11-877 Advanced Multimodal Machine Learning **Spring 2022**

Week 2: Cross-Modal Interactions

Instructors: L.-P. Morency, A. Zadeh, P. Liang *Synopsis Leads: Zhe Chen, Yuchen Xu*

Edited by Paul Liang *Scribes: Kelly Shi, David Lin*

Follow the rest of the class here: https://cmu-multicomp-lab.github.io/adv-mmml-course/spring2022/

Summary: Multimodal machine learning is the study of computer algorithms that learn and improve through the use and experience of multimodal data. In week 2's discussion session, the class aimed to formalize a taxonomy of cross-modal interactions: various ways in which elements from different modalities can relate with each other and the types of new information possibly discovered as a result of these relationships. The following was a list of provided research probes:

1. What are the different ways in which modalities can interact with each other in multimodal tasks? Can we formalize a taxonomy of such cross-modal interactions, which will enable us to compare and contrast them more precisely?
2. What are the design decisions (aka inductive biases) that can be used when modeling these cross-modal interactions in machine learning models?
3. 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, and so on.
4. Given an arbitrary dataset and prediction task, how can we systematically decide what type of cross-modal interactions exist, and how can that inform our modeling decisions?
5. Given trained multimodal models, how can we understand or visualize the nature of cross-modal interactions?

As background, students read the following papers:

1. (Required) Additive interactions: Does my multimodal model learn cross-modal interactions? It's harder to tell than you might think! [Hessel and Lee, 2020]
2. (Required) Grounding interactions: What Does BERT with Vision Look At? [Li et al., 2020]
3. (Suggested) Multiplicative interactions: Multiplicative Interactions and Where to Find Them [Jayakumar et al., 2019]
4. (Suggested) Cooperative interactions: Cooperative Learning for Multi-view Analysis [Ding and Tibshirani, 2021]
5. (Suggested) Visualizations and ablations: Vision-and-Language or Vision-for-Language? On Cross-Modal Influence in Multimodal Transformers [Frank et al., 2021]
6. (Suggested) Visualizations and ablations: Seeing past words: Testing the cross-modal capabilities of pretrained V&L models on counting tasks [Parcabalescu et al., 2020]

We summarize several main takeaway messages from group discussions below:

1 Taxonomy of Cross-Modal Interactions

In Table 1, we summarize an initial taxonomy of cross-modal interactions. These interactions depend on the input modalities, model of choice, and task at hand. Furthermore, each category may not be mutually exclusive (e.g., additive interactions are a special case of multiplicative ones) and this taxonomy is not exhaustive. Future work should further refine and formalize these preliminary definitions, with the eventual

Table 1: An initial taxonomy of cross-modal interactions.

Type	Definition	Reference
Additive	addition of unimodal predictions/representations	[Hessel and Lee, 2020]
Multiplicative	higher-order interactions (vector, matrix, tensor products) between unimodal representations	[Jayakumar et al., 2019]
Cooperative	agreement between unimodal predictions/representations	[Ding and Tibshirani, 2021]
Grounding	matching semantic information between unimodal sub-units	[Li et al., 2020]
Equivalence	modalities produce the same information such that one can replace the others	[Martin, 2002]
Specialization	modalities produce unique information not contained in others	[Martin, 2002]
Transfer	using information from one modality as input to another	[Martin, 2002]
Complementarity	combining information to enhance or emphasize the qualities of each other	[Martin, 2002]

goal of summarizing the inductive biases suitable for capturing each type of cross-modal interaction, while considering the advantages and drawbacks of each design decision (tradeoffs in prediction performance, time/space complexity, interpretability, and so on).

2 Measuring/visualizing cross-modal interactions in data

Given an arbitrary dataset and prediction task, how can we systematically decide what type of cross-modal interactions exist, and how can that inform our modeling decisions?

Generally, human annotation can help us understand the nature of cross-modal interactions in multimodal tasks. We may be able to ask annotators how they make specific annotation decisions (how they think), which can tell us about modality importance and cross-modal interactions used in the data. However, it may be hard for humans to verbalize their thinking process. Since datasets are typically created with humans in the loop, we can ask annotators to make the minimal modifications to each modality to result in a different interaction/prediction, which gives insights on the nature of cross-modal interactions in data.

Connecting to the study of human reasoning in neuroscience, we can perform a brain scan (e.g., fMRI) to observe neural activity while annotators complete the tasks, which could shed light on how humans are integrating information from different modalities. However, as a caveat, the workings of the human brain are still very poorly understood. In contrast to the human brain where vision occupies the most space, language (text) is currently the dominant modality in multimodal datasets.

Furthermore, it is typically also a good idea to also start with strong unimodal baselines on multimodal tasks to check for biases in the task and whether cross-modal interactions are needed in the first place.

3 Measuring/visualizing cross-modal interactions in models

Given trained multimodal models, how can we understand or visualize the nature of cross-modal interactions? The class came up with several ideas:

3.1 Extending EMAP

EMAP [Hessel and Lee, 2020] was proposed to decompose black-box multimodal models into the closest additive model approximating the original model's predictions. EMAP could be extended to detect cross-modal interactions: since the original approach measures the difference in output logits between unimodal components and multimodal components, the difference between these 2 logits contains the cross-modal information. We may be able to to cluster these vectors to obtain a categorization of different cross-modal interactions that exist in the original data, and visualize the corresponding original datapoints to identify the type of cross-modal interactions they belong to.

3.2 Studying model-specific attention heads

Existing models like Multimodal Transformers [Vaswani et al., 2017, Li et al., 2019, Tsai et al., 2019] have the capacity of capturing arbitrary types of interactions using self-attention mechanisms between 2 modality sequences. We can treat attention heads as a full matrix and use pruning to discover the most

informative attention weights between modality sub-units which summarizes additive/multiplicative/grounding interactions, in a manner similar to Li et al. [2020]. Another idea is to discretize the interaction space (e.g. [Li et al., 2020]) which can help us in further visualizations.

3.3 Data manipulation

Some datasets can be solved without modeling cross-modal interactions, while others require models to do so [Hessel and Lee, 2020, Liang et al., 2021]. With this insight, we can try to create focused datasets capturing specific types of interactions as a benchmark for whether models indeed discover them. These datasets can be designed adversarially, such as ablating the input data and measuring the impact on model outputs which can help visualize the nature of cross-modal interactions (e.g., modality importance and symmetry). Ablating the input data also has nice interpretations in causal and counterfactual reasoning as a step towards measuring robustness in multimodal models. Inspired by dynamic benchmarks in NLP [Kiela et al., 2021], we can add noise to (perturb) the data and see how humans respond in the same task. As a further step, we can see the minimal amount of perturbation required to push the model back to random to study cross-modal interactions.

References

Daisy Yi Ding and Robert Tibshirani. Cooperative learning for multi-view analysis. *arXiv preprint arXiv:2112.12337*, 2021.

Stella Frank, Emanuele Bugliarello, and Desmond Elliott. Vision-and-language or vision-for-language? on cross-modal influence in multimodal transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9847–9857, 2021.

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, 2020.

Siddhant M Jayakumar, Wojciech M Czarnecki, Jacob Menick, Jonathan Schwarz, Jack Rae, Simon Osindero, Yee Whye Teh, Tim Harley, and Razvan Pascanu. Multiplicative interactions and where to find them. In *International Conference on Learning Representations*, 2019.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, et al. Dynabench: Rethinking benchmarking in nlp. In *NAACL-HLT*, 2021.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. What does bert with vision look at? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5265–5275, 2020.

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.

Jean-Claude Martin. On the use of the multimodal clues in human behaviour for the modelling of agent co-operative behaviour. *Connection Science*, 14(4):297–309, 2002.

L Parcabalescu, A Gatt, A Frank, I Calixto, Juliette Faille, Albert Gatt, Claire Gardent, Lucie Gianola, Ēriks Ajausks, Victoria Arranz, et al. Seeing past words: Testing the cross-modal capabilities of pretrained v&l models on counting tasks. In *Proceedings of the Workshop Beyond Language: Multimodal Semantic Representations (MMSR'21)*, volume 67, pages 398–411. Association for Computational Lingustics, 2020.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017.