Generative Adversarial Imitation from Observation

Faraz Torabi\textsuperscript{1}, Garrett Warnell\textsuperscript{2}, and Peter Stone\textsuperscript{1}

\textsuperscript{1}The University of Texas at Austin, \textsuperscript{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:

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

- Inverse Reinforcement Learning (IRL)  
  - Guided Cost Learning.

- Generative Adversarial Imitation Learning.
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.
- 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.\textsuperscript{4}

\textsuperscript{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

⁴Scott 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:
  - $D_{demo} = (s_0, s_1, \ldots)$
- Learn:
  - $\pi : S \rightarrow A$
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.\(^7\)

---

\(^5\) 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.
Recover expert policy by

\[ \tilde{c} = \arg \max_{c \in \mathbb{R}^{S \times S}} -\psi(c) + \min_{\pi \in \Pi} E_{\pi} \left[ c(s, s') \right] - E_\pi E_{s, s'} \left[ c(s, s') \right] \]

\[ \tilde{\pi} = \arg \min_{\pi \in \Pi} E_{\pi} \left[ \tilde{c}(s, s') \right] \]

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

Recover expert policy by

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

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

Recover expert policy by

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

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

- \(c(s, s')\): cost as a function of state transition
- \(\pi_E\): expert policy
- \(\Pi\): set of all possible policies
- \(\psi(c)\): regularizer
Using a specific regularizer $\psi(c)$ results in:

$$\tilde{c} = \arg \max_{D \in (0, 1)^{S \times S}} E_{\pi} \left[ \log(D(s, s')) \right] + E_{\pi} E_{\pi} \left[ \log(1 - D(s, s')) \right]$$

$$\tilde{\pi} = \arg \min_{\pi \in \Pi} \max_{D \in (0, 1)^{S \times S}} E_{\pi} \left[ \log(D(s, s')) \right] + E_{\pi} E_{\pi} \left[ \log(1 - D(s, s')) \right]$$

- $D$: classifier (discriminator)
Generative Adversarial Imitation from Observation

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

$$\tilde{c} = \arg \max_{D \in (0,1)^{S \times S}} E_\pi [\log(D(s, s'))] + E_{\pi_E} [\log(1 - D(s, s'))]$$

- $D$: classifier (discriminator)
Generative Adversarial Imitation from Observation

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

$$\tilde{c} = \arg \max_{D \in (0,1)^{S \times S}} E_\pi [\log(D(s, s'))] + E_{\pi_E} [\log(1 - D(s, s'))]$$

$$\tilde{\pi} = \arg \min_{\pi \in \Pi} \max_{D \in (0,1)^{S \times S}} E_\pi [\log(D(s, s'))] + E_{\pi_E} [\log(1 - D(s, s'))]$$

- $D$: classifier (discriminator)
Algorithm

Low-dimensional States

- Initialize policy $\pi$
- While “Policy Improves”:
  - Execute $\pi$ and collect $\tau = \{(s, s')\}$
  - Update $D_\theta$ using loss
    $$- \left( E_\tau [\log(D_\theta(s, s'))] + E_\tau E_\tau [\log(1 - D_\theta(s, s'))] \right)$$
  - Update $\pi$ by TRPO with $r$
    $$- \left( E_{\tau E} [\log(1 - D_\theta(s, s'))] \right)$$
Experiments

Comparison against other IfO approaches and GAIL:

![Graph showing comparison between different methods](image)

- Random
- TCN
- GAlfO
- Expert
- BCO
- GAIL

Faraz Torabi

Generative Adversarial Imitation from Observation

UT Austin
Experiments

Comparison against other IfO approaches and GAIL:

![Comparison Graph]

- Final Avg Normalized Score for Hopper
- Number of demonstrated trajectories: 1, 5, 10, 15, 20
- Comparison methods: Random, TCN, BCO, GAIfO, GAIL
- Graph shows performance over different numbers of trajectories for each method.
Experiments

Comparison against other IfO approaches and GAIL:

![Bar Chart]

- **Hopper**
- **Number of demonstrated trajectories**
- **Final Avg Normalized Score**
- **Random**
- **Expert**
- **TCN**
- **BCO**
- **GAIfO**
- **GAIL**

Faraz Torabi
Generative Adversarial Imitation from Observation
UT Austin
Algorithm

Visual States

Policy

Discriminator
Experiments

Demonstration:
Experiments

Demonstration:

Learned Policy:
Experiments

Comparison against other IfO approaches:

![Graph showing comparison of different approaches]
Experiments

Comparison against other IfO approaches:

![Graph comparing different approaches for Hopper tasks](chart.png)

- Number of demonstrated trajectories: 1, 5, 10, 15
- Final Avg Normalized Score
- Approaches: Random, Expert, TRPO, TCN, BCO, GAIFO
Summary

Collaborators:

Peter Stone  Garrett Warnell