Traffic Simulation & Navigation Environment in RL

Group 4 - Shiwei Yu, Xiaomeng Chen

Introduction

Traffic Simulation and navigation is a popular issue in reinforcement learning that involves simulating traffic scenarios and training agents to navigate in these scenarios. The environment models the behavior of vehicles, and then the agent learns to take actions that maximize the reward or minimize penalties, such as optimizing travel time, reducing accidents or delays, and ensuring safety. This project intends to apply three RL algorithms for discrete action spaces – Proximal Policy Optimization (PPO), DQN, and Actor-Critic (A2C) – on specific traffic environments designed by OpenAI, including highway env[1] and intersection env[2]. Finally all three models are evaluated and compared in terms of effectiveness and drawbacks in this context.

Method

Environment

Highway_env and **Intersection-v0** are two traffic simulation environments created by OpenAI for testing and evaluating reinforcement learning algorithms. The agent navigates through a multi-lane highway scenario and an intersection scenario, respectively. In both environments, the agent is asked to avoid collisions and optimize reward functions, such as minimizing travel time or reducing the number of lane changes. The action space consists of a set of discrete actions, including changing lanes, accelerating, and braking.

RL Algorithms

Proximal Policy Optimization (PPO) is a popular choice for training agents in complex environments such as autonomous driving scenarios due to its stability, robustness, and ease of tuning[3]. In highway_env, PPO model is expected to be trained using stable_baselines3 library with several hyperparameters being tuned, and then generate the plot of cumulative rewards versus episode to assess the overall performance. The mean reward is a metric to measure its efficiency.

Double Q-Network (DQN) can be successful in handling discrete action spaces, which makes it a good choice for highway_env where the agent can choose from a fixed set of

[1] https://pypi.org/project/highway-env/
[2] https://farama-foundation.github.io/HighwayEnv/environments/intersection/
[3] https://arxiv.org/abs/1707.06347

actions. However, compared to PPO, DQN may require more hyperparameter tuning[4] to achieve good performance. Considering that DQN may be more challenging to train and suffer from instability and slow convergence, we mainly focus on faster convergence to an optimal policy when tuning hyperparameters, such as decreasing buffer_size and learning_starts.

Intersection-v0 is also applied with the DQN model, with further training and evaluation by using reward-episode plots and mean reward as the measurements.

Actor-Critic (A2C) combines the advantages of both value-based and policy-based methods[5]. In Actor-Critic, there are two components – the actor & the critic. A2C can be used in conjunction with the intersection-v0 environment by defining the A2C agent from stable_baselines3.

Result

highway_env		intersection-v0	
PPO	DQN	DQN	Actor-Critic (A2C)
18388.01	54037.98	423.26	371.72
21.17	9.55	4.86	4.37
0.62	5.07	4.82	4.89

Table.1 Cumulative, mean, and std reward of each approach

Highway_env

In the context of the highway_env, PPO and DQN were compared in terms of cumulative, mean, and the standard deviation of reward. PPO was found to achieve a better performance in terms of both mean and standard deviation of reward, with values of 21.17 and 0.62 respectively. This suggests that PPO consistently performed better in each episode and is more stable and reliable. DQN, on the other hand, had higher cumulative rewards (54037.98) as it may be able to learn a better long-term strategy, even if individual episodes do not perform as well.

Intersection-v0

The overall performance of intersection-v0 was significantly weaker than that of highway_env. Both DQN and Actor-Critic models yielded very low mean reward and relatively high standard deviation. Cumulative rewards were notably bad as well.

---

[4] https://medium.datadriveninvestor.com/which-reinforcement-learning-rl-algorithm-to-use-where-when-and-in-what-scenario-e3e7617fb0b1

[5] https://huggingface.co/blog/deep-rl-a2c

Possible reasons for the poor performance of models in intersection-v0 include the complexity of the environment and its higher-dimensional state space compared to highway_env. The agent makes decisions on when to start, how fast to move, which lane to take, when to stop, and when to yield to other vehicles and pedestrians. All factors will contribute to the complexity of the task and make it more challenging for the agent to learn an effective policy. Another reason could be that the hyperparameter setting employed for these models may not be suitable for this environment and require further tuning.

Conclusion

In conclusion, DQN outperformed the other two models. Though PPO performed better in terms of mean and standard deviation of rewards in highway_env, DQN had higher cumulative rewards in both environments, suggesting that it may be able to learn a better long-term strategy. However, the overall performance of the models in intersection-v0 was notably worse than that in highway_env, with all measurements being significantly lower. This could be attributed to the increased complexity of the action spaces. It is also possible that the hyperparameter tuning approach used was inappropriate for intersection-v0. Overall, more research is needed to develop more robust and effective models for complex environments with both discrete and continuous action spaces.