Comparing Deep Reinforcement Learning Approaches for Sparse Reward Settings with Discrete State-Action Spaces

Alp Şefik Çapanoğlu (a.s.capanoglu@student.tudelft.nl)[1]

1) Background

- Most Deep Reinforcement Learning (DRL) algorithms struggle to learn in sparse-reward settings.
- Sparse reward environments with discrete state-action spaces are understudied, because their continuous variant overshadows them.
- The bit-flipping environment introduced alongside HER[1] is used as implemented in Stable Baselines 3[2].

2) Research Question

What state-of-the-art DRL Algorithm is the most sample efficient in sparse reward environments with discrete state-action spaces?

3) Algorithms Chosen

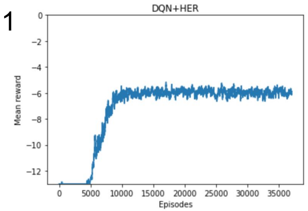
1. Proximal Policy Optimization (PPO)[3] is chosen out of the Maximum Entropy RL approach.
2. Hindsight Experience Replay (HER)[2] is chosen and used with Deep Q-Networks as a baseline.
3. Quantile Regression Deep Q-Networks (QR-DQN)[4] are used out of the Distributional RL approach.

4) Results

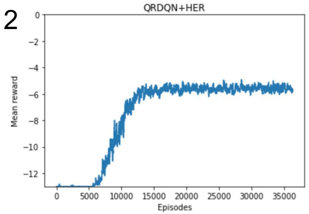
- PPO performs well early on, but falls of rapidly after a certain cardinality of the search space.
- Out of PPO, DQN with and without HER, QR-DQN with and without HER, the most sample efficient approaches are **DQN w/ HER** and **QR-DQN w/ HER**.

Below results are from the last round of experiments.
On y-axis: reward collected, **on x-axis:** training episode

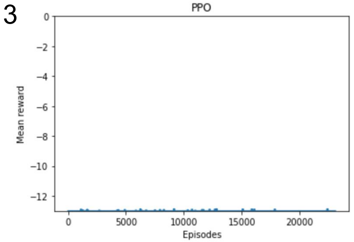
1



2



3



5) Conclusions

- Using HER with off-policy alternatives is the most sample efficient approach out of the candidate algorithms.
- PPO's ability to find sparse reward by exploration falls off in as the cardinality of the state space grows.

6) Future work

- Compare more algorithms from the three approaches, or algorithms that combine them.
- Implement different sparse-reward discrete state-action environments and test them.
- Apply the knowledge gained to a real world problem that has sparse rewards and discrete state-action spaces.

References

1. Andrychowicz, Marcin, et al. "Hindsight experience replay." arXiv preprint arXiv:1707.01495 (2017).
2. Raffin, A., Hill, A., Ernestus, M., Gleave, A., Kanervisto, A., \& Dormann, N.. (2019). Stable Baselines3.
3. Schulman, John, et al. "Proximal policy optimization algorithms." arXiv preprint arXiv:1707.06347 (2017).
4. Dabney, Will, et al. "Distributional reinforcement learning with quantile regression." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 32. No. 1. 2018.

[1]: Supervised by Greg Neustroev (g.neustroev@tudelft.nl) and Matthijs de Weerdt (m.m.deweerdt@tudelft.nl).