Generative Adversarial Imitation from Observation

Faraz Torabi[1], Garrett Warnell[2], and Peter Stone[1]

[1]The University of Texas at Austin, [2]Army Research Laboratory

June 15th, 2019

Our goal?

To develop an **imitation learning from observation** algorithm

Our goal?

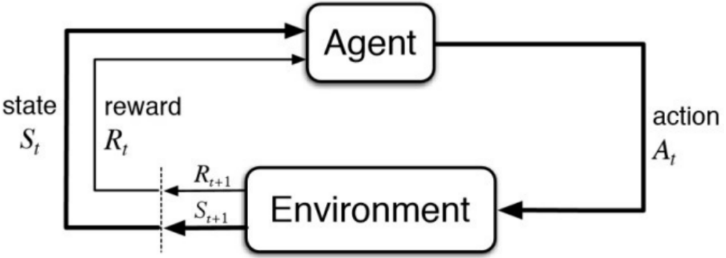
To develop an **imitation learning from observation** algorithm

What is Imitation Learning from Observation?

Reinforcement Learning

Goal:

- Learn how to make decisions in an environment by maximizing some notion of cumulative reward.



Reinforcement Learning

Goal:

- Learn how to make decisions in an environment by maximizing some notion of cumulative reward.

Challenge:

Reinforcement Learning

Goal:

- Learn how to make decisions in an environment by maximizing some notion of cumulative reward.

Challenge:

- Designing reward function for some tasks is hard or very sparse.

Imitation Learning

Goal:

- Learn how to make decisions by trying to imitate another agent.

Imitation Learning

Goal:

- Learn how to make decisions by trying to imitate another agent.

Algorithms:

Imitation Learning

Goal:

- Learn how to make decisions by trying to imitate another agent.

Algorithms:

- Behavioral Cloning (BC)

Imitation Learning

Goal:

- Learn how to make decisions by trying to imitate another agent.

Algorithms:

- Behavioral Cloning (BC)
  - E.g., End to End Learning for Self-Driving Cars.[1]

[1] Jiakai Zhang and Kyunghyun Cho. "Query-Efficient Imitation Learning for End-to-End Simulated Driving.". In: *AAAI*. 2017, pp. 2891–2897.

Imitation Learning

Goal:

- Learn how to make decisions by trying to imitate another agent.

Algorithms:

- Behavioral Cloning (BC)
  - E.g., End to End Learning for Self-Driving Cars.[1]
- Inverse Reinforcement Learning (IRL)

[1] Jiakai Zhang and Kyunghyun Cho. "Query-Efficient Imitation Learning for End-to-End Simulated Driving.". In: *AAAI*. 2017, pp. 2891–2897.

Imitation Learning

Goal:

- Learn how to make decisions by trying to imitate another agent.

Algorithms:

- Behavioral Cloning (BC)
  - E.g., End to End Learning for Self-Driving Cars.[1]
- Inverse Reinforcement Learning (IRL)
  - Guided Cost Learning.[2]

[1] Jiakai Zhang and Kyunghyun Cho. "Query-Efficient Imitation Learning for End-to-End Simulated Driving.". In: *AAAI*. 2017, pp. 2891–2897.

[2] Chelsea Finn, Sergey Levine, and Pieter Abbeel. "Guided cost learning: Deep inverse optimal control via policy optimization". In: *International Conference on Machine Learning*. 2016, pp. 49–58.

Imitation Learning

Goal:

- Learn how to make decisions by trying to imitate another agent.

Algorithms:

- Behavioral Cloning (BC)
  - E.g., End to End Learning for Self-Driving Cars.[1]
- Inverse Reinforcement Learning (IRL)
  - Guided Cost Learning.[2]
- Generative Adversarial Imitation Learning.[3]

[1] Jiakai Zhang and Kyunghyun Cho. "Query-Efficient Imitation Learning for End-to-End Simulated Driving.". In: AAAI. 2017, pp. 2891–2897.

[2] Chelsea Finn, Sergey Levine, and Pieter Abbeel. "Guided cost learning: Deep inverse optimal control via policy optimization". In: International Conference on Machine Learning. 2016, pp. 49–58.

[3] Jonathan Ho and Stefano Ermon. "Generative adversarial imitation learning". In: Advances in Neural Information Processing Systems. 2016, pp. 4565–4573.

Imitation Learning

Conventional Imitation Learning:

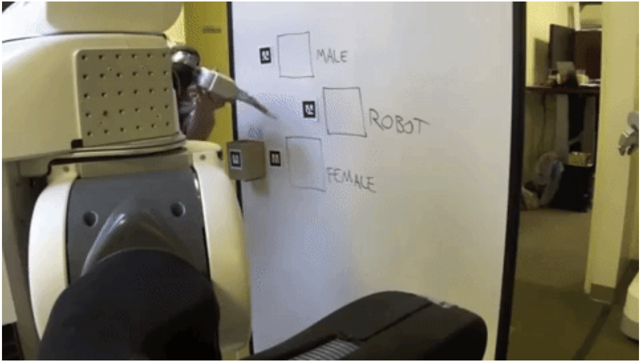
- Observations of other agent (demonstrations) consist of state-action pairs.[4]

[4] Scott Niekum et al. "Learning and generalization of complex tasks from unstructured demonstrations". In: *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on.* IEEE. 2012, pp. 5239–5246.

Imitation Learning

Conventional Imitation Learning:

- Observations of other agent (demonstrations) consist of state-action pairs.[4]



[4] Scott Niekum et al. "Learning and generalization of complex tasks from unstructured demonstrations". In: *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*. IEEE. 2012, pp. 5239–5246.

Imitation Learning

Conventional Imitation Learning:

- Observations of other agent (demonstrations) consist of state-action pairs.[4]

Drawback:

- Precludes using a large amount of demonstration data where action sequences are not given (e.g. YouTube videos).

[4] Scott Niekum et al. "Learning and generalization of complex tasks from unstructured demonstrations". In: *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on.* IEEE. 2012, pp. 5239–5246.

Imitation from Observation

Goal:

- Learn how to perform a task given state-only demonstrations.

Imitation from Observation

Goal:

- Learn how to perform a task given state-only demonstrations.

Formulation:

- Given:
  - $D_{demo} = (s_0, s_1, ...)$
- Learn:
  - $\pi : \mathcal{S} \rightarrow \mathcal{A}$

Imitation from Observation

Previous work:

- Time Contrastive Networks (TCN).[5]
- Imitation from observation: Learning to imitate behaviors from raw video via context translation.[6]
- Learning invariant feature spaces to transfer skills with reinforcement learning.[7]

[5] Pierre Sermanet et al. "Time-contrastive networks: Self-supervised learning from multi-view observation". In: *arXiv preprint arXiv:1704.06888* (2017).

[6] YuXuan Liu et al. "Imitation from observation: Learning to imitate behaviors from raw video via context translation". In: *arXiv preprint arXiv:1707.03374* (2017).

[7] Abhishek Gupta et al. "Learning invariant feature spaces to transfer skills with reinforcement learning". In: *arXiv preprint arXiv:1703.02949* (2017).

Imitation from Observation

Previous work:

- Time Contrastive Networks (TCN).[5]
- Imitation from observation: Learning to imitate behaviors from raw video via context translation.[6]
- Learning invariant feature spaces to transfer skills with reinforcement learning.[7]

Difference:

[5] Pierre Sermanet et al. "Time-contrastive networks: Self-supervised learning from multi-view observation". In: *arXiv preprint arXiv:1704.06888* (2017).

[6] YuXuan Liu et al. "Imitation from observation: Learning to imitate behaviors from raw video via context translation". In: *arXiv preprint arXiv:1707.03374* (2017).

[7] Abhishek Gupta et al. "Learning invariant feature spaces to transfer skills with reinforcement learning". In: *arXiv preprint arXiv:1703.02949* (2017).

Imitation from Observation

Previous work:

- Time Contrastive Networks (TCN).[5]
- Imitation from observation: Learning to imitate behaviors from raw video via context translation.[6]
- Learning invariant feature spaces to transfer skills with reinforcement learning.[7]

Difference:

- Concentrate on perception

[5] Pierre Sermanet et al. "Time-contrastive networks: Self-supervised learning from multi-view observation". In: *arXiv preprint arXiv:1704.06888* (2017).

[6] YuXuan Liu et al. "Imitation from observation: Learning to imitate behaviors from raw video via context translation". In: *arXiv preprint arXiv:1707.03374* (2017).

[7] Abhishek Gupta et al. "Learning invariant feature spaces to transfer skills with reinforcement learning". In: *arXiv preprint arXiv:1703.02949* (2017).

Imitation from Observation

Previous work:

- Time Contrastive Networks (TCN).[5]
- Imitation from observation: Learning to imitate behaviors from raw video via context translation.[6]
- Learning invariant feature spaces to transfer skills with reinforcement learning.[7]

Difference:

- Concentrate on perception
- Hand design a reward function

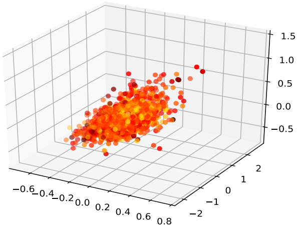
[5] Pierre Sermanet et al. "Time-contrastive networks: Self-supervised learning from multi-view observation". In: *arXiv preprint arXiv:1704.06888* (2017).

[6] YuXuan Liu et al. "Imitation from observation: Learning to imitate behaviors from raw video via context translation". In: *arXiv preprint arXiv:1707.03374* (2017).

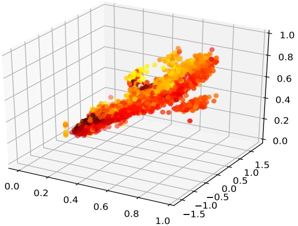
[7] Abhishek Gupta et al. "Learning invariant feature spaces to transfer skills with reinforcement learning". In: *arXiv preprint arXiv:1703.02949* (2017).

Generative Adversarial Imitation from Observation

Intuition:



(a) Random Policy



(b) Expert Policy

Figure: State transition distribution in Hopper domain.

Formulation

Recover expert policy by

- $c(s, s')$: cost as a function of state transition
- π_E: expert policy
- Π: set of all possible policies
- $\psi(c)$: regularizer

Formulation

Recover expert policy by

$$\tilde{c} = \underset{c \in \mathbb{R}^{\mathcal{S} \times \mathcal{S}}}{\arg\max} -\psi(c) + (\min_{\pi \in \Pi} \mathbb{E}_{\pi}[c(s, s')]) - \mathbb{E}_{\pi_E}[c(s, s')])$$

- $c(s, s')$: cost as a function of state transition
- π_E: expert policy
- Π: set of all possible policies
- $\psi(c)$: regularizer

Formulation

Recover expert policy by

$$\tilde{c} = \underset{c \in \mathbb{R}^{\mathcal{S} \times \mathcal{S}}}{\arg\max} -\psi(c) + (\min_{\pi \in \Pi} \mathbb{E}_{\pi}[c(s, s')]) - \mathbb{E}_{\pi_E}[c(s, s')])$$

$$\tilde{\pi} = \underset{\pi \in \Pi}{\arg\min}\ \mathbb{E}_{\pi}[\tilde{c}(s, s')]$$

- $c(s, s')$: cost as a function of state transition
- π_E: expert policy
- Π: set of all possible policies
- $\psi(c)$: regularizer

Generative Adversarial Imitation from Observation

Using a specific regularizer $\psi(c)$ results in:

- D: classifier (discriminator)

Generative Adversarial Imitation from Observation

Using a specific regularizer $\psi(c)$ results in:

$$\tilde{c} = \underset{D\in(0,1)^{\mathcal{S}\times\mathcal{S}}}{\arg\max}\ \mathbb{E}_{\pi}[\log(D(s,s'))] + \mathbb{E}_{\pi_E}[\log(1-D(s,s'))]$$

- D: classifier (discriminator)

Generative Adversarial Imitation from Observation

Using a specific regularizer $\psi(c)$ results in:

$$\tilde{c} = \operatorname*{arg\,max}_{D \in (0,1)^{\mathcal{S} \times \mathcal{S}}} \mathbb{E}_{\pi}[\log(D(s, s'))] + \mathbb{E}_{\pi_E}[\log(1 - D(s, s'))]$$

$$\tilde{\pi} = \operatorname*{arg\,min}_{\pi \in \Pi} \max_{D \in (0,1)^{\mathcal{S} \times \mathcal{S}}} \mathbb{E}_{\pi}[\log(D(s, s'))] + \mathbb{E}_{\pi_E}[\log(1 - D(s, s'))]$$

- D: classifier (discriminator)

Algorithm

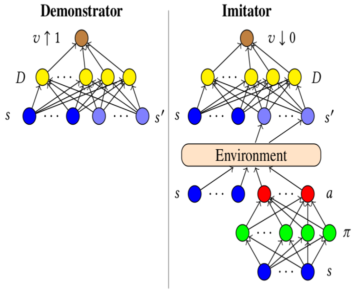
Low-dimensional States

- Initialize policy π
- While "Policy Improves":
- Execute π and collect $\tau = \{(s, s')\}$
- Update D_θ using loss

$$-\Big(\mathbb{E}_\tau[\log(D_\theta(s,s'))] + \mathbb{E}_{\tau_E}[\log(1 - D_\theta(s,s'))]\Big)$$

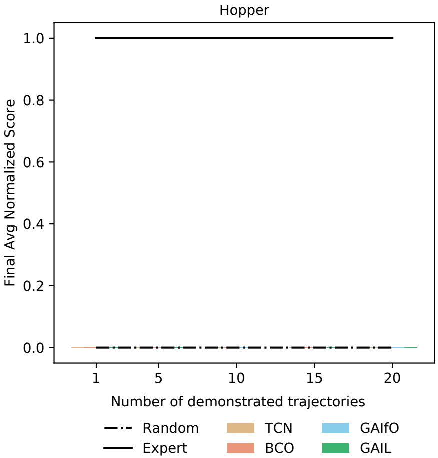
- Update π by *TRPO* with r

$$-\Big(\mathbb{E}_{\tau_E}[\log(1 - D_\theta(s, s'))]\Big)$$



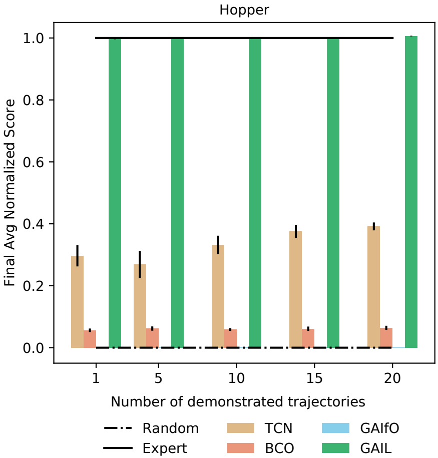
Experiments

Comparison against other IfO approaches and GAIL:



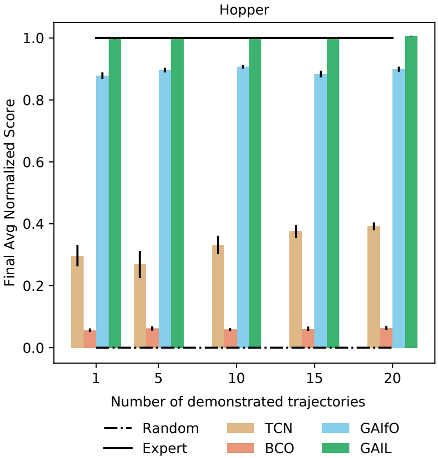
Experiments

Comparison against other IfO approaches and GAIL:



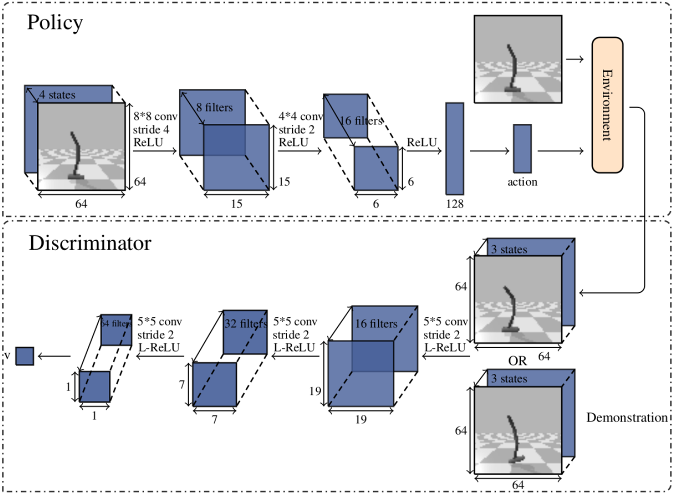
Experiments

Comparison against other IfO approaches and GAIL:



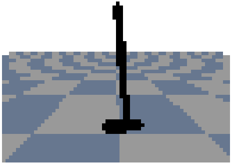
Algorithm

Visual States



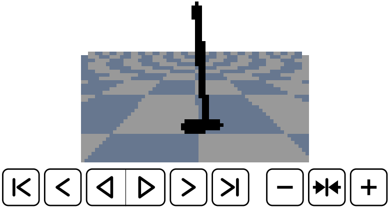
Experiments

Demonstration:

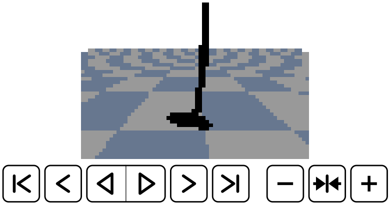


Experiments

Demonstration:

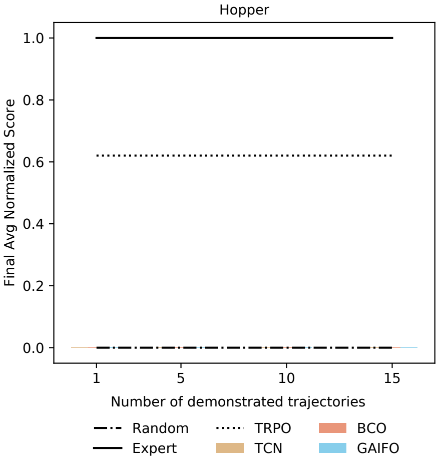


Learned Policy:



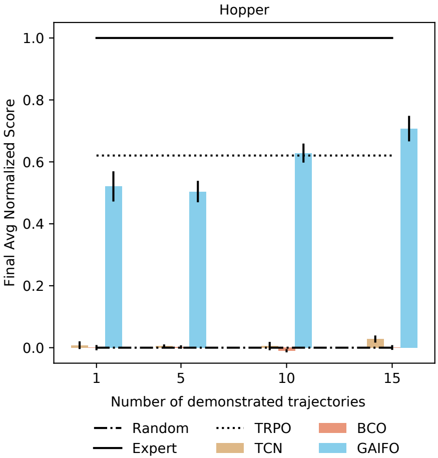
Experiments

Comparison against other IfO approaches:

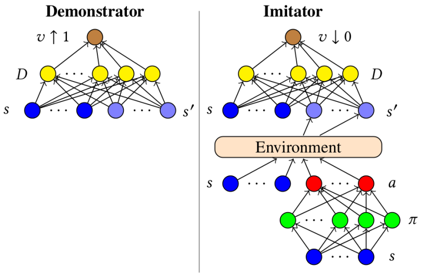


Experiments

Comparison against other IfO approaches:



Summary



Collaborators:

Peter Stone

Garrett Warnell