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

Week 9: Interaction 1: Reasoning and Large Models

Instructors: Paul Liang and Daniel Fried *Synopsis Leads: Simran Khanuja*

Edited by Paul Liang *Scribes: Anwesa Bhattacharya, Simran Khanuja*

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 9's discussion session, we discussed multiple topics revolving around reasoning with multiple modalities. We started with discussing how humans reason over multiple modalities in daily tasks like cooking, disaster management etc., going beyond simply vision and text for these models. Further, we discussed challenges in reasoning with VLMs, the benefit of external knowledge and so on. The following was a list of provided research probes:

1. Currently, most reasoning models are basically limited to vision and language domains. However, in our real world, multimodal reasoning more broadly exists and has more diverse forms. Can you list a few more examples of multimodal reasoning tasks in our daily life that rely on other modalities and how symbolic or unique reasoning methods can be applied to them?
2. Can you create a taxonomy of all potential symbolic systems that can be helpful for different types of multimodal reasoning tasks like AMR graphs, knowledge graphs and programs? What are their unique advantages and disadvantages?
3. Based on [Berglund et al., 2023], are there any other complex reasoning tasks besides reverse logic problems that you think the current foundation models might not handle well? How can neural symbolic models be incorporated to help with those hard cases?
4. Besides [Wang et al., 2024], can you imagine any other potential way to uncover the reasoning capabilities of black-box models, such as large language models and other multimodal foundation models? How can one discover specifically the cross-modal reasoning processes in such a black-box model?
5. 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?
6. What are the main advantages of reasoning-based approaches, when compared to large-scale multimodal models discussed in the previous lectures? What are the potential issues with reasoning? Can we perform reasoning on very large datasets? Why do pre-training methods eventually learn reasoning processes similar to humans? Or will we still need human and domain knowledge to some extent?
7. Are there unique technical challenges that arise when we consider utilizing neural symbolic methods on multimodal data as performed on multimodal data? What are these unique challenges? How can we start studying these challenges in future research?

As background, students read the following papers:

1. (Required) Tree of thoughts: Deliberate problem solving with large language models [Yao et al., 2024]
2. (Required) ViperGPT: Visual Inference via Python Execution for Reasoning [Surís et al., 2023]
3. (Suggested) Understanding the Reasoning Ability of Language Models From the Perspective of Reasoning Paths Aggregation [Wang et al., 2024].
4. (Suggested) Selection-Inference: Exploiting Large Language Models for Interpretable Logical Reasoning [Creswell et al., 2022].

5. (Suggested) Generalization Differences between End-to-End and Neuro-Symbolic Vision-Language Reasoning Systems [Zhu et al., 2022].
6. (Suggested) Multimodal Analogical Reasoning over knowledge graphs [Zhang et al., 2022]
7. (Suggested) The Neuro-Symbolic Concept Learner: Interpreting scenes, words, and sentence from natural supervision [Mao et al., 2019].

We summarize several main takeaway messages from group discussions below:

1 Multimodal reasoning beyond vision and text

Oftentimes when working with multimodal models, we are restricted to reasoning with vision and text since we have digital resources in these modalities to train our models. However, many real-world use-case involve reasoning with multiple modalities beyond simply vision and text. For example, in affect recognition, information from EEG signals, physical sensors, tone of voice, gestures, and so on, can provide for much richer information that is either complementary or helps strengthen the original prediction we make with vision+text. In cooking, we need to use taste and smell to know whether the food is edible or to inform us as to what the dish is missing and so on. In disaster situations like a fire in the building, we need smoke detectors, sensors and this data to be interfaced with external commonsense knowledge, which can inform users which routes they can take, whether they need to run faster and so on. Another application which came up in our discussion was for video models. Typically, one would need to bake in knowledge about the physical properties of objects in the world, when modeling sequential data. We looped back to some of the SORA failure cases we saw with glass shattering.[1]

An interesting proposition that came up in the discussion was whether a model can be trained to predict whether it needs information from these additional modalities (which are harder to obtain and integrate into the learning signal). This would typically be in cases where vision+text don't suffice for disambiguation, or they provide incomplete or contrasting information. While it is certainly appealing to use data from physical sensors for this particular task, we must consider that obtaining learning signals from noisy EEG signal/sensor data is also very challenging. Discussions on training foundation models on EEG data to predict human actions followed. However, the noise can make it hard to do self-supervised learning with such data. They also do not have as much structure as images, and are hard to interpret for humans as well.

However, a counter-argument suggested that despite humans not being able to interpret raw signals, models are equipped to learn signals from high amounts of potentially noisy data, for example, in audio with models like HiFi-GAN [Kong et al., 2020].

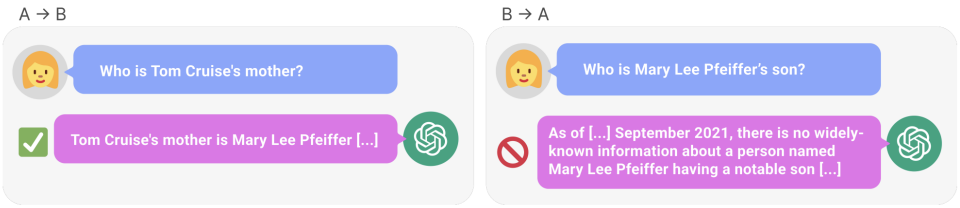

Figure 1: Failure mode of LLM reasoning, the model reversal curse.

2 What are the challenging tasks that remain to be solved in multimodal reasoning?

Here, a reference is made to the paper on reversal curse [Berglund et al., 2023]. This paper shows a major flaw in multimodal reasoning. If a model is trained on a sentence of the form "A is B", it will not automatically

[1] https://openai.com/research/video-generation-models-as-world-simulators

generalize to the reverse direction "B is A", as shown in Figure 1. While the paper provides interesting failure examples, a point of discussion was how is the model expected to deal with QA-pairs where we have a one-to-many mapping or a many-to-many mapping? Training models to perform well on these benchmarks that deal with only a subset of natural phenomena that we observe in the real world, may lead to overfitting and overall lack of generalization.

Another case where multimodal LLMs fail at reasoning is when the reasoning requires external knowledge that constantly changes. For example, cases of humor or sarcasm often include references from political figures or social/cultural situations, which constantly evolve over time and are highly contextual. This can either happen due to lack of knowledge/context or the inability to perform multi-hop reasoning or both. LLMs also generally fail at common-sense reasoning tasks, which may require an understanding of the physics of the world in the multimodal context. These include saying things like, if you empty a mug in the sink, the sink is empty [Zellers et al., 2021], or recent SORA failure examples, which is crucial to get right in robotics/embodiment applications and so on.

3 Imbuing multimodal LLMs with commonsense

From discussions on how multimodal models fail to reason about things which humans terms as "commonsense", we pivoted to discussing how we can teach models the same. The main research question was that humans already know a lot of things, from a nature point of view, through eons of evolution, so how do we teach models these generic skills or reasoning processes? These include things like understanding that red usually means stop in navigation or other contexts. We touched upon the nature/nurture debate and whether commonsense is innate or learnable. If learnable, how can we teach it to models?

Some approaches in the past have attempted to write down all the commonsense knowledge that humans have, made knowledge graphs to reason over, and so on [Sap et al., 2020]. We also discussed how this can be domain-specific and how we can abstract out the more general concepts which don't vary across domains. Related to this, especially if we want to teach models certain skills, recent work shows how doing instruction fine-tuning before domain adaptation primes the model to learn about the task at hand in the domain adaptation phase.

Summary: Overall, we discussed important applications of multimodal reasoning beyond vision+text, what challenges remain to be solved in multimodal reasoning and how we can make models learn commonsense or give them the capability to leverage external knowledge in the reasoning process.

References

Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. The reversal curse: Llms trained on" a is b" fail to learn" b is a". *arXiv preprint arXiv:2309.12288*, 2023.

Antonia Creswell, Murray Shanahan, and Irina Higgins. Selection-inference: Exploiting large language models for interpretable logical reasoning. *arXiv preprint arXiv:2205.09712*, 2022.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in neural information processing systems*, 33:17022–17033, 2020.

Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B Tenenbaum, and Jiajun Wu. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. *arXiv preprint arXiv:1904.12584*, 2019.

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, 2020.

Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11888–11898, 2023.

Xinyi Wang, Alfonso Amayuelas, Kexun Zhang, Liangming Pan, Wenhu Chen, and William Yang Wang. Understanding the reasoning ability of language models from the perspective of reasoning paths aggregation. *arXiv preprint arXiv:2402.03268*, 2024.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36, 2024.

Rowan Zellers, Ari Holtzman, Matthew Peters, Roozbeh Mottaghi, Aniruddha Kembhavi, Ali Farhadi, and Yejin Choi. Piglet: Language grounding through neuro-symbolic interaction in a 3d world. *arXiv preprint arXiv:2106.00188*, 2021.

Ningyu Zhang, Lei Li, Xiang Chen, Xiaozhuan Liang, Shumin Deng, and Huajun Chen. Multimodal analogical reasoning over knowledge graphs. In *The Eleventh International Conference on Learning Representations*, 2022.

Wang Zhu, Jesse Thomason, and Robin Jia. Generalization differences between end-to-end and neuro-symbolic vision-language reasoning systems. *arXiv preprint arXiv:2210.15037*, 2022.