Adversarial Imitation Learning from State-only Demonstrations*

Extended Abstract

Faraz Torabi
The University of Texas at Austin
Austin, Texas
faraztrb@cs.utexas.edu

Garrett Warnell
Army Research Laboratory
Austin, Texas
garrett.a.warnell.civ@mail.mil

Peter Stone
The University of Texas at Austin
Austin, Texas
pstone@cs.utexas.edu

ABSTRACT

Imitation from observation (*IfO*) is the problem of learning directly from state-only demonstrations without having access to the demonstrator's actions. The lack of action information both distinguishes *IfO* from most of the literature in imitation learning, and also sets it apart as a method that may enable agents to learn from a large set of previously inapplicable resources such as internet videos. In this paper, we propose a new *IfO* approach based on generative adversarial networks called *generative adversarial imitation from observation* (*GAIfO*). We demonstrate that our approach performs comparably to classical imitation learning approaches (which have access to the demonstrator's actions) and significantly outperforms existing imitation from observation methods in high-dimensional simulation environments.

KEYWORDS

Reinforcement Learning; Imitation Learning; Control

ACM Reference Format:
Faraz Torabi, Garrett Warnell, and Peter Stone. 2019. Adversarial Imitation Learning from State-only Demonstrations. In *Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019,* IFAAMAS, 3 pages.

1 INTRODUCTION

One well-known way in which artificially-intelligent agents are able to learn to perform tasks is via *imitation learning* [1, 2, 9], where agents attempt to learn a task by observing another, more expert agent perform that task. 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 people performing different tasks [14] – useless. For the demonstrations present in such resources, the actions of the expert are unknown. This limitation has recently motivated work in the area of *imitation from observation (IfO)* [8], in which agents seek to perform imitation learning using state-only demonstrations.

*This is an abbreviated version of the work presented in [13].

Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), N. Agmon, M. E. Taylor, E. Elkind, M. Veloso (eds.), May 13–17, 2019, Montreal, Canada. © 2019 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). All rights reserved.

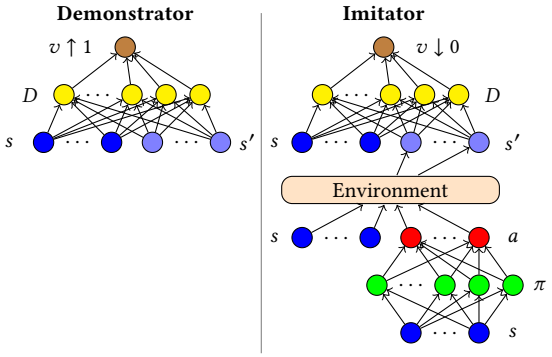

Figure 1: A diagrammatic representation of *GAIfO*.

In this paper, we propose a general framework for the control aspect of *IfO* in which we characterize the cost as a function of state transitions only. Under this framework, the *IfO* problem becomes one of trying to recover the state-transition cost function of the expert. Inspired by the work of Ho and Ermon ([2016]), we introduce a novel, model-free algorithm for *IfO*, called *generative adversarial imitation from observation* (*GAIfO*) and then experimentally evaluate *GAIfO* in high-dimensional simulation environments in two different settings: (1) demonstrations and states of the imitator are manually-defined features, and (2) demonstrations and states of the imitator come exclusively from raw visual observation. We show that the proposed method compares favorably to other recently-developed methods for *IfO* and also that it performs comparably to state-of-the-art conventional imitation learning methods that *do* have access the the demonstrator's actions.

2 ALGORITHM

We consider agents within the framework of Markov decision processes (MDPs). In this framework, $\mathcal{S}$ and $\mathcal{A}$ are the state and action spaces, respectively. An agent at a particular state $s \in \mathcal{S}$, chooses an action $a \in \mathcal{A}$, based on a policy $\pi : \mathcal{S} \times \mathcal{A} \to [0, 1]$ and transitions to state s' with probability of $P(s'|s, a)$ that is predefined by the environment transition dynamics. In our setting, the agent has access to state-only expert demonstration $\tau_E = s$ and the goal is to learn a policy π that results in a similar behavior.

Now we describe our algorithm, generative adversarial imitation from observation (*GAIfO*). In order to imitate the expert, the algorithm solves the following optimization problem:

$$\min_{\pi \in \Pi} \max_{D \in (0,1)^{\mathcal{S} \times \mathcal{S}}} \mathbb{E}_{\pi}[\log(D(s, s'))] + \mathbb{E}_{\pi_E}[\log(1 - D(s, s'))] \quad (1)$$

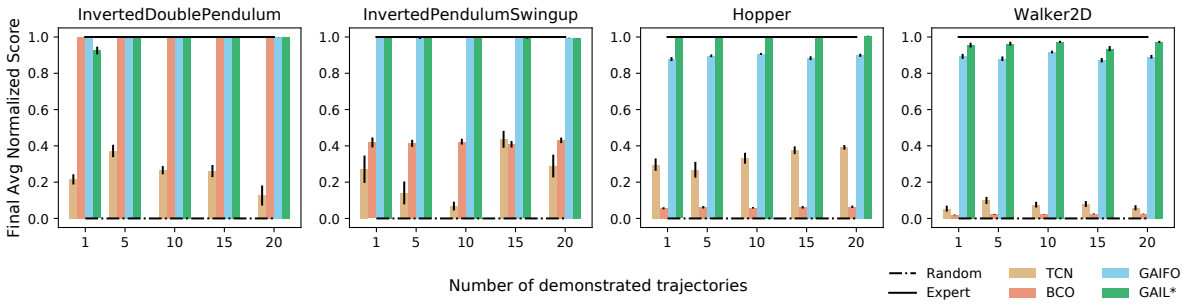

Figure 2: Performance of algorithms with respect to the number of demonstration trajectories. Rectangular bars and error bars represent mean return and standard deviations, respectively. For comparison purposes, we have scaled all the performances such that a random and the expert policy score 0.0 and 1.0, respectively. *GAIL has access to action information.

where $D : \mathcal{S} \times \mathcal{S} \rightarrow (0, 1)$ is a discriminative classifier. we can see that the loss function in (1) is similar to the generative adversarial loss [5]. We can connect this to general *GAN*s 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 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 now specify our practical implementation of the *GAIfO* algorithm (as shown in Figure 1). We represent the discriminator, D, using a multi-layer perceptron with parameters θ that takes as input a state transition and outputs a value between 0 and 1. We represent the policy, π, using a multi-layer perceptron with parameters ϕ that takes as input a state and outputs an action. We begin by randomly initializing each of these networks, after which the imitator selects an action according to π_ϕ and executes that action. This action leads to a new state, and we feed both this state transition and the entire set of expert state transitions to the discriminator. The discriminator is updated using the Adam optimization algorithm [7], 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 perform trust region policy optimization (*TRPO*) [10] 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.

3 EXPERIMENTAL SETUP AND RESULTS

We evaluate our algorithm in domains from OpenAI Gym [3] based on the Pybullet simulator [4]. In each of the domains, we used trust region policy optimization (*TRPO*) [10] to train the expert agents, and we recorded the demonstrations using the resulting policy.

The results shown in Figure 2 are the average over ten independent trials. We compare our algorithm against three baselines (1) **Behavioral Cloning from Observation (*BCO*)[12]**, (2) **Time Contrastive Networks (*TCN*)[11]**, and (3) **Generative Adversarial Imitation Learning (*GAIL*) [6]**

Algorithm 1 *GAIfO*

1: Initialize parametric policy π_ϕ with random ϕ
2: Initialize parametric discriminator D_θ with random θ
3: Obtain state-only expert demonstration trajectories $\tau_E = \{(s, s')\}$
4: **while** Policy Improves **do**
5: Execute π_ϕ and store the resulting state transitions $\tau = \{(s, s')\}$
6: Update D_θ using loss
$$-\Big(\mathbb{E}_\tau[\log(D_\theta(s, s'))] + \mathbb{E}_{\tau_E}[\log(1 - D_\theta(s, s'))]\Big)$$
7: Update π_ϕ by performing *TRPO* updates with reward function
$$-\Big(\mathbb{E}_{\tau_E}[\log(1 - D_\theta(s, s'))]\Big)$$
8: **end while**

Figure 2 compares the final performance of the imitation policies learned by different algorithms. We can clearly see that, for the domains considered here, *GAIfO (a)* performs very well compared to other IfO techniques, and *(b)* is surprisingly comparable to *GAIL* even though *GAIfO* lacks access to explicit action information.

4 CONCLUSION

In this paper, we presented an imitation from observation algorithm (*GAIfO*) that uses a *GAN*s like architecture to bring the state-transition distribution of the imitator to that of the expert. This algorithm is able to find policies without the need for action information, and is able to find imitation policies that perform very close to those found by techniques that do have access to this information.

ACKNOWLEDGMENTS

This work has taken place in the Learning Agents Research Group (LARG) at UT Austin. LARG research is supported in part by NSF (IIS-1637736, IIS-1651089, IIS-1724157), ONR (N00014-18-2243), FLI (RFP2-000), ARL, DARPA, Intel, Raytheon, and Lockheed Martin. Peter Stone serves on the Board of Directors of Cogitai, Inc. The terms of this arrangement have been reviewed and approved by the University of Texas at Austin in accordance with its policy on objectivity in research.

REFERENCES

[1] Brenna D Argall, Sonia Chernova, Manuela Veloso, and Brett Browning. 2009. A survey of robot learning from demonstration. *Robotics and Autonomous Systems* 57, 5 (2009), 469–483.

[2] Aude Billard, Sylvain Calinon, Ruediger Dillmann, and Stefan Schaal. 2008. Robot programming by demonstration. In *Springer handbook of robotics*. Springer, 1371–1394.

[3] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. 2016. OpenAI Gym. (2016). arXiv:arXiv:1606.01540

[4] Erwin Coumans and Yunfei Bai. 2016-2017. Pybullet, a python module for physics simulation for games, robotics and machine learning. http://pybullet.org/

[5] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems*.

[6] Jonathan Ho and Stefano Ermon. 2016. Generative adversarial imitation learning. In *Advances in Neural Information Processing Systems*. 4565–4573.

[7] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[8] YuXuan Liu, Abhishek Gupta, Pieter Abbeel, and Sergey Levine. 2017. Imitation from observation: Learning to imitate behaviors from raw video via context translation. *arXiv preprint arXiv:1707.03374* (2017).

[9] Stefan Schaal. 1997. Learning from demonstration. In *Advances in Neural Information Processing Systems*. 1040–1046.

[10] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. 2015. Trust region policy optimization. In *International Conference on Machine Learning*. 1889–1897.

[11] Pierre Sermanet, Corey Lynch, Jasmine Hsu, and Sergey Levine. 2017. Time-contrastive networks: Self-supervised learning from multi-view observation. *arXiv preprint arXiv:1704.06888* (2017).

[12] Faraz Torabi, Garrett Warnell, and Peter Stone. 2018. Behavioral cloning from observation. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*. AAAI Press, 4950–4957.

[13] Faraz Torabi, Garrett Warnell, and Peter Stone. 2018. Generative adversarial imitation from observation. *arXiv preprint arXiv:1807.06158* (2018).

[14] Luowei Zhou, Chenliang Xu, and Jason J Corso. 2017. Towards Automatic Learning of Procedures from Web Instructional Videos. *arXiv preprint arXiv:1703.09788* (2017).