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

Week 10: Interaction 2: Embodiment and planning

Instructors: Paul Liang and Daniel Fried *Synopsis Leads: Anwesa Bhattacharya*

Edited by Paul Liang *Scribes: Anwesa Bhattacharya, Ashwin Pillay*

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.

During week 10's discussion session, the class focused on identifying potential sensory modalities applicable to embodied tasks, devising intuitive action spaces, addressing challenges in synthesizing high-quality embodied task data, and exploring additional application areas for embodied AI.

The following was a list of provided research probes:

1. For what sort of embodied tasks might code be a good representation for? For what tasks would it be a poor representation for?
2. When generating embodied plans from natural language, several dimensions of difficulty are the ambiguity in the language, the difficulty of grounding language in the environment, and the difficulty of carrying out the plan in the environment. Consider one or two of the papers from this week – which of these dimensions (or others) do they mainly address?
3. When building an embodied AI, one key challenge is to define an easy and clear action space to ground. Given any particular task like cooking and housekeeping, how to design an appropriate action space that can be easily and accurately grounded? Provide a task and its corresponding designed action space for grounding.
4. Robotics requires more broad multisensory machine learning techniques besides well-studied vision/language multimodal techniques. What are the potential sensory modalities for embodied agent tasks that are not well studied now? What specific embodied tasks require the information from sensory modality to be completed?
5. Based on the release of Figure01 (https://www.figure.ai/), what are the three potential main technical challenges for the next steps of embodied AI and why? What are three potential applications like automatic housekeeping that are still not achievable for robotics now?
6. For embodied AI training, embodied data is widely considered a serious bottleneck.There are a lot of data synthesis works based on virtual or physical environments like https://arxiv.org/pdf/2403.08629.pdf. What are the key challenges for embodied tasks data synthesis and how to make sure that synthesized data are high-quality?
7. What challenges do social settings add beyond standard embodied/robotics tasks?

As background, students read the following papers:

1. **(Required)** Voyager: An Open-Ended Embodied Agent with Large Language Models [Wang et al., 2023]
2. **(Required)** Habitat 3.0: A Co-Habitat for Humans, Avatars and Robots [Puig et al., 2023]
3. (Suggested) ProgPrompt: Generating Situated Robot Task Plans using Large Language Models [Singh et al., 2022]
4. (Suggested) Code as Policies: Language Model Programs for Embodied Control [Liang et al., 2023]

5. (Suggested) Eureka: Human-Level Reward Design via Coding Large Language Models [Ma et al., 2023]
6. (Suggested) Learning adaptive planning representations with natural language guidance [Wong et al., 2023]
7. (Suggested) Skill Induction and Planning with Latent Language [Sharma et al., 2022]
8. (Suggested) Building Cooperative Embodied Agents Modularly with Large Language Models [Zhang et al., 2024]

We summarize several main takeaway messages from group discussions below:

1 Modalities for embodied agent tasks

Tasks performed by embodied agents, such as those in robotics, necessitate the utilization of comprehensive multisensory machine learning techniques, in addition to the extensively studied vision/language multimodal techniques. Below are a few potential sensory modalities, their applications, and methods for acquiring data pertaining to these modalities, as derived from our discussion:

- The "Chat with the Environment" paper [Zhao et al., 2023] illustrates how a Large Language Model (LLM) can effectively interact with its surroundings to gather information for a given task, enabling it to accomplish objectives. It begins with a broad task, such as 'Pick up the plastic box,' and utilizes visual cues to perceive the scene. Subsequently, the LLM employs other modalities, such as audio, by tapping on objects to discern which ones sound like plastic. This exemplifies the strategic use of an LLM for initial planning, followed by the integration of multiple sensory inputs to make informed decisions.
- AutoRT [Ahn et al., 2024] endeavors to explore a domain while adhering to Constitutional AI principles. The LLM is prompted to generate a series of manipulation tasks based on foundational, ethical, and embodiment rules. Specifically, embodiment rules dictate the agent's capabilities, such as limitations on lifting objects heavier than a book. Tactile sensing could greatly enhance the functionality of embodied agents operating in industrial environments by enabling them to categorize items based on pressure, thereby determining appropriate handling methods (e.g., treating fragile glass objects with care). Additional modalities like temperature sensing could aid in task prioritization, such as avoiding objects with excessively high temperatures to mitigate potential hazards or triggering alerts.
- Analogously, in culinary contexts, olfactory senses play a crucial role in assessing the condition of food, distinguishing between cooked and uncooked items, or detecting spoilage.
- Another potential sensory modality lies in physiological signals such as EEG, ECG, and EMG, which hold significant importance in healthcare applications like robot-assisted surgeries.
- Incorporating multiple modalities presents challenges in procuring datasets containing such diverse sensory inputs. One suggested approach for data collection involves utilizing public platforms like Recaptcha, capable of capturing user-centric data such as mouse movements and keystrokes, which often reflect emotional states or cognitive load. However, challenges persist, as Recaptcha is not open source and raises privacy and ethical concerns.
- Robotic tasks, including navigation and door opening, entail subtasks like measuring the pressure required to manipulate objects. Pressure sensors offer one means of gathering such data, while EMG signals also provide valuable insight into physical interactions.

2 Designing action spaces

When constructing an embodied AI system, a significant challenge lies in delineating a straightforward and comprehensive action space. This challenge becomes particularly pronounced when tackling tasks such as housekeeping and cooking. The goal is to establish an action space that is both intuitive for humans and accurately aligned with the task at hand.

- Housekeeping encompasses various tasks such as dusting, doing laundry, cleaning, and organizing spaces. The Housekeep benchmark [Kant et al., 2022] challenges embodied agents to make decisions about

placing new objects, navigate unseen environments using only visual cues, and explore areas prone to clutter. It mirrors typical housekeeping duties, which often involve common subtasks like lifting objects, finding appropriate locations for items, moving them, and tidying up spaces. This task structure resembles web navigation challenges, where agents must click on elements, navigate multiple pages, and compile lists. Web navigation benchmarks, like WebArena [Zhou et al., 2023], define action spaces by considering small actions such as clicking or scrolling. These actions are composed to perform more complex tasks such as listing recent orders. Similarly, the action space for housekeeping tasks could involve actions like lifting objects, wiping surfaces, and multimodal actions such as identifying and listing items in a space.

- While most existing reasoning approaches rely on language as an intermediate medium, humans often reason using multiple modalities, particularly vision. This is evident when individuals imagine or dream about potential scenarios, often based on visual cues. Video generation models like SORA [Liu et al., 2024] can aid in this process by envisioning the next state of the environment based on current actions. Reinforcement learning agents like MuZero [Schrittwieser et al., 2020] and Dreamer [Hafner et al., 2019] aim to perform such imagination in latent space. The Dreamer paper highlights the advantage of learning low-memory representations of actions, enabling the generation of numerous possible trajectories crucial for long-horizon tasks.
- A common concern regarding video generation models like SORA [Liu et al., 2024] is their lack of understanding of real-world physics, as demonstrated by examples such as the glass shattering video. One potential solution is to train these models on generated video data while grounding them in the laws of physics, such as Newton's laws or the conservation of mass. Reasoning benchmarks like CLEVRER [Yi et al., 2020] aim to enhance models' ability to reason about temporal and causal structures of events, such as predicting the next event following a collision between objects. These benchmarks provide guidance in learning the physical laws governing real-world interactions.
- The combination of reinforcement learning and LLMs raises intriguing possibilities. While reinforcement learning agents may have limited action spaces, LLMs excel in generating multiple potential actions or instructions. Moreover, LLMs can serve as the reward or scoring model themselves, using preference data as demonstrated in the Direct Preference Optimization [Rafailov et al., 2023] paper. This approach eliminates the need for a separate reward model and aims to enhance reinforcement learning with LLMs.
- The concept of imitation learning prompts consideration of whether LLMs can be fine-tuned to explore environments under direct supervision. Instead of relying on reinforcement learning, direct supervision could guide LLMs in determining the best actions at each step. This approach offers the advantage of allowing LLMs to explore environments and receive feedback, thereby accumulating more data for supervised learning.

3 Data Synthesis for Embodied Tasks

For embodied AI training, the availability of embodied data is often identified as a significant challenge. Numerous efforts have focused on synthesizing data derived from virtual or physical environments. However, ensuring the quality of synthesized data remains a formidable challenge in the field of embodied tasks.

- Synthetic embodied task data poses a challenge in ensuring that agents trained on such data can generalize effectively to real-world environments. While agents may demonstrate proficiency within simulated settings, their performance might falter when deployed in actual scenarios. To mitigate this issue, data augmentation approaches can be employed to diversify the training data, ensuring comprehensive coverage of all possible actions and environmental conditions. For example, in training a robotic arm for household tasks, data augmentation techniques could include varying object placements, lighting conditions, and background clutter to enhance the agent's adaptability.
- Similar to pre-trained models utilized in other areas, pre-training embodied task models could prove beneficial. By pre-training on a diverse range of environments and tasks, these models can acquire a foundational understanding. This pre-training enables the agent to adapt more efficiently to new environments that share similarities with those encountered during pre-training. A robot pre-trained

on various simulated household tasks may exhibit improved performance when deployed in a real home environment due to its prior experience and knowledge.
- An alternative approach involves integrating simulated environments with real-world counterparts, allowing embodied agents to fine-tune its decision-making abilities through a hybrid training approach.

4 Embodied AI applications

The field of embodied AI continues to expand, reaching into various new domains. Several potential application areas were explored during the discussion:

- Embodied agents hold the potential to "read the room," discerning emotions and tailoring responses accordingly. In the healthcare sector, this capability could prove invaluable as doctors and nurses utilize agents to engage with patients. For instance, an agent equipped with emotion-sensing capabilities could detect anxiety in a patient awaiting surgery and provide comforting interactions to alleviate stress.
- Embodied AI offers promising prospects for personalized education, enabling agents to adapt learning pathways to individual learners' preferences and cognitive styles. For example, an educational agent could assess a student's preferences, strengths and weaknesses and tailor exercises and explanations accordingly, optimizing the learning experience for each student.
- The gaming industry is a potential area, particularly in multiplayer games where agents can analyze teammates' responses and adjust interactions accordingly. In a cooperative multiplayer game, an agent could assess teammates' strategies and dynamically adapt its gameplay to complement their actions, enhancing teamwork and overall gaming experience.
- For robotics-assisted surgeries, embodied agents can contribute by interpreting physiological signals from the environment. For instance, during a surgical procedure, an agent could monitor vital signs and surgical instruments' data, providing real-time feedback to assist the surgeon in making precise movements and optimizing surgical outcomes.
- Embodied agents offer potential benefits for autonomous vehicles by responding to environmental cues and other vehicles' behaviors on the road. For example, an autonomous vehicle equipped with embodied AI could interpret pedestrian gestures, traffic patterns, and road conditions, allowing it to make informed decisions and navigate safely through complex traffic scenarios.

References

Michael Ahn, Debidatta Dwibedi, Chelsea Finn, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Karol Hausman, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, Sean Kirmani, Isabel Leal, Edward Lee, Sergey Levine, Yao Lu, Isabel Leal, Sharath Maddineni, Kanishka Rao, Dorsa Sadigh, Pannag Sanketi, Pierre Sermanet, Quan Vuong, Stefan Welker, Fei Xia, Ted Xiao, Peng Xu, Steve Xu, and Zhuo Xu. Autort: Embodied foundation models for large scale orchestration of robotic agents, 2024.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019.

Yash Kant, Arun Ramachandran, Sriram Yenamandra, Igor Gilitschenski, Dhruv Batra, Andrew Szot, and Harsh Agrawal. Housekeep: Tidying virtual households using commonsense reasoning, 2022.

Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. Code as policies: Language model programs for embodied control, 2023.

Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, Lifang He, and Lichao Sun. Sora: A review on background, technology, limitations, and opportunities of large vision models, 2024.

Yecheng Jason Ma, William Liang, Guanzhi Wang, De-An Huang, Osbert Bastani, Dinesh Jayaraman, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Eureka: Human-level reward design via coding large language models, 2023.

Xavier Puig, Eric Undersander, Andrew Szot, Mikael Dallaire Cote, Tsung-Yen Yang, Ruslan Partsey, Ruta Desai, Alexander William Clegg, Michal Hlavac, So Yeon Min, Vladimír Vondruš, Theophile Gervet, Vincent-Pierre Berges, John M. Turner, Oleksandr Maksymets, Zsolt Kira, Mrinal Kalakrishnan, Jitendra Malik, Devendra Singh Chaplot, Unnat Jain, Dhruv Batra, Akshara Rai, and Roozbeh Mottaghi. Habitat 3.0: A co-habitat for humans, avatars and robots, 2023.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2023.

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy Lillicrap, and David Silver. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588(7839):604–609, December 2020. ISSN 1476-4687. doi: 10.1038/s41586-020-03051-4. URL http://dx.doi.org/10.1038/s41586-020-03051-4.

Pratyusha Sharma, Antonio Torralba, and Jacob Andreas. Skill induction and planning with latent language, 2022.

Ishika Singh, Valts Blukis, Arsalan Mousavian, Ankit Goyal, Danfei Xu, Jonathan Tremblay, Dieter Fox, Jesse Thomason, and Animesh Garg. Progprompt: Generating situated robot task plans using large language models, 2022.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models, 2023.

Lionel Wong, Jiayuan Mao, Pratyusha Sharma, Zachary S. Siegel, Jiahai Feng, Noa Korneev, Joshua B. Tenenbaum, and Jacob Andreas. Learning adaptive planning representations with natural language guidance, 2023.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. Clevrer: Collision events for video representation and reasoning, 2020.

Hongxin Zhang, Weihua Du, Jiaming Shan, Qinhong Zhou, Yilun Du, Joshua B. Tenenbaum, Tianmin Shu, and Chuang Gan. Building cooperative embodied agents modularly with large language models, 2024.

Xufeng Zhao, Mengdi Li, Cornelius Weber, Muhammad Burhan Hafez, and Stefan Wermter. Chat with the environment: Interactive multimodal perception using large language models, 2023.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. Webarena: A realistic web environment for building autonomous agents, 2023.