Learning Representations for Robust Human-Robot Interaction

Yen-Ling Kuo

University of Virginia
Department of Computer Science
ylkuo@virginia.edu

For robots to robustly and flexibly interact with humans, they need to acquire skills to use across scenarios. One way to enable the generalization of skills is to learn representations that are useful for downstream tasks. Learning a representation for interactions requires an understanding of what (e.g., objects) as well as how (e.g., actions, controls, and manners) to interact with. However, most existing language or visual representations mainly focus on objects, for example, aligning language with world states via CLIP-like method (Myers et al. 2023). These representations enable robots to interpret intents such as which object to be picked up. But such representations cannot be used to understand sentences like "move slower so you don't spill water" or "help that person move the box." To enable robust human-robot interactions (HRI), we need a representation that is not just grounded at the object level but to reason at the action level. The ability to reason about an agent's own actions and other's actions will be particularly important for long-tail interactions, e.g., interactions that are safety-critical or novel interaction scenarios, as collecting large datasets for these interactions will be costly or even impossible.

Specifically, my research focuses on leveraging the compositional nature of language and reward functions to learn representations that generalize to novel scenarios. Together with the information from multiple modalities, the learned representation can reason about task progress, future behaviors, and the goals/beliefs of an agent. The above ideas have been demonstrated in my research on language understanding and social interactions.

To build representations for language understanding, I utilize compositional linguistic structures like parse trees to combine representation learned for different words to guide motion planners to follow natural language commands (Kuo, Katz, and Barbu 2020a,b). The structure of sentences is reflected in the structure of policy networks to enable robots to understand novel commands and act rationally in new scenarios. When associating sentences with environment features and an agent's past trajectory, the linguistic representation can scaffold the hypothesis of future trajectories (Kuo et al. 2022). To build representations for social interactions, I show how composing an agent's own reward and the reward estimation of other agents can give rise to novel social interactions between agents (Tejwani et al. 2022, 2023). The recursive reward estimation improves an agent's Theory of Mind reasoning capability to enable robots to accurately reason about sequences of actions.

While continuing my research on building compositional representations for HRI, my next research plan include exploring the desired properties of the learned representations such as interpretability and fast inference (Jha et al. 2024) and exploring solutions to adding these properties, which are ongoing work in my group.

References

Jha, K.; Le, T. A.; Jin, C.; Kuo, Y.-L.; Tenenbaum, J. B.; and Shu, T. 2024. Neural Amortized Inference for Nested Multi-agent Reasoning. In *AAAI Conference on Artificial Intelligence*.

Kuo, Y.-L.; Huang, X.; Barbu, A.; McGill, S. G.; Katz, B.; Leonard, J. J.; and Rosman, G. 2022. Trajectory prediction with linguistic representations. In *International Conference on Robotics and Automation*.

Kuo, Y.-L.; Katz, B.; and Barbu, A. 2020a. Compositional networks enable systematic generalization for grounded language understanding. In *Findings of Empirical Methods in Natural Language Processing*.

Kuo, Y.-L.; Katz, B.; and Barbu, A. 2020b. Deep compositional robotic planners that follow natural language commands. In *2020 IEEE International Conference on Robotics and Automation*.

Myers, V.; He, A.; Fang, K.; Walke, H.; Hansen-Estruch, P.; Cheng, C.-A.; Jalobeanu, M.; Kolobov, A.; Dragan, A.; and Levine, S. 2023. Goal Representations for Instruction Following: A Semi-Supervised Language Interface to Control. In *Conference on Robot Learning*.

Tejwani, R.; Kuo, Y.-L.; Shu, T.; Katz, B.; and Barbu, A. 2022. Social interactions as recursive MDPs. In *Conference on Robot Learning*.

Tejwani, R.; Kuo, Y.-L.; Shu, T.; Stankovits, B.; Gutfreund, D.; Tenenbaum, J. B.; Katz, B.; and Barbu, A. 2023. Zero-shot linear combinations of grounded social interactions with Linear Social MDPs. In *AAAI Conference on Artificial Intelligence*.