Generative Adversarial Imitation Learning (GAIL)

NIPS 2016

Stanford University
Jonathan Ho, Stefano Ermon

Qi Liu (ql8va), Hyun Jae Cho (hc2kc), Peng Wang (pw7nc)

November 3rd 2020

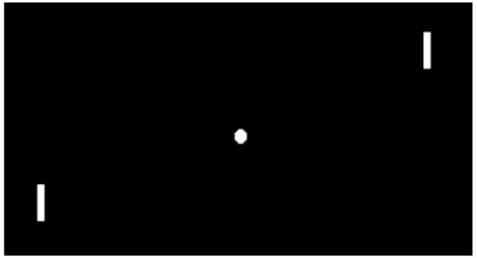
University of Virginia
Reinforcement Learning
Fall 2020

Background

- **Imitation Learning**
  - Learning to perform a task from expert demonstrations without a reward function.

Input: expert behavior generated by π_E

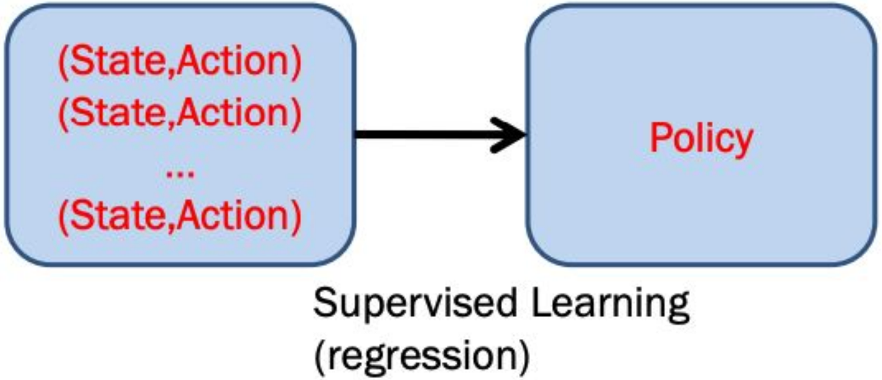
$$\{(s_0^i, a_0^i, s_1^i, a_1^i, \ldots)\}_{i=1}^n \sim \pi_E$$



Goal: learn *cost function (reward) or policy*

Background

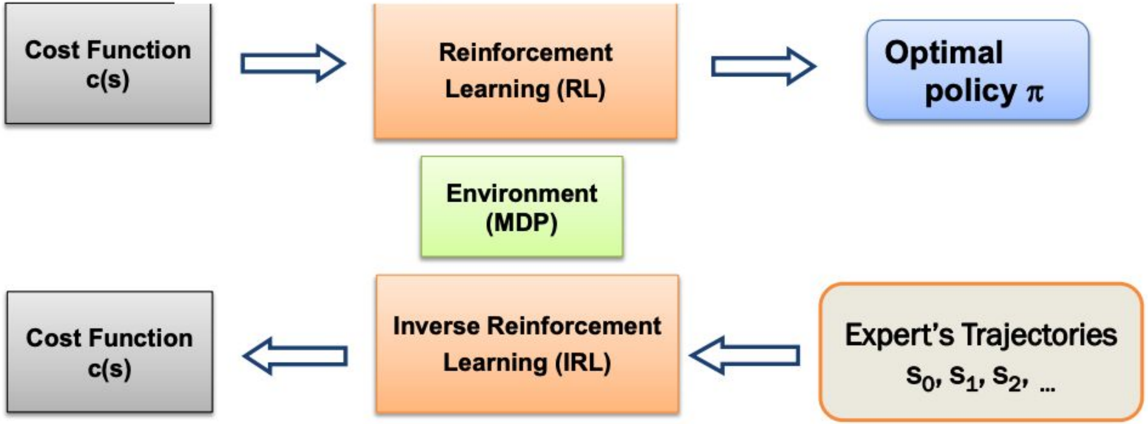
- **Behavioral Cloning**
  - Learning a policy as a supervised learning problem over state-action pairs from the expert trajectories.
- **Problems with Behavioral Cloning**
  - Small errors compound over time (cascading errors)
  - Decisions are purposeful (require planning)



Background

- **Inverse Reinforcement Learning**

$$\mathrm{RL}(c) = \arg\min_{\pi \in \Pi} -H(\pi) + \mathbb{E}_{\pi}[c(s,a)]$$



$$\underset{c \in \mathcal{C}}{\text{maximize}} \left(\min_{\pi \in \Pi} -H(\pi) + \mathbb{E}_{\pi}[c(s,a)] \right) - \mathbb{E}_{\pi_E}[c(s,a)]$$

(Ziebart et al., 2010;
Rust 1987)

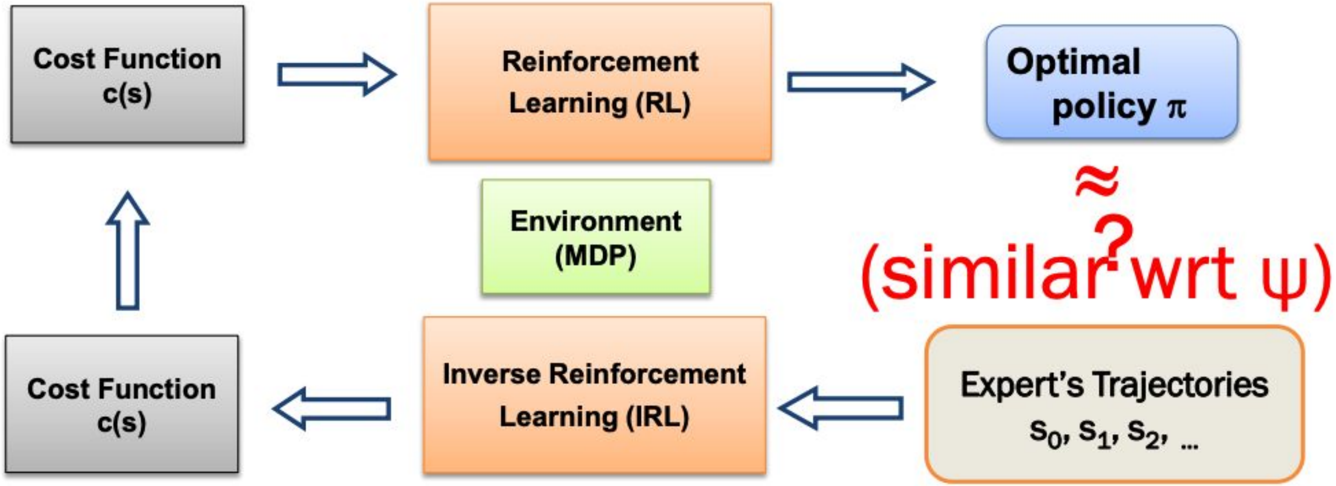
Everything else has high cost

Expert has small cost

15

Background

- **Inverse Reinforcement Learning**



$$\mathrm{IRL}_{\psi}(\pi_E) = \underset{c \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}}{\arg\max} \boxed{-\psi(c)} + \left(\min_{\pi \in \Pi} -H(\pi) + \mathbb{E}_{\pi}[c(s,a)] \right) - \mathbb{E}_{\pi_E}[c(s,a)]$$

Convex cost regularizer

Background

- **Problem with Inverse Reinforcement Learning**
  - Extremely expensive to run
  - Does not directly tell the learner how to act

- **So, the question is**
  - can we learn directly from the expert trajectories?

Review of Imitation Learning

- **Goal of imitation learning**
  - Let the agent behaves like the expert

- Can we do supervised learning over trajectories instead of (s,a) pair?
  - **Introduction of Generative Adversarial Imitation Learning (GAIL)**
  - Solution: use a more expressive class of cost functions

Can we directly compute the policy?

- IRL (**expert trajectories -> cost**) finds a cost function such that

$$IRL_{\psi}(\pi_E) = argmax_c - \psi(c) + \min_{\pi \in \Pi} (-H(\pi) + E_{\pi}[c(s,a)]) - E_{\pi_E}[c(s,a)]$$

- Reinforcement learning (**cost -> policy**)

- Expert trajectories -> policy?

Characterizing the policy

$$\mathrm{RL} \circ \mathrm{IRL}_{\psi}(\pi_E) = \arg\min_{\pi \in \Pi} -H(\pi) + \psi^*(\rho_\pi - \rho_{\pi_E})$$

ρ = occupancy measure

H = causal entropy (avoids overfitting ρ_{π} to ρ_{π_E})

Seeks a policy whose occupancy measure is close to the expert's, as measured by regularizer Ψ^*.

Apprenticeship learning

Def. The process of learning by observing an expert.

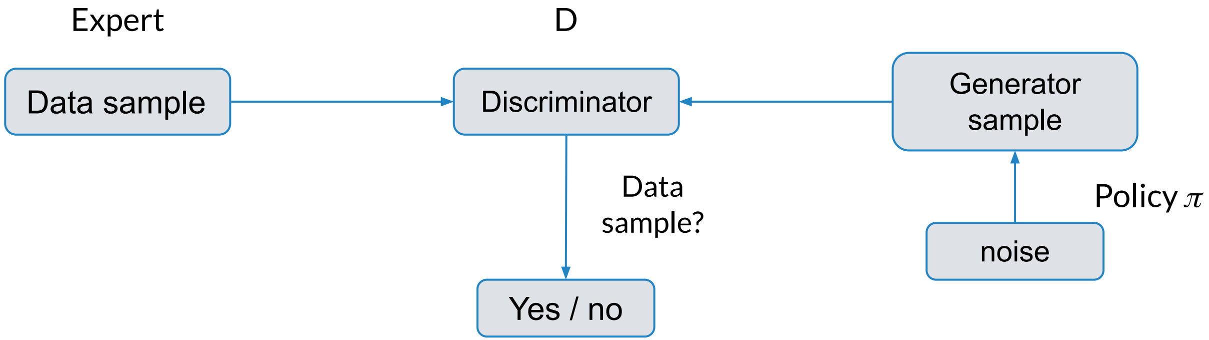
$$\underset{\pi}{\text{minimize}} \ \max_{c \in \mathcal{C}} \mathbb{E}_{\pi}[c(s,a)] - \mathbb{E}_{\pi_E}[c(s,a)]$$

Unless the true expert cost function lies in C, no guarantee that AL will recover the expert policy.

Generative Adversarial Imitation Learning (GAIL)

Discriminative classifier D tries to distinguish state-action pairs from the trajectories generated by π and π_E. Optimized by gradient descent.

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



GAIL Algorithm

Algorithm 1 Generative adversarial imitation learning

1: **Input:** Expert trajectories $\tau_E \sim \pi_E$, initial policy and discriminator parameters θ_0, w_0
2: **for** $i = 0, 1, 2, \dots$ **do**
3: Sample trajectories $\tau_i \sim \pi_{\theta_i}$
4: Update the discriminator parameters from w_i to w_{i+1} with the gradient

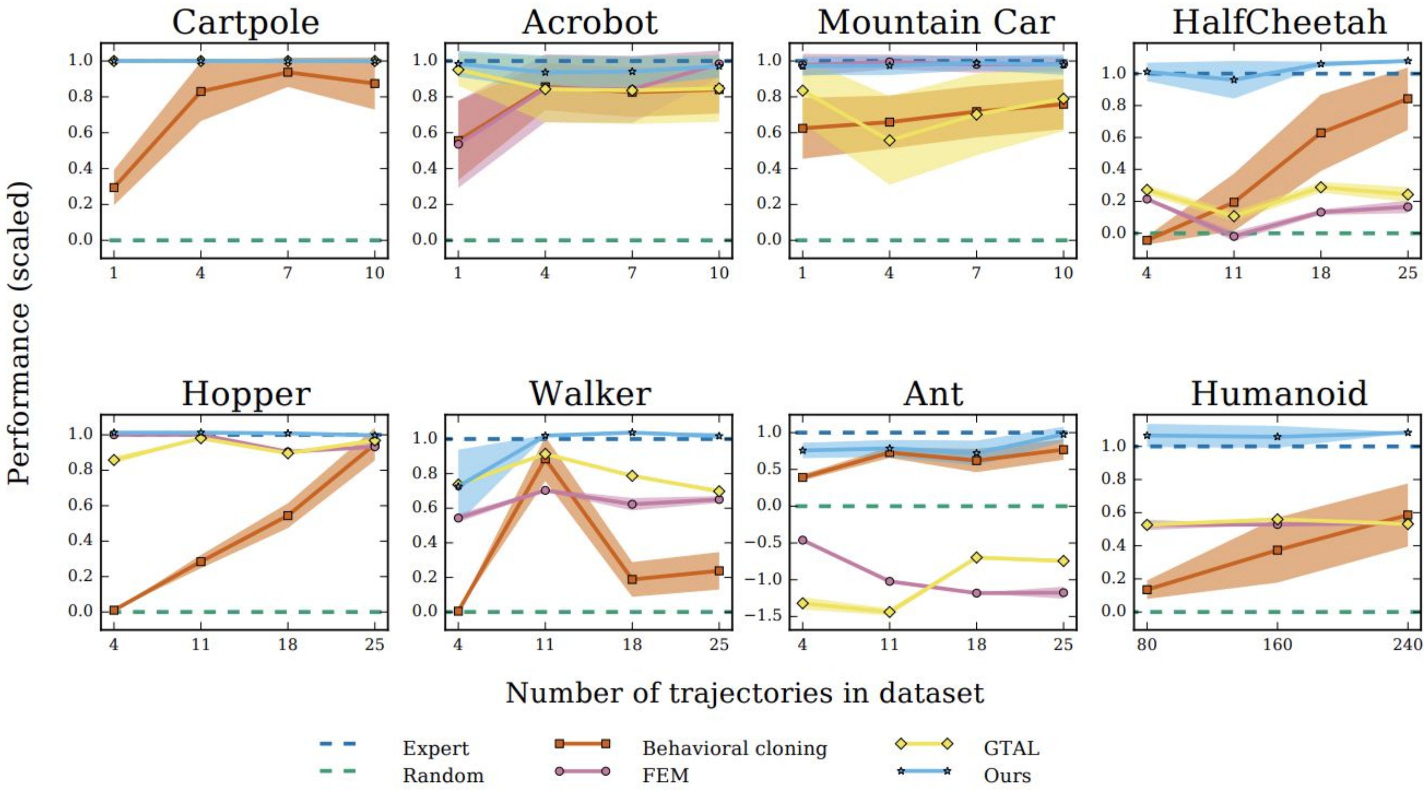
$$\hat{\mathbb{E}}_{\tau_i}[\nabla_w \log(D_w(s,a))] + \hat{\mathbb{E}}_{\tau_E}[\nabla_w \log(1 - D_w(s,a))] \tag{17}$$

5: Take a policy step from θ_i to θ_{i+1}, using the TRPO rule with cost function $\log(D_{w_{i+1}}(s,a))$. Specifically, take a KL-constrained natural gradient step with

$$\begin{aligned} &\hat{\mathbb{E}}_{\tau_i}\left[\nabla_\theta \log \pi_\theta(a|s) Q(s,a)\right] - \lambda \nabla_\theta H(\pi_\theta), \\ &\text{where } Q(\bar{s},\bar{a}) = \hat{\mathbb{E}}_{\tau_i}[\log(D_{w_{i+1}}(s,a)) \mid s_0 = \bar{s}, a_0 = \bar{a}] \end{aligned} \tag{18}$$

6: **end for**

Performance of GAIL



Shortcomings of GAIL

1. Assumes all demonstrations come from a single expert.
2. Needs lots of environment interactions.

InfoGAIL: Interpretable Imitation Learning from Visual Demonstrations

NIPS 2017

MIT, Stanford University
Yunzhu Li, Jiaming Song, Stefano Ermon

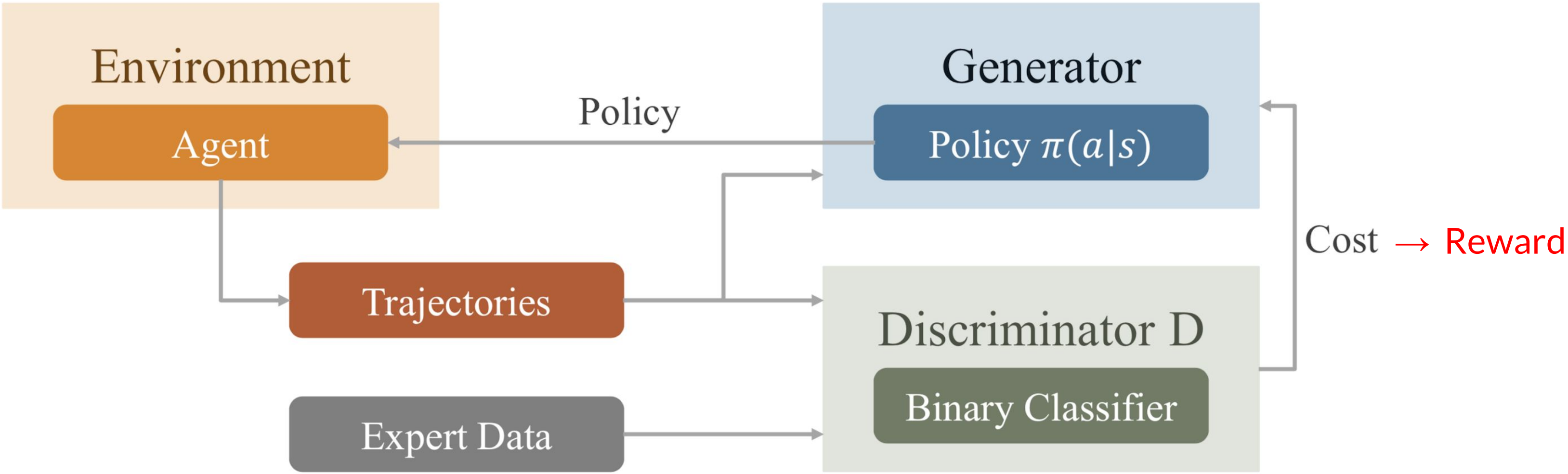
Qi Liu (ql8va), Hyun Jae Cho (hc2kc), Peng Wang (pw7nc)

November 3rd 2020

University of Virginia
Reinforcement Learning
Fall 2020

Recap: GAIL

A generator producing a policy π competes with a discriminator distinguishing π and the expert.



From GAIL to InfoGAIL

- GAIL
  - Expert demonstrations can show significant ***variability***.
  - Lack of ***external latent factors***.
- InfoGAIL
  - The goal of this paper is to develop an imitation learning framework that is able to autonomously discover and disentangle the latent factors of variation underlying human decision making.
  - Combines GAIL, InfoGAN (and Wasserstein GAN).

Motivation: From GAN to InfoGAN

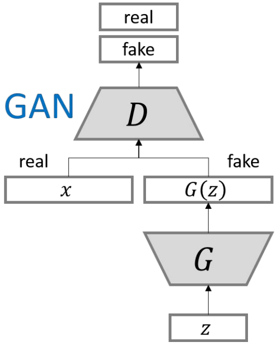
- Objective function of original GAN

$$\min_G \max_D V(D,G) = \mathbb{E}_{\boldsymbol{x} \sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z} \sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))]. \quad (1)$$

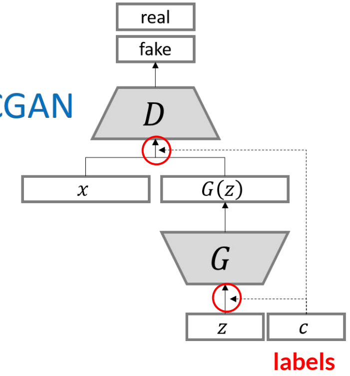
- Add conditional informations: CGAN (Conditional GAN)

$$\min_G \max_D V(D,G) = \mathbb{E}_{\boldsymbol{x} \sim p_{\text{data}}(\boldsymbol{x})}[\log \boxed{D(\boldsymbol{x}|\boldsymbol{y})}] + \mathbb{E}_{\boldsymbol{z} \sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1 - \boxed{D(G(\boldsymbol{z}|\boldsymbol{y}))})]. \quad (2)$$

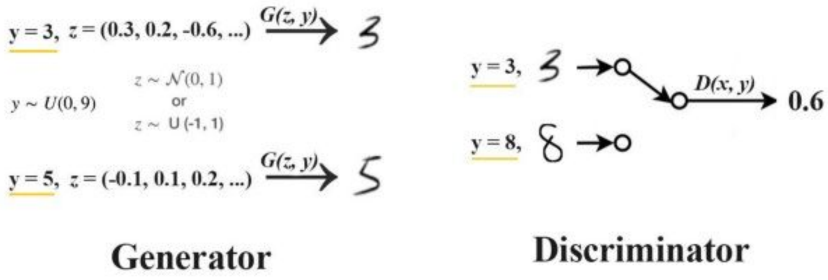
GAN



CGAN

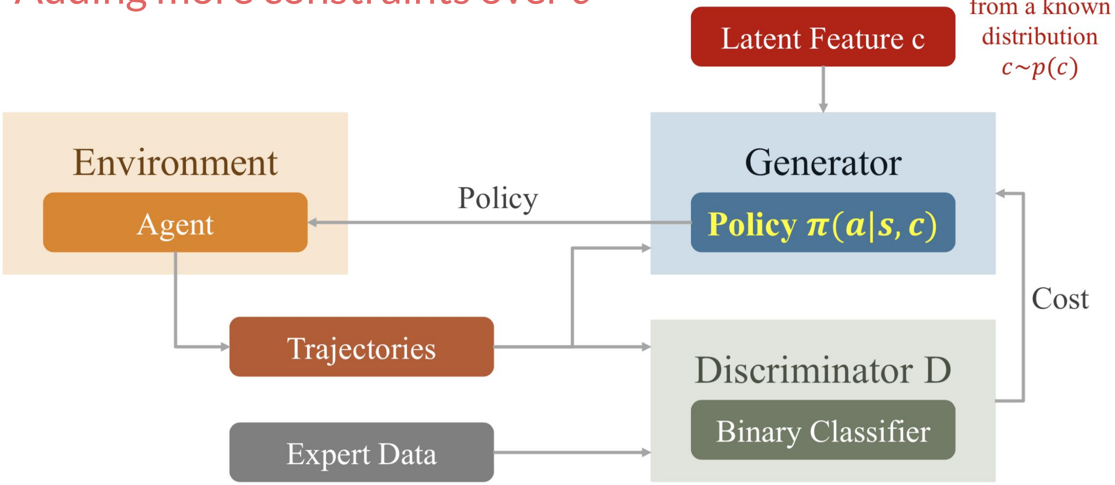


NOTE:
1. ***y*** comes from the labels provided in the training set, a supervised learning setting
2. ***y*** is fed into both the generator and the discriminator



Modified GAIL

- Try: Add latent feature c into policy π (Generator)
  - $\pi \rightarrow \pi(a \mid s, c)$
- Problem: GAIL could simply ignore c and fail to separate different types of behaviors present in the expert trajectories
  - Adding more constraints over c



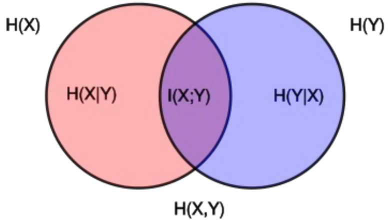
Motivation: From GAN to InfoGAN

- Objective function of original GAN

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim P_{\text{data}}}[\log D(x)] + \mathbb{E}_{z\sim \text{noise}}[\log(1 - D(G(z)))] \qquad (1)$$

- Mutual Information

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \qquad (2)$$



- Objective function of infoGAN
    - Theoretically

$$\min_G \max_D V_I(D,G) = V(D,G) - \lambda I(c; G(z,c)) \qquad (3)$$

Motivation: From GAN to InfoGAN

- Objective function of infoGAN
  - Theoretically

$$\min_G \max_D V_I(D,G) = V(D,G) - \lambda I(c; G(z,c)) \quad (3)$$

  - Variational mutual information maximization

$$\begin{aligned} I(c;G(z,c)) &= H(c) - H(c|G(z,c)) \\ &= \mathbb{E}_{x\sim G(z,c)}[\mathbb{E}_{c'\sim P(c|x)}[\log P(c'|x)]] + H(c) \\ &= \mathbb{E}_{x\sim G(z,c)}[\underbrace{D_{\mathrm{KL}}(P(\cdot|x) \parallel Q(\cdot|x))}_{\geq 0} + \mathbb{E}_{c'\sim P(c|x)}[\log Q(c'|x)]] + H(c) \\ &\geq \mathbb{E}_{x\sim G(z,c)}[\mathbb{E}_{c'\sim P(c|x)}[\log Q(c'|x)]] + H(c) \end{aligned} \quad (4)$$

Prior, Easy!

Lower bound

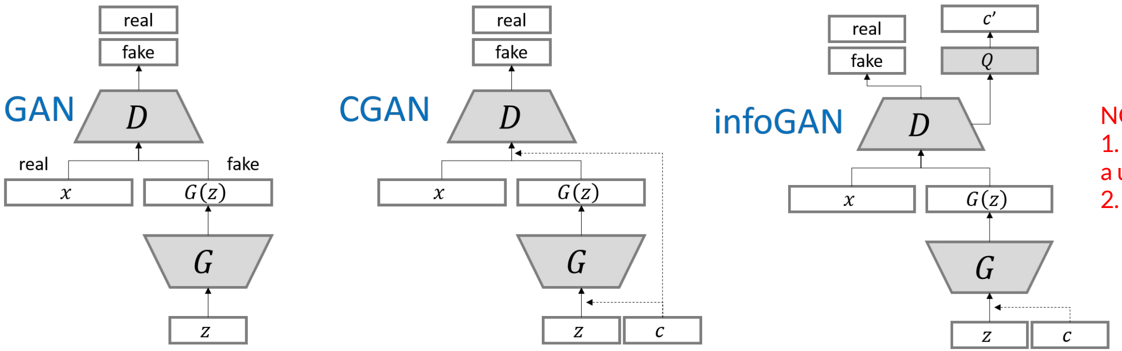
$$\begin{aligned} L_I(G,Q) &= \boxed{E_{c\sim P(c)}}{}_{,x\sim G(z,c)}[\log Q(c|x)] + \boxed{H(c)} \\ &= E_{x\sim G(z,c)}[\boxed{\mathbb{E}_{c'\sim P(c|x)}}[\log Q(c'|x)]] + H(c) \\ &\leq I(c; G(z,c)) \end{aligned} \quad (5)$$

Constant, Trivial

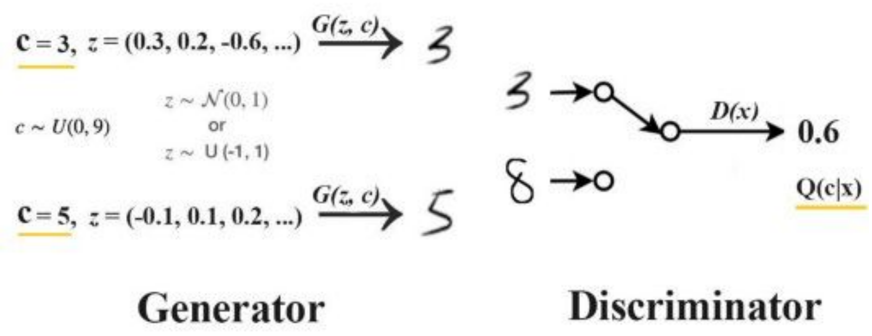
Posterior, Hard

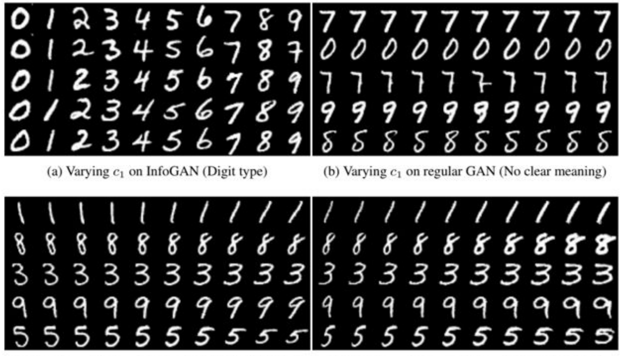
$$\min_{G,Q} \max_D V_{\mathrm{InfoGAN}}(D,G,Q) = V(D,G) - \lambda L_I(G,Q) \quad (6)$$

Motivation: From GAN to InfoGAN



NOTE:
1. ***c*** comes from prior distribution of latent codes, a unsupervised learning setting
2. ***c*** is fed only to the generator





(a) Varying c_1 on InfoGAN (Digit type)

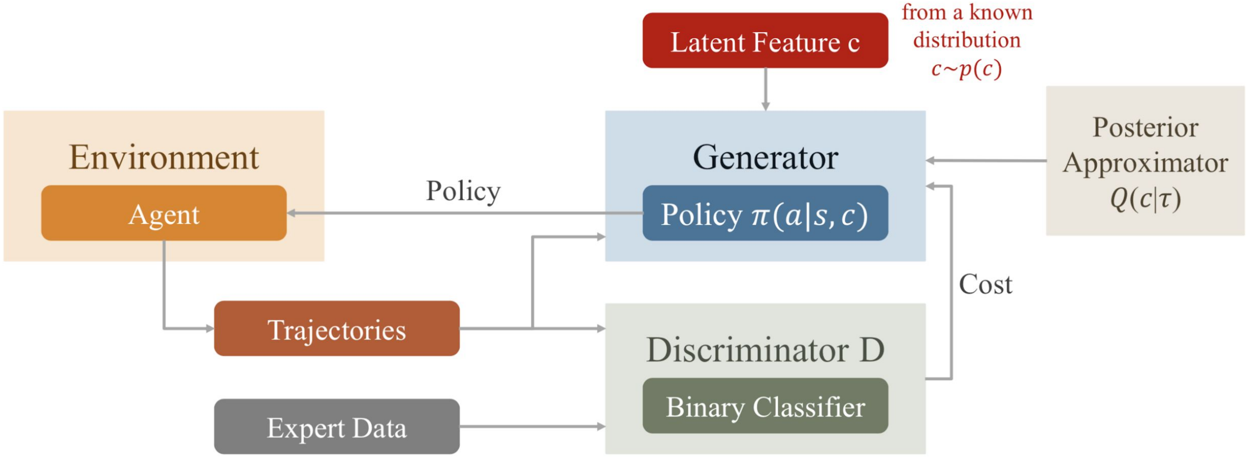
(b) Varying c_1 on regular GAN (No clear meaning)

(c) Varying c_2 from -2 to 2 on InfoGAN (Rotation)

(d) Varying c_3 from -2 to 2 on InfoGAN (Width)

InfoGAIL

- Similar to InfoGAN we applied these 2 extensions to GAIL:
  - Add latent feature c into policy $\boldsymbol{\pi}$.
  - Add $Q\,(c|\tau)$ to compute the mutual information.



Objective Functions

- GAIL

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

- InfoGAIL

$$\min_{\pi, Q} \max_{D} \underbrace{\boxed{\mathbb{E}_{\pi}[\log D(s,a)] + \mathbb{E}_{\pi_E}[\log(1 - D(s,a))]}}_{\text{Same as GAIL}} - \underbrace{\boxed{\lambda_1 L_I(\pi, Q)}}_{\text{Mutual information}} - \underbrace{\boxed{\lambda_2 H(\pi)}}_{\text{Same as GAIL}}$$

InfoGAIL Algorithm

Algorithm 1 InfoGAIL

Input: Initial parameters of policy, discriminator and posterior approximation $\theta_0, \omega_0, \psi_0$; expert trajectories $\tau_E \sim \pi_E$ containing state-action pairs.
Output: Learned policy π_θ
for $i = 0, 1, 2, ...$ **do**

Sample a batch of latent codes: $c_i \sim p(c)$
Sample trajectories: $\tau_i \sim \pi_{\theta_i}(c_i)$, with the latent code fixed during each rollout.

Sample data similar to InfoGAN

Sample state-action pairs $\chi_i \sim \tau_i$ and $\chi_E \sim \tau_E$ with same batch size.

Update ω_i to ω_{i+1} by ascending with gradients

$$\Delta_{\omega_i} = \hat{\mathbb{E}}_{\chi_i}[\nabla_{\omega_i} \log D_{\omega_i}(s, a)] + \hat{\mathbb{E}}_{\chi_E}[\nabla_{\omega_i} \log(1 - D_{\omega_i}(s, a))]$$

Update D similar to GAIL

Update ψ_i to ψ_{i+1} by descending with gradients

$$\Delta_{\psi_i} = -\lambda_1 \hat{\mathbb{E}}_{\chi_i}[\nabla_{\psi_i} \log Q_{\psi_i}(c|s, a)]$$

Update Q similar to InfoGAN

Take a policy step from θ_i to θ_{i+1}, using the TRPO update rule with the following objective:

$$\hat{\mathbb{E}}_{\chi_i}[\log D_{\omega_{i+1}}(s, a)] - \lambda_1 L_I(\pi_{\theta_i}, Q_{\psi_{i+1}}) - \lambda_2 H(\pi_{\theta_i})$$

Update policy using TRPO

end for

Experiments
Learning to Distinguish Trajectories

- Unsupervised learning task, similar to clustering
- Experiment details:
  - The observations at time *t* are positions from *t - 4* to *t*
  - The latent code is a one-hot encoded vector with 3 dimensions and a uniform prior

(a) Expert (b) Behavior cloning (c) GAIL (d) Ours

Experiments

Self-driving car in the TORCS Environment

- Run ***pass*** with different latent codes (left: 0, right: 1)
    - Pass on the left/right side

- Run ***turn*** with different latent codes (left: 0, right: 1)
    - Turn on the inside/outside lane

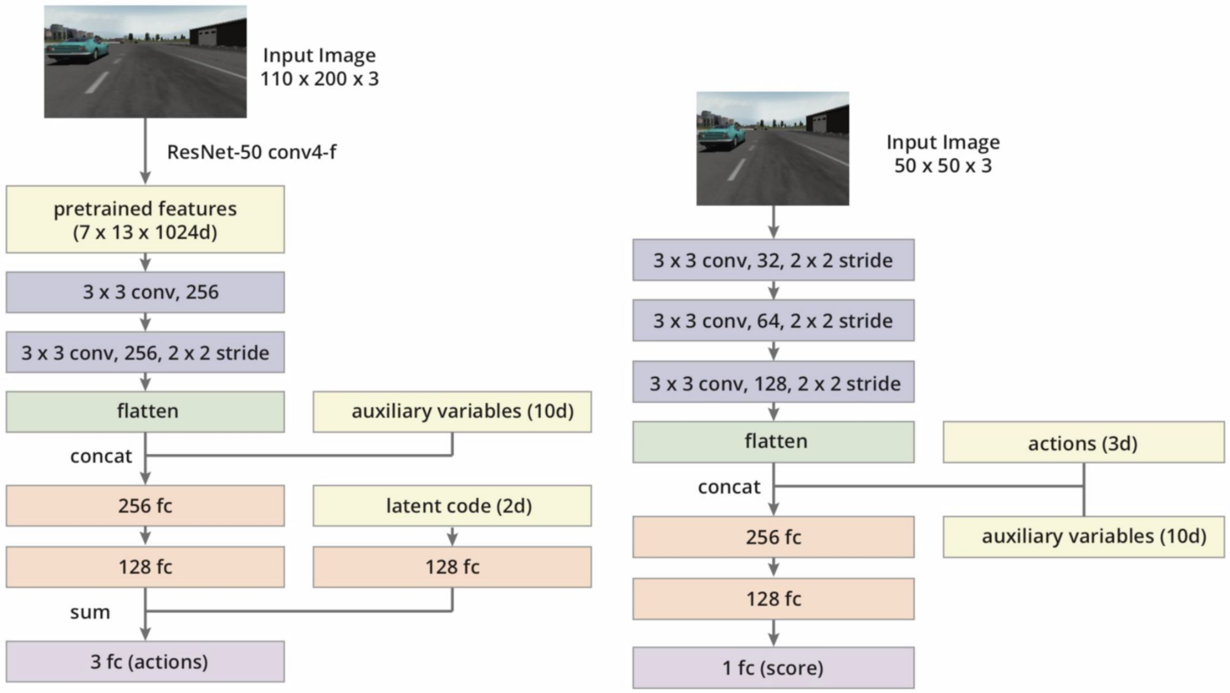
Experiments

Self-driving car in the TORCS Environment

- The demonstrations collected by manually driving
- 3-dimensional continuous action composed of *steering*, *acceleration* and *braking*
- Raw visual inputs as the only external inputs for the state
- Auxiliary information as internal input, including velocity at time *t*, actions at time *t-1* and *t-2*, and damage of the car
- Pre-trained ResNet on ImageNet

Experiments

Self-driving car in the TORCS Environment



(a) Network architecture for the policy/generator π_θ.   (b) Network architecture for the discriminator D_ω.

Experiments

Self-driving car in the TORCS Environment

Thanks!

Any questions?

References

GAIL:

- Ho, Jonathan and Ermon, Stefano. (2016). Generative Adversarial Imitation Learning.
- B. D. Ziebart, J. A. Bagnell and A. K. Dey, "The Principle of Maximum Causal Entropy for Estimating Interacting Processes," in IEEE Transactions on Information Theory, vol. 59, no. 4, pp. 1966-1980, April 2013, doi: 10.1109/TIT.2012.2234824.
- Calafiore, Giuseppe and El Ghaoui, Laurent. (2014). Optimization Models. Cambridge University Press.
- Generative Adversarial Imitation Learning, CVPR, 2018

InfoGAIL:

- Yunzhu Li, Jiaming Song, Stefano Ermon. (2017). InfoGAIL: Interpretable Imitation Learning from Visual Demonstrations.
- Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., & Abbeel, P. (2016). InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. In Advances in neural information processing systems (pp. 2172-2180).