Obstacle Avoidance and Navigation Utilizing PPO

OSCAR Lab

Daniel Zhang and Colleen P. Bailey
{sz0160, Colleen.Bailey}@unt.edu

Department of Electrical Engineering, College of Engineering, UNT EST. 1890

Problem

Unmanned Ground Vehicles (UGVs) are widely used in variety of applications in both civil and military-based scenarios. Controlling UGVs autonmously using reinforcement learning(RL) or Deep RL is a big challenge in academia and industry. This is a difficult problem due to two aspects:

1. Continuous control task.
2. Training speed is relatively slow.

Contributions

We considered a reinforcement learning approach to solve obstacle avoidance and navigation problem from robotics control field. The method is based on the proximal policy optimization algorithm (PPO) [1]. Our main contributions are

1. Implement both PPO and Deep Deterministic Policy Gradient (DDPG).
2. First to apply the new reward shaping technique [2] in RL for robotics control problems.

PPO with reward shaping

for *episode* $= 0, 1, \ldots$ **do**
  for *iteration* $= 0, 1, \ldots, N$ **do**
    Run policy π_{old} in environment for T timesteps;
    Compute advantage estimates $\hat{A}_1, \ldots, \hat{A}_T$;
  end
  Optimize $L^{CLIP}(\theta)$ with respect to θ, with K epochs and minibatch size $M \leq NT$;
  Update $\theta_{old} \leftarrow \theta$;
end

The clipped surrogate objective function $L^{CLIP}(\theta)$ is given as:

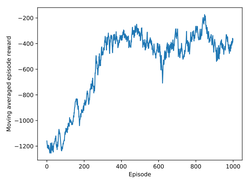
$$\hat{\mathbb{E}}_t[\min(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t)]$$

We utilize the reward shaping method in Eq. 1 to calculate the estimator of advantage function $\hat{A}_t$.

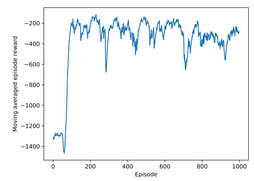
$$R^{lyap} = R(s,a) + \eta(\gamma R(s',a') - R(s,a)) \qquad (1)$$

Comparison and Results

DDPG reward trajectory

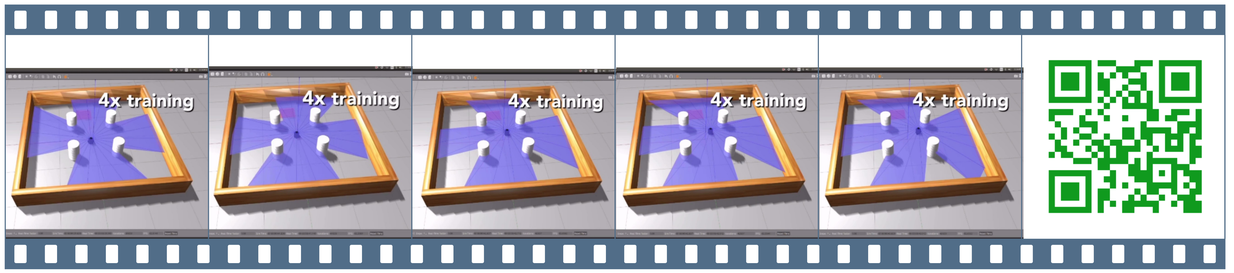


PPO reward trajectory

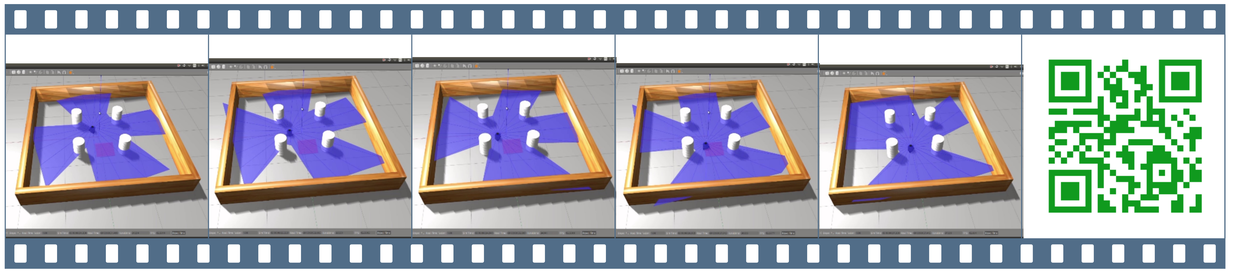


We show the reward trajectory for both DDPG and PPO algorithms separately. The PPO results in the figure above achieve a higher reward value faster than DDPG. Also, PPO achieves the ideal reward level within around 100 episodes while DDPG reaches such a level after 300 episodes. By taking advantages of the new reward shaping technique, both of DDPG and PPO converge significantly faster than before.

Demo



A training episode

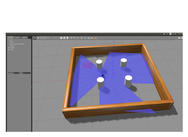


After training

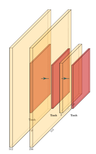
The above demos demonstrate the training session and the performance of the RL agent after training.

Please feel free to scan the QR codes to watch our demonstration movies on YouTube.

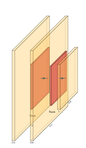
Implementation



Simulation Setup



Actor Network



Critic Network

We use Gazebo, ROS, and Turtlebot 3 Burger® to demonstrate both DDPG and PPO separately. The training environment we set up for demonstrating obstacle avoidance and navigation task is shown in the left figure. The Actor network and Critic network of PPO is shown in the middle and right figure separately.
The 4 white cylinders are the obstacles, the red square is the target for the robot, the blue lines demonstrate the LiDAR scanning from the robot. All the details of training is given in Table 2 of our paper in details.

Details

This work has been accepted as *Obstacle Avoidance and Navigation Utilizing Proximal Policy Optimization* by *SPIE Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications II, California, United States, 2020*.

Conclusion and Future Directions

We consider solving the obstacle avoidance and navigation problem for unmanned ground vehicles by applying proximal policy optimization algorithm equipped with a reward shaping technique. We compare DDPG and PPO in the same learning settings, and see that PPO has better performance than DDPG. For the future directions, we will investigate the performance of PPO applied to multi-agent robots systems.

References

[1] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, O. Klimov Proximal Policy Optimization Algorithms In *arXiv:1707.06347*

[2] Y. Dong, X. Tang, Y. Yuan Principled Reward Shaping for Reinforcement Learning via Lyapunov Stability Theory In *Neurocomputing*