MMOE: Mixture of Multimodal Interaction Experts

Haofei Yu, Paul Pu Liang, Ruslan Salakhutdinov, Louis-Philippe Morency
Carnegie Mellon University
{haofeiy, pliang, rsalakhu, morency}@cs.cmu.edu

Abstract

Multimodal machine learning, which studies the information and interactions across various input modalities, has made significant advancements in understanding the relationship between images and descriptive text. However, this is just a portion of the potential multimodal interactions seen in the real world and does not include new interactions between conflicting utterances and gestures in predicting sarcasm, for example. Notably, the current methods for capturing shared information often do not extend well to these more nuanced interactions, sometimes performing as low as 50% in binary classification. In this paper, we address this problem via a new approach called MMOE, which stands for a mixture of multimodal interaction experts. Our method automatically classifies data points from unlabeled multimodal datasets by their interaction type and employs specialized models for each specific interaction. Based on our experiments, this approach improves performance on these challenging interactions by more than 10%, leading to an overall increase of 2% for tasks like sarcasm prediction. As a result, interaction quantification provides new insights for dataset analysis and yields simple approaches that obtain state-of-the-art performance.

1 Introduction

Recent advances in vision-language multimodal architecture and pretraining [Zhu et al., 2023, Li et al., 2023, Liu et al., 2023] have predominantly centered on enhancing the representation interplay between modalities [Huang et al., 2021]. This enhancement often stems from the principle of multi-view redundancy Liang et al. [2023b]. Parallel to this, contemporary multimodal research has expanded its scope, categorizing a broader spectrum of multimodal interactions. These go beyond mere *redundancy* to capture instances where task-relevant information does not merely reside in the overlapping region of both modalities Liang et al. [2023c]. Instead, it might hinge on *unique* details from either modality (e.g. detecting laughter from someone not observable), or the result of a *synergistic* fusion of both modalities, producing insights absent when either modality is considered in isolation (e.g. sarcasm discerned from incongruent speech and gestures). [Williams and Beer, 2010, Liang et al., 2023a, Marsh and Domas White, 2003] This body of research also delves into discerning whether predictions from different modalities *align* or *contrast* with one another. [Bateman, 2014, Kruk et al., 2019, Zhang et al., 2018] It still remains a challenge for state-of-the-art multimodal models to tackle multiple types of interactions. Furthermore, these interactions can vary significantly even within the same dataset for different data points, for example, sometimes sarcasm is evident from language alone and other times due to synergistic disagreement between language and gestures.

Making progress on these different interactions requires new definitions and modeling paradigms that existing models are not designed to capture. To model these diverse interactions, our key insight is that **single model may not be necessarily suitable for capturing all kinds of interactions at the same time**. Indeed, we find that today's state-of-the-art models struggle significantly with disagreement and synergistic multimodal interactions, with reductions in performance up to 40% on sarcasm detection as compared to redundant cases.

Preprint. Under review.

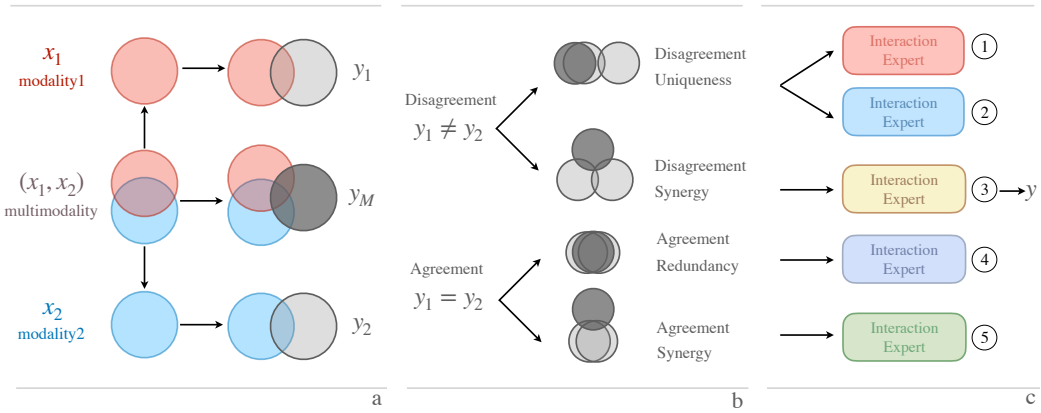

Figure 1: Our proposed *mixture of multimodal interaction expert* framework includes three components: **a**.partial label collection **b**. datapoint-level multimodal interaction categorization **c**.mixture of experts prediction.

Based on this insight, we propose a new approach called MMOE (mixture of multimodal interaction experts), which classifies the different types of multimodal interactions in a given task and utilizes multiple expert models specifically designed for each type of interaction. These expert models can differ based on in-context learning, parameter-based tuning, or an entirely different architecture. Using our multimodal interaction experts, we raise the performance on some interaction types by more than 10% compared with using a single model, resulting in overall improvements over the state-of-the-art models. The resulting approach is easy to implement and scales easily with large-scale models.

2 Mixture of Multimodal Interaction Experts

In this section, we first provide problem formulation for our target multimodal prediction task. After that, we explain three key components in our framework mentioned in Fig 1 in detail.

2.1 Problem formulation

In the multimodal prediction task, let $\mathcal{X}_i$ and $\mathcal{Y}$ be finite sample spaces for modality-specific features and target labels. Let $\Delta_{\mathcal{Y}} = \{r \in \mathbb{R}_+^{|\mathcal{Y}|} \mid \|r\|_1 = 1\}$ be the probability simplex over labels $\mathcal{Y}$. Given modality-specific features X_1, X_2 (with support $\mathcal{X}_i$), for one multimodal prediction task, our goal is to learn a multimodal model $\mathcal{F}_M \ni f_M : \mathcal{X}_1 \times \mathcal{X}_2 \to \Delta_{\mathcal{Y}}$ to predict Y (with support $\mathcal{Y}$).

2.2 Predicting partial labels and full labels

Utilizing labeled unimodal data $\mathcal{D}_i = \{(x_i, y) : \mathcal{X}_i \times \mathcal{Y}\}$ and labeled multimodal data $\mathcal{D}_M = \{(x_1, x_2, y) : \mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{Y}\}$, we can obtain the set of unimodal classifiers $\mathcal{F}_i \ni f_i : \mathcal{X}_i \to \Delta_{\mathcal{Y}}$ and multimodal classifiers $\mathcal{F}_M \ni f_M : \mathcal{X}_1 \times \mathcal{X}_2 \to \Delta_{\mathcal{Y}}$. For each unlabeled multimodal datapoint $d \in \tilde{\mathcal{D}}_M = \{(x_1, x_2) : \mathcal{X}_1 \times \mathcal{X}_2\}$, we can get probability δ_1 and δ_2 from unimodal classifier f_1 and f_2. These two labels are partial label estimations based on unimodal information. Similarly, we can get δ_M from multimodal classifier f_M, which is considered as full label estimation based on multimodal data. Practically, apart from predicting discrete labels y_1, y_2, and y_M from $\mathcal{Y}$, we calibrate pretrained models by utilizing probability from models as δ_1, δ_2, and δ_M.

2.3 Categorizing datapoints based on multimodal interactions

Each labeled multimodal datapoint $d \in \mathcal{D}_M = \{(x_1, x_2, y) : \mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{Y}\}$ can be assigned with different multimodal interaction type, including redundancy, uniqueness, and synergy. Interaction can be estimated by finding data points in which individual predictions from each modality agree or disagree with each other. Moreover, interaction can be detected where the joint distribution differs the most from the marginals.

Definition 1. *(Modality disagreement) Given $x_1 \sim \mathcal{X}_1$, $x_2 \sim \mathcal{X}_2$, as well as unimodal classifiers $f_1 : \mathcal{X}_1 \to \delta_{\mathcal{Y}} \in \mathcal{F}_1$ and $f_2 : \mathcal{X}_i \to \delta_{\mathcal{Y}} \in \mathcal{F}_2$, we define modality disagreement as $d(f_1, f_2)$ where $d : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^{\geq 0}$ is a distance function in label space scoring the disagreement of f_1 and f_2's predictions.*

Converting partial labels to agreement and disagreement We convert model-predicted partial labels into agreement and disagreement by comparing the output of unimodal classifiers against each

other. Based on the modality disagreement definition, when $d(f_1, f_2) > \gamma$, where γ is a pre-defined threshold, two modalities are considered as disagreed with each other. Otherwise, they are considered as agreed. Practically, we define $d(f_1, f_2) = \|\delta_1 - \delta_2\|_1$ as our distance function. Based on our experiments, we found that agreement and disagreement cases distinguish multimodal performance well. Utilizing this definition, unlabeled multimodal data $\tilde{\mathcal{D}}_M = \{(x_1, x_2) : \mathcal{X}_1 \times \mathcal{X}_2\}$ is classified as non-overlapping disagreed ones and agreed ones.

Definition 2. *(Modality **R**edundancy, **U**niqueness, **S**ynergy [**RUS**]) Given $x_1 \sim \mathcal{X}_1$, $x_2 \sim \mathcal{X}_2$, as well as unimodal classifiers $f_1 : \mathcal{X}_1 \to \delta_{\mathcal{Y}} \in \mathcal{F}_1$ and $f_2 : \mathcal{X}_2 \to \delta_{\mathcal{Y}} \in \mathcal{F}_2$ and a multimodal classifier $f_M : \mathcal{X}_1 \times \mathcal{X}_2 \to \delta_{\mathcal{Y}}$, we define modality redundancy as $R = -d(f_1, f_M) - d(f_1, f_2) - d(f_2, f_M)$, modality uniqueness for $\mathcal{X}_1$ as $U_1 = d(f_2, f_M) + d(f_1, f_2) - d(f_1, f_M)$, modality uniqueness for $\mathcal{X}_2$ as $U_2 = d(f_1, f_M) + d(f_1, f_2) - d(f_2, f_M)$, modality synergy as $S = d(f_1, f_M) + d(f_2, f_M)$, where $d : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^{\geq 0}$ is a distance function in label space scoring the disagreement of any two classifiers' predictions.*

Converting partial and full labels to redundancy, uniqueness, and synergy Instead of simply utilizing f_1 and f_2 to measure modality disagreement, we would like to introduce more fine-grained multimodal interaction types by adding f_M and measuring distances between partial labels δ_1, δ_2 and full labels δ_M. Since we are dealing with unlabeled multimodal data $\tilde{\mathcal{D}}_M = \{(x_1, x_2) : \mathcal{X}_1 \times \mathcal{X}_2\}$, we have no access to ground-truth labels $y^* \in \mathcal{Y}$ but only have full labels $\delta_M \in \Delta_{\mathcal{Y}}$ as a proxy metric. Our hypothesis is that while predicted δ_M could be hard and inaccurate for multiple multimodal interactions like synergy, it is easier to classify a datapoint as one interaction type just by looking at how the label changes from unimodal to multimodal, versus actually predicting the datapoint itself. Intuitively, for each datapoint $d \in \tilde{\mathcal{D}}_M = \{(x_1, x_2) : \mathcal{X}_1 \times \mathcal{X}_2\}$, its modality information is redundant if δ_1, δ_2, and δ_M are all close to each other; $\mathcal{X}_1$ information is unique if δ_1 is close to δ_2 but δ_2 is far from δ_M; $\mathcal{X}_2$ information is unique if δ_2 is close to δ_1 but δ_1 is far from δ_M; modality information is synergistic if δ_1 and δ_2 are both far from δ_M. Based on the definition of **RUS** (standing for **R**udandancy, **U**niqueness, and **S**ynergy), unlabeled multimodal dataset $\tilde{\mathcal{D}}_M = \{(x_1, x_2) : \mathcal{X}_1 \times \mathcal{X}_2\}$ can be classified into three non-overlapping parts. One of the main limitations of this method is that it is limited to two modalities and is not able to expand to multiple modalities yet.

2.4 Multimodal Interaction Experts

In section 2.3, we have introduced the definition of modality disagreement, redundancy, uniqueness, and synergy. Moreover, we mentioned that modality disagreement measurement can split unlabeled multimodal datasets into two non-overlapping sub-parts. RUS measurement can split the dataset into three non-overlapping sub-parts. To push a step further, we can combine both disagreement and RUS measurement to create a comprehensive multimodal interaction standard. Typically, when modality agrees with each other, there are no uniqueness-dominant cases. Similarly, when modality disagrees with each other, there should be no redundancy-dominant cases. Therefore, in general, as illustrated in Fig 1, we create a 5-type multimodal interaction standard when facing two modalities $\mathcal{X}_1$ and $\mathcal{X}_2$: ① disagreement $\mathcal{X}_1$ uniqueness; ② disagreement $\mathcal{X}_2$ uniqueness; ③ disagreement synergy; ④ agreement redundancy; ⑤ agreement synergy.

For each multimodal interaction type, we designed an expert model to handle. Motivated by the Socratic model Zeng et al. [2022], we first project information from different modalities into text modality and make sure the text description loses minimal modality-specific features. Secondly, we implement few-shot in-context learning based on large language models (LLMs) using data points that are categorized as the same multimodal interaction type.

3 Experiments

3.1 Experimental Setting

For the dataset, we use `MUSTARD` dataset to perform our experiments and choose speaker-dependent training and testing split. Each sample in `MUSTARD` includes character scripts as text modality, video clip as video modality, and speaking audio as audio modality. For baseline models, we compare with multiple supervised multimodal baselines including LF-DNN [Ding et al., 2022], MULT [Liang et al., 2021], LMF [Liu et al., 2018], MFN [Zadeh et al., 2018], and EF-LSTM [**?**]. Additionally, we compare with multiple unimodal baselines based on face emotion recognition [Xiang and Zhu, 2017] and speech emotion recognition [Hsu et al., 2021]. For our proposed method, We select `gpt-3.5-turbo` as our model for experts. We use few-shot prompting on `gpt-3.5-turbo` to categorize data points and construct few-shot examples under the same category.

MUSTARD	Precision↑	Recall↑	F1↑
Multimodal Model			
SVM	$71.9_{\pm 0.00}$	$71.4_{\pm 0.00}$	$71.5_{\pm 0.00}$
MFN	$72.65_{\pm 1.02}$	$72.46_{\pm 0.92}$	$72.41_{\pm 0.9}$
LMF	$71.81_{\pm 2.25}$	$71.3_{\pm 1.81}$	$71.16_{\pm 1.7}$
MULT	$73.07_{\pm 1.61}$	$72.32_{\pm 0.96}$	$72.11_{\pm 0.90}$
LF-DNN-v2	$\mathbf{73.80}_{\pm 1.85}$	$73.62_{\pm 1.75}$	$73.58_{\pm 1.73}$
Unimodal Model			
MTCNN [vision]	$62.10_{\pm 0.69}$	$58.69_{\pm 1.35}$	$66.49_{\pm 2.10}$
HuBERT [audio]	$62.36_{\pm 4.02}$	$75.28_{\pm 0.57}$	$67.55_{\pm 3.13}$
GPT-3.5 [text]	$65.56_{\pm 0.27}$	$87.40_{\pm 0.13}$	$74.78_{\pm 0.19}$
Multimodal Extended Language Models			
GPT-3.5 [text+audio]	$61.52_{\pm 0.28}$	$91.24_{\pm 0.63}$	$73.34_{\pm 0.12}$
+MoE	$63.47_{\pm 0.03}$	$\mathbf{91.69}_{\pm 0.21}$	$\mathbf{74.93}_{\pm 0.04}$

Table 1: Test performance on sarcasm dataset with speaker-dependent setup. Each score is an average score of three runs with three random seeds.

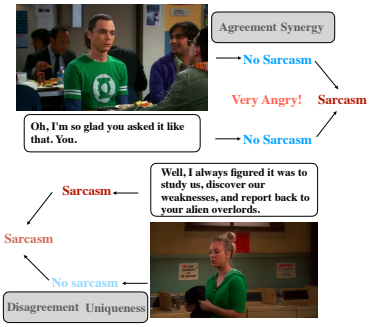

Figure 2: Case study for agreement synergy and disagreement uniqueness.

3.2 Results

Performance Results Based on Table 1, we can see that we are able to get the state-of-the-art F1 result of 74.93% compared with a few-shot unimodal baseline, supervised multimodal baseline, and Socratic model. Additionally, we can conclude that by only adding multiple mixtures of experts based on the Socratic model, the model's performance can have a 1.5% gain. Moreover, we observed that recall increased from 87.40 to 91.69 after introducing multimodal information into the detection process. It indicates that more sarcastic situations can be detected.

Case Study To prove that our proposed mixture of multimodal interaction experts is able to capture appropriate multimodal interaction and fix errors, we provide two concrete case studies to show the fixed uniqueness and synergy type. Fig 2 shows one agreement synergy case where multimodal features motivate new information and one disagreement text uniqueness case where the video does not include a clear sarcastic signal.

Expert Model Analysis We also do analysis based on the whole MUSTARD dataset. We classified all data points by their detected multimodal interaction type and found that redundancy-dominant data points, which have 364 data points, are the most common cases. Text modality uniqueness-dominant cases include 172 data points. It indicates that text modality is crucial for sarcasm detection. Based on Table 2, we can see that with appropriate in-context learning, both agreement and disagreement synergy have much better performance (gain 61.47% and 22.81% separately) compared with a single model. For redundancy-type data points, even though the F1 score remains almost the same, it has a higher confidence score (increasing from 3.49 to 3.77 out of 5) when using specific experts for this interaction type. How to design a more useful and effective multimodal expert for each interaction type is left as a future work.

MUSTARD	#Example	F1↑ (Single Model)	F1↑ (Expert Model)	Improvement↑ (%)
Agreement & Redundancy	364	$\mathbf{85.98}_{3.49}$	$85.33_{\mathbf{3.77}}$	-0.75%
Agreement & Synergy	37	$18.75_{2.33}$	$30.30_{2.52}$	**+61.47%**
Disagreement & Unique [text]	172	$69.19_{3.59}$	$69.19_{3.57}$	+0.00%
Disagreement & Unique [audio]	62	$54.23_{1.88}$	$53.52_{1.31}$	-1.31%
Disagreement & Synergy	55	$47.05_{2.36}$	$57.78_{2.50}$	**+22.81%**

Table 2: An analysis of the full MUSTARD dataset based on few-shot prompting, focusing on both text and audio modalities. For each multimodal interaction F1 score, we present an average confidence score to indicate model's confidence towards its prediction, which ranges from 0 to 5. Each F1 score listed in the table is an average score of three runs.

4 Conclusion

This paper proposed an approach for quantifying multiple types of multimodal interactions in multimodal datasets and designing a new approach called MMOE. Experimentally, using multimodal interaction experts raises the performance on these interactions up to a 74.93% F1 score. As a result, not only does interaction quantification provide new insights for dataset analysis, but also simple approaches to obtain state-of-the-art performance.

References

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

Ning Ding, Sheng-wei Tian, and Long Yu. A multimodal fusion method for sarcasm detection based on late fusion. *Multimedia Tools and Applications*, 81(6):8597–8616, 2022.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021.

Yu Huang, Chenzhuang Du, Zihui Xue, Xuanyao Chen, Hang Zhao, and Longbo Huang. What makes multi-modal learning better than single (provably). *Advances in Neural Information Processing Systems*, 34:10944–10956, 2021.

Julia Kruk, Jonah Lubin, Karan Sikka, Xiao Lin, Dan Jurafsky, and Ajay Divakaran. 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 (EMNLP-IJCNLP)*, pages 4622–4632, 2019.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

Paul Pu Liang, Yiwei Lyu, Xiang Fan, Zetian Wu, Yun Cheng, Jason Wu, Leslie Yufan Chen, Peter Wu, Michelle A Lee, Yuke Zhu, et al. Multibench: Multiscale benchmarks for multimodal representation learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021.

Paul Pu Liang, Yun Cheng, Xiang Fan, Chun Kai Ling, Suzanne Nie, Richard Chen, Zihao Deng, Faisal Mahmood, Ruslan Salakhutdinov, and Louis-Philippe Morency. Quantifying & modeling feature interactions: An information decomposition framework. *arXiv preprint arXiv:2302.12247*, 2023a.

Paul Pu Liang, Zihao Deng, Martin Ma, James Zou, Louis-Philippe Morency, and Ruslan Salakhutdinov. Factorized contrastive learning: Going beyond multi-view redundancy. *arXiv preprint arXiv:2306.05268*, 2023b.

Paul Pu Liang, Chun Kai Ling, Yun Cheng, Alex Obolenskiy, Yudong Liu, Rohan Pandey, Alex Wilf, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal learning without labeled multimodal data: Guarantees and applications. *arXiv preprint arXiv:2306.04539*, 2023c.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. Efficient low-rank multimodal fusion with modality-specific factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2247–2256, 2018.

Emily E Marsh and Marilyn Domas White. A taxonomy of relationships between images and text. *Journal of documentation*, 2003.

Paul L Williams and Randall D Beer. Nonnegative decomposition of multivariate information. *arXiv preprint arXiv:1004.2515*, 2010.

Jia Xiang and Gengming Zhu. Joint face detection and facial expression recognition with mtcnn. In *2017 4th international conference on information science and control engineering (ICISCE)*, pages 424–427. IEEE, 2017.

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Memory fusion network for multi-view sequential learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, et al. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv preprint arXiv:2204.00598*, 2022.

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

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.