PUCRS
Pontifícia Universidade Católica
do Rio Grande do Sul

GPIN

Behavioral Cloning from Image Observation

Nathan S. Gavenski

Problem

The problem

- Given an environment, how can an agent learn how to interact and play it well?
  - Reinforcement Learning
  - Imitation Learning
  - Planner

Imitation Learning

- Given a set of expert unlabeled data, how can I train my model to mimic the movements and perform well in an environment?
- Learning from experience is commonly formulated as a Markov decision process

$$M = \{S, A, T, r, \gamma\}$$

- S: State ($[s_i, s_{i+1}, \dots, s_n]$)
- A: Action (a)
- T: The function denoting the probability of the agent transitioning from the state s_i to s_{i+1} after taking action a
- r: The function specifying the immediate reward after a specific action (a)
- γ: A discount factor for the reward

Hussein, Ahmed, et al. "Imitation learning: A survey of learning methods." *ACM Computing Surveys (CSUR)* 50.2 (2017): 21.

Imitation Learning

- Objective is to learn a Policy (π_θ) that is able to map states to actions and to discover the state dynamics

$$Policy : \pi_\theta(s) \rightarrow a$$
$$StateDynamics : P(s'|s, a)$$

- An episode will consist in a Rollout denoted by:

$$r = [(s_0, a_0), (s_1, a_1), \ldots, (s_n, a_n)]$$

Can be used as a classification and a regression problem

Yisong Yue and Hoang M. Le "Imitation Learning Tutorial", ICML 2018

Behavioral Cloning

- Can be seen as a reduction from Imitation Learning to Supervised Learning
- Now the policy will be our expert experiences and be denoted as π^*
- And our State Dynamics will be:

$$P^* = P(s|\pi^*)$$

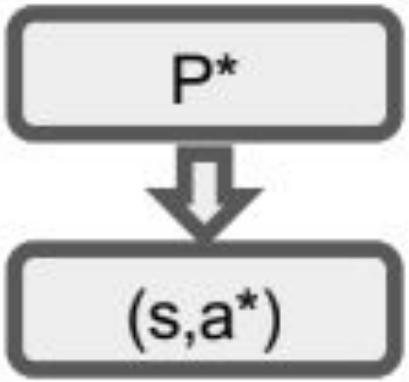
- And our learning objective will be:

$$argmin_\theta E_{(s,a^*)\ P} * L(a^*, \pi_\theta(s))$$

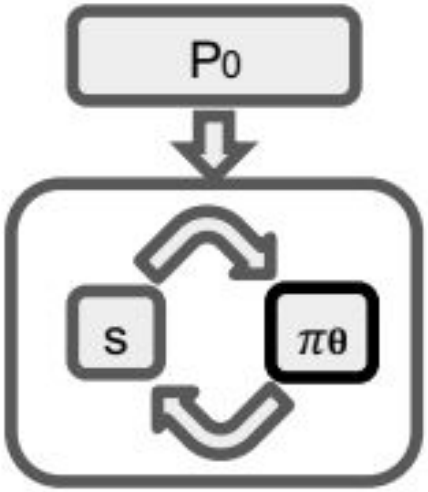
Yisong Yue and Hoang M. Le "Imitation Learning Tutorial", ICML 2018

Difference between BC vs IL

Behavioral Cloning

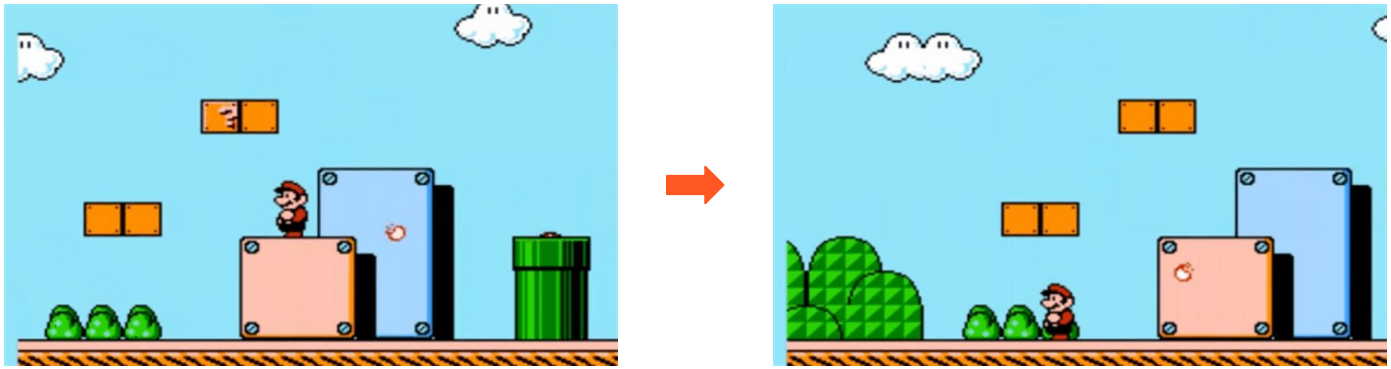


Imitation Learning

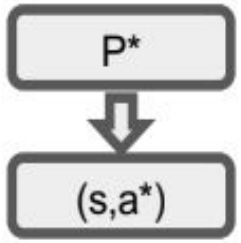


Yisong Yue and Hoang M. Le "Imitation Learning Tutorial", ICML 2018

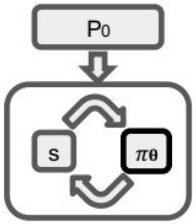
Difference between BC vs IL



BC



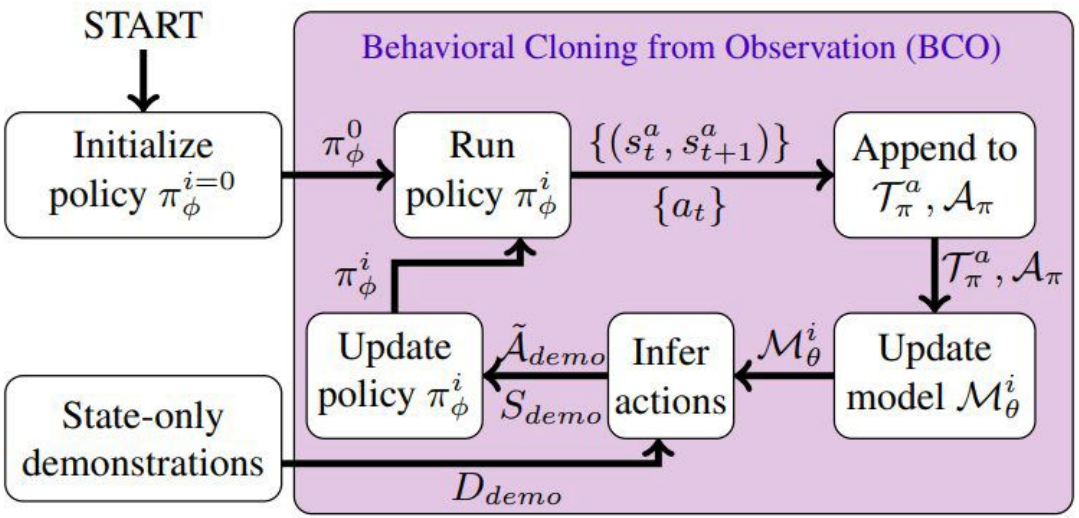
IL



If π_θ makes a mistake, the new state will not be at P*

Behavioral Cloning from Observation

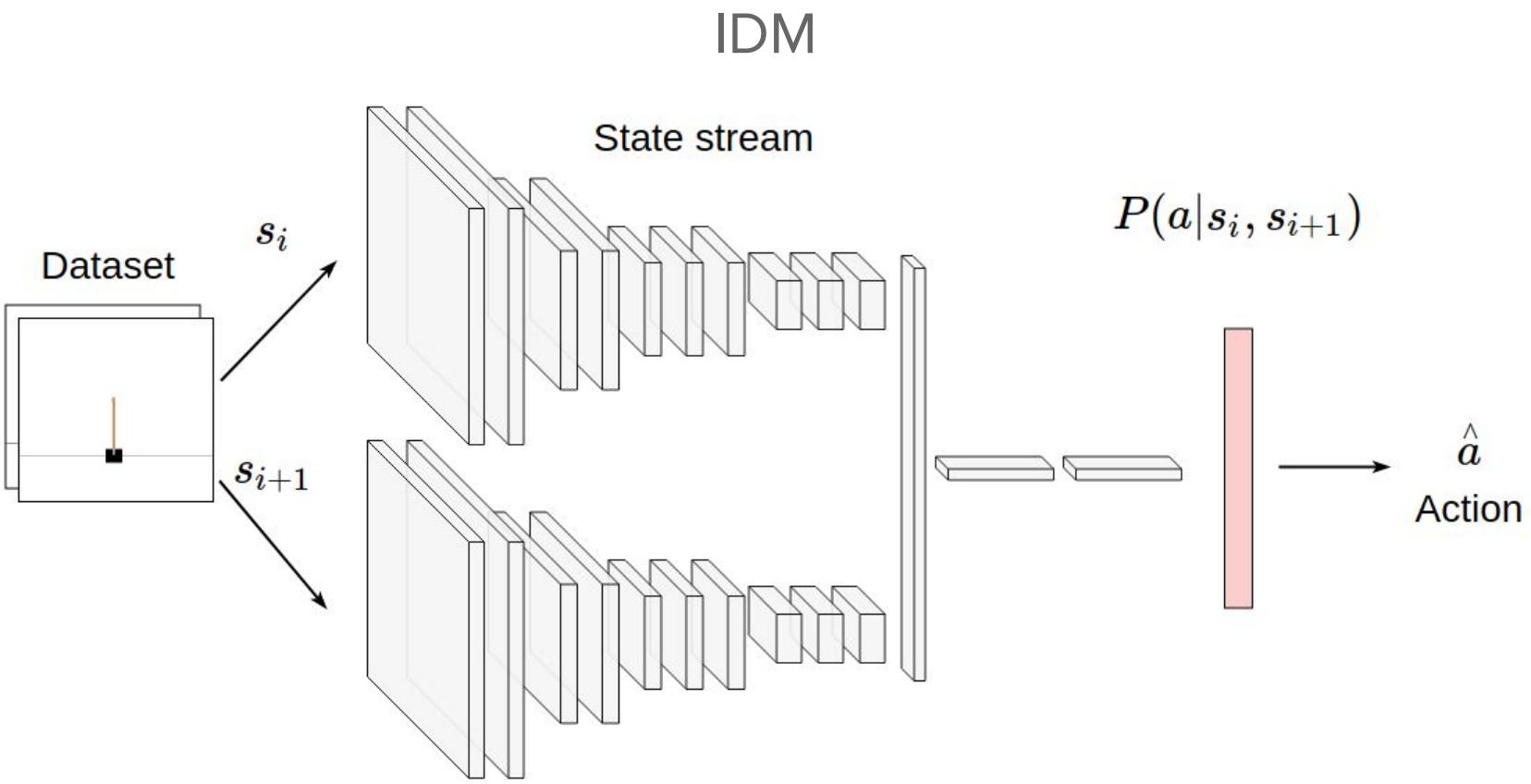
- Two different models:
  - Inverse Dynamic Model (IDM): Task-independent model
  - Policy Model: model responsible for mapping a state to an action



Torabi, Faraz, Garrett Warnell, and Peter Stone. "Behavioral cloning from observation." *arXiv preprint arXiv:1805.01954* (2018).

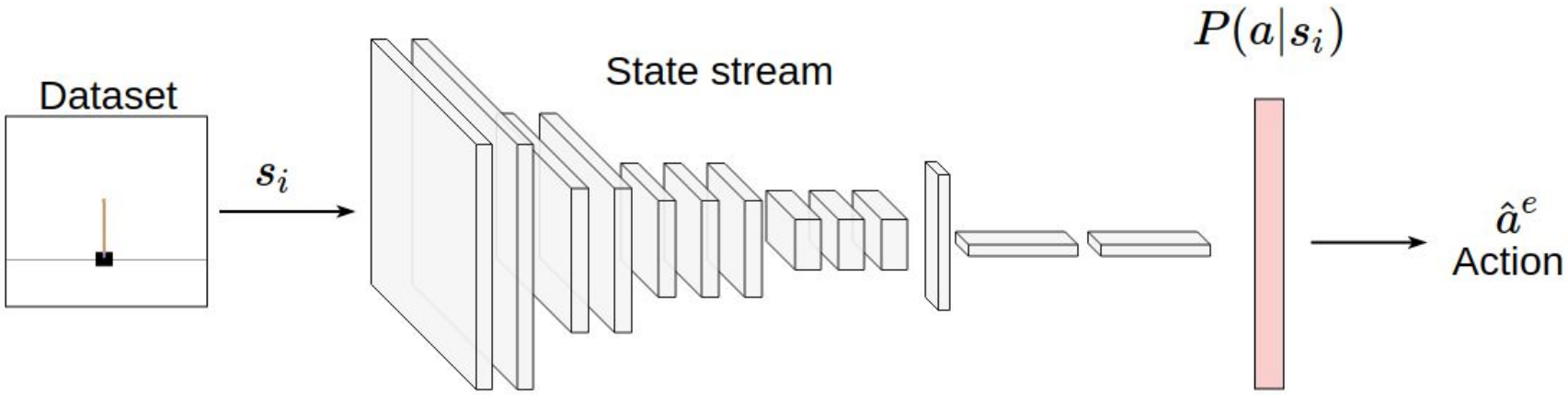
Proposal

The proposed model

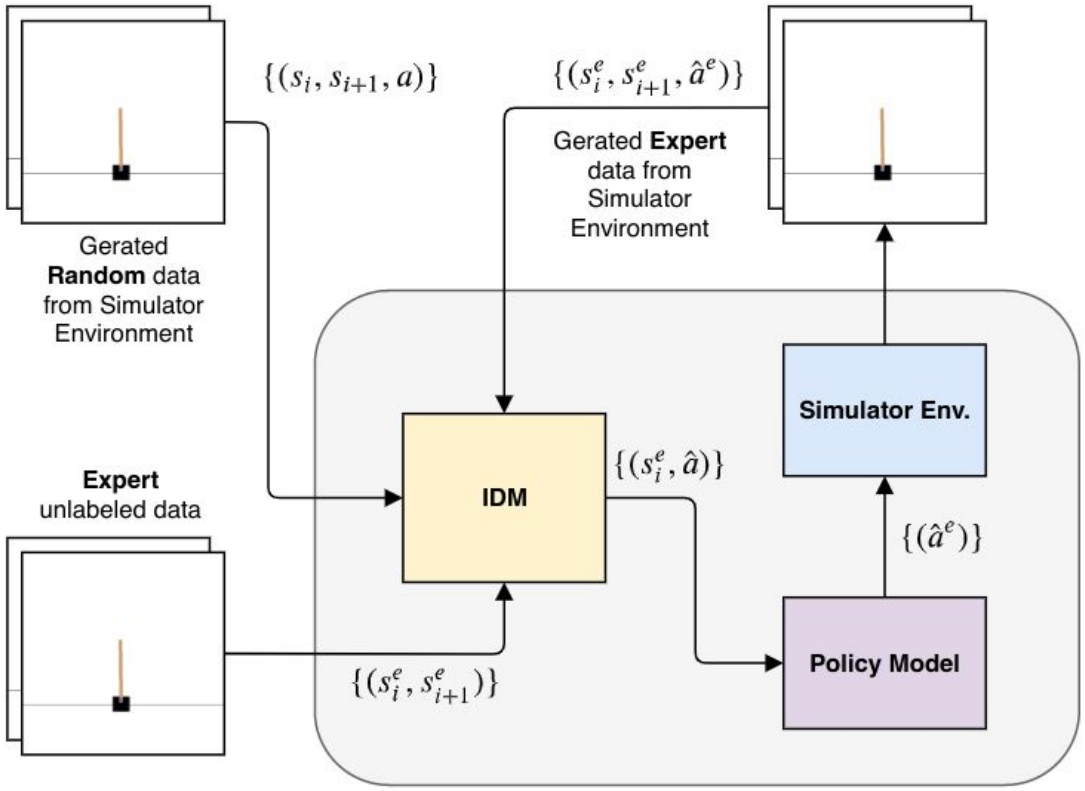


The proposed model

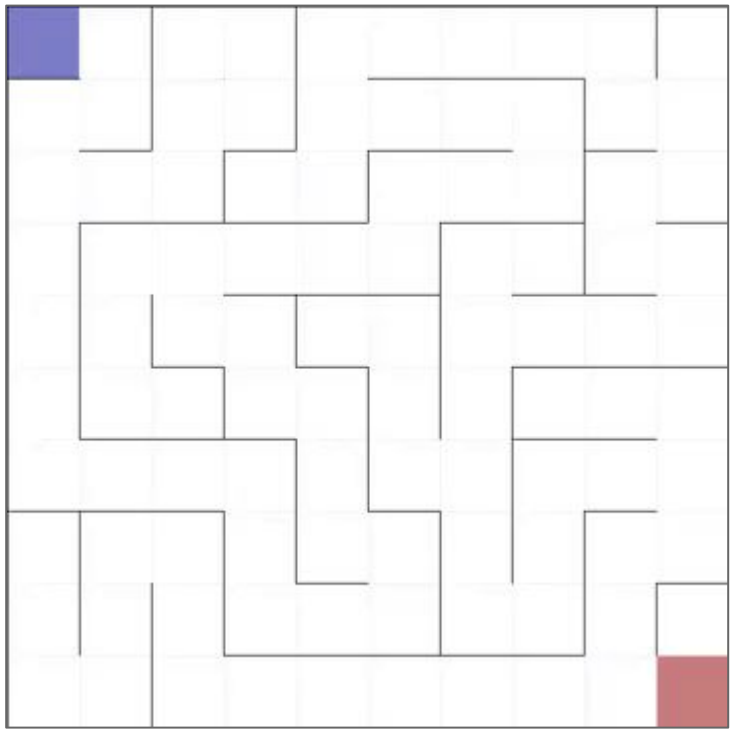
Policy



The proposed model

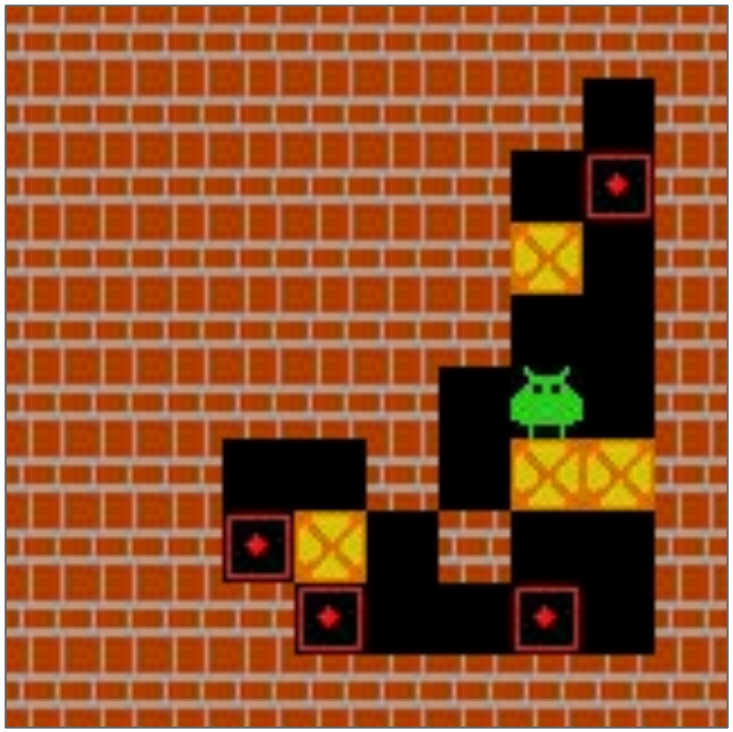


The environments



Maze

https://github.com/MattChanTK/gym-maze



Sokoban

https://github.com/mpSchrader/gym-sokoban

Measurements

- Performance: the reward is scaled so that the expertachieves one and a random policy achieves zero

- Average Episodic Reward: the average reward from *n* runs for each episode

Both measurements don’t show how well the model learned to imitate!

Project Management

Weeks

- Week 1: We need to create expert demonstrations foreach environment we intend to use and create one expert dataset using an algorithmic approach (e.g., Genetic).
- Week 2 and 3: We will train our models with the Maze and Sokoban environments and see how it performs with the same maps that the expert interacts with, and afterward, we will test it with randomly generated mazes and warehouses to see how well it generalized.
- Week 4: We will try to use the same IDM from the Maze environment to learn the Sokoban movement, and how well its knowledge of task-independent movement transfers from one game to another. Domain adaptation is not the objective of this project, but we will this as bonus.
- Week 5: We will analyze the results obtained and write the final essay for this class based on the findings

Thank you!