Imitation Learning from Observation

Faraz Torabi[1]
[1]The University of Texas at Austin
Austin, Texas 78712-0233
faraztrb@cs.utexas.edu

Introduction

One well-studied way in which artificially-intelligent agents are currently able to learn to perform tasks autonomously is via *reinforcement learning (RL)* (Sutton and Barto 1998) techniques. Using these techniques, if agents are able to interact with the world and receive feedback (known as *reward*) based on how well they are performing with respect to a particular task, they are able to use their own experience to improve their future behavior. However, designing a proper feedback mechanism for complex tasks can sometimes prove to be extremely difficult for system designers. Moreover, learning based solely on one's own experience can be exceedingly slow.

Concerns such as the ones above have given rise to the study of *imitation learning* (Argall et al. 2009), where agents instead attempt to learn a task by observing another, more expert agent perform that task. Because the information about how to perform the task is communicated to the imitating agent via a demonstration, this paradigm does not require the explicit design of a reward function. Furthermore, because the demonstrations directly provide rich information regarding how to perform the task correctly, imitation learning is typically faster than *RL*. Imitation learning methods have been used to successfully learn a variety of tasks such as navigation for quadrotors (Giusti et al. 2016) and autonomous ground vehicles (Bojarski et al. 2016), dish placement, and pouring (Finn, Levine, and Abbeel 2016).

Importantly, most of the imitation learning literature has thus far concentrated only on situations in which the imitator not only has the ability to observe the demonstrating agent's *states* (e.g., observable quantities such as spatial location), but also the ability to observe the demonstrator's *actions* (e.g., internal control signals such as motor commands). While this extra information can make the imitation learning problem easier, requiring it is also limiting. In particular, requiring action observations makes a large number of valuable learning resources – e.g., vast quantities of online videos of humans performing different tasks (Zhou, Xu, and Corso 2017) – useless. For the demonstrations present in such resources, the actions of the expert are unknown. This limitation has recently motivated increased interest in the area of *imitation from observation (IfO)* (Liu et al. 2017), in which agents seek to perform imitation learning using state-only demonstrations.

In this thesis, we decompose the imitation from observation problem into two main components: (1) perception of the demonstration, and (2) learning an autonomous control policy. The goal of this thesis is to address each of these parts and integrate them in order to create an agent with the capability of imitation from raw video inputs.

Previous Contributions

Thus far, I have developed two algorithms that learn control policies from state features-only demonstration trajectories which are explained in the following in more details. These works are done with the collaboration of Garrett Warnell of Army Research Laboratory (ARL) and advised by Peter Stone of UT Austin.

Behavioral Cloning from Observation (*BCO*)

For humans to imitate a task, they do not require access to the actions performed by an expert. They watch the expert performing a task, infer the actions that are executed and practice them themselves. Moreover, this process of learning is typically done very quickly. This ability comes from the fact that humans have had experience in the world through interaction and have the knowledge of the effects of each action on the environment. The proposed algorithm is specifically designed for the following goal: given a set of state-only demonstration trajectories, D, find a good imitation policy using a minimal number of interactions after observing the demonstration (post-demonstration interactions).

We were motivated by the fact that humans have access to a large amount of prior experience knowledge about themselves, and so we aimed to also provide an autonomous agent with this same prior knowledge. To do so, before any demonstration information is observed, we allow the agent to explore its own action space by performing random actions. During this step, we collect states and actions of the agent and learn a mapping from state-transitions to the responsible actions for those transitions. This mapping is called inverse dynamics model. Then, upon observation

of a demonstration without action information, *BCO* uses the learned model to infer the missing actions. Finally, *BCO* uses the demonstration and the inferred actions to find a policy via behavioral cloning. With this approach, the agent would perform a reasonably good policy as soon as it observes the demonstrations.

On the other hand, if one is willing to tolerate post-demonstration environment interaction, a modified version of our algorithm can further improve both the learned model and the resulting imitation policy. This modified algorithm proceeds as follows. After the behavioral cloning step, the agent executes the imitation policy in the environment for a short period of time. Then, the newly-observed state-action sequences are used to update the model, and, accordingly, the imitation policy itself. The above procedure is repeated until there is no more improvement in the imitation policy.

Thus far, this algorithm is evaluated on some domains available in OpenAI Gym with different levels of difficulty and the results represent that in most cases, the final performance of the agent is very close to the vanilla behavioral cloning where the agent does have access to actions.

Generative Adversarial Imitation from Observation(*GAIfO*)

From a high-level perspective, in imitation from observation, the goal is to enable the agent to extract what the task is by observing some state sequences. Intuitively, this extraction is possible because we expect the beneficial state transitions for any given task to form a low-dimensional manifold within the $\mathcal{S} \times \mathcal{S}$ space (where $\int$ represents the state space).

Motivated by the mentioned fact, we attempted to get the state-transitions distribution of the imitator and the expert closer together. To do so, inpired by the work of Ho and Ermon (2016), we developed the Generative Adversarial Imitation from Observation(*GAIfO*) algorithm which includes a policy network and a discriminator. The policy takes as input a state and outputs an action and the discriminator takes a state transition and outputs a value between 0 and 1. We begin by randomly initalizing each of the networks, after which the imitator selects an action according to the policy and executes that action. This action leads to a new state, and we feed both this state transtion and the entire set of expert state transitions to the discrimintator. The discriminator is updated using the Adam optimization algorithm (Kingma and Ba 2014), with cross-entropy loss that seeks to push the output for expert state transitions closer to 1 and the imitator's state transitions closer to 0. After the discriminator update, we then perform trust region policy optimization (*TRPO*) (Schulman et al. 2015) to improve the policy using a reward function that encourages state transitions that yield large outputs from the discriminator (i.e., those that appear to be from the demonstrator). This process continues until convergence. If we interpret the expert's demonstrations as the real data, and the data coming from the imitator as the generated data, the discriminator seeks to distinguish the source of the data, and the imitator policy (i.e., the generator) seeks to fool the discriminator to make it look like the state transitions it generates are coming from the expert. The entire process can be interpreted as bringing the distribution of the imitator's state transitions closer to that of the expert. We call this process Generative Adversarial Imitation from Observation (*GAIfO*).

We have evaluated *GAIfO* in simulation environments both over the low-level manually defined state features and the cases that the states are raw visual data in OpenAI Gym, where the results are not only compares favorably to other recently-developed methods for *IfO*, but they also are comparably to state-of-the-art conventional imitation learning methods that *do* have access to the demonstrator's actions.

Ongoing research

In this section, I discuss my ongoing research.

Application of the algorithms on a physical robot

Learning on physical robots is more challenging compared to the simulation environments because of the stochasticity of the environment dynamics, noisiness of the sensors, etc. We plan to test our proposed algorithms on a physical robot, UR5, which is an industrial arm robot that can be utilized in research to learn different types of object manipulation tasks such as reaching, grasping, etc.

References

Argall, B. D.; Chernova, S.; Veloso, M.; and Browning, B. 2009. A survey of robot learning from demonstration. *Robotics and Autonomous Systems* 57(5):469–483.

Bojarski, M.; Del Testa, D.; Dworakowski, D.; Firner, B.; Flepp, B.; Goyal, P.; Jackel, L. D.; Monfort, M.; Muller, U.; Zhang, J.; et al. 2016. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*.

Finn, C.; Levine, S.; and Abbeel, P. 2016. Guided cost learning: Deep inverse optimal control via policy optimization. In *International Conference on Machine Learning*, 49–58.

Giusti, A.; Guzzi, J.; Cireşan, D. C.; He, F.-L.; Rodríguez, J. P.; Fontana, F.; Faessler, M.; Forster, C.; Schmidhuber, J.; Di Caro, G.; et al. 2016. A machine learning approach to visual perception of forest trails for mobile robots. *IEEE Robotics and Automation Letters* 1(2):661–667.

Ho, J., and Ermon, S. 2016. Generative adversarial imitation learning. In *Advances in Neural Information Processing Systems*, 4565–4573.

Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Liu, Y.; Gupta, A.; Abbeel, P.; and Levine, S. 2017. Imitation from observation: Learning to imitate behaviors from raw video via context translation. *arXiv preprint arXiv:1707.03374*.

Schulman, J.; Levine, S.; Abbeel, P.; Jordan, M.; and Moritz, P. 2015. Trust region policy optimization. In *International Conference on Machine Learning*, 1889–1897.

Sutton, R. S., and Barto, A. G. 1998. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge.

Zhou, L.; Xu, C.; and Corso, J. J. 2017. Towards automatic learning of procedures from web instructional videos. *arXiv preprint arXiv:1703.09788*.