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

Week 5: Multimodal Reasoning

Instructors: L.-P. Morency, A. Zadeh, P. Liang *Synopsis Leads: Yuanxin Wang, Dong Won Lee*

Edited by Paul Liang *Scribes: Kelly Shi, Hyukjae Kwark*

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/spring2022/

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 5's discussion session, the class aimed to define a taxonomy of the main processes in multimodal reasoning and the key technical challenges that arise from reasoning-based approaches. Specifically, the class were asked to think about how multimodal reasoning (1) affects how we model cross-modal interactions, (2) can be embedded in a machine learning system, or (3) studied in a large-scale pre-training models via post-hoc methods. The following was a list of provided research probes:

1. What are the various reasoning processes required in multimodal tasks, where data comes from heterogeneous sources? What could be a taxonomy of the main processes involved in multimodal reasoning?
2. Are there unique technical challenges that arise because reasoning is performed on multimodal data? What are these unique challenges? How can we start studying these challenges in future research?
3. How should we model cross-modal interactions when performing reasoning over multimodal data? Grounding words with visual objects could be an example of a reasoning step required with multimodal data. Other reasoning involved in modeling the different types of cross-modal interactions (e.g., additive, multiplicative)?
4. What are the main advantages of reasoning-based approaches, when compared to the large-scale pre-training methods discussed last week? What are the potential issues with reasoning? Can we perform reasoning on very large datasets? Can pre-training methods eventually learn reasoning processes similar to humans? Or will we still need human and domain knowledge to some extent?
5. Can you imagine a way to uncover the reasoning capabilities of black-box model, such as a large-scale pre-trained model? How can one discover specifically the cross-modal reasoning processes in such a black-box model?
6. To what extent do we need external knowledge when performing reasoning, specifically multimodal reasoning? What type of external knowledge is likely to be needed to succeed in multimodal reasoning?

As background, students read the following papers:

1. (Required) Clevrer: Collision events for video representation and reasoning [Yi et al., 2019]
2. (Required) Neuro-Symbolic Visual Reasoning: Disentangling "Visual" from "Reasoning" [Amizadeh et al., 2020]
3. (Suggested) Learning to Compose and Reason with Language Tree Structures for Visual Grounding [Hong et al., 2019]
4. (Suggested) Heterogeneous Graph Learning for Visual Commonsense Reasoning [Yu et al., 2019]
5. (Suggested) Multimodal Logical Inference System for Visual-Textual Entailment [Suzuki et al., 2019]
6. (Suggested) A Closer Look at the Robustness of Vision-and-Language Pre-trained Models [Li et al., 2020a]
7. (Suggested) CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual

Reasoning [Johnson et al., 2017]

8. (Suggested) VQA-LOL: Visual Question Answering under the Lens of Logic [Gokhale et al., 2020]
9. (Suggested) Deep Compositional Question Answering with Neural Module Networks [Andreas et al., 2016]
10. (Suggested) Towards Causal VQA: Revealing and Reducing Spurious Correlations by Invariant and Covariant Semantic Editing [Agarwal et al., 2020]
11. (Suggested) Building a Large-scale Multimodal Knowledge Base System for Answering Visual Queries [Zhu et al., 2015]
12. (Suggested) KAT: A Knowledge Augmented Transformer for Vision-and-Language [Gui et al., 2021]

We summarize several main takeaway messages from group discussions below:

1 Definition and Taxonomy of Reasoning

Reasoning in either unimodal or multimodal tasks can be defined as follows: given a set of concepts expressed through atomic evidences, reasoning defines a process where atomic evidences are composed, possibly via multiple hierarchical steps, to form more abstract concepts useful for making high-level predictions. The reasoning process can be classified along (but not limited to) the following different dimensions:

- What constitutes an atomic evidence and what constitutes a step in reasoning? These are often defined by the user designing a model, typically through the selection of a data structure to represent the reasoning process. Different choices of data structures represent different parameterizations of atoms and relations. Commonly used data structures include logical operators [Gokhale et al., 2020, Amizadeh et al., 2020] (each step is a logical operation such as AND/OR), trees [Hong et al., 2019] and graphs [Zhu et al., 2015, Yu et al., 2019] (each step is an edge between nodes), or modules [Andreas et al., 2016].
- Single-step vs multi-step reasoning: for example, when reasoning through a knowledge graph, how many edges/facts should be traversed in order to answer a question?
- Deterministic vs probabilistic reasoning: deterministic reasoning refers to the paradigm where a fixed model is used to reason based on multiple facts while probabilistic reasoning takes a different path on coming to a conclusion based on the frequency or probability of a set of known facts.
- Temporal-based vs non-temporal reasoning: temporal reasoning could involve reasoning over atomic evidences distributed across long-range sequences of video frames while non-temporal reasoning is limited to atomic evidences within static frames.

2 Cross-Modal Interactions

Reasoning requires an initial set of atomic evidences. In multimodal settings, these atomic evidences typically include both unimodal evidences as well as multimodal correspondences. The latter requires one to model cross-modal interactions: 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 (e.g., additive [Hessel and Lee, 2020], multiplicative [Jayakumar et al., 2019]). Little work has been done in exploring the construction of cross-modal interactions as part of reasoning processes, which leaves avenues for future exploration.

3 How do Pretrained Models Reason?

Large-scale pretrained models are known for learning powerful representations despite not explicitly modeling the reasoning process. While powerful, the lack of a reasoning process causes them to be less interpretable as compared to models integrating symbolic reasoning processes. Furthermore, pretrained models could be less robust by picking up on unimodal biases [Hessel and Lee, 2020], which results from not explicitly modeling the reasoning process.

Given their black-box nature, a natural question to ask is: how can we uncover the reasoning capabilities of a pretrained model? In Table 1, we summarize several initial methods to answer this question.

Method	Description
Latent space visualization [Radford et al., 2021, Itkina et al., 2020]	In a joint vision and language embedding space, visualize whether the visual objects and corresponding text phrases are close to each other.
Attention head inspection [Huang et al., 2019, Li et al., 2020b]	For transformer based models, observe the attention scores in each head between a pre-selected keyword in text and a pre-selected object patch in image.
Stress-testing [Ma et al., 2021, Naik et al., 2018]	Inspired by how the perturbation of inputs (e.g., negation, random order) affects the performance of NLI models, we can also perturb the input to the multimodal reasoning model and observe the changes in output. Similarly, we can completely remove one modality during training and see if the model is still able to reason. We can also add counterfactual inputs and see if the model is reasoning based on its performance.
Perception system degradation [Amizadeh et al., 2020]	Sometimes we can deliberately apply a deficient perception system (e.g., a not well-performing ResNet-50 model) and see if the model is still able to reason well.
Reasoning paths [Hong et al., 2022, Yu et al., 2019, Wei et al., 2022]	Is it possible to generate a series of short sentences that mimic the reasoning process a person might have when responding to a question? Or can we construct a graph/tree architecture to understand the reasoning paths?

Table 1: Attempts in understanding the reasoning capabilities of pretrained models.

4 Perception vs Reasoning

There is a need to design perception systems specifically catered to the task of visual reasoning, via embedding inductive biases that consider reasoning during the perception stage. For example, in the case of VQA, as we (humans) answer questions, the attention map of the visual input should change across time. In current literature, there are end-to-end neural models that incorporate attention structures and object relations [Hudson and Manning, 2018, Santoro et al., 2017, Hu et al., 2017]. Similarly, Chen et al. [2021] attempts to move closer to human visual reasoning by modeling a sequence of atomic regions-of-interest and using human eye-tracking data to learn compositions of these atomic evidences.

5 Insights from Human Reasoning

A way to coordinate reasoning modules is through perceiving it as a “Shared Global Workspace”, an idea that originated from cognitive science [Baars, 2017]. Global Workspace Theory (GWT) can be perceived as the mind’s stage, where conscious contents (actors) are brightly lit up (attended to). The rest of the stage is dark (audience watching the play), which represents unconscious contents. In neural networks, interactions between different elements or contents are typically modelled via pairwise interactions. Transformers make use of self-attention to capture interactions from elements in other positions, object-centric architectures make use of graph neural networks to model interactions between entities. Meta-architectures such as those inspired by the GWT or other computational models of human reasoning could also present a path towards more accurate and interpretable reasoning models [Goyal et al., 2021, Blum and Blum, 2021].

6 Leveraging External Knowledge

There are several directions in which external knowledge bases could be used to augment existing reasoning models: (1) In low-resource settings, where the model does not have enough data to reason under a specific domain, specialized knowledge bases would be helpful in facilitating the reasoning process. (2) Common sense knowledge bases could also be helpful in imparting world knowledge to reasoning models.

References

Vedika Agarwal, Rakshith Shetty, and Mario Fritz. Towards causal vqa: Revealing and reducing spurious correlations by invariant and covariant semantic editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9690–9698, 2020.

Saeed Amizadeh, Hamid Palangi, Oleksandr Polozov, Yichen Huang, and Kazuhito Koishida. Neuro-symbolic visual reasoning: Disentangling ”visual” from ”reasoning”. *ArXiv*, abs/2006.11524, 2020.

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 39–48, 2016.

Bernard J Baars. The global workspace theory of consciousness: Predictions and results. *The blackwell companion to consciousness*, pages 227–242, 2017.

Manuel Blum and Lenore Blum. A theoretical computer science perspective on consciousness. *Journal of Artificial Intelligence and Consciousness*, 8(01):1–42, 2021.

Shi Chen, Ming Jiang, Jinhui Yang, and Qi Zhao. Attention in reasoning: Dataset, analysis, and modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. Vqa-lol: Visual question answering under the lens of logic. In *European conference on computer vision*, pages 379–396. Springer, 2020.

Anirudh Goyal, Aniket Didolkar, Alex Lamb, Kartikeya Badola, Nan Rosemary Ke, Nasim Rahaman, Jonathan Binas, Charles Blundell, Michael Mozer, and Yoshua Bengio. Coordination among neural modules through a shared global workspace. *arXiv preprint arXiv:2103.01197*, 2021.

Liangke Gui, Borui Wang, Qiuyuan Huang, Alex Hauptmann, Yonatan Bisk, and Jianfeng Gao. Kat: A knowledge augmented transformer for vision-and-language. *arXiv preprint arXiv:2112.08614*, 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.

Richang Hong, Daqing Liu, Xiaoyu Mo, Xiangnan He, and Hanwang Zhang. Learning to compose and reason with language tree structures for visual grounding. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

Richang Hong, Daqing Liu, Xiaoyu Mo, Xiangnan He, and Hanwang Zhang. Learning to compose and reason with language tree structures for visual grounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44:684–696, 2022.

Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Kate Saenko. Learning to reason: End-to-end module networks for visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 804–813, 2017.

Po-Yao Huang, Xiaojun Chang, and Alexander Hauptmann. Multi-head attention with diversity for learning grounded multilingual multimodal representations. In *EMNLP*, 2019.

Drew A Hudson and Christopher D Manning. Compositional attention networks for machine reasoning. *arXiv preprint arXiv:1803.03067*, 2018.

Masha Itkina, B. Ivanovic, Ransalu Senanayake, Mykel J. Kochenderfer, and Marco Pavone. Evidential sparsification of multimodal latent spaces in conditional variational autoencoders. *ArXiv*, abs/2010.09164, 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.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017.

Linjie Li, Zhe Gan, and Jingjing Liu. A closer look at the robustness of vision-and-language pre-trained models. *arXiv preprint arXiv:2012.08673*, 2020a.

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, 2020b.

Mengmeng Ma, Jian Ren, Long Zhao, S. Tulyakov, Cathy Wu, and Xi Peng. Smil: Multimodal learning with severely missing modality. *ArXiv*, abs/2103.05677, 2021.

Aakanksha Naik, Abhilasha Ravichander, Norman M. Sadeh, Carolyn Penstein Rosé, and Graham Neubig. Stress test evaluation for natural language inference. In *COLING*, 2018.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. A simple neural network module for relational reasoning. *Advances in neural information processing systems*, 30, 2017.

Riko Suzuki, Hitomi Yanaka, Masashi Yoshikawa, Koji Mineshima, and Daisuke Bekki. Multimodal logical inference system for visual-textual entailment. *arXiv preprint arXiv:1906.03952*, 2019.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *ArXiv*, abs/2201.11903, 2022.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. Clevrer: Collision events for video representation and reasoning. *arXiv preprint arXiv:1910.01442*, 2019.

Weijiang Yu, Jingwen Zhou, Weihao Yu, Xiaodan Liang, and Nong Xiao. Heterogeneous graph learning for visual commonsense reasoning. In *NeurIPS*, 2019.

Yuke Zhu, Ce Zhang, Christopher Ré, and Li Fei-Fei. Building a large-scale multimodal knowledge base system for answering visual queries. *arXiv preprint arXiv:1507.05670*, 2015.