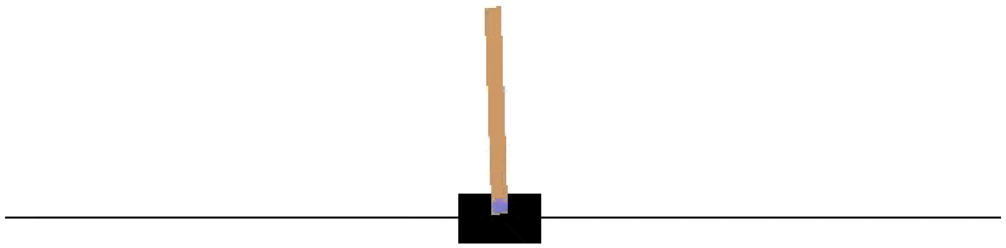


Deep Reinforcement Learning: REINFORCE vs A2C

Bryan Van Draanen

github.com/bryanvandraanen/Deep-Reinforcement-Learning

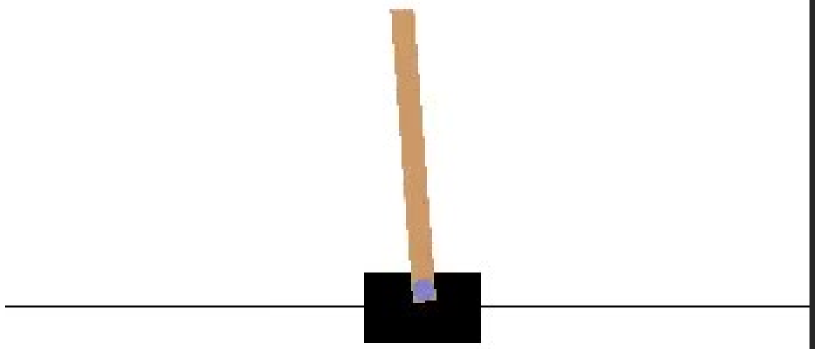
Motivation

- OpenAI Gym provides perfect abstraction to create this agent
- REINFORCE and A2C policy-based reinforcement learning methods[1]
  - Directly optimizes the policy without relying on a state value function
- REINFORCE optimizes policy based on total reward from episode
- A2C combines REINFORCE with “critic”
- Critic approximates the value function to evaluate each action taken by the agent using “advantage”

OpenAI

Definition

- Implement REINFORCE and A2C algorithms[2,3]
  - *"We have not seen any evidence that the noise introduced by [A3C] provides any performance benefit"*[4]
- Train and test agents on OpenAI Gym "CartPole-v1"[5]



Observation (input data)
Cart Position
Cart Velocity
Pole Angle
Pole Velocity at Tip

Actions (output space)
Push cart left
Push cart right

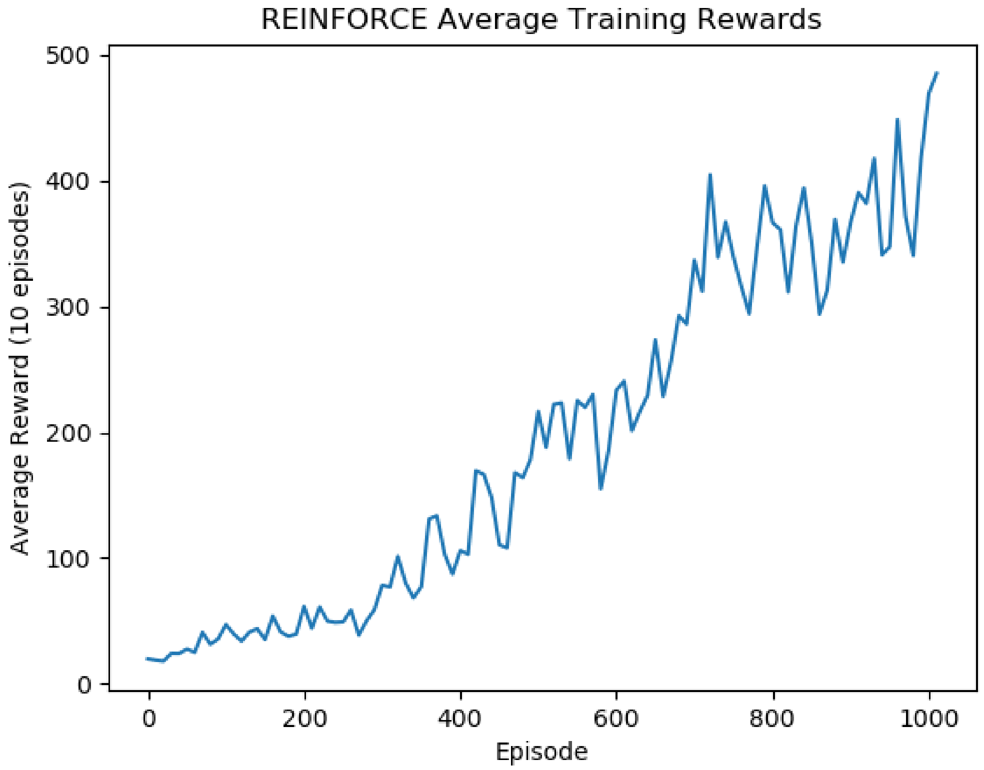
  - Reward is 1 for every transition taken
  - Episode ends if 500 steps taken, pole angle > 12˚, or cart travels out of bounds
- Generalize agents to common reinforcement learning abstraction

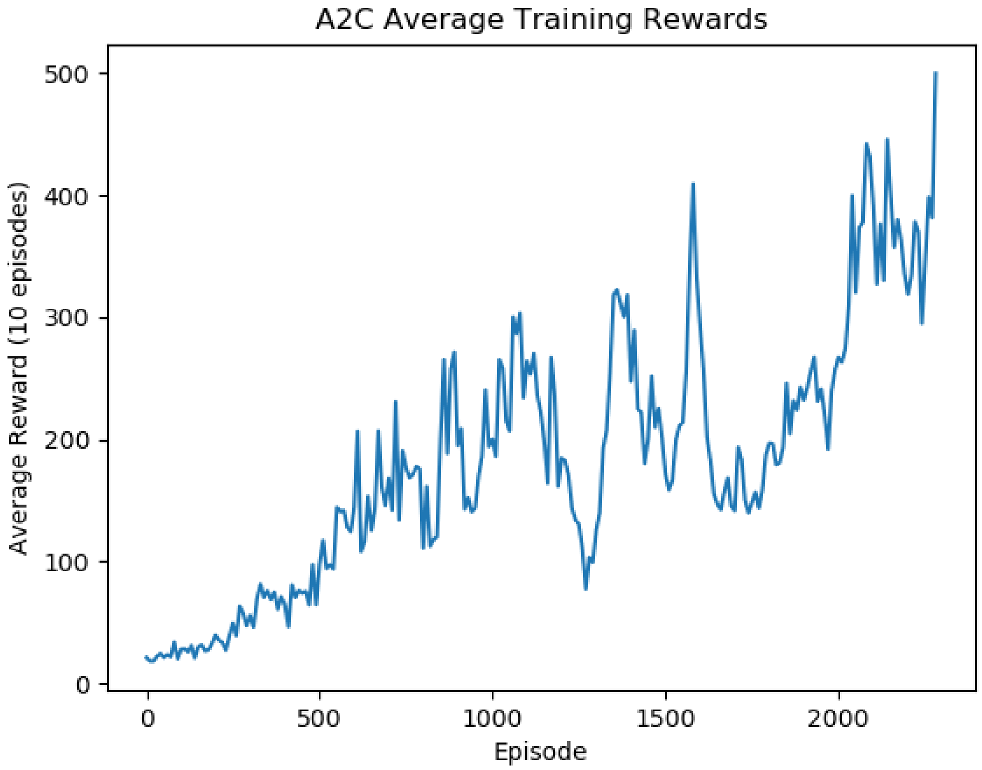
Approach

- Researched REINFORCE and A2C algorithms
- Implemented REINFORCE then A2C using PyTorch
  - Designed implementation to share same “actor” network between algorithms
- Evaluated each algorithm on OpenAI Gym CartPole-v1
- Trained each agent until reached proficiency
  - Proficiency reached when average reward over 10 episodes exceeds environment reward threshold
- Performed final performance evaluation of each agent acting on optimal policy for 100 episodes

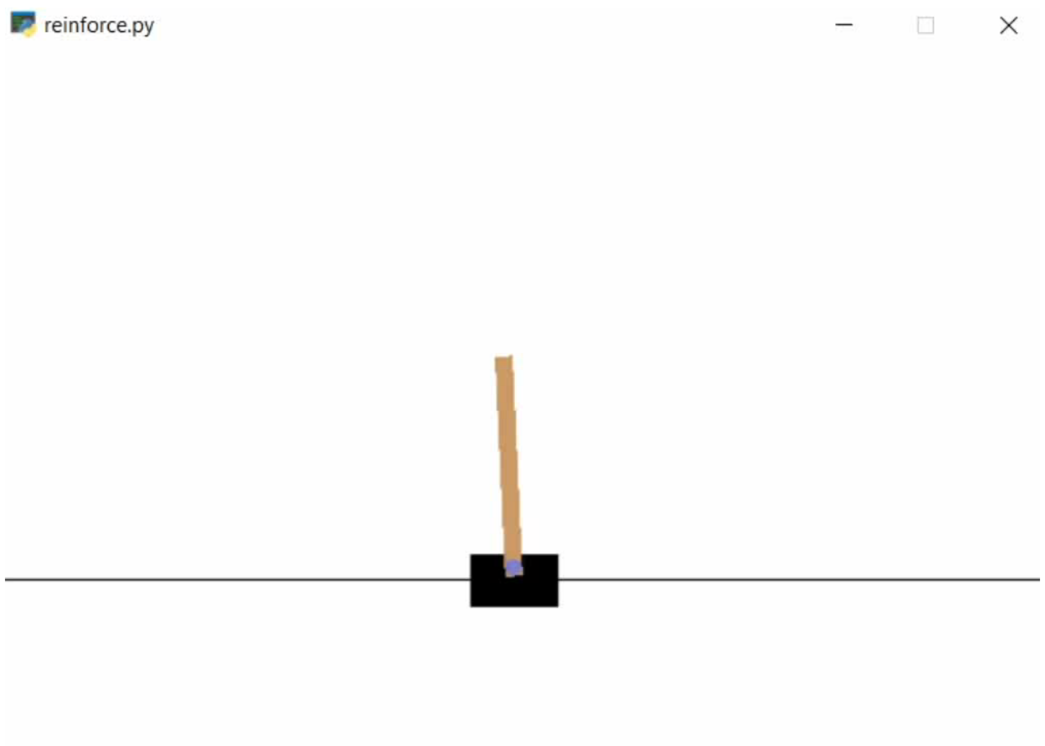
Results

Average reward over past 10 episodes during training until proficient

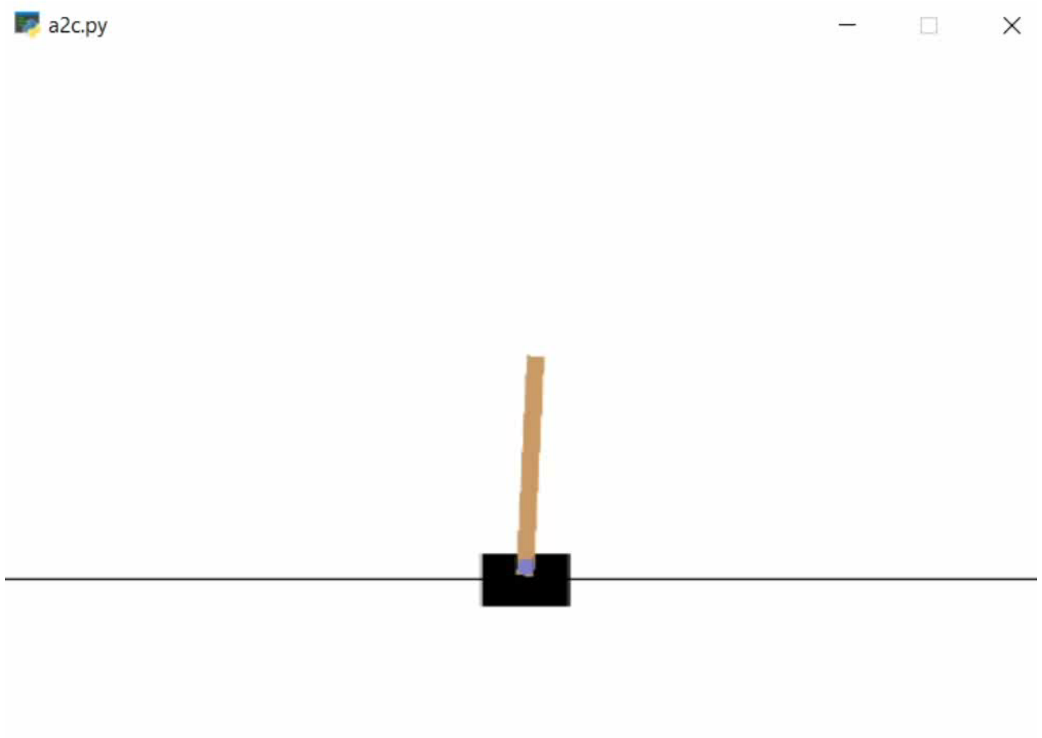




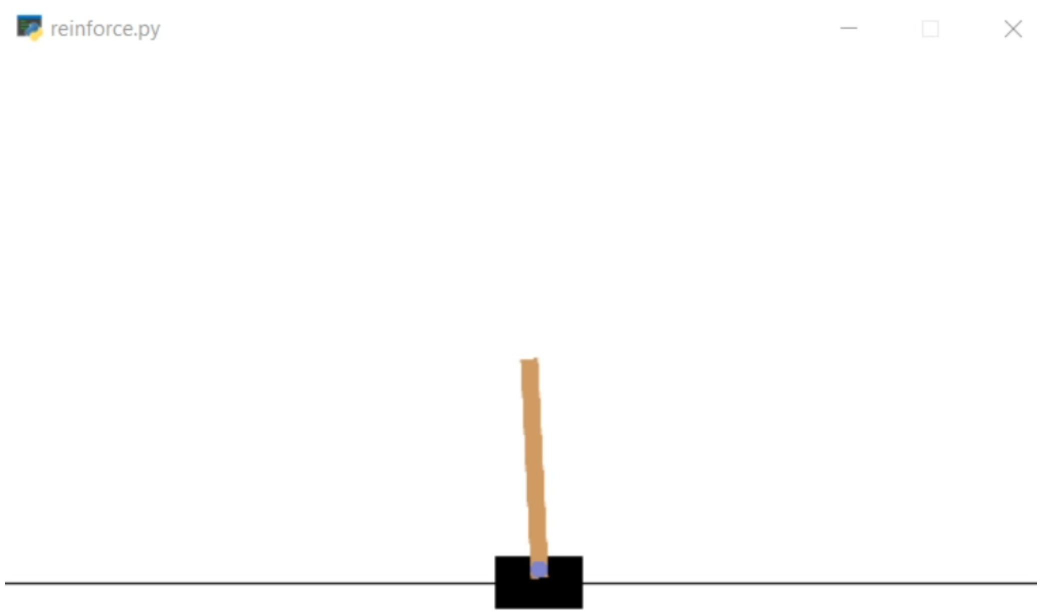
REINFORCE Start Training



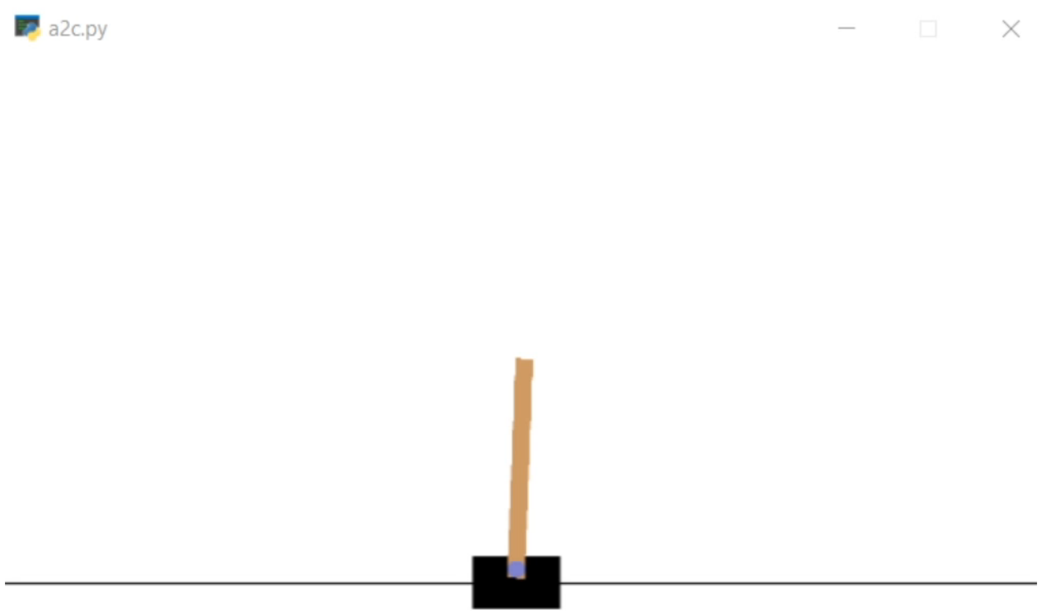
A2C Start Training



REINFORCE once Proficient



A2C once Proficient



Analysis

- Both REINFORCE and A2C perform flawlessly during final evaluation
- A2C requires longer to train to achieve proficiency
  - Trains same REINFORCE network in addition to critic network
  - Further hyperparameter tuning may give more effective training configuration
- Policy gradient methods are better for continuous action spaces
  - CartPole has two discrete actions
  - Plans to incorporate both deep reinforcement learning methods into continuous environment for comparison in future

References

1. medium.com/free-code-camp/an-intro-to-advantage-actor-critic-methods-lets-play-sonic-the-hedgehog-86d6240171d
2. papers.nips.cc/paper/1713-policy-gradient-methods-for-reinforcement-learning-with-function-approximation.pdf
3. arxiv.org/pdf/1602.01783.pdf
4. openai.com/blog/baselines-acktr-a2c/
5. gym.openai.com/envs/CartPole-v1/