Beyond Additive Fusion: Learning Non-Additive Multimodal Interactions

Torsten Wörtwein
Language Technologies Institute
Carnegie Mellon University
twoertwe@cs.cmu.edu

Lisa B. Sheeber
Oregon Research Institute
lsheeber@ori.org

Nicholas Allen
Department of Psychology
University of Oregon
nallen3@uoregon.edu

Jeffrey F. Cohn
Department of Psychology
University of Pittsburgh
jeffcohn@pitt.edu

Louis-Philippe Morency
Language Technologies Institute
Carnegie Mellon University
morency@cs.cmu.edu

Abstract

Multimodal fusion addresses the problem of analyzing spoken words in the multimodal context, including visual expressions and prosodic cues. Even when multimodal models lead to performance improvements, it is often unclear whether bimodal and trimodal interactions are learned or whether modalities are processed independently of each other. We propose Multimodal Residual Optimization (MRO)[1] to separate unimodal, bimodal, and trimodal interactions in a multimodal model. This improves interpretability as the multimodal interaction can be quantified. Inspired by Occam's razor, the main intuition of MRO is that (simpler) unimodal contributions should be learned before learning (more complex) bimodal and trimodal interactions. For example, bimodal predictions should learn to correct the mistakes (residuals) of unimodal predictions, thereby letting the bimodal predictions focus on the remaining bimodal interactions. Empirically, we observe that MRO successfully separates unimodal, bimodal, and trimodal interactions while not degrading predictive performance. We complement our empirical results with a human perception study and observe that MRO learns multimodal interactions that align with human judgments.

1 Introduction

Multimodal fusion integrates information from what we say, how we speak, and how we visually express ourselves. While multimodal models have led to performance improvements (Zadeh et al., 2017; Tsai et al., 2019; Zellers et al., 2021), they often have the downside of being difficult to interpret: it is unclear whether interactions between two modalities (bimodal) or three modalities (trimodal) are learned, whether modalities are processed independently of each other, or whether these models focus on only one modality (Wu et al., 2021).

[1]Code available at https://github.com/twoertwein/MultimodalResidualOptimization.

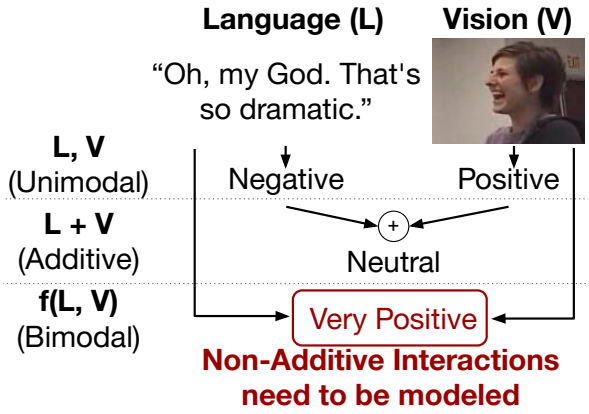

Figure 1: The joint assessment of language and vision (denoted as $f(L, V)$) is different from the sum of unimodal assessments (additive). This is an example for valence from the IEMOCAP dataset (Busso et al., 2008).

Quantifying multimodal interactions is an essential building block for future research: in model debugging as a step to better understand models and improve their performance (Du et al., 2019) as well as in AI applications as a step to be more interpretable (Goodman and Flaxman, 2017).

Seminal work (Hessel and Lee, 2020) observed that many multimodal models function like the sum of unimodal models, so-called additive models. In other words, these models might not be learning as many non-additive (bimodal and trimodal) interactions as expected. The non-additive interaction example in Figure 1 exemplifies how humans perceive the whole multimodal example as more than the sum of the two modalities. While the current approach of separating additive and non-additive interactions highlighted the problem of models primarily learning additive contributions, it did not provide solutions to learn non-additive interactions explicitly (Hessel and Lee, 2020). However, many multimodal tasks, such as visual question answering (Cadene et al., 2019), require learning unimodal, bimodal, and trimodal interactions.

In this paper, we introduce Multimodal Residual

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 4710 - 4725
December 7-11, 2022

Optimization (MRO) to explicitly learn and decompose predictions into the sum of unimodal, bimodal, and trimodal interactions. Inspired by Occam's razor, to prefer simpler solutions, the main intuition of MRO is that (simpler) unimodal contributions should be learned before learning (more complex) bimodal and trimodal interactions. For example, the bimodal predictions should learn to correct the mistakes (residuals) of the unimodal predictions, thereby letting the bimodal predictions focus on the remaining bimodal interactions. Similarly, trimodal predictions should learn what is not modeled by unimodal and bimodal predictions.

We evaluate MRO on six multimodal language datasets, including tasks for intent, sentiment, and emotion recognition. MRO aims to separate multimodal interactions (unimodal, bimodal, and trimodal) without degrading predictive performance. As part of evaluating MRO, we propose a new evaluation metric that extends prior work to three modalities (Hessel and Lee, 2020). We complement our empirical results with a human perceptions study to evaluate whether MRO learns non-additive interactions that align with human judgment.

2 Related Work

We review previous research on four aspects related to multimodal interactions: the prevalence of additive interactions, model-specific and model-agnostic quantification of modality interactions, and taxonomies of multimodal interactions.

Prevalence of Additive Interactions: Growing empirical evidence (Hessel and Lee, 2020) and annotation studies (Provost et al., 2015; Kruk et al., 2019; Wörtwein et al., 2021) highlight that additive interactions are prevalent especially on datasets that are not carefully balanced, e.g., not having the same image contextualized with different captions (Hessel and Lee, 2020). An empirical approach highlights that multimodal models can be factorized into additive models without significant loss in performance (Hessel and Lee, 2020), indicating that the examined models primarily relied on additive interactions. Similarly, multimodal perception studies indicate the importance of additive interactions: unimodal ratings of emotions are predictive of multimodal ratings (Provost et al., 2015). Further, annotations of the semiotic mode, how the multimodal meaning emerges from individual modalities (Bateman, 2014), of text-image pairs found that modalities provide mostly the same meaning (Kruk et al., 2019). Moreover, modality importance annotations for affective states found that a single modality often contains sufficient information to confirm an affective state (Wörtwein et al., 2021). While additive interactions are sufficient in many cases, non-additive interactions are still needed, especially when datasets contain the same unimodal representation in different multimodal contexts (Provost et al., 2015; Hessel and Lee, 2020).

Model-specific quantification: Models can indicate how much they rely on potentially non-additive interactions (Zadeh et al., 2018; Tsai et al., 2020). Multimodal routing (Tsai et al., 2020) was recently proposed to interpret the relative importance of multimodal interactions. It uses the routing-by-agreement algorithm (Sabour et al., 2017) to focus more on modalities whose embedding is similar to other modalities' embeddings. The performance gains of the routing model hint at modalities containing partially redundant information (De Gelder and Bertelson, 2003) for emotion and sentiment prediction. While most model-specific approaches cannot rule out that a multimodal model potentially uses only one modality (Wu et al., 2021), MRO encourages that a bimodal model focuses on bimodal interactions.

Model-agnostic quantification: Multimodal interactions can be quantified after a model has been trained (Hessel and Lee, 2020; Tsang et al., 2020; Wang et al., 2021; Lyu et al., 2022). EMAP (Hessel and Lee, 2020) is based on the idea of factorizing any trained model into additive and non-additive interactions. Unfortunately, this marginalizing is very costly: with m modalities and a dataset of N samples, it requires N^m forward passes. Compared to EMAP, MRO learns a model that directly separates multimodal interactions.

Taxonomy of Multimodal Interactions: Many categorizations have been proposed to quantify the relationship between modalities (Kloepfer, 1976; Zhang et al., 2018; Wang et al., 2021). A recent study (Kruk et al., 2019) uses Koepfer's parallel, amplifying, and divergent. Parallel signals that only one modality is needed for prediction as they all provide the same meaning. Amplifying is sometimes also referred to as "additive" in a non-mathematical sense: modalities provide similar information but their combined meaning is either amplified or diminished. Finally, divergent indi-

cates that modalities provide opposing information. Figure 1 is an example of opposing information.

3 Quantifying Multimodal Interactions

To learn a multimodal model that separates unimodal, bimodal, and trimodal interactions, we first define how to quantify these three types of multimodal interactions. Further, the work presented in this section extends prior work (Hessel and Lee, 2020), which defined evaluation metrics to quantify multimodal interactions in the bimodal case.

Consider three modalities T (text), V (vision), and A (acoustic) with corresponding features x_T, x_V, x_A. A bimodal function f is *additive* when it can be factorized into the sum of two unimodal functions, $\forall x_T, x_V : f(x_T, x_V) = g(x_T) + h(x_V)$. Further, f contains unimodal contributions when parts of the prediction depend on only one modality: $\exists x_T : \mathbb{E}_v f(x_T, v) \neq 0$ (Lyu et al., 2022). This equation is illustrated for the language modality but has the same formulation for the vision modality. Prior work (Hessel and Lee, 2020) proposed EMAP to quantify unimodal contributions (UC) in the context of two modalities. In this paper, we generalize UC to three modalities.

Claim 1. *A trimodal function f contains unimodal contributions when $UC(f, x_T, x_V, x_A) \neq 0$ with*

$$\begin{aligned} UC(f, x_T, x_V, x_A) = \\ \mathop{\mathbb{E}}_{v,a} f(x_T, v, a) + \mathop{\mathbb{E}}_{t,a} f(t, x_V, a) \\ + \mathop{\mathbb{E}}_{t,v} f(t, v, x_A) - 2 \mathop{\mathbb{E}}_{t,v,a} f(t, v, a) \,. \end{aligned} \quad (1)$$

The idea of UC is to evaluate the model with all possible combinations of unimodal features (even feature combinations that are not in a dataset) so that the model cannot use non-additive interactions between modalities. Similarly, we can formulate a function BI to quantify bimodal interactions.

Claim 2. *A trimodal function f contains bimodal interactions (BI) when $BI(f, x_T, x_V, x_A) \neq 0$ with*

$$\begin{aligned} BI(f, x_T, x_V, x_A) = \\ \mathop{\mathbb{E}}_{t}[f(t, x_V, x_A) - UC(f, t, x_V, x_A)] \\ + \mathop{\mathbb{E}}_{v}[f(x_T, v, x_A) - UC(f, x_T, v, x_A)] \\ + \mathop{\mathbb{E}}_{a}[f(x_T, x_V, a) - UC(f, x_T, x_V, x_a)] \,. \end{aligned} \quad (2)$$

The remaining trimodal interactions (TI) are then simply what is not covered by the unimodal contributions and bimodal interactions:

$$\begin{aligned} TI(f, x_T, x_V, x_A) = f(x_T, x_V, x_A) \\ - UC(f, x_T, x_V, x_A) - BI(f, x_T, x_V, x_A) \,. \end{aligned} \quad (3)$$

When computational feasible[2], UC, BI and TI are valuable tools to evaluate whether a trimodal model contains unimodal, bimodal, and trimodal interactions. We will use these metrics to evaluate our proposed approaches.

4 Multimodal Residual Optimization

The main contribution of this paper is Multimodal Residual Optimization (MRO) which has the goal of learning and decomposing predictions into unimodal, bimodal and trimodal interactions to quantify them. Inspired by Occam's razor, the intuition of MRO is that (simpler) unimodal interactions should be prioritized before learning (more complex) bimodal and trimodal interactions. MRO has two components to separate modality interactions: an architecture and loss-function component.

4.1 MRO Architecture

Instead of using a single trimodal function to make a prediction $\hat{y} = f(x_T, x_V, x_A)$, the goal of MRO is to make predictions as $\hat{y} = UC(f, x_T, x_V, x_A) + BI(f, x_T, x_V, x_A) + TI(f, x_T, x_V, x_A)$ without having to compute UC, BI and TI. Therefore, MRO makes predictions $\hat{y}$ based on three components:

$$\hat{y} = \hat{y}_{\text{uni}} + \hat{y}_{\text{bi}} + \hat{y}_{\text{tri}} \quad (4)$$

where $\hat{y}_{\text{uni}}, \hat{y}_{\text{bi}}$ and $\hat{y}_{\text{tri}}$ model the unimodal, bimodal, and trimodal interactions respectively. It is important to note that $\hat{y}_{\text{bi}}$ and $\hat{y}_{\text{tri}}$ are intended to model only non-additive interactions, while $\hat{y}_{\text{uni}}$ is designed to model only additive interactions. $\hat{y}_{\text{uni}}$ is defined as

$$\hat{y}_{\text{uni}} = f_{\theta_T}(x_T) + f_{\theta_V}(x_V) + f_{\theta_A}(x_A) \quad (5)$$

[2] UC, BI, and TI can computationally be demanding given the expectation terms. While this is not as much of an issue when used as evaluation metrics, the computational cost prohibits us from using them as part of an iterative optimization process, e.g., in the loss function of neural networks.

where $f_{\theta_T}, f_{\theta_V}$ and f_{θ_A} are models, e.g., neural networks that use only one modality as an input. Each model has its own set of parameters (θ_T, θ_V, and θ_A). We parameterize the bimodal and trimodal models in a similar manner:

$$\hat{y}_{\text{bi}} = f_{\theta_{TV}}(x_T, x_V) + f_{\theta_{TA}}(x_T, x_A) + f_{\theta_{AV}}(x_A, x_V) \quad (6)$$

$$\hat{y}_{\text{tri}} = f_{\theta_{TVA}}(x_T, x_V, x_A) \quad (7)$$

where $f_{\theta_{TV}}, f_{\theta_{TA}}$ and $f_{\theta_{AV}}$ are the bimodal models that take only two modalities as input, and $f_{\theta_{TVA}}$ takes all three modalities as input. The whole MRO model is parameterized with $\Theta = (\theta_T, \theta_V, \theta_A, \theta_{TV}, \theta_{TA}, \theta_{AV}, \theta_{TVA})$.

This architecture already enforces that $\hat{y}_{\text{uni}}$ can only contain unimodal contributions. While dedicating unimodal, bimodal, and trimodal models was explored in prior work (Zadeh et al., 2016, 2019; Tsai et al., 2020), they did not explicitly encourage $\hat{y}_{\text{bi}}$ and $\hat{y}_{\text{tri}}$ not to contain unimodal contributions and similarly $\hat{y}_{\text{tri}}$ not to contain bimodal interactions. The MRO loss function described in the next section addresses this issue.

4.2 MRO Loss Function

We first explain MRO for two modalities (language and vision) before presenting the more general formulation for three and more modalities.

Bimodal case: To encourage $\hat{y}_{\text{bi}}$ to not contain unimodal contributions, MRO prioritizes $\hat{y}_{\text{uni}}$. MRO defines the loss function as

$$L(y, \hat{y}) = L(y, \hat{y}_{\text{uni}}) + L(y, sg\,(\hat{y}_{\text{uni}}) + \hat{y}_{\text{bi}}) \quad (8)$$

where sg refers to stop-gradient (Razavi et al., 2019), which prevents back-propagation through sg's arguments. The first part of Equation 8 updates θ_T and θ_V to predict y using only unimodal contributions $\hat{y}_{\text{uni}} = f_{\theta_T}(x_T) + f_{\theta_V}(x_V)$. The second part of Equation 8 updates θ_{TV} so that $L(y, \hat{y}_{\text{uni}} + \hat{y}_{\text{bi}})$ is smaller; i.e., $\hat{y}_{\text{bi}}$ corrects mistakes that $\hat{y}_{\text{uni}}$ makes. We do not backpropagate again to θ_T and θ_V so that $\hat{y}_{\text{bi}}$ does not influence $\hat{y}_{\text{uni}}$; i.e., $\hat{y}_{\text{uni}}$ is optimized independently of $\hat{y}_{\text{bi}}$.

Figure 2 summarizes MRO in the bimodal case.

m-modal case: In the case of m modalities, we have m types of interactions: unimodal, bimodal, trimodal, ..., m-modal. Instead of separating just additive from all non-additive interactions, we want to separate these m types of interactions. MRO

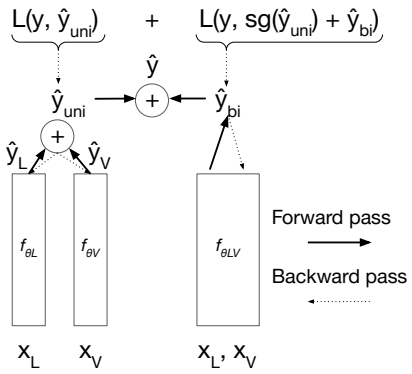

Figure 2: Overview of MRO: bimodal model learns what cannot be predicted by the unimodal contributions.

defines the loss function as

$$L(y, \hat{y}) = \sum_{i=1}^{m} L\left(y, sg\left(\sum_{j=1}^{i-1} \hat{y}_j\right) + \hat{y}_i\right) \quad (9)$$

where $\hat{y}_i$ refers to the i-modal predictions, i.e., $\hat{y}_1 = \hat{y}_{\text{uni}}, \hat{y}_2 = \hat{y}_{\text{bi}}, \hat{y}_3 = \hat{y}_{\text{tri}}$. For the trimodal case, $\hat{y}_{\text{uni}}$, $\hat{y}_{\text{bi}}$, and $\hat{y}_{\text{tri}}$ were defined in subsection 4.1. When m is large than three, the models can be defined following the same approach. Similar to the bimodal case, $\hat{y}_{\text{bi}}$ is optimized independently of $\hat{y}_{\text{tri}}$ as the gradient of $\hat{y}_{\text{bi}}$ is stopped by sg when optimizing $\hat{y}_{\text{tri}}$.

4.3 Sequential MRO

An alternative to MRO's approach of simultaneously optimizing all prediction components ($\hat{y}_{\text{uni}}$, $\hat{y}_{\text{bi}}$, $\hat{y}_{\text{tri}}$), the sequential MRO (sMRO) proposes to optimize them sequentially.

First, sMRO optimizes the parameters of $\hat{y}_{\text{uni}}$ using the loss $L(y, \hat{y}_{\text{uni}})$ until convergence and then freezes its parameters θ_L, θ_V, and θ_A before optimizing $\hat{y}_{\text{bi}}$ and $\hat{y}_{\text{tri}}$. Next, sMRO optimizes the parameters of $\hat{y}_{\text{bi}}$ using the loss $L(y, \hat{y}_{\text{uni}} + \hat{y}_{\text{bi}})$ until convergence and then freeze the bimodal parameters θ_{LV}, θ_{LA} and θ_{VA}. The trimodal $\hat{y}_{\text{tri}}$ can then be optimized using the loss $L(y, \hat{y}_{\text{uni}} + \hat{y}_{\text{bi}} + \hat{y}_{\text{tri}})$. For cases with more than three modalities, sMRO can optimize the parameters of $\hat{y}_m$ for $L(y, \sum_{i=1}^{m} \hat{y}_i)$ until convergence and then freeze the parameters of $\hat{y}_m$.

sMRO has similarities with gradient boosting (GB) (Friedman, 2001) when GB has, in the trimodal case, three learners that correspond to the prediction components $\hat{y}_{\text{uni}}$, $\hat{y}_{\text{bi}}$, and $\hat{y}_{\text{tri}}$. Unlike sMRO, GB is not suitable for some loss functions, such as the mean absolute error (MAE; its gradient is not proportional to the residual), as each learner

in GB estimates the gradient of the errors from the previous learners. In the case of MAE, learners will predict -1 or 1, which leads to a poor fit with only three learners.

5 Experimental Methodology

We evaluate whether we can train a model that separates unimodal, bimodal, and trimodal interactions while not degrading predictive performance.

Datasets: We focus on five sentiment- and emotion-annotated datasets for which prior work used multimodal models, see Table 1. We also include a sixth Instagram dataset (Kruk et al., 2019) as it has modality interaction annotations (semiotic modes), which we can use to evaluate MRO.

We use the same features across all sentiment and emotion datasets: RoBERTa (Liu et al., 2020) as a representation of transcribed utterances; OpenFace 2.0 (Baltrusaitis et al., 2018) to summarize face-related features, and openSMILE's eGeMAPS (Eyben et al., 2015) to summarize acoustic features. For the Instagram dataset, we use the author-provided ResNet features (He et al., 2016) to summarize the image content and use RoBERTa to represent captions.

Evaluation: We want that the prediction components $\hat{y}_{\text{uni}}, \hat{y}_{\text{bi}}$ and $\hat{y}_{\text{tri}}$ correspond to $UC(\hat{y}), BI(\hat{y})$, and $TI(\hat{y})$ so that the prediction components represent only unimodal, only bimodal, and only trimodal interactions. To test this, we use $|UC(\hat{y}_{\text{bi}} + \hat{y}_{\text{tri}})|$ to evaluate whether the bimodal and trimodal predictions contain unimodal contributions and $|BI(\hat{y}_{\text{tri}})|$ whether the trimodal prediction contains bimodal contributions. Given the MRO-architecture, $\hat{y}_{\text{uni}}$ cannot include bimodal and trimodal interactions and $\hat{y}_{\text{bi}}$ cannot include trimodal interactions. This means, if $|UC(\hat{y}_{\text{bi}} + \hat{y}_{\text{tri}})| + |BI(\hat{y}_{\text{tri}})|$ is 0, the model perfectly separates unimodal, bimodal, and trimodal interactions, i.e., $\hat{y}_{\text{uni}} = UC(\hat{y}), \hat{y}_{\text{bi}} = BI(\hat{y})$, and $\hat{y}_{\text{tri}} = TI(\hat{y})$. We use 5-fold test setup for all datasets.

Models: We compare the MRO-architecture when optimized in different manners: with $L(y, \hat{y}_{\text{uni}} + \hat{y}_{\text{bi}} + \hat{y}_{\text{tri}})$ (referred to as **Joint**), **sMRO**, and **MRO**. For performance comparison, we include the routing model (Tsai et al., 2020) (referred to as **Routing**), a recently proposed model with the goal of modality interpretability. Lastly, we compare the performance against a single trimodal model $\hat{y} = f_{\theta_{TVA}}(x_T, x_V, x_A)$ (referred to as **Tri**) to evaluate whether the larger MRO-architecture has two many parameters for smaller datasets.

Implementation Details: The functions f of Equation 4 are instantiated as multi-layer perceptrons. For each multimodal model, e.g., $f_{\theta_{TV}}$, we implement two popular types of fusion: early fusion (concatenating the modalities) and tensor fusion (Zadeh et al., 2017) (outer product between modalities after learning unimodal embeddings). The type of fusion is a hyper-parameter together with the number of layers, their width, learning rate, learning rate decay, L2 weight decay, dropout, and with/without prior feature selection. As a loss function, we use the mean absolute error for regression tasks and the cross-entropy loss for classification tasks.

6 Multimodal Perception Study

We conduct a multimodal perception study to evaluate whether MRO learns non-additive interactions, when humans also require non-additive interactions. We choose arousal and valence on the IEMOCAP dataset for this study as arousal and valence are two fundamental dimensions to describe emotional states (Munezero et al., 2014).

Study Design: Crowd workers[3] are asked to rate arousal and valence of video segments when being exposed to only a subset of modalities. The four subsets are: 1) the transcript of what the person says (T); 2) the muted video (V); 3) the low-pass filtered audio (A), and 4) the transcripts, the video, and the original audio (TVA_O). IEMOCAP has ten speakers. We randomly select ten segments for each speaker, i.e. 100 segments.

Audio Processing: It is challenging to disentangle speech content and how we speak (Bhargava and Başkent, 2012). Similar to previous work, we low-pass filter the audio signal (Yang et al., 2012). Instead of using 850 Hz as a cut-off (Yang et al., 2012), we use a lower cut-off frequency, as we could understand spoken words at 850 Hz. We choose 660 Hz[4] as it is the mean of the maximum pitch in an empirical study (Li and Yiu, 2006) and it also closely coincides with the maximum pitch of contralto singers (E_5 at 659.25 Hz). We choose this pitch-focused definition as we believe that prosodic

[3] We recruited 40 US-based crowed works from the platform prolific https://www.prolific.co/ whose first language is English.

[4] We use ffmpeg for low-pass filtering with the following filter configuration: firequalizer=gain='if(lt(f,660), 0, -INF)':min_phase=1

Original Paper	Tasks	Abbreviation	Samples	Modalities
(Zadeh et al., 2016)	Sentiment (regression)	MOSI	2.2k	3
(Zadeh et al., 2018)	Sentiment, Polarity, Happiness (regression)	MOSEI	22.9k	3
(Busso et al., 2008)	Arousal and Valence (regression)	IEMOCAP	4.8k	3
(Valstar et al., 2016)	Arousal and Valence (regression)	SEWA	1.9k	3
(Nelson et al., 2021)	Affect categories (4-way classification)	TPOT	17.3k	3
(Kruk et al., 2019)	Intent of Instagram posts (7-way classification)	Instagram	1.3k	2

Table 1: Dataset overview.

information will predict arousal and valence.

Avoiding learning effects: Raters might be able to infer the missing multi-modal context after having rated some of the unimodal subsets for a specific segment. We therefore use two mechanisms to address learning effects across the modalities. First, each of the raters annotates only 20 randomly selected segments for each modality subset (we have eight raters per segment and modality subset). Second, we structurally randomize the order of the modality subsets by first presenting all unimodal subsets in a random order and in the end the trimodal segments.

Ratings and reliability: Following the annotation setup from IEMOCAP, we use the ordinal arousal and valence manikins scale consisting of five levels (Bradley and Lang, 1994) to rate the two emotional dimensions. The effective reliability (Rosenthal, 2005) over k raters as measured by the Intra-class Correlation Coefficient ICC(2, k-1) is excellent (above 0.9) (Koo and Li, 2016) for all modality subsets. Further, our new trimodal ratings (TVA_O) correlate highly with the existing annotations on IEMOCAP $r(98) = 0.88, p < 0.001$ for arousal and $r(98) = 0.92, p < 0.001$ for valence, indicating that we can use our new annotations to inspect models trained on the original annotations.

Evaluation: To evaluate when humans require non-additive interactions, we train a linear regression model (an additive model) that predicts TVA_O given T, V, and A. We refer to this model as $\hat{y}^{\text{human}}_{\text{uni}}$. The model fit of $\hat{y}^{\text{human}}_{\text{uni}}$ shows how important the missing non-additive interactions are (Provost et al., 2015). Further, the absolute error $|\text{TVA}_O - \hat{y}^{\text{human}}_{\text{uni}}|$ measures how important the missing non-additive interactions are to humans for each segment. We use $|\text{TVA}_O - \hat{y}^{\text{human}}_{\text{uni}}|$ to answer the question: does MRO learn more non-additive interactions when $|\text{TVA}_O - \hat{y}^{\text{human}}_{\text{uni}}|$ is larger, i.e., when humans require non-additive interactions?

	Arousal	Valence
Min. age	19	21
Mean age	36	37
Max. age	79	62
Female	20	19
Male	20	21

Table 2: Basic demographic information about the annotators.

7 Results and Discussion

Sanity Check: Before evaluating MRO on more complex datasets, we conduct a sanity check on two simpler datasets: $x_T + x_V + x_A$ which requires only unimodal contributions (we refer to it as **Sanity Check Unimodal**) and $x_T x_V + x_T x_A + x_V x_A$ which requires only bimodal interactions (we refer to it as **Sanity Check Bimodal**). Figure 3 shows that the joint and the routing model do not separate unimodal, bimodal, and trimodal interactions well as $|UC(\hat{y}_{\text{bi}} + \hat{y}_{\text{tri}})| + |BI(\hat{y}_{\text{tri}})|$ is high. As expected, sMRO and MRO separate the interacts almost perfectly as $|UC(\hat{y}_{\text{bi}} + \hat{y}_{\text{tri}})| + |BI(\hat{y}_{\text{tri}})|$ is very close to 0.

To test how many epochs are needed to minimize $|UC(\hat{y}_{\text{bi}} + \hat{y}_{\text{tri}})| + |BI(\hat{y}_{\text{tri}})|$, we evaluate it after each epoch. The results in Figure 4 show that the separation during the first epochs becomes worse as $\hat{y}_{\text{uni}}$ has not yet learned much, meaning $\hat{y}_{\text{bi}}$ and $\hat{y}_{\text{tri}}$ try to predict unimodal contributions which increases $|UC(\hat{y}_{\text{bi}} + \hat{y}_{\text{tri}})|$. However, after a few epochs the separation becomes better and

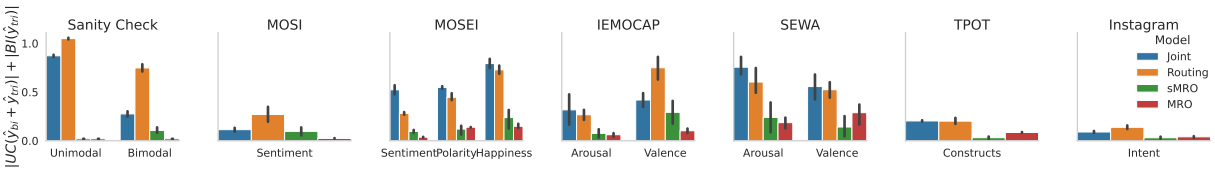

Figure 3: Average $|UC(\hat{y}_{\text{bi}} + \hat{y}_{\text{tri}})| + |BI(\hat{y}_{\text{tri}})|$ for all models and datasets. Lower values indicate a better separation of unimodal, bimodal, and trimodal contributions.

$|UC(\hat{y}_{\text{bi}} + \hat{y}_{\text{tri}})| + |BI(\hat{y}_{\text{tri}})|$ reaches 0. The same can be observed for the bimodal sanity check in Figure 4.

MRO significantly reduces $|UC(\hat{y}_{\textbf{bi}} + \hat{y}_{\textbf{tri}})| + |BI(\hat{y}_{\textbf{tri}})|$. Similar to the sanity check on simpler dataset, we want that $|UC(\hat{y}_{\text{bi}} + \hat{y}_{\text{tri}})| + |BI(\hat{y}_{\text{tri}})|$ is as small as possible. For easier comparison across datasets, we normalize $|UC(\hat{y}_{\text{bi}} + \hat{y}_{\text{tri}})| + |BI(\hat{y}_{\text{tri}})|$ by the standard deviation of the ground truth from the training set. Figure 3 shows that sMRO and MRO significantly reduce $|UC(\hat{y}_{\text{bi}} + \hat{y}_{\text{tri}})| + |BI(\hat{y}_{\text{tri}})|$ compared to models optimized with $L(y, \hat{y}_{\text{uni}} + \hat{y}_{\text{bi}} + \hat{y}_{\text{tri}})$ (Joint) and the routing model.

As it is computationally very expensive to evaluate $|UC(\hat{y}_{\text{bi}} + \hat{y}_{\text{tri}})| + |BI(\hat{y}_{\text{tri}})|$ after each epoch, we plot it only for arousal and valence on IEMOCAP in Figure 4. We focus on IEMOCAP as we also conduct the perception study on it, see section 6. While the plot for arousal in Figure 4 is a bit noisy, MRO quickly reduces $|UC(\hat{y}_{\text{bi}} + tri)| + |BI(\hat{y}_{\text{tri}})|$. The same can be observed for valence in Figure 4.

MRO does not degrade performance. The secondary goal of MRO is not degrading performance. Table 3 lists the models' performance. Models optimized with MRO are in no case significantly worse than any other model. However, they are statistically significantly better than the joint model for valence on SEWA and happiness on MOSEI.

MRO might generalizes slightly better because, similar to structural risk minimization (Vapnik, 1999), it prioritizes simpler models and relies on more complex multimodal models only when needed. Another reason is that MRO has similar effects as having auxiliary unimodal loss functions which seems beneficial for multimodal models (Wang et al., 2020; Zeng et al., 2021).

Ablating $\hat{y}_{\textbf{bi}} + \hat{y}_{\textbf{tri}}$ decreases performance. We quantify the average performance impact of post-hoc removing $\hat{y}_{\text{bi}} + \hat{y}_{\text{tri}}$ across datasets, i.e., $\hat{y} = \hat{y}_{\text{uni}}$. When comparing Table 4 with Table 3, we observe that removing $\hat{y}_{\text{bi}} + \hat{y}_{\text{tri}}$ (the non-additive predictions), hurts performance. While additive contributions are very important, non-additive interactions are needed for best performance.

	Tri	Routing	Joint	sMRO	MRO
MOSI (Pearson's r)					
Sentiment	0.662	0.658	0.657	0.656	0.661
MOSEI (Pearson's r)					
Sentiment	0.723	0.727	0.727	0.726	0.727
Polarity	0.599	0.597	0.606	0.593	0.605
Happiness	0.637	0.642	0.637	0.630	0.641
IEMOCAP (Concordance Correlation Coefficient)					
Arousal	0.588	0.613	0.622	0.624	0.611
Valence	0.647	0.655	0.624	0.603	0.634
SEWA (Concordance Correlation Coefficient)					
Arousal	0.317	0.263	0.293	0.292	0.304
Valence	0.268	0.335	0.268	0.310	0.337
TPOT (Accuracy)					
Constructs	0.565	0.554	0.566	0.566	0.574
Instagram (macro ROC AUC)					
Intent	0.876	0.731	0.891	0.888	0.891
Mean	0.588	0.595	0.589	0.589	**0.599**

Table 3: Average performance over the test folds. Higher is better.

	sMRO	MRO
Mean	0.577	0.587

Table 4: Average performance when post-hoc removing $\hat{y}_{\text{bi}} + \hat{y}_{\text{tri}}$, i.e., $\hat{y} = \hat{y}_{\text{uni}}$.

MRO learns more non-additive interactions when two modalities are informative. The TPOT dataset has human judgments for how important modalities are to confirm the current affective state (Wörtwein et al., 2021). Three importance levels were annotated: 1) a modality is sufficient to confirm the affective state (while ignoring other modalities), 2) a modality contains relevant information for the affective state (information from a second modality is needed), and 3) a modality contains no information for the current affective state.

We hypothesize that MRO uses more non-additive interactions ($\hat{y}_{\text{bi}} + \hat{y}_{\text{tri}}$) for samples with

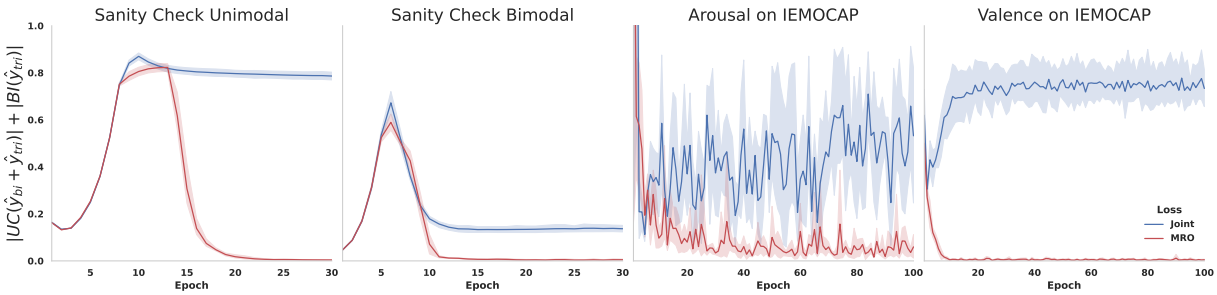

Figure 4: $|UC(\hat{y}_{\text{bi}} + \hat{y}_{\text{tri}})| + |BI(\hat{y}_{\text{tri}})|$ for the same model optimized with either $L(y, \hat{y}_{\text{uni}} + \hat{y}_{\text{bi}} + \hat{y}_{\text{tri}})$ (Joint, in blue) or with MRO (in red). Lower values indicate a better separation of unimodal, bimodal, and trimodal interactions.

at least two informative modalities (relevant or sufficient) compared to samples with only one informative modality. To measure whether $\hat{y}_{\text{bi}} + \hat{y}_{\text{tri}}$ are used more, we calculate how much the softmax probabilities (TPOT is a classification task) change when removing $\hat{y}_{\text{bi}} + \hat{y}_{\text{tri}}$, i.e., $\sum_{k=1}^{4} |\text{softmax}(\hat{y})^{(k)} - \text{softmax}(\hat{y}_{\text{uni}})^{(k)}|$ where k indexes the probability vector for the four classes. The means of samples with two informative modalities (0.299) and only one informative modality (0.264) are significantly different according to an independent t-test, $t(2671) = 5.059, p < 0.001$. This suggests that MRO not only mathematically separates unimodal, bimodal, and trimodal interactions but that its separation also correlates with human assessments. Further, this observation provides evidence that models are more likely to learn non-additive interactions when several modalities are themselves informative.

MRO learns more non-additive interactions when modalities amplify each other. We included the Instagram dataset (Kruk et al., 2019) because it has modality interaction annotations (semiotic modes) that are inspired by Kloepfer (Kloepfer, 1976). To test whether $\hat{y}_{\text{bi}}$ (this dataset has only two modalities) contributes more depending on the semiotic mode (parallel, amplifying, and divergent), we conduct a one-way ANOVA on the probability changes when removing $\hat{y}_{\text{bi}}$. The means between the semiotic modes are significantly different, $F(2, 1296) = 5.059, p = 0.006$, with the highest absolute average change for amplifying (0.317), followed by parallel (0.272), and then divergent (0.256). The means between amplifying and parallel are significantly different $t(1297) = 2.432, p = 0.015$ as well as between amplifying and divergent $t(1297) = 2.874, p = 0.004$. Similar to the results on TPOT, it is confirming that MRO learned significantly larger non-additive contributions ($\hat{y}_{\text{bi}}$) for amplifying than for parallel. A possible explanation why diverging seems to require the least non-additive interactions is that the definition of diverging requires that only the meaning of the modalities is opposing but it does not specify how the combined meaning is formed. Even if the combined meaning of Figure 1 was neutral (additive), the semiotic mode is still divergent.

MRO learns non-additive interaction when humans need non-additive interactions. The additive model $\hat{y}_{\text{uni}}^{\text{human}}$ of predicting the mutlimodal ratings TVA_O given the uni-modal ratings, fits very well ($r^2 = 0.85$ for arousal and $r^2 = 0.85$ for valence) which is inline with similar prior work (Provost et al., 2015). Even though our multimodal model is not on par with $\hat{y}_{\text{uni}}^{\text{human}}$ ($r^2 = 0.68$ for arousal and $r^2 = 0.66$ for valence), we observe a significant correlation of $r(98) = 0.202, p = 0.043$ for valence between $|\text{TVA}_O - \hat{y}_{\text{uni}}^{\text{human}}|$ (the missing non-additive interactions) and $|\hat{y}_{\text{bi}} + \hat{y}_{\text{tri}}|$ (non-additive contributions). This indicates that $\hat{y}_{\text{bi}} + \hat{y}_{\text{tri}}$ learned non-additive interactions that cannot be explained by $\hat{y}_{\text{uni}}^{\text{human}}$. For arousal, we do not observe a significant correlation, potentially because the optimization seems far nosier for arousal then for valence, see Figure 4.

8 Conclusion

We proposed MRO to explicitly learn and separate unimodal, bimodal, and trimodal interactions in a multimodal model. This separation is essential for quantifying how much a model uses multimodal interactions and is a step towards more interpretable models. Based on prior work (Hessel and Lee, 2020) we proposed a new evaluation metrics to quantify whether a trimodal models uses unimodal, bimodal, and trimodal interactions. Empirically, we observed that MRO successfully separated unimodal, bimodal, and trimodal interactions while not degrading predictive performance. Beyond the

empirical evaluation, MRO learns non-additive interactions in accordance with human judgments on three datasets.

Limitations

We evaluated MRO in the context of language, vision, and acoustic modalities. Future work could explore MRO's performance on different modalities. Exploring MRO beyond three modalities will also be interesting. To address a potentially growing number of parameters for models with more than three modalities, sharing modality representation could be explored: the bi-modal models could be given access to intermediate representations from uni-modal models for modalities they have in common. Sharing representations could reduce the overall model size.

While we evaluated MRO on many sentiment and emotion annotated datasets, these datasets are primarily in English, and one is in German (SEWA). More research is needed to work with a more diverse set of languages.

It will also be interesting to study MRO in tasks that require multimodal fusion and translation: generating a modality given a set of different modalities.

Ethics Statement

Emotional states can provide insights into mental health, especially into mood disorders like depression. While emotion recognition systems can be part of medical pre-screening tools to facilitate care (DeVault et al., 2014), the same technology can be part of job interview tools (Naim et al., 2016) potentially leading to discrimination against people with mood disorders. More work on visualizing and interpreting model predictions are tools to highlight potential biases and help better understand the internal decision process of multimodal models.

Acknowledgements

This material is based upon work partially supported by the National Science Foundation (Awards #1722822 and #1750439), and National Institutes of Health (Awards #R01MH125740, #R01MH096951, #U01MH116925, and #U01MH116923). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or National Institutes of Health, and no official endorsement should be inferred.

References

Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 59–66. IEEE.

John A Bateman. 2014. *Text and image: A critical introduction to the visual/verbal divide*. Routledge.

Pranesh Bhargava and Deniz Başkent. 2012. Effects of low-pass filtering on intelligibility of periodically interrupted speech. *The Journal of the Acoustical Society of America*, 131(2):EL87–EL92.

Margaret M Bradley and Peter J Lang. 1994. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, 25(1):49–59.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359.

Remi Cadene, Corentin Dancette, Matthieu Cord, Devi Parikh, et al. 2019. Rubi: Reducing unimodal biases for visual question answering. *Advances in neural information processing systems*, 32.

Beatrice De Gelder and Paul Bertelson. 2003. Multisensory integration, perception and ecological validity. *Trends in cognitive sciences*, 7(10):460–467.

David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jon Gratch, Arno Hartholt, Margaux Lhommet, et al. 2014. Simsensei kiosk: A virtual human interviewer for healthcare decision support. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pages 1061–1068.

Mengnan Du, Ninghao Liu, and Xia Hu. 2019. Techniques for interpretable machine learning. *Communications of the ACM*, 63(1):68–77.

Paul Ekman. 1982. Methods for measuring facial action. *Handbook of methods in nonverbal behavior research*, pages 45–90.

Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. 2015. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE transactions on affective computing*, 7(2):190–202.

Jerome H Friedman. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

Bryce Goodman and Seth Flaxman. 2017. European union regulations on algorithmic decision-making and a "right to explanation". *AI magazine*, 38(3):50–57.

Michael Grimm and Kristian Kroschel. 2005. Evaluation of natural emotions using self assessment manikins. In *IEEE Workshop on Automatic Speech Recognition and Understanding, 2005.*, pages 381–385. IEEE.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Jack Hessel and Lillian Lee. 2020. Does my multimodal model learn cross-modal interactions? it's harder to tell than you might think! In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225.

Rolf Kloepfer. 1976. Komplementarität von sprache und bild am beispiel von comic, karikatur und reklame. *Sprache in Technischen Zeitalter Stuttgart*.

Terry K Koo and Mae Y Li. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 15(2):155–163.

Julia Kruk, Jonah Lubin, Karan Sikka, Xiao Lin, Dan Jurafsky, and Ajay Divakaran. 2019. Integrating text and image: Determining multimodal document intent in instagram posts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*.

Nicole YK Li and Edwin M-L Yiu. 2006. Acoustic and perceptual analysis of modal and falsetto registers in females with dysphonia. *Clinical linguistics & phonetics*, 20(6):463–481.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. RoBERTa: A robustly optimized BERT pretraining approach. https://openreview.net/forum?id=SyxS0T4tvS.

Yiwei Lyu, Paul Pu Liang, Zihao Deng, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2022. Dime: Fine-grained interpretations of multimodal models via disentangled local explanations. *arXiv preprint arXiv:2203.02013*.

Myriam Munezero, Calkin Suero Montero, Erkki Sutinen, and John Pajunen. 2014. Are they different? affect, feeling, emotion, sentiment, and opinion detection in text. *IEEE transactions on affective computing*, 5(2):101–111.

Iftekhar Naim, Md Iftekhar Tanveer, Daniel Gildea, and Mohammed Ehsan Hoque. 2016. Automated analysis and prediction of job interview performance. *IEEE Transactions on Affective Computing*, 9(2):191–204.

Benjamin W Nelson, Lisa Sheeber, Jennifer Pfeifer, and Nicholas B Allen. 2021. Psychobiological markers of allostatic load in depressed and nondepressed mothers and their adolescent offspring. *Journal of Child Psychology and Psychiatry*, 62(2):199–211.

Emily Mower Provost, Yuan Shangguan, and Carlos Busso. 2015. Umeme: University of michigan emotional mcgurk effect data set. *IEEE Transactions on Affective Computing*, 6(4):395–409.

Ali Razavi, Aaron van den Oord, and Oriol Vinyals. 2019. Generating diverse high-fidelity images with vq-vae-2. In *Advances in neural information processing systems*, pages 14866–14876.

Robert Rosenthal. 2005. Conducting judgment studies: Some methodological issues. *The new handbook of methods in nonverbal behavior research*, pages 199–234.

Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. 2017. Dynamic routing between capsules. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*.

Yao-Hung Hubert Tsai, Martin Q Ma, Muqiao Yang, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. Multimodal routing: Improving local and global interpretability of multimodal language analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*.

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

Aman Tyagi, Anjalie Field, Priyank Lathwal, Yulia Tsvetkov, and Kathleen M Carley. 2020. A computational analysis of polarization on indian and pakistani social media. In *International Conference on Social Informatics*, pages 364–379. Springer.

Michel Valstar, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Denis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic. 2016. Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the 6th international workshop on audio/visual emotion challenge*, pages 3–10.

Vladimir N Vapnik. 1999. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Weiyao Wang, Du Tran, and Matt Feiszli. 2020. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12695–12705.

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

Torsten Wörtwein, Lisa B Sheeber, Nicholas Allen, Jeffrey F Cohn, and Louis-Philippe Morency. 2021. Human-guided modality informativeness for affective states. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, pages 728–734.

Zhiyong Wu, Lingpeng Kong, Wei Bi, Xiang Li, and Ben Kao. 2021. Good for misconceived reasons: An empirical revisiting on the need for visual context in multimodal machine translation. In *"Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 6153–6166.

Ying Yang, Catherine Fairbairn, and Jeffrey F Cohn. 2012. Detecting depression severity from vocal prosody. *IEEE transactions on affective computing*, 4(2):142–150.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.

Amir Zadeh, Chengfeng Mao, Kelly Shi, Yiwei Zhang, Paul Pu Liang, Soujanya Poria, and Louis-Philippe Morency. 2019. Factorized multimodal transformer for multimodal sequential learning. In *Elsevier Information Fusion Journal (IF 11.21)*.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. 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*, pages 2236–2246.

Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. 2021. Merlot: Multimodal neural script knowledge models. In *Advances in Neural Information Processing Systems 34*.

Ying Zeng, Sijie Mai, and Haifeng Hu. 2021. Which is making the contribution: Modulating unimodal and cross-modal dynamics for multimodal sentiment analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1262–1274. Association for Computational Linguistics.

Mingda Zhang, Rebecca Hwa, and Adriana Kovashka. 2018. Equal but not the same: Understanding the implicit relationship between persuasive images and text. In *Proceedings of the British Machine Vision Conference (BMVC)*.

A Experimental Details

A.1 IEMOCAP

IEMOCAP has multiple recording conditions of two speakers interacting: acted interactions, improvised interactions, and spontaneous interactions (Busso et al., 2008). In this paper, we use the improvised interactions as they cover a diverse range of emotional expressions and are not tied to a set of fixed utterances as is the case of the acted interactions.

A.2 MOSEI

Polarity is an established dimension in ethics research (Tyagi et al., 2020) and is typically defined as the absolute value of the sentiment intensity (Hutto and Gilbert, 2014). We use this definition and apply it to MOSEI's sentiment ratings.

A.3 SEWA

Instead of using SEWA's time continuous ratings of valence and arousal, we take the average of the ratings for each utterance to make SEWA similar to the other datasets.

A.4 Features

We use the openSMILe configuration eGeMAPS v0.1b (Eyben et al., 2015) which extracts instantaneous low-level descriptors and summaries over a moving window. For the low-level descriptors, we calculate the median and interquartile range for each segment. For the summary features, we take the median over each segment.

OpenFace 2.0 extracts many face-related features. We summarize OpenFace's facial action units (Ekman, 1982), head pose, and eye gaze features with the mean and standard deviation.

B UC and BI for trimodal Models

Any trimodal function f can be expressed as $f_T + f_V + f_A + f_{TV} + f_{TA} + f_{VA} + f_{TVA}$ such that $UC(f) = f_T + f_V + f_A$ and $BI(f) = f_{TV} + f_{TA} + f_{VA}$ where the bimodal functions do not contains unimodal contributions: $\forall x_T : \mathbb{E}_v[f_{TV}(x_T, v)] = 0$ and similar for x_V and x_A. Further, the trimodal function should not contains unimodal and bimodal interactions: $\forall x_T, x_V : \mathbb{E}_a[f_{TVA}(x_T, x_V, a) = 0$ and similar for the pairs (x_T, x_A) and (x_V, x_A).

Proof. As any trimodal function can be expressed as the above function, we show that the definition of UC returns exactly the unimodal contributions $f_T + f_V + f_A$.

$$
\begin{aligned}
& UC(f) && (10)\\
=& \mathop{\mathbb{E}}_{v,a} f(x_T, v, a) + \mathop{\mathbb{E}}_{t,a} f(t, x_V, a)\\
&+ \mathop{\mathbb{E}}_{t,v} f(t, v, x_A) - 2 \mathop{\mathbb{E}}_{t,v,a} f(t, v, a) && (11)\\
=&\Big(f_T(x_T) + \mathop{\mathbb{E}}_{v} f_V(v) + \mathop{\mathbb{E}}_{a} f_A(a)\\
&+ \mathop{\mathbb{E}}_{v} f_{TV}(x_T, v) + \mathop{\mathbb{E}}_{a} f_{TA}(x_T, a)\\
&+ \mathop{\mathbb{E}}_{v,a} f_{VA}(v, a) + \mathop{\mathbb{E}}_{v,a} f_{TVA}(x_T, v, a)\Big)\\
&+ \Big(\mathop{\mathbb{E}}_{t} f_T(t) + f_V(x_V) + \mathop{\mathbb{E}}_{a} f_A(a)\\
&+ \mathop{\mathbb{E}}_{t} f_{TV}(t, x_V) + \mathop{\mathbb{E}}_{t,a} f_{TA}(t, a)\\
&+ \mathop{\mathbb{E}}_{a} f_{VA}(x_V, a) + \mathop{\mathbb{E}}_{t,a} f_{TVA}(t, x_V, a)\Big)\\
&+ \Big(\mathop{\mathbb{E}}_{t} f_T(t) + \mathop{\mathbb{E}}_{v} f_V(v) + f_A(x_A)\\
&+ \mathop{\mathbb{E}}_{tv} f_{TV}(t, v) + \mathop{\mathbb{E}}_{t} f_{TA}(t, x_A)\\
&+ \mathop{\mathbb{E}}_{v} f_{VA}(v, x_A) + \mathop{\mathbb{E}}_{t,v} f_{TVA}(t, v, x_A)\Big)\\
&- 2\Big(\mathop{\mathbb{E}}_{t} f_T(t) + \mathop{\mathbb{E}}_{v} f_V(v) + \mathop{\mathbb{E}}_{a} f_A(a)\\
&+ \mathop{\mathbb{E}}_{t,v} f_{TV}(t, v) + \mathop{\mathbb{E}}_{t,a} f_{TA}(t, a)\\
&+ \mathop{\mathbb{E}}_{v,a} f_{VA}(v, a) + \mathop{\mathbb{E}}_{t,v,a} f_{TVA}(t, v, a)\Big) && (12)\\
=&\Big(f_T(x_T) + \mathop{\mathbb{E}}_{v} f_V(v) + \mathop{\mathbb{E}}_{a} f_A(a)\Big)\\
&+ \Big(\mathop{\mathbb{E}}_{t} f_T(t) + f_V(x_V) + \mathop{\mathbb{E}}_{a} f_A(a)\Big)\\
&+ \Big(\mathop{\mathbb{E}}_{t} f_T(t) + \mathop{\mathbb{E}}_{v} f_V(v) + f_A(x_A)\Big)\\
&- 2\Big(\mathop{\mathbb{E}}_{t} f_T(t) + \mathop{\mathbb{E}}_{v} f_V(v) + \mathop{\mathbb{E}}_{a} f_A(a)\Big) && (13)\\
=& f_T(x_T) + f_V(x_V) + f_A(x_A) && (14)
\end{aligned}
$$

□

Compared to BI in the bimodal case, we need to also remove trimodal interactions for BI in the trimodal context.

Claim 3. *BI is defined for three modalities as*

$$
\begin{aligned}
& BI(f)\\
=& \mathop{\mathbb{E}}_{t}[f(t, x_V, x_A) - UC(f, t, x_V, x_A)]\\
&+ \mathop{\mathbb{E}}_{v}[f(x_T, v, x_A) - UC(f, x_T, v, x_A)]\\
&+ \mathop{\mathbb{E}}_{a}[f(x_T, x_V, a) - UC(f, x_T, x_V, a)] && (15)\\
=& \mathop{\mathbb{E}}_{t} f(t, x_V, x_A) + \mathop{\mathbb{E}}_{v} f(x_T, v, x_A)\\
&+ \mathop{\mathbb{E}}_{a} f(x_T, x_V, a) - 2 \mathop{\mathbb{E}}_{t,v} f(t, v, x_A)\\
&- 2 \mathop{\mathbb{E}}_{t,a} f(t, x_V, a) - 2 \mathop{\mathbb{E}}_{v,a} f(x_T, v, a)\\
&+ 3 \mathop{\mathbb{E}}_{t,v,a} f(t, v, a) && (16)\\
=& f_{TV}(x_T, x_V) + f_{TA}(x_T, x_A) + f_{VA}(x_V, x_A)
\end{aligned}
$$

The omitted steps are to apply the definition of UC and cancelling terms to get to Equation 16. From there on, we write f as $f_T + f_V + f_A + f_{TV} + f_{TA} + f_{VA} + f_{TVA}$ and use their properties (expected values of the bi/trimodal function are 0).

C Study Details

In addition to the three unimodal and the trimodal combinations we explored bimodal combinations: 1) the muted video with the transcript (TV); 2) the muted video with the low-pass filtered audio (VA); 3) the transcript with the low-pass filtered audio (TA); 4) for comparison the original audio with the transcript (TA_O).

C.1 Reliability

We report two types reliabilities: the averaged pairwise reliability between two random raters

Combination	Avg. ICC(2, 1)		ICC(2, k-1)	
	Arousal	Valence	Arousal	Valence
T	0.36	0.55	0.96	0.98
V	0.52	0.64	0.98	0.99
A	0.57	0.38	0.98	0.96
TV	0.48	0.62	0.97	0.98
VA	0.56	0.61	0.98	0.98
TA	0.60	0.54	0.98	0.98
TA_O	0.55	0.62	0.98	0.98
TVA_O	0.56	0.64	0.98	0.99

Table 5: Pairwise and effective reliability across the eight combinations. ICC is calculated with the R package psych.

(ICC(2,1)) and the effective reliability of the mean over k=8 raters (ICC(2, k-1)). Pairwise and effective reliability address different purposes: pairwise is needed to determine how many raters are needed to achieve a targeted effective reliability (Rosenthal, 2005). Averaging over raters is important as emotional dimensions are subjective and difficult to annotate (especially when modalities are missing). The effective reliability describes how reliable the mean over the raters is, i.e., if we were to draw a new set of ratings and were to average them, how similar is this new mean to our current mean.

Except for transcripts-only (T) on arousal and acoustic-only (A) on valence, all pairwise reliabilities are moderate (between 0.5 and 0.75) (Koo and Li, 2016), see Table 5. The effective reliability (Rosenthal, 2005) of the mean over k raters as measured by ICC(2, k-1) is excellent (above 0.9) for all combinations. Instead of directly taking the mean over the raters, we apply, as common in affective computing, a z-normalization for each rater (Valstar et al., 2016; Busso et al., 2008) and take a weighted mean (Grimm and Kroschel, 2005) over the raters.

C.2 Compensation

All raters are paid the same fixed amount, leading to an average hourly rate of 11.14 USD/h.

D Additional Experiments

MRO reaches $|UC(\hat{y}_{\text{bi}} + \hat{y}_{\text{tri}})| + |BI(\hat{y}_{\text{tri}})| = 0$ when trained long enough. As we have seen in Figure 4, a model might need to be optimized long enough to minimize $|UC(\hat{y}_{\text{bi}} + \hat{y}_{\text{tri}})| + |BI(\hat{y}_{\text{tri}})|$. We therefore also investigate $|UC(\hat{y}_{\text{bi}} + \hat{y}_{\text{tri}})| + |BI(\hat{y}_{\text{tri}})|$ when models are trained without early stopping. While such models are more likely to have poor generalization performance this allows us to test how much we could minimize $|UC(\hat{y}_{\text{bi}} + \hat{y}_{\text{tri}})| + |BI(\hat{y}_{\text{tri}})|$ with MRO and sMRO. As can be seen in Table 6, MRO and sMRO are numerically very close to 0.0 demonstrating that such optimized models almost perfectly separate unimoda, bimodal, and trimodal interactions.

	Joint	sMRO	MRO
SEWA			
Arousal	[0.81, 0.87]	[0.00, 0.00]	[0.00, 0.01]
Valence	[0.29, 0.31]	[0.00, 0.00]	[0.02, 0.03]
IEMOCAP			
Arousal	[0.41, 0.42]	[0.00, 0.00]	[0.03, 0.03]
Valence	[0.20, 0.21]	[0.00, 0.00]	[0.05, 0.06]
MOSI			
Sentiment	[0.14, 0.15]	[0.00, 0.00]	[0.00, 0.00]
MOSEI			
Sentiment	[0.41, 0.42]	[0.00, 0.00]	[0.00, 0.00]
Polarity	[0.44, 0.45]	[0.04, 0.04]	[0.00, 0.00]
Happiness	[0.44, 0.45]	[0.15, 0.15]	[0.01, 0.01]
TPOT			
Constructs	[0.29, 0.29]	[0.01, 0.01]	[0.04,0.04]
Instagram			
Intent	[0.12, 0.13]	[0.00, 0.01]	[0.02, 0.02]

Table 6: Bootstrapped average of normalized $|UC(\hat{y}_{\text{bi}} + \hat{y}_{\text{tri}})| + |BI(\hat{y}_{\text{tri}})|$ on the test folds (1.0 corresponds to a magnitude of one standard deviation) when models are trained without early stopping. Lower is better (ideally 0.0).

Removing stop-gradient leads to a worse separation: Theoretically, we should be able to remove stop-gradient (sg) from Equation 9 as they have the same global minima. In practice, we observe that doing so leads to worse separation of interactions, see Table 7.

MRO applies to transformers as well: When choosing transformers (Vaswani et al., 2017) as a base model instead of multilayer perceptrons, we observe the same trend that a model trained without MRO does not separate the multimodal interactions, whereas when trained with MRO the interactions are far better separated, see Table 8.

Interactions needed for amplifiers, ambiguity, and rare behaviors. Table 9 summarizes the five segments with the largest absolute errors $|\text{TVA}_O - \hat{y}^{\text{human}}_{\text{uni}}|$ separately for arousal and valence (we refer to these segments with A_1 to A_5 and V_1 to V_5). Qualitatively, three groups emerge: amplifiers, ambiguities, and rare behaviors.

Unimodal amplifiers: Amplifiers are essential

	Performance	$\lvert UC(\hat{y}_{\text{bi}} + \hat{y}_{\text{tri}})\rvert + \lvert BI(\hat{y}_{\text{tri}})\rvert$
SEWA		
Arousal	0.599	[0.05, 0.06]
Valence	0.638	[0.13, 0.14]
IEMOCAP		
Arousal	0.316	[0.27, 0.29]
Valence	0.297	[0.30, 0.32]
MOSI		
Sentiment	0.660	[0.03, 0.03]
MOSEI		
Sentiment	0.724	[0.06, 0.06]
Polarity	0.605	[0.17, 0.19]
Happiness	0.644	[0.16, 0.16]
TPOT		
Constructs	0.569	[0.13, 0.13]
Instagram		
Intent	0.890	[0.04, 0.04]

Table 7: Average performance and bootstrapped average of normalized $\lvert UC(\hat{y}_{\text{bi}} + \hat{y}_{\text{tri}})\rvert + \lvert BI(\hat{y}_{\text{tri}})\rvert$ on the test folds (1.0 corresponds to a magnitude of one standard deviation) when models are trained without early stopping.

	Joint	MRO
MOSEI		
Sentiment	[0.64, 0.67]	[0.01, 0.01]

Table 8: Average of normalized $\lvert UC(\hat{y}_{\text{bi}} + \hat{y}_{\text{tri}})\rvert + \lvert BI(\hat{y}_{\text{tri}})\rvert$ on the test folds (1.0 corresponds to a magnitude of one standard deviation) when using transformers as a base model instead of multilayer perceptrons.

for valence and sentiment as an intense expression can be very negative or positive. A modality might contain a strong amplifier (language in case of example V_3 and V_4) but the modality might not provide strong evidence for the directionality. In such cases, non-additive interactions are needed to combine the directionality from one modality with the amplifier from another modality.

Ambiguities: When a modality might not provide information in either direction (language in case V_1, V_2, V_5, A_1, A_2), more contextual information in form of bimodal interactions is needed.

Rare behaviors: When a typically important modality is "missing" (language in case of A_2 and vision in case of A_3) or a typically less important modality contains an important behavior (acoustic in case of V_2) it changes the relative importance of the remaining modalities. Unlike the routing model, additive models have no mechanism to re-weight how important modalities are. When a modality is unexpectedly very (un)important, a bimodal or trimodal model becomes necessary.

E Reproducibility

Computing Resources: All model are implemented in PyTorch and were optimized on servers with consumer-level graphic cards.

Model Information: The validation performance, the training time, and the number of parameters for the best models as chosen based on the validation performance, are listed in Table 10.

Hyperparameter Search: All models and datasets have the same exhaustive hyperparameter search, see Table 11. The gridsearch determined in most cases the same hyperparameter across the different optimization strategies (Joint, sMRO, and MRO), we therefore only highlight the best hyperparameters for MRO in Table 11. The performance metrics reported in Table 3 are also used to selected the best validation model.

Datasplits: MOSI and MOSEI have an established hold-out test set, we use it for testing. SEWA has a private test set: we use the public development set for testing. IEMOCAP and Instagram have an established 5-fold test setup which we use.

	Transcript	Non-verbal behaviors	T	V	A	TVA_O	$\hat{y}_{\text{uni}}^{\text{human}}$
Arousal							
A_1	I'm gonna forget him.	gaze aversion, almost whining	-0.57	0.04	0.32	1.16	0.05
A_2	What?	not attentive, quiet	-0.62	-0.83	-1.67	-2.1	-1.17
A_3	Do you know how long it's gonna take me to start all over and fill out the new form?	little movement, loud	1.11	-0.36	0.1	0.98	0.1
A_4	That's right. That's right. I mean he would want us to, you know, celebrate the life that he- that he lived and, you know, enjoy the rest of ours as much as we can.	gaze aversion, quiet	-0.23	-1.11	-1.14	-0.17	-0.99
A_5	Well, I need–I need you to be able to do this for me 'cause I can't do anything about it.	gaze aversion, almost whining	0.47	-0.43	0.29	0.73	0.03
Valence							
V_1	I mean, it's just as hard for me, but–I know that we can do it, you know	eye-gaze aversion, fidgeting, quiet	0.25	-0.82	-0.08	-1.51	-0.41
V_2	I did exactly what they told me to do.	loud, determined	-0.02	0.30	-1.45	-1.06	-0.06
V_3	Oh, wow. You got in?	smile, loud/staccato-like voice	1.08	1.49	-0.86	2.04	1.08
V_4	Oh, my God. That's so dramatic.	smile, slight laughter, pitch jumps	-0.70	1.70	-0.12	1.68	0.74
V_5	How can you lose my luggage like from like, in –	looking up, hand gestures	-1.39	0.88	-0.37	-0.89	-0.01

Table 9: The five segments with the largest absolute errors when predicting TVA_O with T, V, A.

	Tri			Routing			Joint			sMRO			MRO		
	perf	sec	params	perf	sec	params	perf	sec	params	perf	sec	params	perf	sec	params
MOSI															
Sentiment	0.73	20.3	138263	0.741	68.2	556168	0.724	29.7	138263	0.735	50.8	111623	0.732	32.7	111623
MOSEI															
Sentiment	0.71	97.6	141163	0.71	201.9	567768	0.713	189.9	503347	0.715	254.9	441187	0.71	171.4	114523
Polarity	0.611	99.1	441187	0.613	295.9	567768	0.617	181.8	441187	0.614	247.3	86647	0.617	290.9	441187
Happiness	0.636	87.4	441187	0.642	358.3	823208	0.638	252.6	441187	0.644	372.3	21317	0.642	380.0	503347
IEMOCAP															
Arousal	0.64	41.6	111623	0.644	301.2	811608	0.643	60.3	84327	0.662	165.8	491747	0.65	69.6	4303
Valence	0.67	35.5	138263	0.66	118.8	811608	0.658	27.3	138263	0.642	64.8	22683	0.651	41.4	138263
SEWA															
Arousal	0.367	174.7	22683	0.372	691.9	78518	0.388	266.3	41467	0.396	466.6	43077	0.384	484.1	43077
Valence	0.424	190.3	20283	0.427	527.3	200408	0.424	147.0	20283	0.425	461.5	11693	0.428	144.1	22683
TPOT															
Constructs	0.569	105.2	80038	0.557	359.7	56036	0.568	143.4	180710	0.565	251.7	14298	0.576	243.5	14298
Instagram															
Intent	0.888	9.0	12098	0.757	75.7	13427	0.9	24.7	12098	0.89	24.9	12098	0.896	18.0	12098

Table 10: Validation performance (perf), training time in seconds (sec), and the number of parameters of the best validation model (params).

Learning rate	$\{0.05^{5}, 0.01^{1}, 0.005^{2,6,10}, 0.001^{7,8}, 0.0001^{3,4}, 0.00001^{9}\}$
L2 weight decay	$\{0.1, 0.01^{1,3,4,5,6,7,8,9,10}, 0.001^{2}, 0.0\}$
Hidden layers	$\{(5,), (10,)^{10}, (20, 10)^{5,8,9}, (10, 20)^{7}, (100, 20, 10)^{1,2,3}, (100, 100, 10)^{4,6}\}$
Feature Selection	$\{\text{yes}^{5,10}, \text{no}^{1,2,3,4,6,7,8,9}\}$
Fusion	$\{\text{concatenation}^{3,4,7,10}, \text{tensor fusion}^{1,2,5,6,8,9}\}$

Table 11: Parameter search. Parameters that were determined to be best for MRO and task x are indicated by x in the superscript. We enumerate tasks in the same order as they are presented in Table 3.