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

# Week 7: Multimodal Reasoning

*Instructors: Louis-Philippe Morency and Paul Liang* *Synopsis Leads: Sean Chang and Pratik Joshi*

*Edited by Paul Liang* *Scribes: Yiqing Xie and Haofei Yu*



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

**Summary:** Multimodal reasoning is a very relevant research direction in today's world, looking at how models can reason about the world in an accurate and interpretable manner.

In week 7's discussion session, the class aimed to understand **multimodal reasoning** and recent works around it. The following was a list of provided research probes:

1. What is reasoning, and what are its subchallenges? What could be a taxonomy of the main processes involved in reasoning? What are some potential formal definitions for reasoning?
2. Are there unique technical challenges that arise because reasoning is performed on multimodal versus unimodal data? How can we start studying these challenges in future research? Try to link it back to our previous definition of heterogeneity, connections, and interactions when thinking about multimodal reasoning challenges.
3. 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-based methods? Can we come up with a research agenda that combines the best of both worlds?
4. 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. What are some ways to uncover the reasoning capabilities of multimodal models? What additional techniques do we need over measuring reasoning of unimodal models?
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 (REQ = required, SUG = suggested, REL = relevant):

1. (REQ) Multimodal Chain-of-Thought Reasoning in Language Models [Zhang et al., 2023]
2. (REQ) What Can Neural Networks Reason About? [Xu et al., 2020]
3. (SUG) The Curious Case of Commonsense Intelligence
4. (SUG) Towards Reasoning in Large Language Models: A Survey [Huang and Chang, 2022]
5. (SUG) Multimodal Analogical Reasoning over Knowledge Graphs [Zhang et al., 2022]
6. (SUG) Generalization Differences between End-to-End and Neuro-Symbolic Vision-Language Reasoning Systems [Zhu et al., 2022]
7. (REL) Socratic Models: Composing Zero-Shot Multimodal Reasoning with Language [Zeng et al., 2023]
8. (REL) Winoground: Probing Vision and Language Models for Visio-Linguistic Compositionality [Thrush et al., 2022]
9. (REL) Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering [Lu et al., 2022]
10. (REL) Introductory Tutorial: Commonsense Reasoning for Natural Language Processing [Sap et al., 2020]
11. (REL) WebQA: A Multimodal Multihop NeurIPS Challenge [Chang and Bisk, 2022]

We summarize several main takeaway messages from group discussions below:

# 1 What is reasoning?

One definition that we came up with is that reasoning is following a logic chain to arrive at a conclusion. When we talk about reasoning in models, we focus on the model's ability to show us reasoning, which mainly comes in the form of how they arrived at the answer. There are multiple ways to represent the model's reasoning process, and it is critical to have finegrained metrics beyond downstream task accuracy to measure performance over components/lines of reasoning. Other than the standard chain of thought, future research could look at more structured representations such as code or a graph. These structured representations could potentially make evaluation easier as well.

Another direction is to create a formal taxonomy for reasoning. The aim would be to understand and reason within the information that is available, and not categories not present to our understanding or in the data. The taxonomy should ideally contribute to several over-arching subchallenges of reasoning, such as guarantee of consistent reasoning, and sufficient evidence. For example, in math proving and reasoning tasks [Koncel-Kedziorski et al., 2016], reasoning should ensure that lemmas are used accurately to contribute to the solving. The taxonomy should also be composably used over multi-hop inferences.

# 2 Challenges in Multimodal Reasoning

We identified the following as some of the challenges in multimodal reasoning

- Model and training design: LLMs have been shown to be good at natural language, but there is no architecture that is known to be the best for multimodal learning. It is also necessary to consider things such as early fusion vs late fusion. One training paradigm to consider is reinforcement learning (RL). A strong advantage of RL is the exploration phase, where, without explicit supervision but only defined rewards, the model learns to come to a solution. This could allow steps of reasoning to be applied more naturally and with more guarantee than supervision.
- Evaluation: It's difficult to tell which modalities are necessary for reasoning, and the reasoning path may not be able to be shown naturally for some modalities. It is also not clear how we can evaluate a model's reasoning process. Model-free scores such as BLEU can unfairly penalize correct processes while model-based scores can induce an extra layer of bias.
- Dataset: It's difficult to make multimodal reasoning datasets at scale, especially if we want N-way aligned data. One way to address this is socratic models [Zeng et al., 2023], which efficiently combine pairwise multimodal models to perform reasoning on a wide range of tasks. They use the language model as a semantic "core", to link outputs from different pretrained models, through chain prompting. This method benefits from ease-of-use and no required extra training. One downside is that getting a reward for an unobserved state during simulation/exploration is costly (can require human annotation) [Goyal et al., 2019]. A future research direction is to combine RL with pretraining, with the intuition that pretraining gets the model to a more rich, learned state, and sparse RL rewards may be sufficient to learn a strong policy.

# 3 Going beyond deductive reasoning

Deductive reasoning is using previous information, abductive reasoning may require the model to come up with new ideas. LMs are more used to generating things that appear before, such as ChatGPT being really good at summarizing. When trying to prompt ChatGPT to reason about why there might be empty soup bowls on the table, it can come up with plausible theories. It can also assign "likelihoods" to each theory that sum to 100, but it is unknown if it actually understands what those numbers are supposed to represent.

Evaluation also becomes harder for abductive reasoning since there are multiple conclusions. Future research direction could focus on first solving evaluation and then go towards universal representation.

## 4 Should models reason similarly to humans?

There is a desire, in the longer term goal of generally intelligent reasoners, that models simulate how humans reason. Currently, it is tricky to determine whether models reason similarly to humans, and if pretraining methods help do so. One way of investigating this is by looking at whether models simulate designated functions that parts of our brain carry out for overall cognition. For example, it has long been shown that convolutional neural networks behave similarly to the visual cortex in the brain, and the ventral pathway [Hubel and Wiesel, 1959, Guclu and van Gerven, 2015]. One research agenda can be to align models to follow functions of sections in the brain. However, another perspective is to move in the direction of creating models that match human-level reasoning performance, rather than simulating human-like reasoning itself. In other words, do we really need to model artificial reasoning ability to be similar to humans in order to get the best reasoning models?

## 5 How are reasoning models useful to us?

Currently, the search paradigm that we use helps us primarily by allowing us to enter queries (of particular formats) in order to go through articles that will give us an answer and a rationale as to why the answer is so. ChatGPT-style conversational reasoning allows us to ask specific, unique (answer possibly not on the internet) questions and get an answer, and an explanation for why this is so. The aspect of getting curated, logical rationales in reasoning questions is crucial. In the context of multimodal question-answering, especially commonsense reasoning, ChatGPT still only takes text input, and gives text answers and rationales. Having a multimodal rationale could be very useful depending on the problem. For example, a visual rationale could be very useful for reasoning in complex scenes with a clutter of objects. Rather than the text rationale describing every object in the scene, a visual rationale highlighting a point of focus could communicate the point more concisely. GradCAM [Selvaraju et al., 2017] provided a strong motivation for this argument. Potentially, a combined visual and text rationale could be the most informative to understand.

Also, an important aspect of usefulness is retrieving evidence in reasoning, which will provide sources to claims, allowing us to cross-verify the reasoning model and explore further. External knowledge graphs may be useful as evidences or support to the model. Similar to how humans may know an answer without definite proof, but Google for proof and sources to confirm, a pretrained model could similarly generate an answer, and then query over a multimodal knowledge graph to provide evidence. This is also a very relevant problem in conversational search, and we see current efforts in BingChat API to add references to generations.

## References

Yingshan Chang and Yonatan Bisk. Webqa: A multimodal multihop neurips challenge. In Douwe Kiela, Marco Ciccone, and Barbara Caputo, editors, *Proceedings of the NeurIPS 2021 Competitions and Demonstrations Track*, volume 176 of *Proceedings of Machine Learning Research*, pages 232–245. PMLR, 06–14 Dec 2022. URL https://proceedings.mlr.press/v176/chang22a.html.

Prasoon Goyal, Scott Niekum, and Raymond J. Mooney. Using natural language for reward shaping in reinforcement learning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 2385–2391. International Joint Conferences on Artificial Intelligence Organization, 7 2019. doi: 10.24963/ijcai.2019/331. URL https://doi.org/10.24963/ijcai.2019/331.

U. Guclu and M. A. J. van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27):10005–10014, jul 2015. doi: 10.1523/jneurosci.5023-14.2015. URL https://doi.org/10.1523%2Fjneurosci.5023-14.2015.

Jie Huang and Kevin Chen-Chuan Chang. Towards reasoning in large language models: A survey, 2022. URL https://arxiv.org/abs/2212.10403.

David H. Hubel and Torsten N. Wiesel. Receptive fields of single neurones in the cat's striate cortex. *The Journal of Physiology*, 148, 1959.

Rik Koncel-Kedziorski, Subhro Roy, Aida Amini, Nate Kushman, and Hannaneh Hajishirzi. MAWPS: A math word problem repository. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1152–1157, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1136. URL https://aclanthology.org/N16-1136.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering, 2022. URL https://arxiv.org/abs/2209.09513.

Maarten Sap, Vered Shwartz, Antoine Bosselut, Yejin Choi, and Dan Roth. Commonsense reasoning for natural language processing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 27–33, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-tutorials.7. URL https://aclanthology.org/2020.acl-tutorials.7.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017. doi: 10.1109/ICCV.2017.74.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality, 2022. URL https://arxiv.org/abs/2204.03162.

Keyulu Xu, Jingling Li, Mozhi Zhang, Simon S. Du, Ken ichi Kawarabayashi, and Stefanie Jegelka. What can neural networks reason about? In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rJxbJeHFPS.

Andy Zeng, Maria Attarian, brian ichter, Krzysztof Marcin Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael S Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, and Pete Florence. Socratic models: Composing zero-shot multimodal reasoning with language. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=G2Q2Mh3avow.

Ningyu Zhang, Lei Li, Xiang Chen, Xiaozhuan Liang, Shumin Deng, and Huajun Chen. Multimodal analogical reasoning over knowledge graphs, 2022. URL https://arxiv.org/abs/2210.00312.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models, 2023. URL https://arxiv.org/abs/2302.00923.

Wang Zhu, Jesse Thomason, and Robin Jia. Generalization differences between end-to-end and neuro-symbolic vision-language reasoning systems, 2022. URL https://arxiv.org/abs/2210.15037.