Exploring Highway Environments with Reinforcement Learning

Reinforcement Learning Course

Chater Oumaima

For more details visit the repository:
HighwayRLExplorer Repository

Course professor:
HÉDI HADIJI
hedi.hadiji@centralesupelec.fr

CentraleSupélec

April 2024

Contents

1 Introduction

This report documents our exploration of Reinforcement Learning (RL) methodologies applied to autonomous driving scenarios within the Highway-env collection [1]. Our investigation focuses on three distinct environments: Highway, Parking, and Roundabout navigation.

Initially, we tackle the challenges of highway navigation, detailing the training performance of our Deep Q-Network (DQN) implementation. Subsequently, we transition to the Parking environment, examining the behaviour of our RL agent with continuous actions compared to discrete actions used in the Highway scenario. Finally, we examine roundabout navigation, employing existing RL algorithms from the Stable Baselines library. Through our experimentation, we seek to uncover insights into the effectiveness of RL techniques for autonomous driving applications. In particular, we address questions of algorithm generalization, hyperparameter influence, and implicit modelling within the Markov Decision Process (MDP). Our findings contribute to the advancement of safer and more efficient transportation systems.

2 Highway Environment

The DQN was deployed to tackle the challenges of the Highway environment, leveraging its capacity to handle high-dimensional observations and discrete actions. With a neural network approximating the optimal action-value function, our model empowers the agent to navigate safely and execute lane changes effectively within a discrete action space.

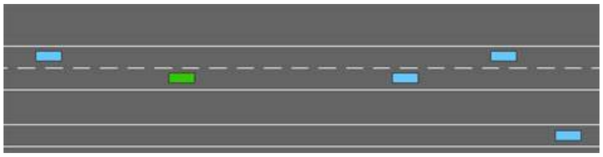

Figure 1: Highway environment

2.1 Application of DQN on the Highway Environment

2.1.1 Algorithm Implementation

The DQN was implemented to address the challenges posed by the Highway environment. Given the complexity of dynamic traffic scenarios, DQN's ability to manage high-dimensional observation spaces and discrete action spaces is invaluable. Our model uses a neural network to approximate the optimal action-value function, enabling the agent to make informed decisions for safe navigation and effective lane changes.

2.1.2 Training Process

The training process involved numerous episodes, each contributing to the iterative refinement of the agent's policy. The agent interacted with the environment, taking actions based on the current policy and observing the outcomes to adjust its approach. Over time, this process aimed to maximize the cumulative reward, a combination of factors including safety, efficiency, and adherence to traffic regulations.

2.1.3 Results

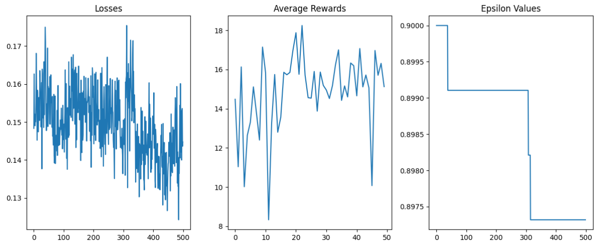

Figure 2: Training metrics for the DQN agent in the Highway environment. The plots show the progression of losses, average rewards, and epsilon values over the episodes.

Losses The plot on the left in Figure 2 depicts the agent's loss over 500 episodes. Loss values indicate the difference between predicted Q-values and the target Q-values, reflecting the agent's learning progress. A decreasing trend is observed, implying that the DQN agent progressively improved in estimating the action-value function.

Average Rewards The middle plot of Figure 2 shows the average rewards obtained per episode. This metric directly relates to the agent's performance, with higher rewards indicating better policy outcomes. Fluctuations in the plot suggest the explorative nature of the learning process, yet an overall upward trend demonstrates that the agent is learning to navigate the environment more effectively.

Epsilon Values The right plot illustrates the epsilon decay in the epsilon-greedy policy, signifying the trade-off between exploration and exploitation. As training progresses, epsilon decreases, indicating the agent's increasing reliance on the learned policy over random actions. The stepped decay of epsilon values is a result of the agent's strategy to explore initially and gradually shift towards exploitation as it gains confidence in its policy.

Conclusion The application of DQN to the Highway environment illustrates the algorithm's strength in managing the challenges of autonomous driving. The training metrics reveal a learning agent that not only improves its decision-making capabilities over time but also adapts its exploration strategy to optimize for the complex tasks at hand. This progression is crucial for developing autonomous agents that can operate safely and efficiently in real-world driving scenarios.

3 Parking Environment

The Parking environment within the HighwayRLExplorer project provides a realistic simulation for the task of parking a vehicle. Here, agents are required to master the intricacies of navigating into parking spots within confined spaces, exemplifying a crucial aspect of autonomous driving.

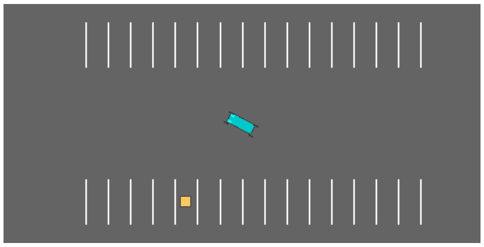

Figure 3: Parking environment

3.1 Model-Based Agent

Our Model-Based Agent incorporates a predictive model that enables the simulation of potential action outcomes for strategic planning. This foresight is analogous to a driver mentally planning maneuvers before execution, allowing for efficient and calculated navigation.

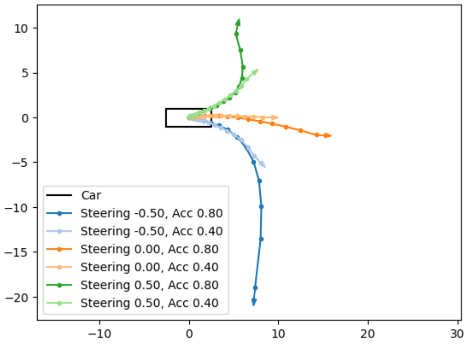

Figure 4: Trajectory visualization based on different steering and acceleration commands.

Key observations of Figure 4 include the sharper turns at lower accelerations for identical steering inputs, evident from the tighter curves of the dashed trajectories. Without steering input, the car advances straight, and the disparity in acceleration is evident in the differing trajectory lengths. This plot is crucial for predicting the vehicle's future state given various control inputs, serving as an invaluable tool in autonomous driving, path planning, and control systems design.

3.2 DDPG Agent

The Deep Deterministic Policy Gradient (DDPG)[2] agent introduces a model-free, continuous control paradigm to the parking challenge. The actor-critic architecture underlying DDPG is well-suited for the fine-tuned control necessary for parking, promoting stable learning through the segregation of policy and value estimation.

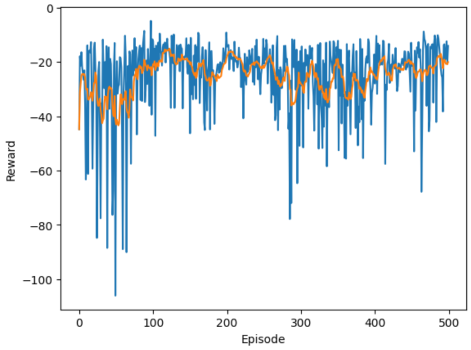

Figure 5: Reward progression of the DDPG agent across episodes. The blue line indicates the reward per episode, and the orange line represents a moving average, highlighting the learning trend over time.

Figure 5 presents the reward progression for the DDPG agent over 500 episodes. The plot shows the rewards received per episode (in blue) and a moving average (in orange), which smooths out the volatility and indicates the overall trend in performance. While the episode rewards fluctuate, reflecting the exploration-exploitation dynamics inherent in reinforcement learning, the moving average demonstrates a gradual improvement in the agent's ability to accumulate higher rewards over time. This trend suggests that the DDPG agent is learning to optimize its policy for the parking task, successfully navigating the complexities of the environment with continuous action decisions.

3.3 DDPG with HER

Extending the capabilities of DDPG, we incorporate Hindsight Experience Replay (HER) [3] to enhance learning from all experiences, successful or otherwise. By reinterpreting unsuccessful parking attempts as alternate successful outcomes, the DDPG with HER agent displays increased efficiency and adaptability.

3.3.1 Continuous Control Efficacy

DDPG's continuous action space capability allows the agent to execute more refined maneuvers compared to discrete actions. This is evident in the agent's performance, where smoother and more precise actions are observable, especially in making slight adjustments when aligning with the parking space.

3.3.2 Advantages and Implementation

HER's main advantage is the utilization of all experiences, redefining the agent's learning scope. This strategy proves especially beneficial in sparse-reward environments, improving policy robustness and generalization. Our implementation showcases the training process from the ground up, delineating the agent's interaction and learning within the environment.

3.3.3 Training Performance of DDPG with HER

In the application of the DDPG algorithm with Hindsight Experience Replay (HER) to the Parking environment, we observed a notable trend in the agent's performance as it learned over time. The following plot illustrates the agent's reward across episodes.

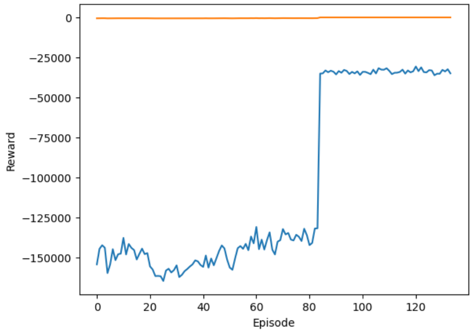

Figure 6: Reward progression of the DDPG agent with HER across episodes. A significant increase in the reward indicates substantial learning gains.

Initially, the agent's cumulative reward per episode shows fluctuations, which is typical in the early stages of training due to exploration. As episodes progress, we observe a dramatic increase in the reward, suggesting that the agent has begun to consistently apply an effective strategy for the parking task. This sudden surge in performance can be attributed to the agent's ability to extract useful learning signals from previously unsuccessful attempts, thanks to the HER technique.

The plateauing of the reward at a high level indicates that the agent has achieved a stable policy that performs well in the given task. This outcome reinforces the efficacy of HER in improving the sample efficiency of DDPG, allowing the agent to make the most out of each learning experience and effectively navigate the complexities of the Parking environment.

4 Roundabout Environment

The Roundabout environment in our HighwayRLExplorer project presents a sophisticated simulation challenge where autonomous agents learn efficient navigation through roundabouts—a task that is critical in real-world driving. In this environment, agents must make strategic decisions to merge, yield, and exit the roundabout safely and efficiently.

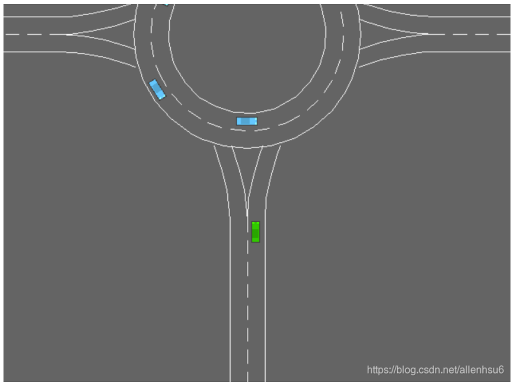

Figure 7: Roundabout environment

For this particular scenario, we employ the Deep Q-Network (DQN) algorithm, which is well-suited for environments with discrete action spaces. Our implementation leverages the Stable Baselines3 (sb3) library, which provides an optimized and user-friendly version of DQN that is capable of handling the discrete decision-making required in the Roundabout environment.

4.1 Hyperparameter Fine-Tuning

In the process of training our DQN agent, hyperparameter fine-tuning emerged as a crucial step to enhance the agent's performance. The selection of hyperparameters was meticulously conducted, with a focus on two pivotal factors: the discount factor γ and the learning rate.

4.1.1 Discount Factor γ

The discount factor γ influences the agent's consideration of future rewards. By adjusting γ, we can control the agent's foresight, emphasizing either immediate or long-term benefits.

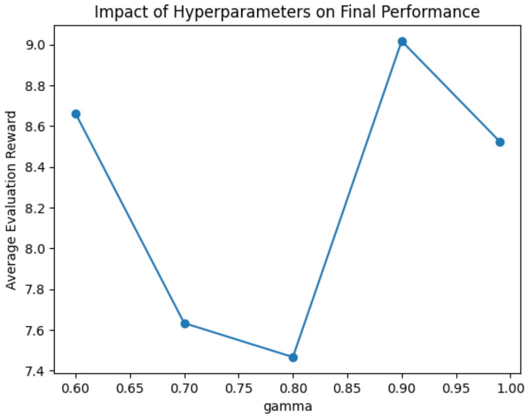

Figure 8: The relationship between the discount factor γ and the agent's performance, measured in average evaluation reward.

Figure 8 illustrates the correlation between varying γ values and the agent's average evaluation reward, emphasizing the significance of this hyperparameter in the training process.

4.1.2 Learning Rate

The learning rate determines the size of the updates to the agent's policy at each step. An optimal learning rate ensures a balance between rapid learning and the stability of the convergence process.

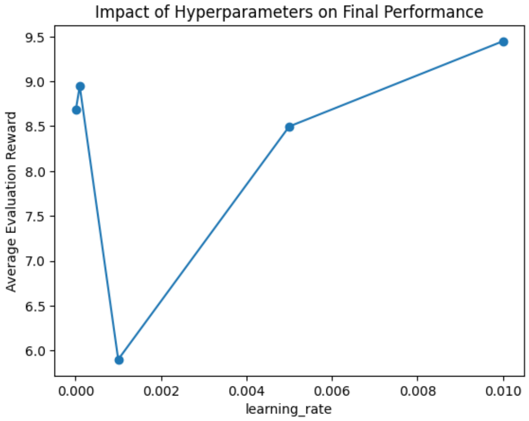

Figure 9: The impact of the learning rate on the agent's performance in the Roundabout environment.

As shown in Figure 9, the learning rate's fine-tuning played a pivotal role in the agent's ability to acquire a high-performing policy.

5 Conclusion

This report has presented an in-depth analysis of the application of reinforcement learning techniques across three different driving scenarios: Highway, Parking, and Roundabout navigation, each with its unique set of challenges and complexities. Through these environments, we have showcased the adaptability and effectiveness of RL methods such as DQN, DDPG, and the innovative use of HER.

Our findings reveal the nuanced behaviors that reinforcement learning agents can develop and the profound impact of hyperparameter tuning on their performance. In the Highway environment, we highlighted the DQN's ability to navigate complex traffic with discrete actions. In the Parking scenario, we illustrated how DDPG with HER could achieve precise vehicle control. Lastly, in the Roundabout environment, we detailed the success of an agent utilizing discrete actions via the sb3 library's DQN implementation, emphasizing the careful calibration of hyperparameters like the discount factor and learning rate.

While this report provides a thorough overview of our project and findings, it only scratches the surface of the extensive research and experimentation conducted. For a more detailed account of our methodologies, implementations, and results, we encourage readers to visit our GitHub repository.

As we conclude, we reflect on the transformative potential of reinforcement learning in advancing the field of autonomous driving. The progress documented here fuels our optimism for the future, where RL continues to evolve and contribute to the development of autonomous vehicles that are not only intelligent but also safe and reliable contributors to our transportation ecosystems.

References

[1] Leurent, Edouard, *An Environment for Autonomous Driving Decision-Making*, *GitHub repository*, 2018. `https://github.com/eleurent/highway-env`.

[2] Timothy P. Lillicrap and Jonathan J. Hunt and Alexander Pritzel and Nicolas Heess and Tom Erez and Yuval Tassa and David Silver and Daan Wierstra *Continuous control with deep reinforcement learning*, *arXiv*, 2019. `https://doi.org/10.48550/arXiv.1509.02971`.

[3] Marcin Andrychowicz and Filip Wolski and Alex Ray and Jonas Schneider and Rachel Fong and Peter Welinder and Bob McGrew and Josh Tobin and Pieter Abbeel and Wojciech Zaremba *Hindsight Experience Replay*, *arXiv*, 2018. `https://arxiv.org/pdf/1707.01495.pdf`.