11-877 Advanced Multimodal Machine Learning **Spring 2023**

Week 4: Modality Interactions

Instructors: Louis-Philippe Morency and Paul Liang *Synopsis Leads: Jingyi Zhang, Neehar Peri*

Edited by Paul Liang *Scribes: Aditya Srikanth Veerubhotla*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

Follow the rest of the class here: https://cmu-multicomp-lab.github.io/adv-mmml-course/spring2023/

Summary: Multimodal machine learning is the study of computer algorithms that learn and improve through the use and experience of multimodal data. It brings unique challenges for both computational and theoretical research given the heterogeneity of various data sources. In week 4's discussion session, the class discussed modality interactions - the various ways in which modalities combine with each other to give rise to new information for task inference. The following was a list of provided research probes:

- What are the different ways in which modalities can interact with each other when used for prediction tasks? Think across both semantic and statistical perspectives. Can we formalize a taxonomy of such interactions, which will enable us to compare and contrast them more precisely? In fact, should we even try creating such a taxonomy?
- Can you think of ways modalities could interact with each other, even if there is no prediction task? How are modalities interacting during cross-modal translation? During multimodal generation?
- Linking back to last week's discussion, are there cases where modalities are connected but do not interact? Or interact but are not connected? Can we design formal experiments to test either hypothesis?
- What mathematical or empirical frameworks can be used to formalize the meaning of interactions? How can we subsequently define estimators, where we can accurately quantify the presence of each type of interaction given a dataset?
- Some definitions (from the semantic category) typically require human interactions to detect and quantify interactions. What are some opportunities and limitations of using human judgment to analyze interactions? Can we potentially design estimators to automate the human labeling process?
- What are the design decisions (aka inductive biases) that can be used when modeling each type of interaction in machine learning models?
- What are the advantages and drawbacks of designing models to capture each type of cross-modal interaction? Consider not just prediction performance, but tradeoffs in time/space complexity, interpretability, etc.

As background students read the following papers:

- (Required) [Partan and Marler, 2005]. This paper studies semantic definitions of interactions in communicative modalities used by humans and animals. Think about how we can operationalize these intuitions in machine learning.
- (Required) [Williams and Beer, 2010]. This paper introduces the PID framework and estimates the different ways in which 2 or more features can interact to make a prediction, using an information theory perspective.
- (Recommended) [Tsakona, 2009] Study of image-text interactions in multimedia, specifically humor. See if you can find similar references discussing what types of interactions exist, and what they mean for other modalities and tasks too.
- (Recommended) [Jewitt, 2011] Chapter 11 of *The Routledge Handbook of Multimodal Analysis*. This chapter discusses potential interactions between image and text that affect how we prescribe meaning to both modalities together.

- (Recommended) [Ruiz et al., 2006] A case study of classifying interactions through how humans use the modalities.
- (Recommended) [Hessel and Lee, 2020] A quantification method to detect whether a model learns non-additive interactions.
- (Recommended) [Liang et al., 2023], Scalable estimators for PID for modern multimodal datasets and models, including using PID estimates for model selection.
- (Relevant) [Tsang et al., 2020] Detecting interactions in black-box models using gradient-based methods.
- (Relevant) [Jakulin and Bratko, 2003] Detecting interactions in black-box models, an information theory perspective.
- (Relevant) [Baron and Kenny, 1986] Landmark paper from statistics looking at building models capturing various interactions through moderator and mediator variables

1 A Taxonomy of Multimodal Interactions

Redundant	Non-Redundant
Equivalence	Independent
Enhancement	Dominance
	Modulation
	Emergence
	⋆ Destructive
	⋆ Logical
	⋆ Temporal
	⋆ Translation

Table 1: Modalities can interact through redundant (e.g equivalence, enhancement) and non-redundant signals (e.g. independence, modulation, dominance, emergence).

What are interactions? Interactions are task-specific, and investigate how modalities are combined to bring about new information for a task. We can consider modeling each unimodal feature in one node and build a bipartite graph between nodes to model the interactions. If edge weights are affecting the update of the representations of the nodes, then we can say there's an interaction between modalities. Otherwise, they are just connected. Also, we can define interaction to be how parameter gradients flow within the deep neural network. Different tasks can require different interactions at different scales. For example, in VQA, there can be dominance between modalities in some cases while in other cases the modalities are contributing equally.

We summarize a categorization of multimodal interactions in Table 1, building upon an initial taxonomy by Partan and Marler [2005] (denoted with a star). We define new terms in the taxonomy defined above:

- Destructive: Two modalities can be considered to be destructive if the presence of one diminishes the signal present in another. For example, if we combine an image with noise, the noise destroys the original signal.
- Logical. Logical interactions between two modalities can occur if one signal undertakes, contradicts, or contains another signal. These logical interactions can also include expressions such as NOT, AND, OR, NOR, XOR, XNOR, etc. For example, two signals may represent the logical signal AND if the full information content is only present when both signals are present.
- Temporal: Temporal interactions are characterized by the order and rate of different modalities appearing as input to a system. For example, neurons fire at different rates, which affects the encoded signal. Similarly, repeating the same action over time may have a magnifying effect on the output.

Human interactions often convey multimodal interactions [Ruiz et al., 2006]. For example, humor is a multimodal signal [Tsakona, 2009] that leverages the multimodal interactions described above. In the context of memes (e.g humorous text-image pairs), we describe three interactions.

- Emergence: The text description may be humorous and have a particular meaning. The image may be humorous and have a particular meaning. The meaning of the text-image pair may be entirely different than either the meaning of the image or text when examined together. For example, a contradicting text-image pair is considered humorous.
- Enhancement: The text description may be humorous, and the image may be humorous. When combined, the humor is greater than that provided by each individual component.

We can further divide these interactions into categories based on how the interactions happen in the model. For example, it can happen through attention layers, additive layers, and multiplicative layers. Different mechanisms can capture different interactions. Interaction can also be classified by their co-occurrence. Two features from two different modalities can either always co-occur or sometimes co-occur. Another way is to classify by the output dependence on input features, which is whether the output of the model is dependent on one modality or both of them.

Interactions can be implicit or explicit. Explicit interactions come out when both modalities are presented as input while retaining. For an image-to-text generation task, more implicit interactions are revealed. Through training, a shared representation space will be created and reasoning could happen in that latent space, which causes implicit interaction.

2 Identifying Interactions

We identify four methods for identifying multimodal interactions within our data. Importantly, we hypothesize that the interactions we care about should be task dependent. Assuming we can identify the multimodal interactions, we want to remove the independent, redundant, or destructive components. Ideally, we want to combine data with modulation and emergent behavior as this makes the most use of the multimodal input.

- Analyzing model predictions: Given two input signals, we can measure the relative interaction and importance of these signals by first evaluating the output with each signal independently, and then again with these signals combined. If the prediction is more “correct” (according to some ground truth label) using one modality instead of both, we know that one of the input signals is destructive. Similarly, if the prediction is unchanged using two modalities instead of one, we know that the second modality provides redundant information.
- Analysis by disagreement: We can evaluate whether two modalities share a significant amount of redundant information by asking human annotators to predict the content of one modality given another. We can compare the human prediction with the ground truth. If the human prediction and the ground truth are very similar, this means that the first modality is enough to make a prediction. If the human prediction and the second modality disagree significantly (as rated by a third human), we know the two modalities have little shared information.
- Unsupervised clustering: We can decompose the data set into features to analyze the PCA and tSNE projections using a pre-trained model. We hypothesize that this will uncover clusters of data that hint at the underlying multimodal interaction.

3 Capturing Interactions

One possible improvement is to add modules to quantify if interaction occurs and define the type of interaction. For example, using an adversarial system that classifies interaction between modalities (redundant/synergistic) while maximizing mutual information could potentially increase the performance by learning to select a signal based on what type of interaction it thinks it should be. Another possible direction is to regularize the interaction by using the cost function to attach weights to cross-modal interactions.

References

R M Baron and David A. Kenny. The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *Journal of personality and social psychology*, 51 6: 1173–82, 1986.

Jack Hessel and Lillian Lee. 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 (EMNLP)*, pages 861–877, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.62. URL https://aclanthology.org/2020.emnlp-main.62.

Aleks Jakulin and Ivan Bratko. Quantifying and visualizing attribute interactions, 2003. URL https://arxiv.org/abs/cs/0308002.

Carey Ed Jewitt. *The Routledge handbook of multimodal analysis.* Routledge/Taylor & Francis Group, 2011.

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

Sarah R. Partan and Peter Marler. Issues in the classification of multimodal communication signals. *The American Naturalist*, 166(2):231–245, 2005. doi: 10.1086/431246.

Natalie Ruiz, Ronnie Taib, and Fang Chen. Examining the redundancy of multimodal input. In *Proceedings of the 18th Australia Conference on Computer-Human Interaction: Design: Activities, Artefacts and Environments*, OZCHI '06, page 389–392, New York, NY, USA, 2006. Association for Computing Machinery. ISBN 1595935452. doi: 10.1145/1228175.1228254. URL https://doi.org/10.1145/1228175.1228254.

Villy Tsakona. Language and image interaction in cartoons: Towards a multimodal theory of humor. *Journal of Pragmatics*, 41(6):1171–1188, 2009. ISSN 0378-2166. doi: https://doi.org/10.1016/j.pragma.2008.12.003.

Michael Tsang, Sirisha Rambhatla, and Yan Liu. How does this interaction affect me? interpretable attribution for feature interactions, 2020. URL https://arxiv.org/abs/2006.10965.

Paul L. Williams and Randall D. Beer. Nonnegative decomposition of multivariate information. *CoRR*, abs/1004.2515, 2010. URL http://arxiv.org/abs/1004.2515.