Embodiment Adaptation from Interactive Trajectory Preferences

Michael Walton, Ben Migliori, John Reeder

Space and Naval Warfare Systems Center Pacific
http://www.public.navy.mil/spawar/Pacific
{michael.walton, benjamin.migliori, john.d.reeder}@navy.mil

Keywords: Imitation Learning · Preference Learning · Reinforcement Learning

1 Introduction

Imitation learning provides an attractive approach to communicate complex goals to autonomous systems in domains where explicit reward functions are unavailable, tedious to specify or rely on substantial or high-cost expert knowledge. Standard Imitation Learning implicitly assumes that the embodiment of the learning agent and the teacher are either the same or intuitively compatible from the perspective of the demonstrator. In this work, we consider control tasks which violate these assumptions and propose a framework for estimating *embodiment adaptors* using human feedback expressed through pairwise preferences over control trajectories.

2 Background

Recent advances in reinforcement learning (RL) have largely been driven by scaling algorithms well understood in simple task domains to complex, high-dimensional problems using deep neural networks for value function approximation [6] and policy learning [5]. In the standard formulation of a reinforcement learning problem, often posed as a Markov Decision Process (MDP), one assumes access to a reward function $R : S \times A \rightarrow \mathbb{R}$ which associates a scalar reward with agent actions $a \in A$ taken in states $s \in S$. The agents' objective, therefore, is to maximize it's cumulative reward. In many well posed control tasks, this objective may be straightforward to specify: the score of a game, the goal configuration in robotic manipulation tasks, forward velocity for walking or crawling.

Complementary to RL, Imitation Learning provides an approach for learning a control policy without an explicit reward function. This approach is desirable in problems domains where a concise goal statement may be challenging to express [1], [2]. Prior work has also explored imitation learning to improve the sample efficiency of reinforcement learning [3], [4]. Conventional approaches to imitation learning, however, fundamentally rely on the availability of demonstrations of expert control in the form of observation, action tuples. Demonstration data may

be acquired through teleoperation [1] or kinesthetic teaching [2]. In the former case, the imitator and the demonstrator are assumed to have the same embodiment, eg. their state and action spaces are assumed to be consistent. In the latter, the demonstrator must inhabit the same physical space as the embodied agent and must be able to efficiently pose and manipulate its effectors.

Many complex control tasks may exhibit incompatibilities between the embodiments of the demonstrator and the imitating agent. Consider for instance a robotic arm we may wish to train to perform household tasks such as preparing food; pose estimates of a human demonstrator's arm will yield sequences of actions with different degrees of freedom and dynamics than the imitating arm.

3 Methods

Our proposed approach takes two stages: In the first stage the human demonstrator provides undirected feedback to the agent to optimize a policy $\pi_\alpha : A_H \to A_\ell$ which translates between the demonstrators action space A_H and agent's action space A_ℓ. This is achieved through trajectory preference learning [1], however in our formulation preferences are assigned to the trajectory that best matched the demonstrators' desired action. Formally, we state that a trajectory τ^1 is preferred, denoted $\succ$ to τ^2 following a reward function r known only to the demonstrator if:

$$\tau^1 \succ \tau^2 \equiv \sum_t r(a_t^1, \pi_\alpha(a_t^1)) > \sum_t r(a_t^2, \pi_\alpha(a_t^2)) \tag{1}$$

After each interaction, a pairwise preference is assigned between the two trajectories and an reward function approximation $\hat{r}$ is estimated using the method specified in [1]. The embodiment adaptation policy is then subsequently trained to maximized $\hat{r}$ using standard reinforcement learning. After learning an embodiment adaptation policy, the second phase uses this mechanism to learn a behavior policy π_β from translated demonstrations using (for instance) behavioral cloning. In this simple formulation, the optimal policy given expert demonstrations D is the policy that minimizes the divergence between π_β and the expert actions translated by π_α; assuming continuous actions, we may define this objective in terms of the quadratic loss:

$$\pi_\beta^* = \arg\min_{\pi_\beta \in \Pi_\beta} \mathbb{E}_{(s,a)\sim D}[(\pi_\alpha(a) - \pi_\beta(s))^2] \tag{2}$$

We propose two proof of concept embodiment translation tasks to demonstrate the utility of our method: a classic gridworld with discrete state and action

[1] The demonstrator directly controls the agent which records action selections for imitation

[2] The demonstrator physically manipulates an embodied agent by applying force to its effectors; demonstration in these scenarios may be, for instance, resultant torques on the joints of a robotic arm

spaces and the continuous control problem *lunar lander*. In the lunar lander task, for instance, the human demonstrator must select thrust directions using the up, left and right keys; it is observed in [7] that humans tend to fail on this task. Distinct from previous work, we hypothesize that this is an unintuitive interface for a human operator to demonstrate correct behavior. A more natural interface, perhaps, may be a joystick-like interface. We apply our method to learn an embodiment adaptor policy π_α which translates continuous forces applied to a joystick to sequences of discrete thruster pulses which are compatible with the imitator's embodiment.

References

1. Christiano, P., Leike, J., Brown, T.B., Martic, M., Legg, S., Amodei, D.: Deep reinforcement learning from human preferences (2017)
2. Hadfield-Menell, D., Dragan, A., Abbeel, P., Russell, S.: The off-switch game (2016)
3. Hester, T., Vecerik, M., Pietquin, O., Lanctot, M., Schaul, T., Piot, B., Horgan, D., Quan, J., Sendonaris, A., Dulac-Arnold, G., Osband, I., Agapiou, J., Leibo, J.Z., Gruslys, A.: Deep q-learning from demonstrations (2017)
4. Hester, T., Vecerik, M., Pietquin, O., Lanctot, M., Schaul, T., Piot, B., Sendonaris, A., Dulac-Arnold, G., Osband, I., Agapiou, J., Z. Leibo, J., Gruslys, A.: Learning from demonstrations for real world reinforcement learning (04 2017)
5. Lillicrap, T.P., Hunt, J.J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., Wierstra, D.: Continuous control with deep reinforcement learning (2015)
6. Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., Riedmiller, M.: Playing atari with deep reinforcement learning (2013)
7. Reddy, S., Dragan, A.D., Levine, S.: Shared autonomy via deep reinforcement learning (2018)