Inverse reinforcement learning for video games

NIPS2019　B4 水谷純暉

概要

- AIRL(敵対的逆強化学習)にCNNを組み込んだ
- discriminatorの学習安定化のために
  - 報酬正規化
  - データセットサイズの変更
- AutoEncoderでの低次元状態表現
  - エキスパートデモンストレーションからのサンプリング効率の向上
- Catcherでは高性能
- Enduroなどの複雑な環境であるといまだ低性能

AIRLにCNNを組み込んだ

- AIRL
  - Adversarial Inverse Reinforcement Learning
  - 元論文：LEARNING ROBUST REWARDS WITH ADVERSARIAL INVERSE REINFORCEMENT LEARNING

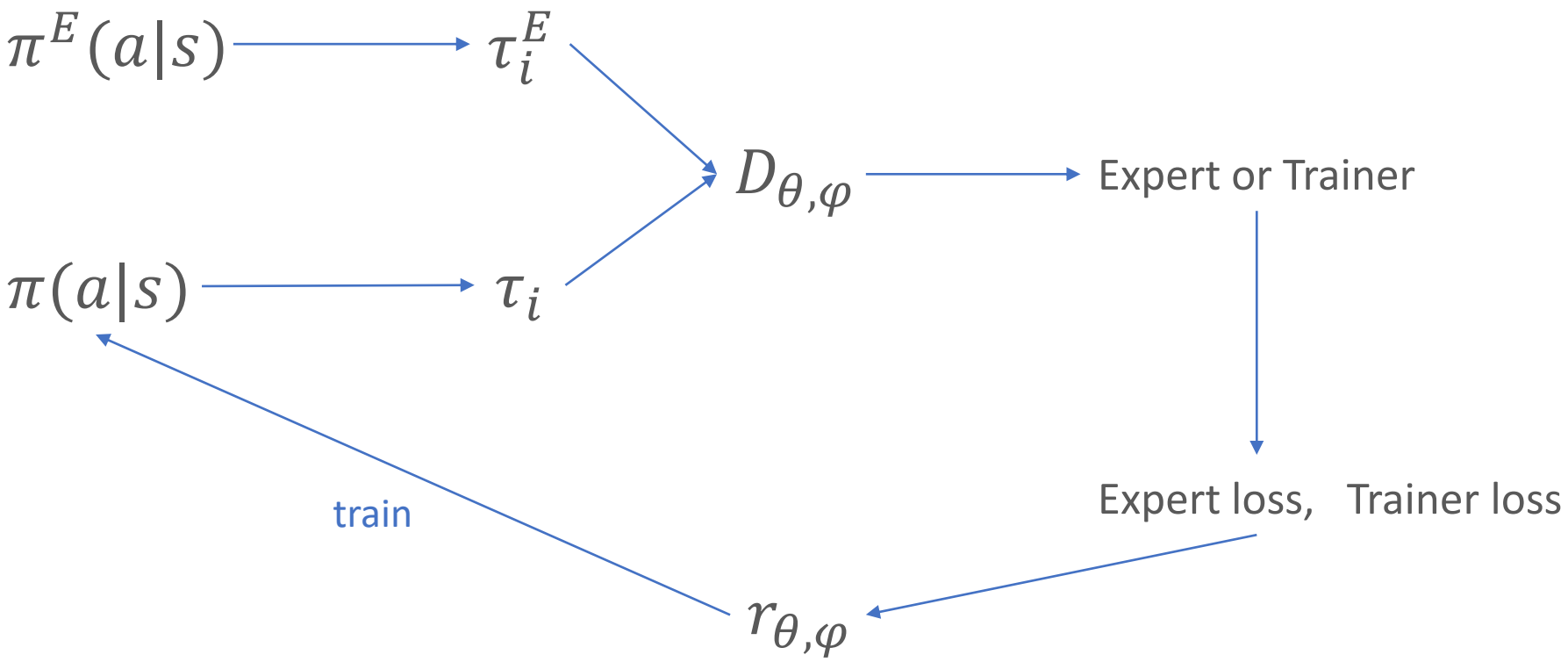
Algorithm 1 Adversarial inverse reinforcement learning

1: Obtain expert trajectories τ_i^E
2: Initialize policy π and discriminator $D_{\theta,\phi}$.
3: **for** step t in $\{1, \ldots, \mathrm{N}\}$ **do**
4: Collect trajectories $\tau_i = (s_0, a_0, ..., s_T, a_T)$ by executing π.
5: Train $D_{\theta,\phi}$ via binary logistic regression to classify expert data τ_i^E from samples τ_i.
6: Update reward $r_{\theta,\phi}(s, a, s') \leftarrow \log D_{\theta,\phi}(s, a, s') - \log(1 - D_{\theta,\phi}(s, a, s'))$
7: Update π with respect to $r_{\theta,\phi}$ using any policy optimization method.
8: **end for**

$$D_{\theta,\phi}(s, a, s') = \frac{\exp\{f_{\theta,\phi}(s, a, s')\}}{\exp\{f_{\theta,\phi}(s, a, s')\} + \pi(a|s)},$$

AIRLにCNNを組み込んだ

- AIRL

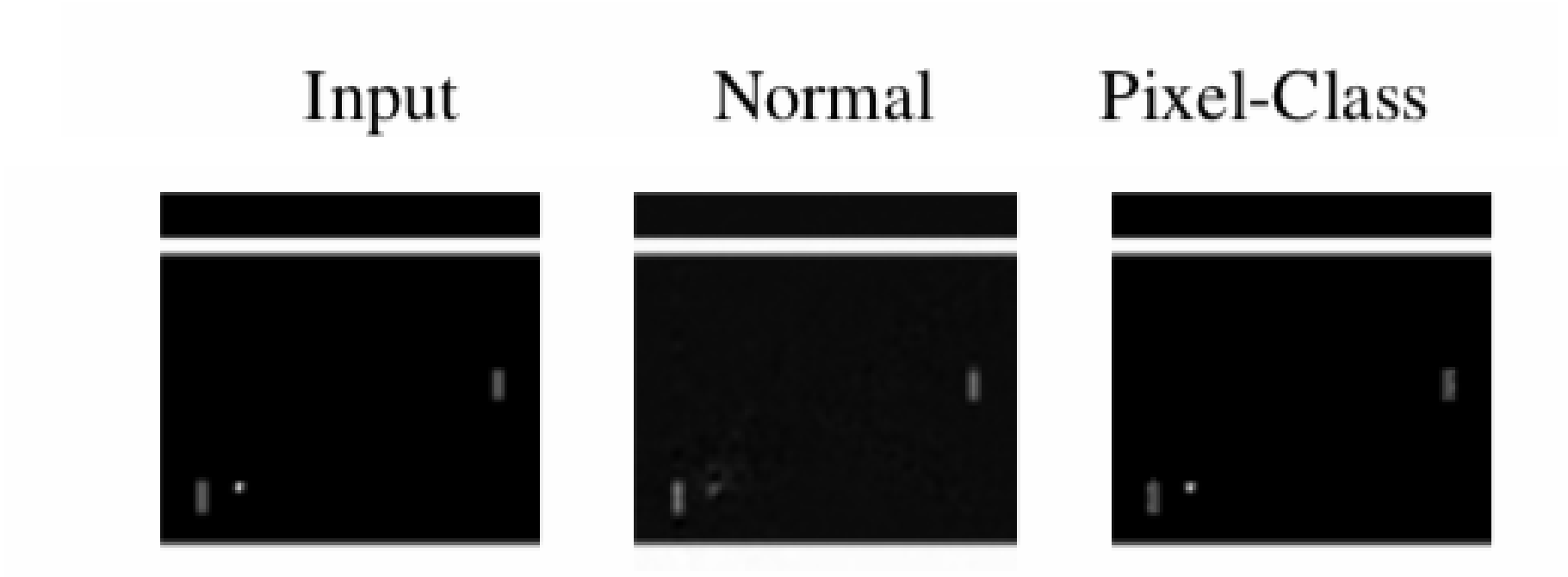


discriminatorの学習安定化のために

- discriminatorの学習に使われた軌跡の
  最後のサンプルに対する
  報酬の平均と標準偏差によってリスケーリング
- データセットサイズは、最小1024フレーム、最大16834フレーム

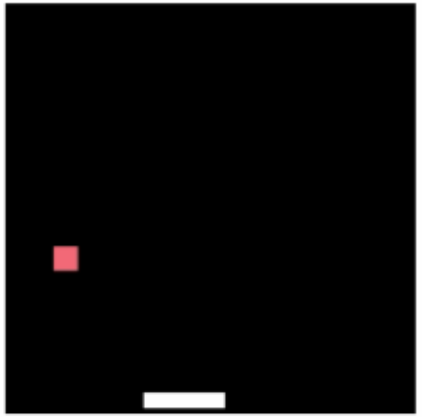
AutoEncoderでの低次元状態表現

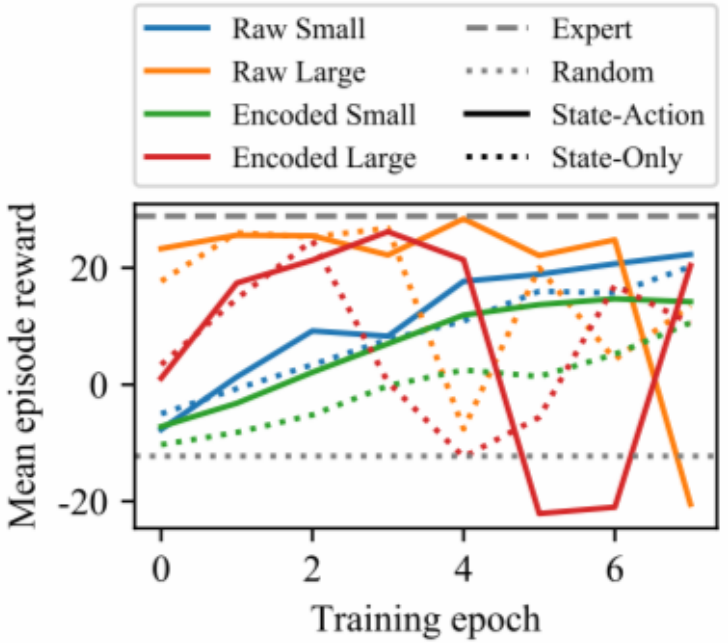
- 逆強化学習で利用するエキスパートの軌跡は、通常大量に用意することが難しい
- エキスパートの軌跡からのサンプリング効率を上げることを目的にAutoEncoderを利用する
- 低次元表現を例えると、Pongの場合、パドルとボールの座標と速度、現在のスコアを表すだけで十分
- pixel-class CNNを用いたAutoEncoderを使用



Catcherでは高性能

- Pixel-class AE によって、標準なAEよりもシャープに再構成できている
- 環境が単純すぎるため、低レベルな概念検証くらいにしか利用できない





(a) Screenshot from random exploration.

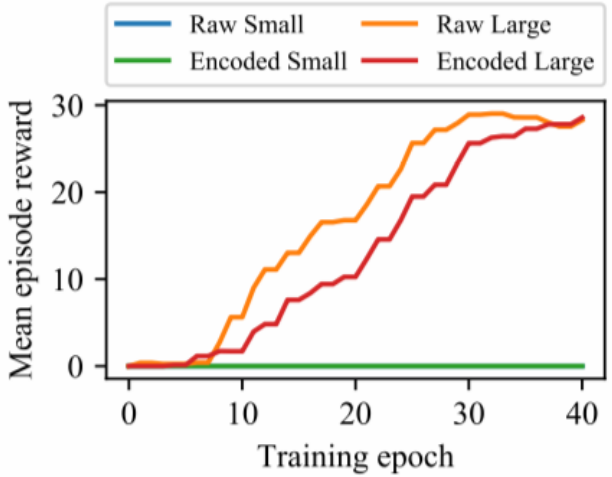
(b) Mean episode reward of IRL policy. Colors denote whether the discriminator input was raw images or an encoding, and whether the batch size was small or large. Solid or dashed line indicates if the discriminator also received an action input or just the state.

Figure 1: IRL on Catcher.

Enduroなどの複雑な環境であるといまだ低性能

- 複雑さ
  - 加速、ステアリング、衝突、道路の曲率
  - 背景と車両の両方が時間経過にとともに色が変わる
- AIRLの安定化にはさらなる改善が必要

(a) Screenshot from random exploration.



(b) Mean episode reward of IRL policy. Colors denote whether the discrimina tor input was raw images or an encoding, and whether the batch size was smal or large. In this test, the discriminator always receives an action input.

Figure 2: IRL on Enduro.