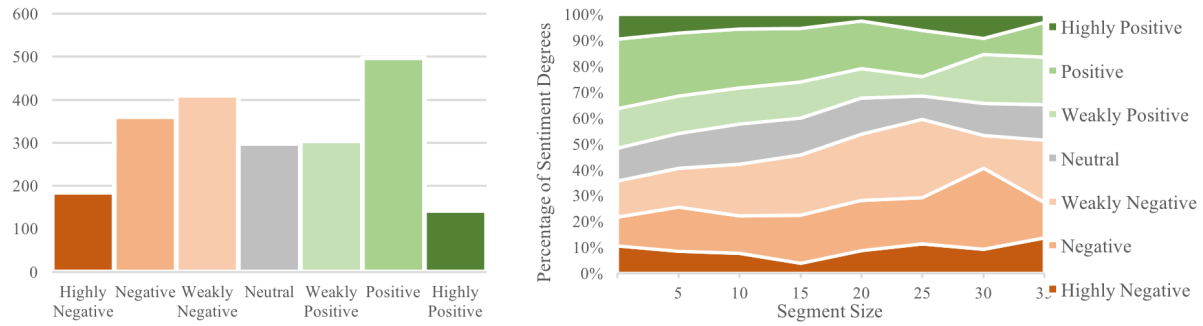

Figure 4.1: A figure from MOSI paper [22]: Left: the distribution of sentiment over MOSI. Right: percentage of each sentiment degree per segment size.

therefore, for the purely additive dataset, we define $v = l + a$. For the strictly non-additive dataset, we define $v = l \times a$.

4. We also want to see how the model responds to different levels of additive information overlap. The additive dataset we created in step 3 has 100% information overlap. We also created additive datasets with 75% and 50% information overlap by randomly choosing 25% and 50% datapoints and Sample v of those datapoints randomly.

4.1.2 MOSI

We use the Multimodal Opinion-level Sentiment Intensity dataset(MOSI) [22] to study the sentiment in real online opinion videos. This dataset contains three modalities: language(L), acoustic(A), and visual(V) modality, and it is annotated with sentiment intensity. The sentiment intensity ranges from -3 to $+3$ with a linear scale to denote the sentiment from strongly negative to strongly positive.

Table 4.1 shows statistics of the MOSI dataset, and figure 4.1 shows the distribution of sentiment over the entire dataset on the left and the percentage of each sentiment degree per segment size(number of words in opinion segment) on the right.

Total number of opinion segments	2199
Total number of videos	93
Total number of distinct speakers	89
Average length of opinion segments	4.2 sec
Average word count per opinion segment	12

Table 4.1: MOSI dataset statistics.

4.2 Evaluation

4.2.1 Sanity Check

BIT-MRO prioritizes learning additive interactions

Dataset	Model	Branch	Cosine Similarity
Additive Synthetic (100% information overlap)	BIT	Both	1.00
		Unimodal	0.0
		Bimodal	1
	BIT-MRO	Both	1.00
		Unimodal	1
		Bimodal	0.0
Non-additive Synthetic (100% information overlap)	BIT	Both	0.93
		Unimodal	0.08
		Bimodal	0.59
	BIT-MRO	Both	0.79
		Unimodal	-0.01
		Bimodal	0.49

Table 4.2: A sanity check on how well our model decomposes additive and non-additive information.

First, we run BIT and BIT-MRO on the strictly additive synthetic dataset with 100% information overlap. Although both BIT and BIT-MRO are able to predict the differences in v from l and a, from results reported in table 4.2, we can see that in the BIT model, all contributions come from the bimodal branch, and we are not utilizing the unimodal branch at all. However, the BIT-MRO model is able to capture all the additive interactions used for predicting v. The model was able to recognize that the dataset only contains additive interactions, therefore not using the bimodal branch at all.

Then we run both models on the synthetic dataset that contains only non-additive information. As reported in table 4.2, neither model was able to achieve 1 on the cosine similarity metric. This implies that both BIT and BIT-MRO are not able to capture all non-additive interactions. However, the unimodal branch of BIT-MRO contributes less than that of the BIT model. This implies that the BIT-MRO model is able to recognize there is little to no additive information in the dataset.

This experiment verifies that BIT-MRO effectively decomposes the additive and non-additive interactions. The unimodal branch is able to capture all additive interactions, and BIT-MRO is able to recognize there is no additive interactions present and stop using unimodal contributions when all interactions are non-additive.

Sanity Check 2: The bimodal branch does not contain additive information

To further confirm that the bimodal branch of BIT-MRO only contains non-additive interactions, we run EMAP on the bimodal branch of our models trained on MOSI.

Empirical Multimodally-Additive Function Projection(EMAP) [7] projects the model onto the set of multimodally-additive functions. If the bimodal learns nothing additive, then the EMAP score computed from the bimodal branch output will be close to zero. As shown in table 4.3, the EMAP scores of the bimodal branches trained on the MOSI dataset are all very close to zero, with the baseline model having a slightly larger EMAP score. This verifies that the bimodal branch does not learn additive interactions.

Dataset	Target modality	Model	EMAP score
MOSI	language	BIT	0.04
		BIT-MRO	0.01
	acoustic	BIT	9.99×10^{-6}
		BIT-MRO	0.001
	visual	BIT	3.13×10^{-6}
		BIT-MRO	0.00

Table 4.3: We use EMAP to verify that the bimodal branch do not learn additive information.

Sanity check 3: Quantifying the amount of information overlap

In table 4.4, we report the model performance on our additive synthetic dataset with different levels of information overlap. As shown in figure 4.2, the amount of additive information overlap is best reflected by the cosine similarity of the output from the unimodal branch of the BIT-MRO model.

4.2.2 A proxy metric for information overlap

Now that we have shown that cosine similarity could be a proxy metric for quantifying information overlap, we report the performance on the three settings we are interested in learning(mentioned in 3.2).

We performed a grid search on each setting and chose the best model based on the performance on the validation set. The metric function g is defined as follows:

$$g(m_3^{d_1} - m_3^{d_2}, \hat{y}^{d_1} - \hat{y}^{d_2}) = -\text{CosineSimilarity}(m_3^{d_1} - m_3^{d_2}, \hat{y}^{d_1} - \hat{y}^{d_2}) \tag{4.1}$$

We also report the performance of the unimodal and bimodal branches of the BIT-MRO model to see how much contribution comes from the additive part.

As shown in table 4.5, among all three settings, we were able to best predict the acoustic modality from language and visual modalities, and it was the hardest to predict L from V and A. This result does confirm our intuition that language is a more abstract and complex modality than the acoustic and visual modalities. As a result, we are able to transfer more information through learning unimodal contributions from the language modality.

Moreover, BIT-MRO barely uses non-additive interactions to predict the target modality, as the cosine similarity reported on the unimodal branch is very close to that reported from both branches. We suspect that the bimodal branch is too simple to pick up the complex non-additive

Dataset	Model	Branch	Cosine Similarity
Additive Synthetic (100% information overlap)	BIT	Both	1.00
		Unimodal	0.00
		Bimodal	1.00
	BIT-MRO	Both	1.00
		Unimodal	1.00
		Bimodal	0.00
Additive Synthetic (75% information overlap))	BIT	Both	0.84
		Unimodal	0.49
		Bimodal	0.63
	BIT-MRO	Both	0.83
		Unimodal	0.82
		Bimodal	0.76
Additive Synthetic (50% information overlap))	BIT	Both	0.72
		Unimodal	0.20
		Bimodal	0.55
	BIT-MRO	Both	0.71
		Unimodal	0.45
		Bimodal	0.67

Table 4.4: Cosine similarity of BIT(-MRO) on the synthetic dataset with different levels of additive information overlap

Source Modalities	Target modality	Model	Branch	Cosine Similarity
Visual+Acoustic ($V + A$)	Language (L)	BIT	Both	0.0079
		BIT-MRO	Both	0.0059
			Unimodal	0.0060
			Bimodal	0.0058
Language+Acoustic($L + A$)	Visual (V)	BIT	Both	0.0935
		BIT-MRO	Both	0.1165
			Unimodal	0.1153
			Bimodal	0.01953
Language+Visual($L + V$)	Acoustic (A)	BIT	Both	0.126
		BIT-MRO	Both	0.1308
			Unimodal	0.1308
			Bimodal	0.0

Table 4.5: Use cosine similarity to quantify information overlap captured by the BIT-MRO model in three settings.

interactions since BIT-MRO could not reach the value of 1 on the cosine similarity metric on the strictly non-additive synthetic dataset.

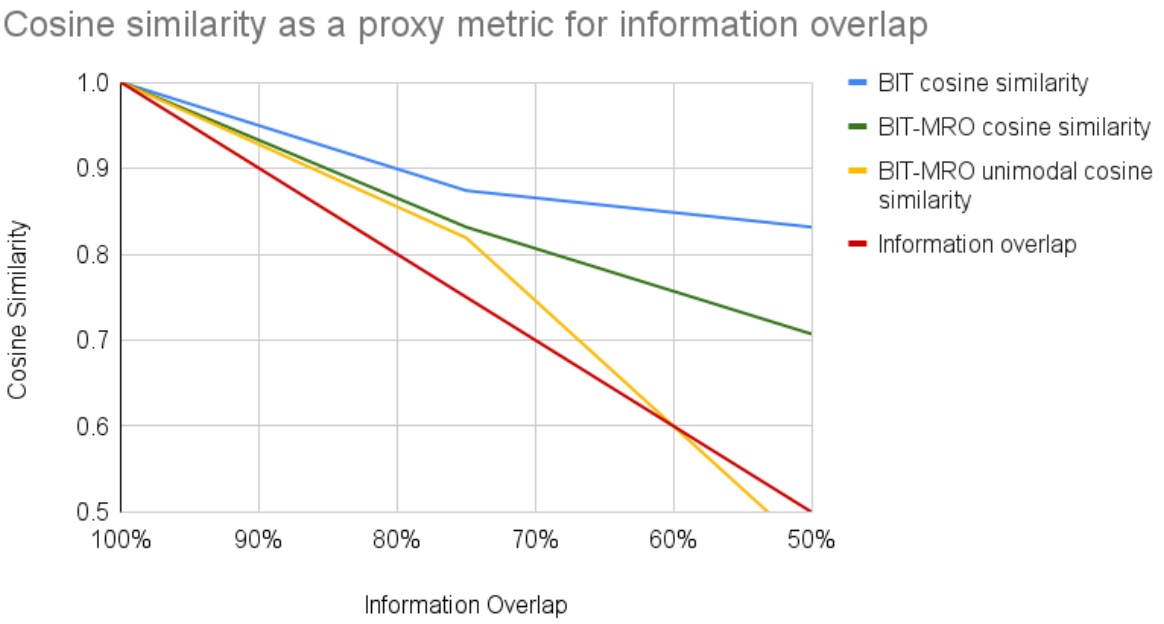

Figure 4.2: Cosine similarity vs. information overlap: cosine similarity of embeddings learned in the unimodal branch of BIT-MRO effectively quantifies additive information overlap. The red line is the information overlap we would like to approximate.

4.2.3 What is learned besides information overlap

We are also interested in knowing if any information useful for emotion prediction can be learned through this self-supervised task. We train unimodal models to predict emotion using the unimodal embeddings we learned in the self-supervised task. Moreover, we will compare the performances of the models that are trained on embeddings learned from BIT and BIT-MRO to see if separating additive and non-additive interactions will make a difference in their performance in the downstream task.

For each embedding, we did a grid search on $[0, 1, 2]$ layers of MLP with learning rate $[0.005, 0.001, 0.0001]$ and decay rate $[0.0, 0.01, 0.001]$, and we report the MAE score and correlation between the prediction and true labels in table 4.6.

Interpreting the result

From table 4.6, the MAE scores of models trained on all six embeddings are very close. The correlation is the highest when we use visual embedding learned in BIT to predict emotion. A possible explanation is that since the BIT model did not learn additive and non-additive interactions separately, the model is able to capture more information relevant to emotional cues rather than interactions. Another possible factor could be that more non-additive interactions are captured in the self-supervised task of transforming information in language and acoustic to visual modalities(in table 4.5). The non-additive interactions encode useful cross-modal interactions

that are helpful for predicting emotions.

On the contrary, we observed that the correlation is negative when we use the language embedding learned from the acoustic and visual modalities. The first possible reason is that there was the least information overlap detected when we transformed information from visual and acoustic to language modality(in table 4.5); therefore, the language embedding encodes little information. This result also confirms the fact that the visual and acoustic modalities are the less helpful modalities for predicting emotions(as shown in 4.7). Using the original visual and acoustic already gives a very small correlation score; therefore, the model may not be able to pick up useful information for predicting emotions from our self-supervised task.

Through our experiment of evaluating learned representations for emotion prediction, we have the following observations:

1. when the language modality is missing, it is hard to transform enough useful information from the other two modalities for emotion prediction.
2. Although modeling additive and non-additive interactions separately using MRO loss allow us to approximate the amount of information overlap, we may sacrifice the possibility of learning more useful information for downstream tasks.
3. Capturing non-additive interactions in our self-supervised information transformation task may be helpful for learning cross-model interactions that predict emotion.

Modality	Embedding	MAE	Corr
Language	BIT	0.8897	-0.41
	BIT-MRO	0.8819	-0.23
Visual	BIT	0.8814	0.25
	BIT-MRO	0.8870	0.07
Acoustic	BIT	0.8870	0.04
	BIT-MRO	0.8771	0.14

Table 4.6: Unimodal performances of different embeddings on predicting emotion.

Modality	Model	Corr
Language	SVR	0.68
	Random Forest Regressor	0.52
Visual	SVR	0.12
	Random Forest Regressor	0.17
Acoustic	SVR	0.07
	Random Forest Regressor	0.07

Table 4.7: Unimodal baseline performance of each modality in the MOSI dataset.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

In this work, we aim to explore what we could learn about opinionated videos using self-supervised learning methods. We proposed a self-supervised task to study the cross-modal interactions present in the multimodal language datasets (with language, acoustic and visual modalities).

We defined a proxy metric to quantify the information overlap between modalities through modeling additive and non-additive interactions. Our model detected the most information overlap from language and visual modalities to acoustic modality. Moreover, we observed that our model only utilized additive information overlap to predict the target modality. A possible explanation could be that we need a more complex bimodal branch to capture the non-additive information. It is also possible that non-additive interactions are not very helpful in translating information between modalities.

We also evaluated our model on how much information useful for predicting emotion is learned and gained several insights on under what circumstances we are able to learn more information useful for downstream tasks through our self-supervised task.

In summary, this thesis is a self-supervised study on multimodal interactions in opinionated videos. Our work investigates how much information overlap exists between different modalities, quantifies the amount of cross-modal interactions, and evaluates how much information can be learned from a missing modality given other available modalities.

5.2 Future Work

Our work is a preliminary attempt to study multimodal interactions in sentiment dataset with a self-supervised task. We propose directions of possible future work:

1. Evaluate the model on more multimodal datasets on sentiment such as CMU-MOSEI [3], IEMOCAP [16], POM [5], etc.
2. We can go beyond emotions and try to model multimodal interactions on other types of multimodal datasets.

3. We can transform information between not just language, acoustic and visual modalities. Moreover, the number of source and target modalities can be more than two or one.
4. We can explore other architectures of the bimodal branch to increase its capability of capturing non-additive interactions.
5. We may need to design modality-specific models to transform information, as it is indicated that language is a more coarse-grained modality than acoustic and visual modalities [2]. The model can adapt to the natures of different modalities and could possibly capture more interactions.
6. Future work can be done to further decompose non-additive interactions. For example, multiplicative interaction is a subset of non-additive interactions. We can try to transform information through modeling additive, multiplicative and non-multiplicative non-additive interactions.

Bibliography

[1] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. VATT: Transformers for multimodal self-supervised learning from raw video, audio and text. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=RzYrn625bu8. 1.1, 2.4

[2] Jean-Baptiste Alayrac, Adrià Recasens, Rosalia Schneider, Relja Arandjelovic, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. Self-supervised multimodal versatile networks. *CoRR*, abs/2006.16228, 2020. URL https://arxiv.org/abs/2006.16228. 1.1, 2.4, 5

[3] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1208. URL https://aclanthology.org/P18-1208. 1.1, 1

[4] Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *CoRR*, abs/1705.09406, 2017. URL http://arxiv.org/abs/1705.09406. 1.1

[5] Alexandre Garcia, Slim Essid, Florence d'Alché-Buc, and Chloé Clavel. A multimodal movie review corpus for fine-grained opinion mining. *CoRR*, abs/1902.10102, 2019. URL http://arxiv.org/abs/1902.10102. 1

[6] Yash Goyal, Akrit Mohapatra, Devi Parikh, and Dhruv Batra. Interpreting visual question answering models. *CoRR*, abs/1608.08974, 2016. URL http://arxiv.org/abs/1608.08974. 2.2

[7] Jack Hessel and Lillian Lee. Does my multimodal model learn cross-modal interactions? it's harder to tell than you might think! 2020. doi: 10.48550/ARXIV.2010.06572. URL https://arxiv.org/abs/2010.06572. 1.1, 2.3, 1, 4.2.1

[8] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. A survey on contrastive self-supervised learning. *CoRR*, abs/2011.00362, 2020. URL https://arxiv.org/abs/2011.00362. 1.1

[9] Yiwei Lyu, Paul Pu Liang, Zihao Deng, Ruslan Salakhutdinov, and Louis-Philippe Morency. Dime: Fine-grained interpretations of multimodal models via disentangled lo-

cal explanations. *arXiv preprint arXiv:2203.02013*, 2022. 2.3

[10] Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. Found in translation: Learning robust joint representations by cyclic translations between modalities. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6892–6899, Jul. 2019. doi: 10.1609/aaai.v33i01.33016892. URL https://ojs.aaai.org/index.php/AAAI/article/view/4666. 2.1

[11] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37(C): 98–125, 2017. doi: 10.1016/j.inffus.2017.02.003. 1.1

[12] Emily Mower Provost, Yuan Shangguan, and Carlos Busso. Umeme: University of michigan emotional mcgurk effect data set. *IEEE Transactions on Affective Computing*, 6(4): 395–409, 2015. doi: 10.1109/taffc.2015.2407898. 3.1.1

[13] Verónica Pérez Rosas, Rada Mihalcea, and Louis-Philippe Morency. Multimodal sentiment analysis of spanish online videos. *IEEE Intelligent Systems*, 28(3):38–45, 2013. doi: 10.1109/MIS.2013.9. 1.1

[14] Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2, 2019. URL https://arxiv.org/abs/1906.00446. 3.3.2

[15] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier, 2016. URL https://arxiv.org/abs/1602.04938. 2.2

[16] Samarth Tripathi and Homayoon S. M. Beigi. Multi-modal emotion recognition on IEMOCAP dataset using deep learning. *CoRR*, abs/1804.05788, 2018. URL http://arxiv.org/abs/1804.05788. 1

[17] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1656. URL https://aclanthology.org/P19-1656. 2.1

[18] Michael Tsang, Dehua Cheng, Hanpeng Liu, Xue Feng, Eric Zhou, and Yan Liu. Feature interaction interpretability: A case for explaining ad-recommendation systems via neural interaction detection. *arXiv preprint arXiv:2006.10966*, 2020. 2.2

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017. URL https://arxiv.org/abs/1706.03762. 2.4

[20] Xingbo Wang, Jianben He, Zhihua Jin, Muqiao Yang, Yong Wang, and Huamin Qu. M2lens: Visualizing and explaining multimodal models for sentiment analysis. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):802–812, 2021. 2.2

[21] Torsten Wörtwein, Lisa B. Sheeber, Nicholas Allen, Jeffrey F. Cohn, and Louis-Philippe Morency. Beyond additive fusion: Learning non-additive multimodal interactions. 2022.

under submission. 3.3, 3.3.2

[22] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Mosi: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. 2016. doi: 10.48550/ARXIV.1606.06259. URL https://arxiv.org/abs/1606.06259. (document), 2, 4.1, 4.1.2

[23] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis, 2017. URL https://arxiv.org/abs/1707.07250. 2.1, 3.1

[24] Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Memory fusion network for multi-view sequential learning, 2018. URL https://arxiv.org/abs/1802.00927. 2.1