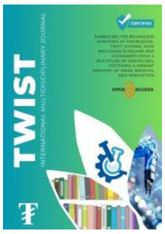
DOI 10.5281/twist.17590418.14

TWIST, 2022, Vol. 17, Issue 1, pp. 14-21

TWIST

Journal homepage: www.twistjournal.net



Environment Interaction of a Bipedal Robot using Model-Free Control Framework Hybrid off-Policy and on-Policy Reinforcement Learning Algorithm

Weiyhi Mahh*
College of Arts and Sciences, Mindanao State University – Buug, Buug, Zamboanga Sibugay, Philippines
[*Corresponding author]

Rosi Jones
College of Arts and Sciences, Mindanao State University – Buug, Buug, Zamboanga Sibugay, Philippines

Abstract

In the reinforcement learning Algorithm, there are different kinds of environments; one of them is a continuous environment. In the continuous environment, the data sources change very rapidly. Due to this capricious in the data, it’s difficult to train the agent using reinforcement learning. We have used the hybrid algorithm of DDPG and PPO. The DDPG is the off-policy algorithm that uses the old policy to train the agent, if we use the past observation to get obtains a new policy is not considered good as it brings instability in learning. DDPG is also data-efficient meaning if the collect number of the past old policy before if update a new current policy which makes them difficult to tune and unstable. PPO is the on-policy algorithm that focuses on keeping the new policy nearly close to the old policy. They have better sample complexity as they update multiple updates of mini batches of the data collected from the environment. And it’s easily tuned. We have combined these two policies (on-policy and offpolicy) by taking the data efficiency from the off-policy algorithms and using the high variance gradient of on policy, which helps in a large number of samples and distributed training the agent. However, combining the on-off policy algorithms is difficult to find the hyper parameters which are suitable to govern this trade-off. In this proposed algorithm we update the number of offpolicy with each on-policy update. We used specialized clipping in objective function epsilon ‘ Ɛ ‘ to remove the incentives to keep the new policy as near as the old policy. In the experiment, we used the open AI gym Box2D benchmark Biped walker and biped walker Hardcore**.**

Keywords

Environment, Hybrid off-Policy, Learning, DDPG

The full length manuscript can be sought from the corresponding author or upon requesting the editorial office with due intensions for usage and implementation.