11-866 Artificial Social Intelligence **Spring 2023**

Week 11: Multimodal Social Understanding

Instructors: Louis-Philippe Morency, Paul Liang *Synopsis Leads: Santiago Benoit, Leena Mathur*

Edited by Paul Liang *Scribes: Cathy Jiao, Alex Wilf*

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/asi-course/spring2023/

Summary: Artificial social intelligence (ASI) research focuses on advancing AI systems that can perceive and interpret human social information (e.g., communication patterns, pragmatics, intent) and engage in seamless human-AI interactions. In week 11's discussion session, the class focused on discussing **multimodal social understanding**, which focused on the challenges and future directions for modeling skills and abilities that underlie social scene understanding. The following was a list of provided research probes:

1. This week, we will discuss AI technologies for social understanding. A popular trend is to learn large-scale pre-trained models, such as the MERLOT Reserve model. What social skills and abilities are likely to be well suited to be modeled and understood by large-pretrained models? Which ones will require a different approach?
2. A different approach is to explicitly represent and model intermediate stages of the social understanding process, including the detection of the unimodal and multimodal behaviors related to social interactions. One example is to explicitly learn and model graphically the behaviors, interactions and relationships between people using graphs or similar representations. When should we use such an approach to better understand social interactions? What are the pros and cons of this approach?
3. Many large-scale pre-trained models are learned through masking. In what situations do you expect masking to succeed, and as importantly, when will it fail? If not using masking, how should we learn models for social understanding?How to integrate theories and knowledge of social intelligence in these large-scale pretrained models? Do we need motion to understand social interactions, or can it be done from images?
4. To succeed in social understanding, do we need agents to experience and interact themselves? Can it be learned from observations only? Can it be learned just from text (e.g., Social IQA dataset)? When is nonverbal and multimodal information helpful? What else would be needed to learn efficiently? What are the differences between social understanding when second-person view (e.g., social agent) vs third-person view (outside observer)?
5. Is question-answering the right way to evaluate multimodal social understanding? Are current benchmarks properly evaluating social understanding, or only a subset such as social perception? What would be an alternative to this benchmarking paradigm? What would it take for you to be convinced that an AI system is socially intelligent and can understand social interactions?

As background, students read the following papers:

1. Moviegraphs: Towards Understanding Human-Centric Situations from Videos [Vicol et al., 2018]
2. Merlot Reserve: Neural Script Knowledge through Vision and Language and Sound [Zellers et al., 2022]
3. Review and Challenges of Technologies for Real-Time Human Behavior Monitoring [Dávila-Montero et al., 2021]
4. Revisiting the "Video" in Video-Language Understanding [Buch et al., 2022]
5. Social signal processing Survey of an emerging domain [Vinciarelli et al., 2009]
6. TVQA: Localized, Compositional Video Question Answering [Lei et al., 2018]

We summarize main takeaway messages from group discussions below:

1 What social skills and abilities are likely to be well suited to be modeled and understood by large-pretrained models?

What are the tradeoffs between MERLOT RESERVE and graph-based approaches?

Our group discussed tradeoffs between approaches such as MERLOT RESERVE [Zellers et al., 2022] and more traditional graph-based approaches with respect to modeling social skills and abilities. We began by categorizing social skills into two categories: **basic social skills** and **complex social skills**. Basic social skills include sentiment analysis, emotion recognition, and context understanding, corresponding to social competencies for perception and understanding. More complex social skills include higher order interactions and relationships among people, as well as counterfactual social reasoning.

Graph-based approaches might be a good way to obtain interaction and relationship skill annotations, since interactions and relationships can be easily represented as edges between nodes representing individuals. MERLOT RESERVE would be useful for drawing inferences from multimodal data, due to its powerful joint encoder over images, text, and audio. MERLOT RESERVE is likely better suited for conducting tasks that require basic social skills, since these are related to drawing inferences from multimodal data. Our group speculated that graph-based approaches would likely be better for tasks requiring complex social skills, which involve reasoning over interactions and relationships, which can be reasoned over in graphs.

What do language models do well and what are areas of limitations?

Language models can perform well at tasks requiring perception and understanding, but not as well at reasoning. Evidence of this phenomena can be seen in the GPT-4 capabilities paper [Bubeck et al., 2023], which found that language models struggled with reasoning that required **planning** and **backtracking** in order to reach optional conclusions (examples of failures in GPT-4 reasoning were drawn from domains such as mathematics). Similarly, our group anticipated that language models would not perform as well in social reasoning tasks that required multiple steps of planning, especially tasks requiring socially-intelligent theory-of-mind [Sap et al., 2022]. We also discussed that language models are prone to hallucinations; addressing and mitigating against these hallucinations is an open research question worth addressing. In addition, language models currently do not have a robust world model that is embodied [Bisk et al., 2020]. We hypothesized that language models would not be able to answer questions requiring a deep physical understanding of the world, which can only be gained by experience and interaction. However, methods such as RLHF may enable language models to develop this "embodied" understanding of the world via interaction with humans at-scale [Christiano et al., 2017].

2 How can large pretrained models be useful for social skills and abilities?

A common challenge with pretrained vision-language models is their capacity to encode harmful biases. These biases can propagate to downstream tasks [Chuang et al., 2023, Berg et al., 2022] and is especially important to consider in social tasks. Aside from being harmful to the humans interacting with AI systems based on these large language models, bias can also skew the model's ability to perform well across domains. LLMs are trained on public information, which might be biased to include more social interaction data from large public groups than small private groups [Dodge et al., 2021]. Small group social norms differ from large groups and, therefore, an LLM's social knowledge may suffer gaps in knowledge regarding underrepresented social groups.

Our group discussed ways we could potentially improve the performance of LLMs for social interactions. To be more useful for social interactions, LLMs could be finetuned on small group data, if there is not enough small group data in the original dataset. Additionally, human-in-the-loop methods can be used to finetune LLMs for social tasks.

Many social norms are context-specific. LLMs can follow an "observe before participating" approach where the model observes an interaction for a while and gathers contextual information. Once the LLM has enough contextual information, it could be better suited for participating in the social interaction. It could be useful to identify flexible group contexts that cover most small group interactions, in order to train a more generalized model, which could then more easily adapt to specific small group contexts.

3 What is reasoning and social reasoning?

Our group discussed distinctions between reasoning (including deductive and inductive reasoning) and social reasoning. Reasoning can be **deductive**, meaning that an agent starts with a premise or set of premises and uses them to progressively work towards a conclusion, step-by-step. [Huang and Chang, 2022]. Reasoning can also be **inductive**, meaning that an agent starts with data and comes up with probabilities for different premises that may have generated or created environments that led to the instantiation of that data. **Social reasoning** is unique from other types of reasoning because social reasoning includes a mix of both deductive and inductive reasoning. Our group also identified a key challenge of social reasoning that makes it distinct from reasoning in physical settings: there are typically **multiple correct answers** for social reasoning, which compounds the difficulty of the problem when evaluating whether an agent (both human or machine) has strong social reasoning capabilities

4 How can we make sure a social agent is doing social understanding and not taking shortcuts?

Transparency and interpretability are often challenges with deep learning models, which sometimes learn to take shortcuts based on artifacts in the data. For building trustworthy social agents, it is important to understand the model's social reasoning, as shortcuts learned from dataset artifacts could lead to incorrect behavior in real-life situations.

One way to approach this challenge is through additional supervision. Having a "fully labeled" dataset can be useful for understanding a model's reasoning, using auxiliary outputs for example, but is difficult to generalize this to different tasks.

One additional form of supervision could be dynamic mask learning. This approach would involve collecting data on what humans are focusing on a scenario, and training a model with this data to output attention masks, demonstrating what data the model is using for social understanding.

Interpretability remains a big challenge for deep learning based social agents, especially LLMs. LLMs distill knowledge from their training data, and it would be useful to find exactly what an LLM "knows". A good human interpretable format for knowledge could be a knowledge graph. Building a LLM to use a human interpretable graph structure in its reasoning would address this problem of interpretability. One example of a potential model for this purpose is a neurosymbolic model, which can combine a symbolic, human interpretable representation of knowledge with deep learning approaches. One example of a neurosymbolic model that builds interpretable knowledge graphs is in Bosselut et al. [2021], which dynamically constructs local commonsense knowledge graphs and then reason on the knowledge graph to answer questions about a situation.

5 In what situations towards artificial social intelligence do you expect masking to succeed and in what situations will it fail?

Traditional approaches for masking, based on random masking, do not require reasoning in the objective; even approaches that select task-specific information to mask [Gu et al., 2020] do not require reasoning. Therefore, we anticipate that current models trained with traditional random masking objectives will not perform well in situations that require reasoning skills. Our group characterized **social reasoning** as a type of reasoning that includes a mix of deductive reasoning and inductive reasoning, centered on constructing

plausible explanations of human socioemotional states and generating these states. More advanced masking techniques, perhaps causally-inspired or in long-range situations, may enable models to learn social intelligence competencies.

References

Hugo Berg, Siobhan Mackenzie Hall, Yash Bhalgat, Wonsuk Yang, Hannah Rose Kirk, Aleksandar Shtedritski, and Max Bain. A prompt array keeps the bias away: Debiasing vision-language models with adversarial learning. *arXiv preprint arXiv:2203.11933*, 2022.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, et al. Experience grounds language. *arXiv preprint arXiv:2004.10151*, 2020.

Antoine Bosselut, Ronan Le Bras, and Yejin Choi. Dynamic neuro-symbolic knowledge graph construction for zero-shot commonsense question answering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(6):4923–4931, May 2021. doi: 10.1609/aaai.v35i6.16625. URL https://ojs.aaai.org/index.php/AAAI/article/view/16625.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

Shyamal Buch, Cristóbal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei, and Juan Carlos Niebles. Revisiting the" video" in video-language understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2917–2927, 2022.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Ching-Yao Chuang, Varun Jampani, Yuanzhen Li, Antonio Torralba, and Stefanie Jegelka. Debiasing vision-language models via biased prompts. *arXiv preprint arXiv:2302.00070*, 2023.

Sylmarie Dávila-Montero, Jocelyn Alisa Dana-Lê, Gary Bente, Angela T Hall, and Andrew J Mason. Review and challenges of technologies for real-time human behavior monitoring. *IEEE Transactions on Biomedical Circuits and Systems*, 15(1):2–28, 2021.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. *arXiv preprint arXiv:2104.08758*, 2021.

Yuxian Gu, Zhengyan Zhang, Xiaozhi Wang, Zhiyuan Liu, and Maosong Sun. Train no evil: Selective masking for task-guided pre-training. *arXiv preprint arXiv:2004.09733*, 2020.

Jie Huang and Kevin Chen-Chuan Chang. Towards reasoning in large language models: A survey. *arXiv preprint arXiv:2212.10403*, 2022.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696*, 2018.

Maarten Sap, Ronan LeBras, Daniel Fried, and Yejin Choi. Neural theory-of-mind? on the limits of social intelligence in large lms. *arXiv preprint arXiv:2210.13312*, 2022.

Paul Vicol, Makarand Tapaswi, Lluis Castrejon, and Sanja Fidler. Moviegraphs: Towards understanding human-centric situations from videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8581–8590, 2018.

Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. Social signal processing: Survey of an emerging domain. *Image and vision computing*, 27(12):1743–1759, 2009.

Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. Merlot reserve: Neural script knowledge through vision and language and sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16375–16387, 2022.