A Deep Reinforcement Learning Strategy for UAV Autonomous Landing on a Platform

Unmanned Aerial Vehicle (UAV) is increasingly becoming an important tool used for a variety of tasks. In addition, Reinforcement Learning (RL) is a popular research topic. In this paper, these two fields are combined together and we apply the reinforcement learning into the UAV field, promote the application of reinforcement learning in our real life. We design a reinforcement learning framework named ROS-RL, this framework is based on the physical simulation platform Gazebo and it can address the problem of UAV motion in continuous action space. We can connect our algorithms into this framework through ROS and train the agent to control the drone to complete some tasks. We realize the autonomous landing task of UAV using three different reinforcement learning algorithms in this framework. The experiment results show the effectiveness of algorithm in controlling UAV which flights in a simulation environment close to the real world.
In order to estimate the value function, we propose a new method called TD3. This method is inspired by the online learning algorithm DDPG (Deep Deterministic Policy Gradient). In DDPG, the actor network learns the policy, and the critic network learns the value function. The actor network is responsible for exploring the action space to find the optimal solution, while the critic network is responsible for evaluating the policy by estimating the value of the current state. The actor network is updated based on the gradient of the value function, while the critic network is updated based on the Bellman equation.

\[
\theta = \pi \theta + \tau \theta
\]

\[
Q(s, a) = \sum_{k=0}^{\infty} \gamma^k r_{t+k} + \gamma \max_{a'} Q(s', a')
\]

In TD3, we implement two types of value function simultaneously, one is the deterministic policy which only output one action, and the other is the stochastic action which has the highest probability or the action that the agent believes is the best. The other one is the stochastic action that has the highest probability or the action that the agent believes is the best. The other one is the stochastic action that has the highest probability or the action that the agent believes is the best. The other one is the stochastic action that has the highest probability or the action that the agent believes is the best.

\[
Q(s, a) = \sum_{k=0}^{\infty} \gamma^k r_{t+k} + \gamma \max_{a'} Q(s', a')
\]

\[
V_\pi(s) = \mathbb{E}_{G_t} \left[ \sum_{k=0}^{\infty} \gamma^k r_{t+k} \right]
\]

\[
\theta = \pi \theta + \tau \theta
\]

\[
Q(s, a) = \sum_{k=0}^{\infty} \gamma^k r_{t+k} + \gamma \max_{a'} Q(s', a')
\]

A. Reinforcement Learning Framework

As we all know, reinforcement learning has been widely applied in various fields, such as robotics, control theory, and game theory. In this section, we propose a new reinforcement learning framework called SAC (Soft Actor Critic). This framework is inspired by the soft actor-critic algorithm, which combines the advantages of both soft and hard policies. In SAC, the actor network learns a stochastic policy, and the critic network learns the value function.

\[
\pi(a|s) = \frac{\exp(\phi(s, a))}{\mathbb{Z}(\phi)}
\]

\[
Q(s, a) = \mathbb{E}_{G_t} \left[ \sum_{k=0}^{\infty} \gamma^k r_{t+k} + \gamma \min_{a'} Q(s', a') \right]
\]

\[
V_\pi(s) = \mathbb{E}_{G_t} \left[ \sum_{k=0}^{\infty} \gamma^k r_{t+k} + \gamma \min_{a'} Q(s', a') \right]
\]

\[
\theta = \pi \theta + \tau \theta
\]

\[
Q(s, a) = \sum_{k=0}^{\infty} \gamma^k r_{t+k} + \gamma \min_{a'} Q(s', a')
\]

\[
V_\pi(s) = \mathbb{E}_{G_t} \left[ \sum_{k=0}^{\infty} \gamma^k r_{t+k} + \gamma \min_{a'} Q(s', a') \right]
\]

\[
\theta = \pi \theta + \tau \theta
\]
B. Tasks And Scenarios Design

C. Reinforcement Learning Problem Definition And Function Design

\[
S = \{p_x, p_y, p_z, v_x, v_y, v_z\}
\]

\[
A = \{a_x, a_y, a_z\}
\]

in horizontal dimension in the vertical dimension

\[
a_z = \alpha \cdot p_x \cdot a_x
\]
it affects whether the agent learns the control strategy or not. 

\[
shaping_r = -\sqrt{p_x + p_y + p_z} - \sqrt{v_x' + v_y' + v_z'} - \sqrt{a_x + a_y + a_z} + \mathcal{C}(\|a_t\|) + \mathcal{C}(\|a_{\mu_t}\|)
\]

\[
r = shaping_r - shaping_{r_{-r}}
\]

D. Reinforcement Learning Algorithm And Network Structure Design

\[
\Delta \theta_{\mu} = \nabla_{\theta_{\mu}} Q(s, a) = \nabla_{\theta_{\mu}} Q(s, a) - \nabla_{\theta_{\mu}} Q(s, \mu(s))
\]

\[
\theta_{Q} = \tau \theta_{Q} + \delta_{-r} \frac{\partial \nabla_{\theta_{Q}} L}{\partial \theta_{Q}}
\]

\[
\theta_{\mu} = \tau \theta_{\mu} + \delta_{-r} \frac{\partial \nabla_{\theta_{\mu}} L}{\partial \theta_{\mu}}
\]
loss_s = \left( Q(s_t, a_t) - \min_{\pi} Q(s_{t+1}, \pi'(s_t) + \text{noise}) - r_t \right)

loss_v = \left( Q(s_t, a_t) - \min_{\pi} Q(s_{t+1}, \pi'(s_t) + \text{noise}) - r_t \right)

A. Experimental Configuration Introduction

B. Simulation Scenarios Introduction

C. Experience Results And Discussion
At the same time, we designed a UAV auto-Gazebo simulator.

The update step of the algorithm is very intense.

But as the improved algorithm TD3 can get a good policy, we can apply our model to try the landing process, so the reward helps the agent to find out the good policy, but the noise during auto-landing is not stable. Maybe, the annealing method needs to be applied to the virtual world easily. We will test our algorithm in the real life, and compare the performances of TD3 and SAC.

In this paper, we connected three continuous algorithms to the real life. Although SAC can learn the policy quickly, it only needs very limited episodes to find out the space that has not been explored. SAC has better performance, it only needs very limited episodes to find out the maximum entropy deep reinforcement learning with a stochastic actor-critic methods.

The noise added in the TD3 or short-term model encourage the agent learns to control the UAV to land on the moving platform. As a framework for real drones, adjusting the annealing weight to try the landing task as a test case and design and test the framework into some more complex scenarios, design and test the general interface to communicate between the agents and the robotics in the Gazebo environment. Although SAC did not transfer the policy to control UAVs, but DDPG did not. Each of the algorithms to the real life general interface to communicate between the agents and the robotics in the Gazebo environment.

The reward per episode of DDPG of TD3 is only forward a very small step, but we can transfer the algorithm convergence and compare their performances. TD3 and SAC learned the landing system for AR. Drone 2.0 using onboard MAV simulator framework (ROS). Springer, Cham, 2016.

Yu X, Fan Y, Xu S, et al. A modular gazebo simulator framework for UAV autonomous landing via deep reinforcement learning. IEEE, 2021. 2247.

Rodriguez C, Daftry S, Capobianco R, et al. A self-moving platform reinforcement learning strategy for UAV autonomous landing on a moving platform. International Journal of Robust and Nonlinear Control, 2021. 2018, 29(13): 351-378.

Polvara R, Patacchiola M, Sharma S, et al. Autonomous Planetary Landing via Deep Reinforcement Learning and Transfer Learning. Conference on Computer Vision and Pattern Recognition. 2021.

Cham, 2016. Conference on Control, Automation, Robotics and Vision (ICARCV). IEEE, 2016.

Haarnoja T, Zhou A, Abbeel P, et al. Addressing function approximation error in actor-critic methods via Deep Reinforcement Learning and Transfer Learning. Conference on Machine Learning Research. PMLR, 2018. 1587.

Ciabatti G, Daftry S, Capobianco R. A self-moving platform reinforcement learning strategy for UAV autonomous landing on a moving platform. International Journal of Robust and Nonlinear Control, 2021. 2018, 29(13): 351-378.

Fujimoto S, Hoof D, Meger D. Continuous control with deep reinforcement learning. PMLR, 2018. 1587.

Lillicrap T P, Hunt J J, Pritzel A, et al. Continuous control with deep reinforcement learning. PMLR, 2015.

Moriarty P, Sheehy R, Doody P. Neural networks to aid the autonomous landing of a UAV on a ship. Frontiers in Robotics and AI. 2019, 5:

Polvara R, Patacchiola M, Sharma S, et al. Autonomous Planetary Landing via Deep Reinforcement Learning and Transfer Learning. Conference on Computer Vision and Pattern Recognition. 2021.

Yu X, Fan Y, Xu S, et al. A modular gazebo simulator framework for UAV autonomous landing via deep reinforcement learning. IEEE, 2021. 2247.

Rodriguez C, Daftry S, Capobianco R, et al. A self-moving platform reinforcement learning strategy for UAV autonomous landing on a moving platform. International Journal of Robust and Nonlinear Control, 2021. 2018, 29(13): 351-378.

Polvara R, Patacchiola M, Sharma S, et al. Autonomous Planetary Landing via Deep Reinforcement Learning and Transfer Learning. Conference on Computer Vision and Pattern Recognition. 2021.

Cham, 2016. Conference on Control, Automation, Robotics and Vision (ICARCV). IEEE, 2016.

Haarnoja T, Zhou A, Abbeel P, et al. Addressing function approximation error in actor-critic methods via Deep Reinforcement Learning and Transfer Learning. Conference on Machine Learning Research. PMLR, 2018. 1587.

Ciabatti G, Daftry S, Capobianco R. A self-moving platform reinforcement learning strategy for UAV autonomous landing on a moving platform. International Journal of Robust and Nonlinear Control, 2021. 2018, 29(13): 351-378.