11-877 Advanced Multimodal Machine Learning **Spring 2024**

Week 4: Modality Interactions

Instructors: Paul Liang and Daniel Fried *Synopsis Leads: Ashwin Pillay*

Edited by Paul Liang *Scribes: Jiya Zhang and Ashwin Pillay*

Follow the rest of the class here: `https://cmu-multicomp-lab.github.io/adv-mmml-course/spring2024/`

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 attempted to formalize a taxonomy of semantic multimodal interactions, compared connections with interactions, explored how to measure these interactions, and discussed integrating Large Language Models (LLMs) to improve understanding of cross-modal interactions. The following was a list of provided research probes:

1. 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?
2. 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?
3. 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?
4. 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 interactions given a dataset?
5. 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?
6. Can you think of ways to utilize large language models or other foundation models to enhance the learning process of multimodal interactions?
7. How to utilize cognitive theory to design a framework that can be used to understand and learn the interactions between multiple modalities that human beings face everyday?

As background, students read the following papers:

1. **(Required)** Training Vision-Language Transformers from Captions [Gui et al., 2023]
2. **(Required)** Ten Myths of Multimodal Interaction [Oviatt, 1999]
3. (Suggested) A Vision Check-up for Language Models [Sharma et al., 2024]
4. (Suggested) Scaling Vision-Language Models with Sparse Mixture of Experts [Shen et al., 2023]
5. (Suggested) Quantifying & Modeling Multimodal Interactions: An Information Decomposition Framework [Liang et al., 2023]
6. (Suggested) Does my multimodal model learn cross-modal interactions? It's harder to tell than you might think! [Hessel and Lee, 2020]
7. (Suggested) Multimodal interaction: A review [Turk, 2014]
8. (Suggested) When Do We Interact Multimodally? Cognitive Load and Multimodal Communication Patterns [Oviatt et al., 2004]

9. (Suggested) A multimodal parallel architecture: A cognitive framework for multimodal interactions [Cohn, 2016]
10. (Suggested) Quantifying and Visualizing Attribute Interactions [Jakulin and Bratko, 2004]
11. (Suggested) The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations [Baron and Kenny, 1986]

We summarize several main takeaway messages from group discussions below:

1 Taxonomy of Semantic Cross-Modal Interactions

While multimodal interactions can be analyzed from both statistical and semantic perspectives, our discussions focus on the latter owing to a larger degree of underlying ambiguity. Table 1 presents some of the dimensions of semantic interactions we analyzed:

Table 1: Dimensions of Interactions between Modalities

Semantic Interaction	Description
Complementary	When modalities by themselves cannot provide all the information they impart post interacting with each other. Examples include: • How text and images interact to deliver a holistic message in comic books. • As discussed in [Oviatt, 1999]: Screen control using speech v/s a combination of mouse-pointing and speech.
Redundant	When modalities interact with each other to deliver information already provided by one of the modalities (here, the redundancy is expressed in terms of the information imparted).
Hierarchical	When interactions are established by a hierarchical combination of relationships. [Otto et al., 2020] described such interactions between the image and text modalities.
Dominance	When one modality dominates over the other during the interaction, even when they contribute different information. An example of such interactions are seen in ViLT (text must adhere to image) as observed by [Gui et al., 2023].
Conditional Existence	When one modality can exist only if the other exists as well, even when they contribute non-redundant information. For example, vibration and sound data from the analysis of construction faults in buildings.

2 Comparing Multimodal Connections and Interactions

1. Connections are inter-modal relations that inherently exist within the dataset. On the other hand, interactions involve learning task-relevant relations between this data. Establishing connections between the modalities involved can be seen as a precursor to establishing interactions.
2. However, it is not impossible for models to learn cross-modal interactions even when the modalities are independent (ie, not connected). A powerful model can learn interactions by force if unconnected or weakly-connected modality representations are simply fused together and provided to them. [Gui et al., 2023] makes this observation during their analysis of systems like ViLT[Kim et al., 2021] and PixelBERT[Huang et al., 2020]. Interactions thus learnt are observed to not be useful for a conceptual understanding of the provided data.
3. Usually, the pre-training step of many foundational models involves establishing available connections in the data. Subsequently, task dependent fine-tuning involves preserving the task-specific interactions. In this regard, some steps to consider would be:
   - Identifying and preserving the set of interactions that are task-specific.
   - Identifying possibilities of non task-relevant connections C_1 and C_2 combining to form a task-

relevant interaction I_3.

3 Measuring and Estimating Interactions

Depending on the nature of the task and the type of data available, we highlight possible ways to measure cross-modal interactions as follows:

1. When data labels are available for the given task (for example, classification tasks):
   (a) Contrastive learning techniques may serve as simple strategies to determine how closely modalities can possibly interact; this can be done by training them on positive examples only and evaluating how close the resulting projections are to each other.
      - However, the effectiveness of contrastive learning could be task-dependent: as shown by [Gui et al., 2023], VLC with its non-contrastive cross-attention mechanism and task-specific pre-training (image-text matching) outperformed systems like CLIP.

   (b) Alternately, other approaches towards estimating interactions may include:
      - Using partial information decomposition techniques.
      - Establishing hierarchical interactions as done in [Otto et al., 2020], and extending them to a multi-label, multi-class classification objective.
      - Directly evaluating multimodal performance with unimodal models of similar tasks for each concerned modality.
      - Developing projections that can be compared directly via techniques like additive composition.
2. When data labels are not available for the given task (for example, generative modelling):
   (a) Attention map analysis on within and out-of-domain data can serve as a good indicator of the kinds of interactions learnt.
   (b) Selective deactivation of modalities; for example, given a visual understanding task, comparing the model outputs for the following inputs:
      - Relevant text combined with the image.
      - Empty text string combined with the image.
3. When the data does not have a one-to-one mapping (for example, sarcasm identification in image-text data like comic books): contrastive learning on a smaller model could be a useful option to derive generalized interactions between underlying abstract structures.

4 Incorporating LLMs into Learning Cross-Modal Interactions

Considering rapid progress in the capabilities of LLMs towards understanding and reasoning, we identify the following ways in integrating them into the study and development of cross-modal interactions. LLMs can serve as:

1. *Reward predictors* to rate how well a model captures semantic interactions, and accordingly guide the model to enhance its performance.
2. *Agents* to generate code that can specifically implement certain statistical interactions.
3. *Evaluators* to critique the model's learnt interactions. For more abstract interactions, this may be achieved via few-shot prompting.
4. Synthetic data *generators*; for example, generating a large volume of image-caption pairs with synergistic relationships and captions that are correspondingly descriptive.

References

Reuben Baron and David Kenny. The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51: 1173–1182, 01 1986. doi: 10.1037//0022-3514.51.6.1173.

Neil Cohn. A multimodal parallel architecture: A cognitive framework for multimodal interactions. *Cognition*, 146:304–323, 2016. ISSN 0010-0277. doi: https://doi.org/10.1016/j.cognition.2015.10.007. URL https://www.sciencedirect.com/science/article/pii/S0010027715300858.

Liangke Gui, Yingshan Chang, Qiuyuan Huang, Subhojit Som, Alex Hauptmann, Jianfeng Gao, and Yonatan Bisk. Training vision-language transformers from captions, 2023.

Jack Hessel and Lillian Lee. Does my multimodal model learn cross-modal interactions? it's harder to tell than you might think! In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *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.

Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers, 2020.

Aleks Jakulin and Ivan Bratko. Quantifying and visualizing attribute interactions, 2004.

Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision, 2021.

Paul Pu Liang, Yun Cheng, Xiang Fan, Chun Kai Ling, Suzanne Nie, Richard Chen, Zihao Deng, Nicholas Allen, Randy Auerbach, Faisal Mahmood, Ruslan Salakhutdinov, and Louis-Philippe Morency. Quantifying modeling multimodal interactions: An information decomposition framework, 2023.

Christian Otto, Matthias Springstein, Avishek Anand, and Ralph Ewerth. Characterization and classification of semantic image-text relations. *International Journal of Multimedia Information Retrieval*, 9(1):31–45, March 2020. ISSN 2192-662X. doi: 10.1007/s13735-019-00187-6. URL https://doi.org/10.1007/s13735-019-00187-6.

Sharon Oviatt. Ten myths of multimodal interaction. *Commun. ACM*, 42(11):74–81, nov 1999. ISSN 0001-0782. doi: 10.1145/319382.319398. URL https://doi.org/10.1145/319382.319398.

Sharon Oviatt, Rachel Coulston, and Rebecca Lunsford. When do we interact multimodally? cognitive load and multimodal communication patterns. In *Proceedings of the 6th International Conference on Multimodal Interfaces*, ICMI '04, page 129–136, New York, NY, USA, 2004. Association for Computing Machinery. ISBN 1581139950. doi: 10.1145/1027933.1027957. URL https://doi.org/10.1145/1027933.1027957.

Pratyusha Sharma, Tamar Rott Shaham, Manel Baradad, Stephanie Fu, Adrian Rodriguez-Munoz, Shivam Duggal, Phillip Isola, and Antonio Torralba. A vision check-up for language models, 2024.

Sheng Shen, Zhewei Yao, Chunyuan Li, Trevor Darrell, Kurt Keutzer, and Yuxiong He. Scaling vision-language models with sparse mixture of experts, 2023.

Matthew Turk. Multimodal interaction: A review. *Pattern Recognition Letters*, 36:189–195, 2014. ISSN 0167-8655. doi: https://doi.org/10.1016/j.patrec.2013.07.003. URL https://www.sciencedirect.com/science/article/pii/S0167865513002584.