Proximal Policy Optimization:
A State-of-the-art Reinforcement Learning Algorithm

By Tokey Tahmid

Test Questions

1. What are policy gradients?

2. What is the key concept of Proximal Policy Optimization (PPO)?

3. What are your thoughts on the PPO algorithm after the presentation?

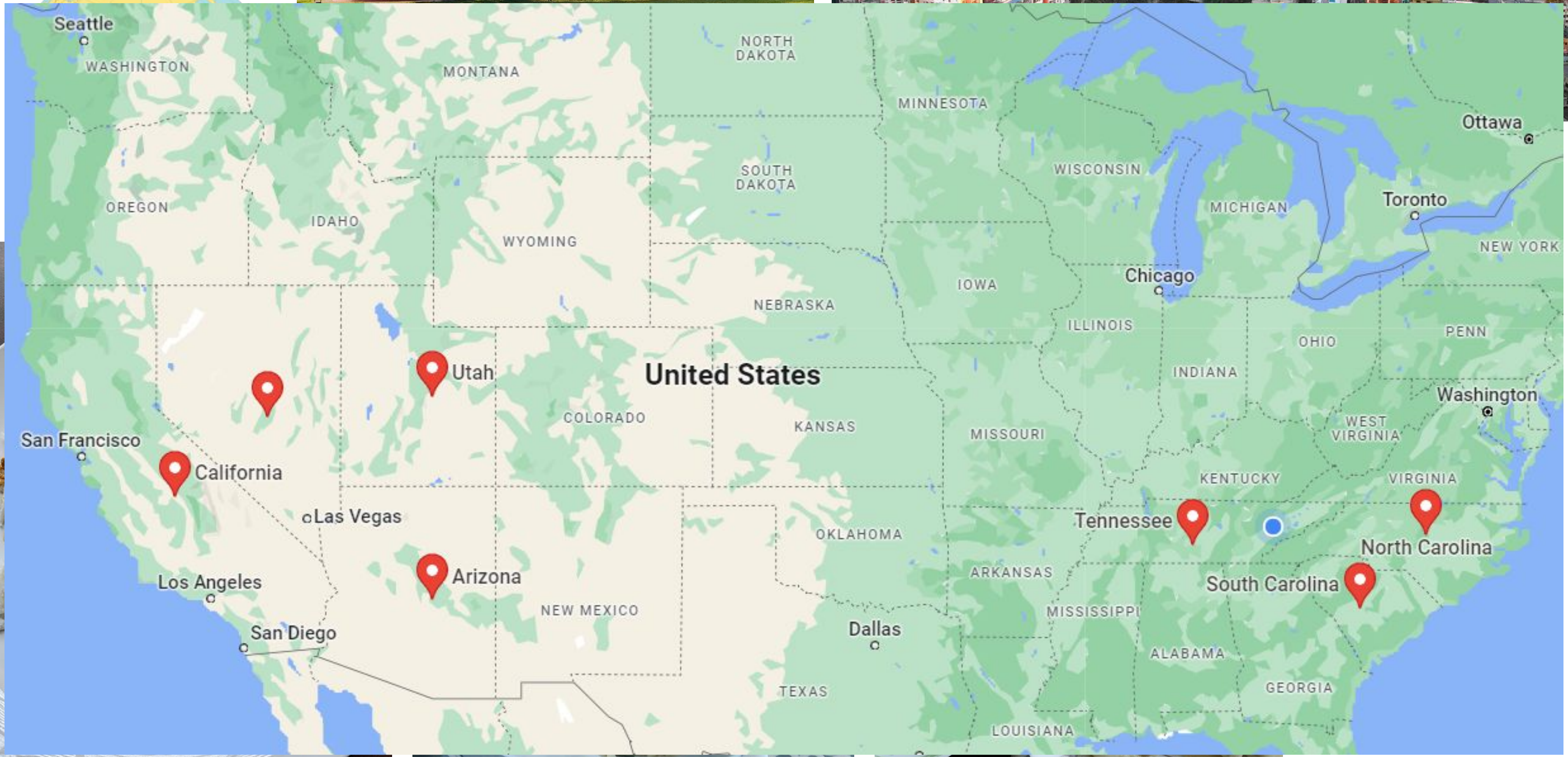
Bio

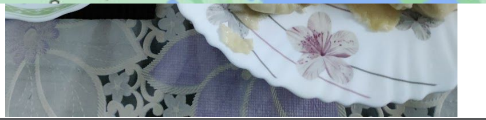
- Master’s in Computer Science at UTK
- Working as a GRA at ICL under Dr. Piotr Luszczek
- My research interests are - Deep Reinforcement Learning (Deep RL), High Performance Computing (HPC), and Artificial General Intelligence (AGI)

OpenAI. (2019, April 15). Openai five defeats dota 2 world champions. OpenAI Five defeats Dota 2 world champions. Retrieved April 17, 2023, from https://openai.com/research/openai-five-defeats-dota-2-world-champions

Not so Formal Bio

- Home Country - Dhaka, Bangladesh
- Religion - Islam
- Language - Bengali, English
- Interests - Travelling, Cooking





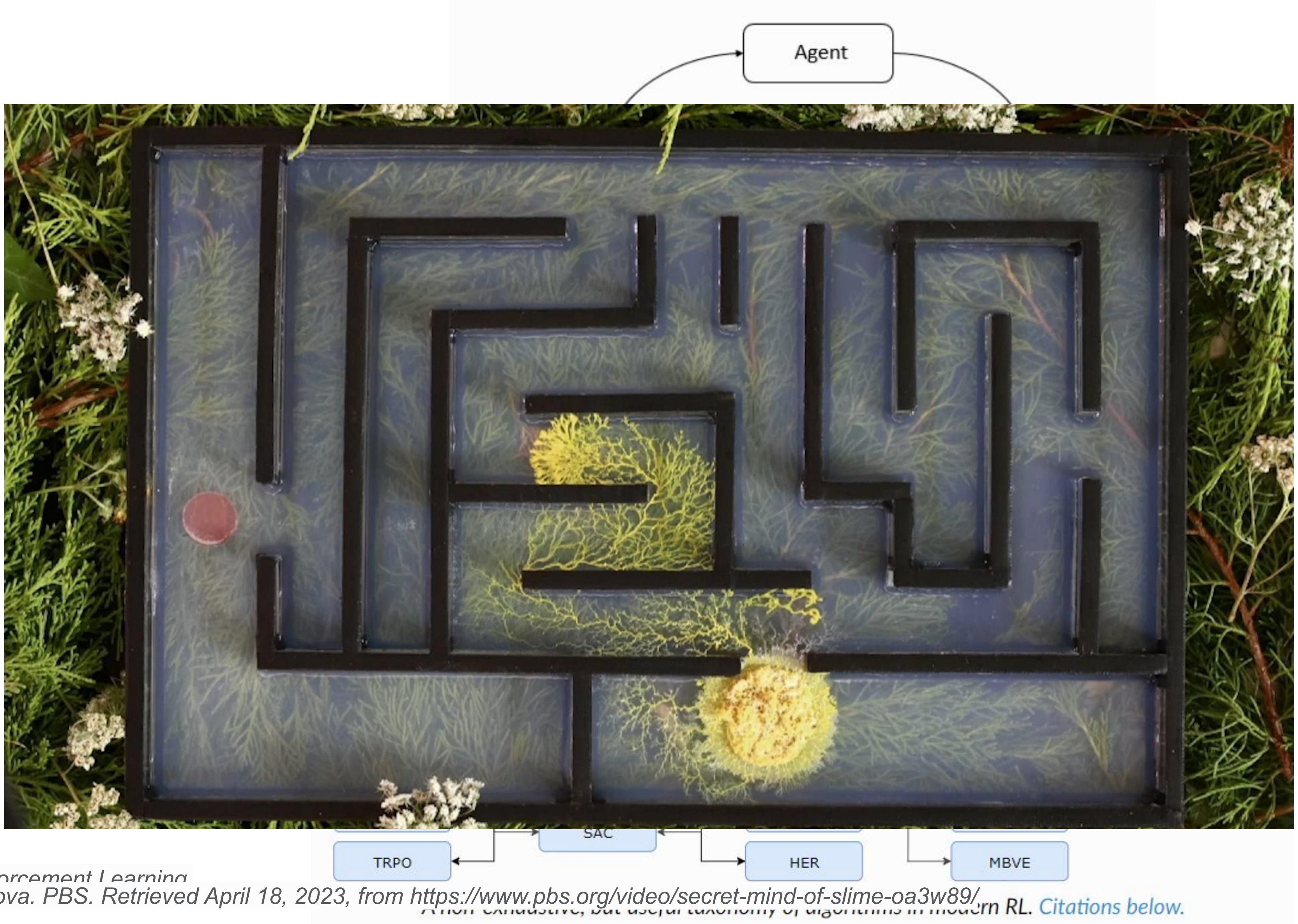
Outline

- Overview of the topic
- Background and history
- Introduction to the Algorithm
- How it works
- Results and comparisons
- Real world applications
- My research work using the Algorithm
- Challenges and future possibilities
- Summary, discussion, and conclusion
- Revisiting Test Questions

Overview

Terminologies -

- Reinforcement Learning,
- States and Observations,
- Action Spaces,
- Policies,
- Trajectories,
- Reward and Return,
- The RL Problem,
- Value Functions,
- Delayed Reward,
- Exploration vs Exploitation.



A non-exhaustive, but useful taxonomy of algorithms in modern RL. Citations below.

Achiam, J. (2018). Spinning Up in Deep Reinforcement Learning

PBS. (2020, September 16). Nova. PBS. Retrieved April 18, 2023, from https://www.pbs.org/video/secret-mind-of-slime-oa3w89/.

Background History

Reinforcement Learning (RL) History:

- Early ideas in 1950s for RL, include trial-and-error learning by Alan Turing and the concept of rewards by Richard Bellman
- In 1980s, development of Temporal Difference (TD) learning algorithms by Richard Sutton, bridging dynamic programming by Richard Bellman and Monte Carlo methods by Claude Shannon
- 1989: Christopher Watkins introduces Q-learning, a popular off-policy TD learning method
- 1990s: Early policy gradient methods by Ronald Williams
- 2000s: Breakthroughs in function approximation techniques, enabling RL to handle large state spaces, with contributions from researchers like Geoffrey Hinton, Richard Sutton, and Andrew Ng

PPO History:

- 2013: Silver et al. propose Deterministic Policy Gradient (DPG) algorithm, combining policy gradient and actor-critic methods for continuous control tasks
- 2015: John Schulman et al. introduce Trust Region Policy Optimization (TRPO)
- 2016: Lillicrap et al. propose Deep Deterministic Policy Gradient (DDPG)
- 2017: John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov develop Proximal Policy Optimization (PPO)

PPO Background

- **Vanilla Policy Gradient (VPG)**
  - A basic form of policy gradient method that aims to optimize an agent's policy directly by estimating the gradient of the expected return
  - Makes small updates to the policy parameters in a direction that maximizes the expected cumulative reward
  - limitations of VPG are high variance in gradient estimation, slow convergence, and instability during training
- **Trust Region Policy Optimization (TRPO)**
  - Advanced policy gradient method that builds upon VPG and aims to address its limitations
  - Uses a trust region approach, which limits the size of policy updates to ensure stability during training
  - Computationally expensive and challenging to implement
- **Proximal Policy Optimization (PPO)**
  - A simplification of TRPO that retains many of its advantages while being easier to implement and computationally more efficient
  - Introduces a clipped surrogate objective function, which penalizes large policy updates and encourages small, stable updates
  - Maintains a good balance between sample efficiency, stability, and ease of implementation.

Introduction to PPO

- Developed by John Schulman and his colleagues at OpenAI in 2017
- Builds upon the foundation of Trust Region Policy Optimization (TRPO)
- An on-policy, model-free algorithm that combines policy gradient and actor-critic methods
- PPO balances exploration and exploitation during training
- Clipping mechanism stabilizes training and prevents overly aggressive updates
- Simplicity, efficiency, and ease of implementation make PPO state-of-the-art for tackling complex RL problems

PPO Algorithm

- **Initialize the policy network:** Initializes policy network that takes the environment's state as input and outputs action probabilities
- **Collect experience:** Interacts with the environment using the current policy to collect a set of trajectories consisting of state-action-reward tuples
- **Estimate the policy gradient:** Computes the policy gradient using the collected trajectories
- **Calculate the surrogate objective:** A surrogate objective function is defined to compare the new policy's action probabilities to the old policy and incorporates a clipping term to penalize large policy updates
- **Update the policy:** By optimizing the clipped surrogate objective function using first-order optimization algorithms like stochastic gradient descent
- **Iterate:** Iterates until convergence or a specified number of iterations

Algorithm 1 PPO-Clip

Input: initial policy parameters θ_0, initial value function parameters ϕ_0
for $k = 0, 1, 2, ...$ **do**
  Collect set of trajectories $\mathcal{D}_k = \{\tau_i\}$ by running policy $\pi_k = \pi(\theta_k)$ in the environment.
  Compute rewards-to-go $\hat{R}_t$.
  Compute advantage estimates, $\hat{A}_t$ (using any method of advantage estimation) based on the current value function V_{ϕ_k}.
  Update the policy by maximizing the PPO-Clip objective:

$$\theta_{k+1} = \arg\max_{\theta} \frac{1}{|\mathcal{D}_k| T} \sum_{\tau \in \mathcal{D}_k} \sum_{t=0}^{T} \min\left(\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_k}(a_t|s_t)} A^{\pi_{\theta_k}}(s_t, a_t),\ \ g(\epsilon, A^{\pi_{\theta_k}}(s_t, a_t)) \right),$$

  typically via stochastic gradient ascent with Adam.
  Fit value function by regression on mean-squared error:

$$\phi_{k+1} = \arg\min_{\phi} \frac{1}{|\mathcal{D}_k| T} \sum_{\tau \in \mathcal{D}_k} \sum_{t=0}^{T} \left(V_\phi(s_t) - \hat{R}_t \right)^2,$$

  typically via some gradient descent algorithm.
end for

Outcome

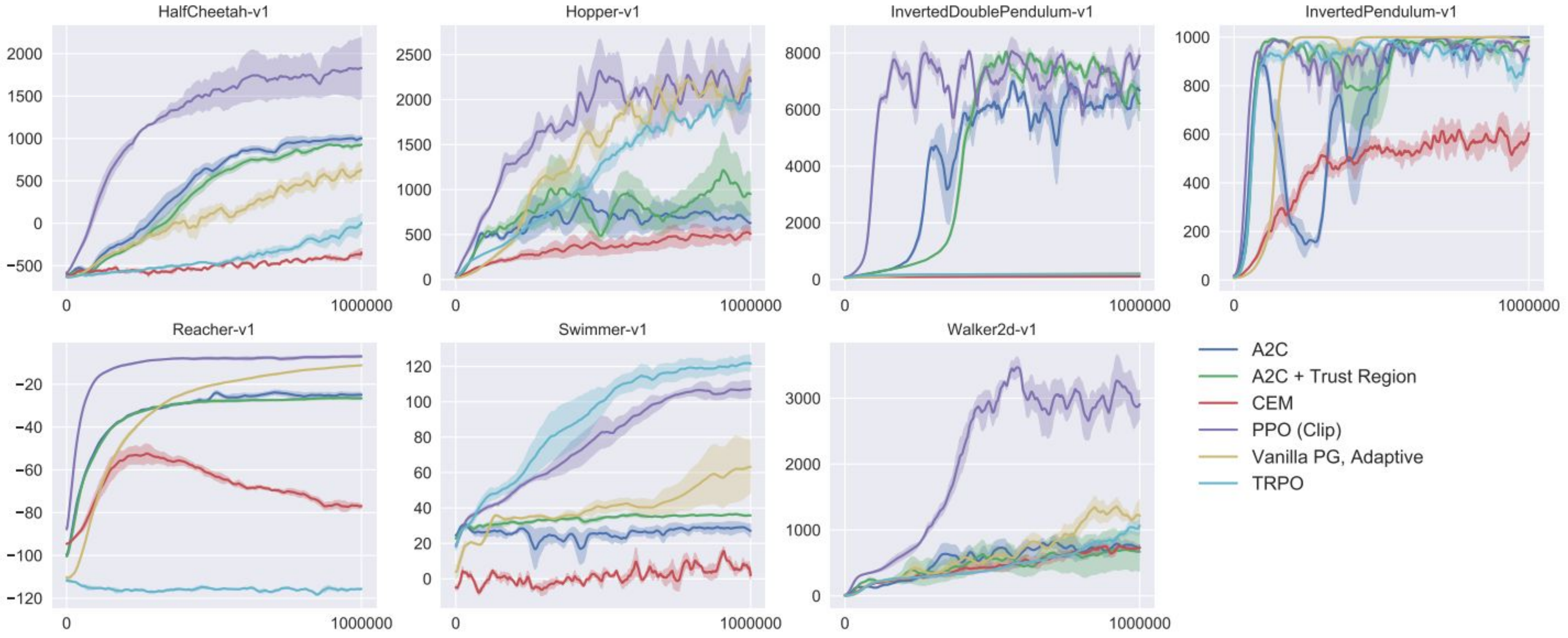


Figure 3: Comparison of several algorithms on several MuJoCo environments, training for one million timesteps.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347

Applications

Robotics: Training robotic arms, manipulators, and legged robots

Continuous control tasks: Excels in continuous control tasks

Game playing: Training AI agents to play complex games like Dota 2

Autonomous vehicles: Train autonomous cars and drones for navigation and obstacle avoidance

Multi-agent environments: Agents learn to cooperate and compete with each other

Limitations

Sample inefficiency: When data is scarce PPO struggles with sample efficiency

Exploration: Struggles with exploration in environments with sparse rewards or large state-action spaces.

Hyperparameter tuning: Finding the right set of hyperparameters can be time-consuming

Model-free approach: Cannot leverage any prior knowledge or structure in the environment to improve learning efficiency

No guarantees of global optimality: Does not provide any guarantees of finding a globally optimal policy

References

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347

Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction. MIT press

Achiam, J. (2018). Spinning Up in Deep Reinforcement Learning.

OpenAI. (2019, April 15). Openai five defeats dota 2 world champions. OpenAI Five defeats Dota 2 world champions. Retrieved April 17, 2023, from https://openai.com/research/openai-five-defeats-dota-2-world-champions

PBS. (2020, September 16). Nova. PBS. Retrieved April 18, 2023, from https://www.pbs.org/video/secret-mind-of-slime-oa3w89/

Questions and Questions

Test Questions Revisited:

1. What are policy gradients?

2. What is the key concept of Proximal Policy Optimization (PPO)?

3. What are your thoughts on the PPO algorithm after the presentation?

… Any Questions?