Showing Utility as an Active Speaker

Zhancun Mu
YuanPei College
Peking University
yhbylch@stu.pku.edu.cn

Abstract

One of the most important assumptions in decision-making theory is that rational agents make decisions to maximize expected utility. Based on this assumption, many methods have been proposed to understand demonstrated behaviors or to learn utility functions from experts. An important field is inverse reinforcement learning (IRL), which aims to learn utility functions from demonstrations. However, little work has investigated how to actively and efficiently demonstrate one's utility to others. In this essay, we will draw inspiration from communication systems and attempt to provide a framework for generating pedagogical demonstrations.

1 Introduction

Many argue that preference, value, or utility drives us in daily life. We choose seats by comfort, food by taste, and clothes by preference. The concept of utility is central in economics and decision theory: a rational agent acts to maximize expected utility based on beliefs and desires [8].

However, utility is important not only for single agents, but also multi-agent systems. Previous works utilize this assumption for understanding demonstrated behaviors or learning expert utility functions. We can view agents in these works as listeners in communication. But little work discusses the speaker role in this processthat is, how to actively and efficiently demonstrate utility to others.

In this essay, we will discuss this problem within the inverse reinforcement learning (IRL) framework [1] and attempt to draw inspiration from communication models.

2 Speaker in communication systems

We live in a cooperative society where communication is vital when agents have asymmetric information and goal misalignment. In these cases, communication helps agents align their individual versions of the group and converge on shared goals or utility functions. A well-informed speaker needs to help listeners align with the speaker's utility function, a concept close to paternalistic helping [6].

One approach is for the speaker to act rationally, expecting the listener to infer the utility function under the assumption of rationality (through Bayesian inference), i.e.

$$P_{\text{Speaker}}(a|s;u) \propto \exp\{\beta U(a|s)\}$$
$$P_{\text{Listener}}(u|a,s) \propto P_{\text{Speaker}}(a|s;u)P(u)$$

The listener's reasoning process is called *inverse planning*.

However, this may not be efficient, as the example in Fig. 1 shows. This inefficiency arises because the speaker does not model the listener's reasoning. In other words, the speaker needs to do *inverse inverse planning* [2]. A similar framework is the Rational Speech Act (RSA) [3], with listeners and speakers built upon one another. However, RSA only considers pragmatics, which is insufficient for complex communication.

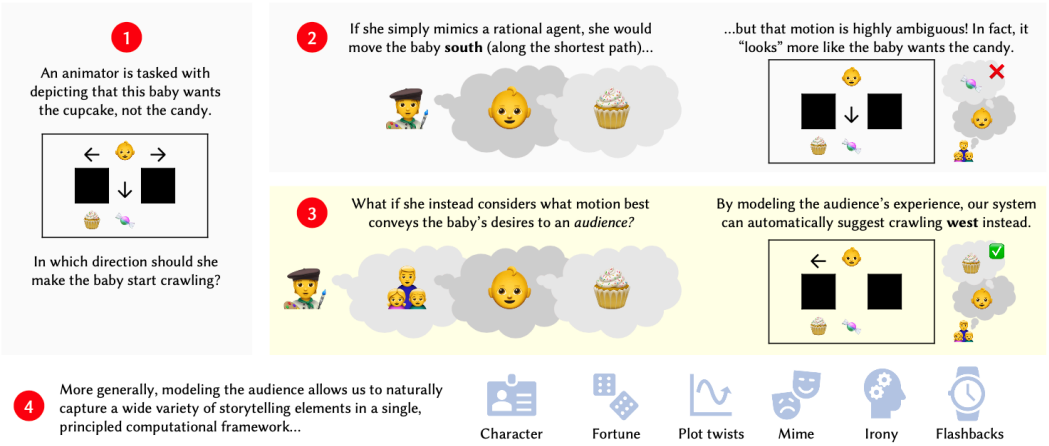

Figure 1: Act as inverse inverse planning. Image from [2]

.

Stacy et al. [7] argue joint utility should also be considered for cooperation. Suppose the multi-agent system has a joint utility function. Instead of modeling the listener's reasoning, the speaker directly models a "we" perspective. We argue this method could potentially generate superior demonstrations for learning an expert's utility function for use in IRL.

3 Speaker in Inverse Reinforcement Learning

First we need to answer three questions:

How to formulate IRL and our problem? We can formulate IRL as learning rewards/utility from demonstrations. That is, given a set of demonstrations $\mathcal{D} = (s_i, a_i)_{i=1}^N$, we want to learn a function R that explains the demonstrations. For our problem, we want to modify the expert's policy to make it more informative and comprehensible to the learner. Formally, given the listener's learning algorithm $\mathcal{A}$, we want to generate N demonstrations that maximize the learner's performance.

Why inverse RL? Compared to imitation learning, IRL can learn more generalizable policies from limited demonstrations. Additionally, learning others' utility functions helps us better understand differing worldviews while integrating our own preferences.

Why need pedagogical demonstrations? Many existing IRL methods assume optimal demonstrations. For instance, Ramachandran and Amir [5] employ Bayesian inference to infer utility function:

$$\Pr_{\mathcal{X}}(\mathbf{R}|O_{\mathcal{X}}) = \frac{1}{Z'} e^{\alpha_{\mathcal{X}} E(O_{\mathcal{X}}, \mathbf{R})} P_R(\mathbf{R}).$$

However, optimal demonstrations are not necessarily pedagogical. In MINECRAFT, for instance, optimal demonstrations may resemble tool-assisted speedruns (TAS), which are incomprehensible to beginners. In contrast, pedagogical demonstrations can comprise step-by-step, distinct actions with clear purposes. These understandable demonstrations can teach utility functions, as shown in [4]. They model utility as a linear combination of state features via some feature function $\phi : R(s, a_S, a_L, \theta) = \phi(s)^T \theta$. Their idea matches the expected feature sum over a trajectory distribution with the feature sum from a single trajectory. However, they only partially address the issue and still face computational complexity.

We now consider this problem from the lens of communication and theory of mind (ToM). While [7] focuses on joint utility, the idea seamlessly transfers to our problem since the listener aims to learn the same utility function. We can adapt the main equations in [7] accordingly:

$$P_S(\tau_S|u) \propto \exp\{\beta \mathbb{E}[U(\tau_S|u)]\}$$
$$\mathbb{E}[U(\tau_S|u)] = \mathbb{E}_{P(\tau_L|\tau_S)}[u(\tau_L)]$$
$$P(\tau_L|\tau_S) \propto \int P(\tau_L|u_L) P(u_L|\tau_S) du_L$$

That is, the speaker generates demonstrations τ_S that maximize the listener's total reward $u(\tau_L)$ under the learned utility u_L and corresponding policy $P(\tau_L|u_L)$.

The main problem is estimating $P(\tau_L|u_L)$ and $P(u_L|\tau_S)$. The first term is the policy requiring a utility-based model. However, this quickly grows complicated, with existing methods only solving goal-conditioned problems where utility is $r = \alpha \cdot \mathbb{1}(s = g_t)$. Still, we believe pre-trained models like decision transformers (DT) could generate trajectories for any utility function. The second term infers utility from demonstrations, i.e. $P(u_L|\tau_S)$. This directly relates to the listener's learning algorithm.

4 Conclusion

In this essay, we discuss the speaker side of the IRL problem. We cast the teacher/expert as the speaker and the learner as the listener. From a communication lens, we attempt to provide a framework for generating pedagogical demonstrations that convey the speaker's utility. However, current method limitations preclude implementing this preliminary framework, including lacking generalized utility-based policies and efficient IRL algorithms. Deeper mathematical analysis alongside a thorough literature survey is needed. In summary, we propose an interactive IRL perspective where a teacher and learner interact to align utility functions. We believe this emerging field warrants formulation and exploration.

References

[1] Saurabh Arora and Prashant Doshi. A survey of inverse reinforcement learning: Challenges, methods and progress. *Artificial Intelligence*, 297:103500, 2021. 1

[2] Kartik Chandra, Tzu-Mao Li, Joshua Tenenbaum, and Jonathan Ragan-Kelley. Acting as inverse inverse planning. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–12, 2023. 1, 2

[3] Michael C Frank and Noah D Goodman. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998, 2012. 1

[4] Dylan Hadfield-Menell, Stuart J Russell, Pieter Abbeel, and Anca Dragan. Cooperative inverse reinforcement learning. *Advances in neural information processing systems*, 29, 2016. 2

[5] Deepak Ramachandran and Eyal Amir. Bayesian inverse reinforcement learning. In *IJCAI*, volume 7, pages 2586–2591, 2007. 2

[6] Stephanie Stacy, Siyi Gong, Aishni Parab, Minglu Zhao, Kaiwen Jiang, and Tao Gao. A bayesian theory of mind approach to modeling cooperation and communication. *Wiley Interdisciplinary Reviews: Computational Statistics*, page e1631. 1

[7] Stephanie Stacy, Chenfei Li, Minglu Zhao, Yiling Yun, Qingyi Zhao, Max Kleiman-Weiner, and Tao Gao. Modeling Communication to Coordinate Perspectives in Cooperation, June 2021. 2

[8] Yixin Zhu, Tao Gao, Lifeng Fan, Siyuan Huang, Mark Edmonds, Hangxin Liu, Feng Gao, Chi Zhang, Siyuan Qi, Ying Nian Wu, et al. Dark, beyond deep: A paradigm shift to cognitive ai with humanlike common sense. *Engineering*, 6(3):310–345, 2020. 1