Intrinsic Motivation and Automatic Curricula via Asymmetric Self-Play

Sainbayar Sukhbaatar[1], Zeming Lin[2], Ilya Kostrikov[1], Gabriel Synnaeve[2], Arthur Szlam[2], Rob Fergus[12]
[1]New York University [2]Facebook AI Research

Motivation

Training an agent requires a lot of reward signals. But often external **rewards are expensive** to obtain. Can an agent learn about its environment without external reward?

Let the agent generate its own task and rewards!

Well defined game rules

No given game rule

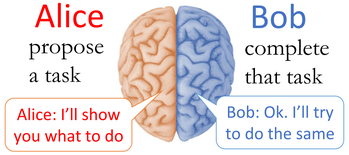
Train by self-play

Can't do self-play

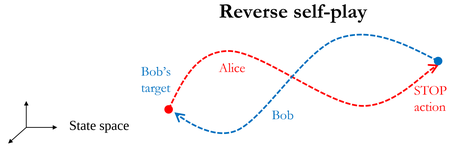
Invent a game rule?

Approach

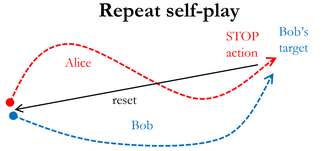
- Let the agent play an **imitation game** with itself
- Single agent, but **two** separate minds



Two versions of the game



First, Alice acts freely in the environment until she stops. Then, Bob takes control and asked to reach the initial state where Alice started.



The environment is **reset** after Alice finished. Then Bob is asked to reach the same state as Alice.

Internal reward during self-play

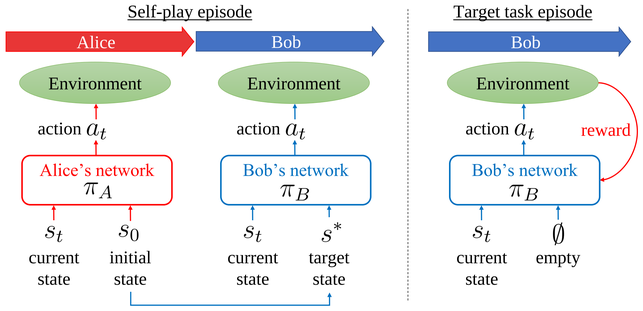
Time spent by Alice and Bob

Alice's reward $R_a = \max(0, t_b - t_a)$ Bob's reward $R_b = -t_b$

Alice wants to find simplest tasks where Bob struggles → Easier for Bob to learn since it's only just beyond his current capabilities. → Automatic curriculum

In the end, we want to learn a certain **target task**

Train on **mixture** of target task and self-play episodes



Related Work

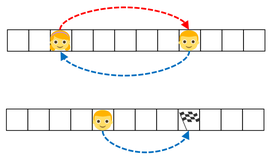
- Robust Adversarial RL (Pinto et al. 2017)
- Automatic Goal Generation (Held et al., 2017)
- Hindsight Experience Replay (Andrychowicz et al., 2017)
- Reverse Curriculum Generation (Florensa et al., 2017)

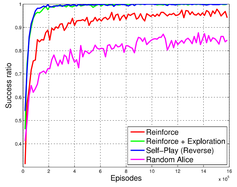
Experiments (code available: http://cims.nyu.edu/~sainbar/selfplay)

Toy example: Long hallway

Baselines: count-based exploration $\alpha/\sqrt{N_s}$

- Learn to navigate in a long corridor
- Reverse self-play
- Simple tabular policies

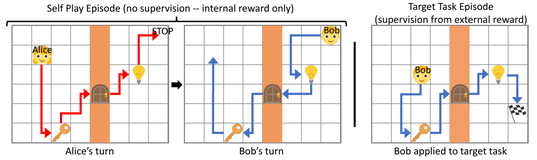


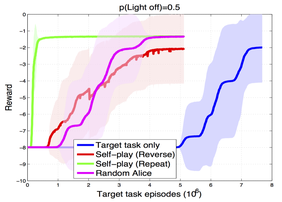


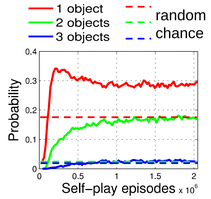
MazeBase: LightKey task

- Small 2D grid separated into two rooms by a wall
- The grid is procedurally generated (object/agent locations are randomized for each episode)

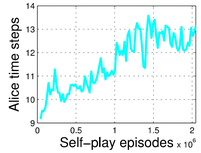
Target task is to reach the goal flag in the opposite room when light is off and door is locked.

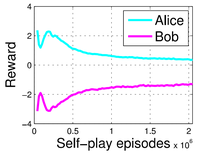






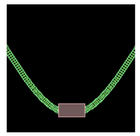
Alice learns to interact with more objects, taking more time.



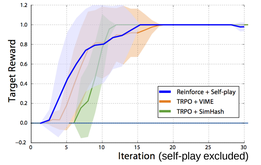


But her reward decreases as Bob learns the proposed tasks.

Box2D: MountainCar

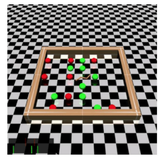


- The target task is to reach mountain top
- Bob's goal in self-play is to reach the same state as Alice (position + velocity)

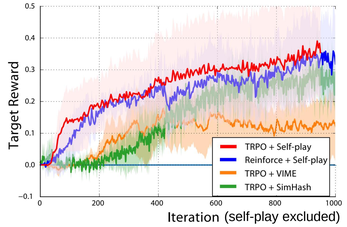


Baselines: VIME (Houthooft et al., 2016), SimHash (Tang et al., 2017)

Mujoco: SwimmerGather

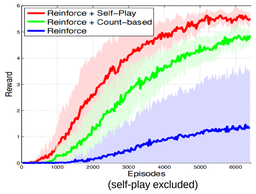


- The target task is to swim and eat green apples
- Bob's goal in self-play is to return to Alice's initial position (reverse self-play)



StarCraft: Build Marines

- Control multiple units
- The target task is to build marines in given time
- Bob's goal in self-play is to build as much stuff as Alice (ignore positions)



Conclusion & Future directions

- An intrinsic motivation method for learning transitions between states
- Works with discrete and continuous environments
- A novel way to use self-play in a single agent environments
- In future: self-play in abstract state space, option discovery, different game