**Reinforcement Learning for Autonomous Driving in Highway-Env**

**Overall Project link:** [**https://drive.google.com/drive/folders/1djC\_2D1YeeZIHfGqngXRkDqqzqLmlGGY?usp=sharing**](https://drive.google.com/drive/folders/1djC_2D1YeeZIHfGqngXRkDqqzqLmlGGY?usp=sharing)

**OVERVIEW:**

This project explores how Reinforcement Learning (RL) can be used to train an autonomous driving agent in Highway-Env. Three RL algorithms DQN, PPO, and A2C were implemented to evaluate their effectiveness in different driving situations, ranging from easy to hard environments. To fine tune the agent’s decision-making various hyperparameters are tested such as learning rate, discount factor and network architecture. The goal was to understand how each model learns, adapts and makes driving decisions over time.

**ENVIRONMENTAL SETUP:**

The below are the three different driving environments we selected for this project:

* **Highway-fast-v0:** This is a straightforward where the agent must move at high speed without colliding with the surrounding cars. The most significant challenge here is managing speed and lane-changing without colliding with other vehicles.
* **Intersection-v0:** This is a more complex environment where the agent moves through intersections and must anticipate and react to cross-traffic.
* **Roundabout-v0:** This environment presents a continuous lane changing challenges, in which the agent must merge into a circular stream of traffic, switch lanes and exit without crashes.

**EVALUATION METRICS:**

To assess the model’s learning progress and overall efficiency, we tracked the following metrics:

* **Mean Episode Reward:** This is used measures how well the agent maximized rewards over training episodes.
* **Success Rate:** Measures how frequently the agent met the environment’s success criteria.

**HYPERPARAMETER CONFIGURATIONS:**

To determine the most effective training setup, we experimented with two different hyperparameters settings learning rates (LR) and discount factors (γ). The learning rate decides how quickly the agent updates its policy based on new experiences, while the discount factor controls how much the agent values future rewards.

* **Lower LR (0.0005):** This allows for slow and stable learning.
* **Higher LR (0.001):** Higher LR enables faster merging but may cause instability.
* **Higher γ (0.95):** It encourages long-term planning.
* **Lower γ (0.8):** This prioritizes short-term rewards which leads to more immediate decision making.

**Analysis of Deep Q-Network (DQN) Performance in Autonomous Driving Scenarios:**

**INTRODUCTION:**

We trained and implemented a Deep Q-Network (DQN) agent to navigate through different driving scenarios using the Highway-env reinforcement learning environment. The overall goal was to train the agent to make optimal driving decisions by prioritizing safety, speed, and minimizing the collisions. The agent was trained using different hyperparameter settings like learning rates (LR) and discount factors (γ). These parameters greatly influence the way the agent learns to change its policy and how much it considers future rewards. The research also discusses how these parameters influenced the outcome of the training on three driving situations: highway, intersection and roundabout. The results in this work provide insight into how DQN-based agents learn with different driving conditions and where they can be improved.

**RESULTS & ANALYSIS FOR DQN:**

For each environment we evaluated the agent’s performance using “Mean Reward” (the average cumulative reward per episode) and “Success Rate” (a metric approximating how well the agent sticks to optimal driving behavior). The results are analyzed across different settings to identify trends and performance variations.

* **Highway-fast-v0:** The agent showed consistent learning progress, with steadily improving performance. The results showed that higher γ values led to better long-term planning, allowing the agent to make lane changes that maximized future rewards.

|  |  |  |
| --- | --- | --- |
| **Configuration** | **Mean Reward** | **Success Rate** |
| γ = 0.95, LR = 0.0005 | 25.95 | 0.42 |
| γ = 0.95, LR = 0.001 | 15.79 | 0.12 |
| γ = 0.8, LR = 0.0005 | 39.07 | 1.00 |
| γ = 0.8, LR = 0.001 | 32.32 | 0.72 |

The agent performed best when using lower discount factor (γ = 0.8) and a lower learning rate (0.0005), with a mean reward 39.07 and success ratio of 1.00. This shows that short-term decision making, and stable learning helped the agent effectively to switch lanes and maintain speed. On the other hand, higher γ values caused unstable policy updates which led to reducing in overall performance.

* **Intersection-v0:** In this environment we observed the performance was significantly lower than the highway environment due to short term decision making due to cross-traffic.

|  |  |  |
| --- | --- | --- |
| **Configuration** | **Mean Reward** | **Success Rate** |
| γ = 0.95, LR = 0.0005 | 0.80 | 0.50 |
| γ = 0.95, LR = 0.001 | -0.40 | 0.00 |
| γ = 0.8, LR = 0.0005 | -0.20 | 0.00 |
| γ = 0.8, LR = 0.001 | 0.40 | 0.02 |

The results shows that long-term planning (γ = 0.95) was slightly more helpful in intersections, it helps to the predict the cross traffic and avoid collisions better than the lower γ configuration.

* **Roundabout-v0:** This environment was the most challenging due to the dynamic lane changing and speed control requirements. The agent struggled with balancing speed and lane changes leading to high variance in rewards across training run.

|  |  |  |
| --- | --- | --- |
| **Configuration** | **Mean Reward** | **Success Rate** |
| γ = 0.95, LR = 0.0005 | 6.81 | 0.72 |
| γ = 0.95, LR = 0.001 | 4.67 | 0.45 |
| γ = 0.8, LR = 0.0005 | 5.11 | 0.42 |
| γ = 0.8, LR = 0.001 | 5.47 | 0.47 |

The best configuration in this environment was γ = 0.95 with LR = 0.0005, suggesting that a higher discount factor helps in planning lane changes and exits more effectively.

**OBSERVATIONS FOR DQN:**

* Across all environments, the lower learning rates led to more stable and better policy performance. Higher learning rates (0.001) resulted in inconsistent policy performance especially in the intersection environment.
* The discount factor γ = 0.95 performed best in intersection and roundabout environments, where long-term planning was crucial for safe navigation and optimizing lane changes.
* DQN agents showed best results in the highway environment with achieving high-speed, low-collision policies.

**Analysis of Proximal Policy Optimization (PPO) Performance in Autonomous Driving Scenarios:**

**INTRODUCTION:**

We implemented and trained an autonomous driving agent using PPO within the highway-env reinforcement learning framework. The objective was to evaluate how different hyperparameter settings plays a critical role in the model’s ability to navigate dynamic traffic conditions. The agent was trained under different hyperparameter settings, specifically learning rates (LR) and discount factors (γ).

**RESULTS & ANALYSIS FOR PPO:**

The PPO algorithm was trained across multiple environments with different hyperparameter configurations. The goal is to evaluate the effect of learning rate and gamma on cumulative rewards and overall performance.

* **Highway-fast-v0:** In this environment primarily tests the agent’s ability to drive at high speeds while executing safe lane changes. It maintains a balance between aggressive overtaking and collision avoidance.

|  |  |  |
| --- | --- | --- |
| **Configuration** | **Mean Reward** | **Success Rate** |
| γ = 0.95, LR = 0.0005 | 27.78 | 0.98 |
| γ = 0.8, LR = 0.0005 | 23.16 | 0.56 |
| γ = 0.95, LR = 0.001 | 29.28 | 0.95 |
| γ = 0.8, LR = 0.001 | 32.16 | 0.95 |

The best performing PPO model is γ = 0.8, LR = 0.001, achieving a reward of 32.16 and a success rate of 0.95. This suggest that prioritizing short term rewards (lower γ) along with a faster learning rate helps the agent to react more efficiently to traffic conditions which enables optimal lane changing and high-speed driving. On the other hand, lower learning rates results in slow policy optimization which leads to lower rewards and reduced driving efficiency.

* **Intersection-v0:** The intersection environment faced a significant challenge for the PPO agent, which requires exact navigation through cross-traffic and strategic decision making. The intersection environment introduces a frequent stopping, turning and reacting to other vehicles moving in multiple directions.

|  |  |  |
| --- | --- | --- |
| **Configuration** | **Mean Reward** | **Success Rate** |
| γ = 0.95, LR = 0.0005 | -0.20 | 0.00 |
| γ = 0.8, LR = 0.0005 | -1.00 | 0.00 |
| γ = 0.95, LR = 0.001 | 0.20 | 0.04 |
| γ = 0.8, LR = 0.001 | -0.20 | 0.00 |

The best performing configuration is γ = 0.95 and LR = 0.001, achieving the highest reward (0.20) and 0.04 success rate which is not best, but better when compared to other configurations. The results tell that long-term planning with higher LR which is important for handling intersections.

* **Roundabout-v0:** In this environment the agent requires exact lane merging, speed control and efficient exit strategies. The roundabouts require continuous adjustments to avoid collisions and successfully exit the loop.

|  |  |  |
| --- | --- | --- |
| **Configuration** | **Mean Reward** | **Success Rate** |
| γ = 0.95, LR = 0.0005 | 6.67 | 0.65 |
| γ = 0.8, LR = 0.0005 | 7.67 | 1.00 |
| γ = 0.95, LR = 0.001 | 5.46 | 0.48 |
| γ = 0.8, LR = 0.001 | 7.67 | 1.00 |

The configurations with γ = 0.8 and LR = 0.0005 or 0.001, achieved the highest reward (7.67) and 100% success rate. These results indicate that short term decision making (γ = 0.8) is the key factor in successful navigation as roundabouts require quick reactive lane changes and speed adjustments rather than long-term route optimization. On other hand γ (0.95) resulted in lower rewards and success rates when combined with high learning rate which confirms that roundabouts require fast decision making.

**OBSERVATIONS FOR PPO:**

* The lower learning rate (0.0005) results in more stable learning, while high learning rates (0.001) improves performance in structured environments like highway but has caused instability in complex situations like intersections.
* On the other hand, lower discount factors (γ = 0.8) performed best in highway and roundabout environments, which enables fast decision making for lane changing and merging. The higher discount factor (γ = 0.95) is best fit for intersections where long term planning is necessary.
* Overall, the PPO performed best in highway and roundabouts with different configuration values but struggled in intersections due to complexity of cross traffic interactions, leading to negative rewards.

**Analysis of Advantage Actor-Critic (A2C) Performance in Autonomous Driving Scenarios:**

**INTRODUCTION:**

In this section, we trained an autonomous driving agent using the Advantage Actor-Critic (A2C) in the highway-env environment. The main goal is to closely examine the influence of two critical hyperparameters learning rate and discount factor across different driving situations. Unlike DQN and PPO, A2C is a policy gradient method because it simultaneously trains both a policy (actor) and a value estimator (critic) making it important but also sensitive to hyperparameter selection.

**RESULTS & ANALYSIS FOR A2C:**

The A2C trained across three driving situations. The goal is to investigate how variations in two hyperparameters settings i.e. learning rate and discount factor affect the agent’s ability to develop driving strategies.

* **Highway-fast-v0:** In this environment the agent faced the challenges of maintaining high speed, overtaking slow vehicles and avoiding collisions while balancing between aggressive driving and safety.

|  |  |  |
| --- | --- | --- |
| **Configuration** | **Mean Reward** | **Success Rate** |
| γ = 0.95, LR = 0.0005 | 8.49 | 0.30 |
| γ = 0.8, LR = 0.0005 | 14.10 | 0.82 |
| γ = 0.95, LR = 0.001 | 14.25 | 0.25 |
| γ = 0.8, LR = 0.001 | 7.80 | 0.48 |

The best configuration is γ = 0.8, LR = 0.0005 achieving mean reward of 14.10 and success rate of 0.82. This allows the agent to effectively balance high speed driving and safe lane changes by focusing on short-term decision making and stable policy updates. On other hand higher discount factor or higher learning rates leads to unstable performance causing frequent collisions.

* **Intersection-v0:** The intersection situation requires the agent to safely navigate cross-traffic situations, which demands exact timing, lane selection and the ability to predict vehicle movements.

|  |  |  |
| --- | --- | --- |
| **Configuration** | **Mean Reward** | **Success Rate** |
| γ = 0.95, LR = 0.0005 | -0.20 | 0.00 |
| γ = 0.8, LR = 0.0005 | -0.20 | 0.00 |
| γ = 0.95, LR = 0.001 | 0.40 | 0.00 |
| γ = 0.8, LR = 0.001 | -0.20 | 0.00 |

Here the A2C model faced slight difficulties because of the low and negative rewards across all configurations. The best performing configuration (γ = 0.95, LR = 0.001) has managed to have only one positive reward (0.40) but still had a success rate of 0.00. The slight advantage seen with higher discount factors and learning rates suggests long term planning.

* **Roundabout-v0:** In the roundabout environment, the agent requires continuous merging into traffic, dynamic lane switching and timely exits. Unlike other situations the roundabout have a particular dynamic challenge as the agent had to adapt to changing traffic conditions.

|  |  |  |
| --- | --- | --- |
| **Configuration** | **Mean Reward** | **Success Rate** |
| γ = 0.95, LR = 0.0005 | 6.73 | 0.72 |
| γ = 0.8, LR = 0.0005 | 4.25 | 0.33 |
| γ = 0.95, LR = 0.001 | 6.33 | 0.58 |
| γ = 0.8, LR = 0.001 | 5.20 | 0.42 |

In this environment the best performing A2C configuration is γ = 0.95, LR = 0.0005, with a mean of 6.73 and success rate of 0.72. This shows that a higher discount factor combined with a low learning rate allows the agent to effectively handle continuous merging, lane changing and timely exists. On other hand configurations focuses on short term rewards or higher learning rates struggled significantly which highlights the importance of stable policy updates.

**OBSERVATIONS FOR A2C:**

* Overall, in this model the lower learning rate (LR=0.0005) consistently led to more stable and effective performance across three different environments.
* The highway environment has more benefit from a lower discount factor (γ = 0.8) which focuses on short term decision making for effective lane changes and avoiding collisions.
* The intersection and roundabout performed better with a higher discount factor (γ = 0.95) which indicates long term planning helps to manage complex interactions and dynamic merging situations.
* The A2C showed strong performances in highway and roundabout achieving high success rates and stable rewards under optimal parameter settings.

**OVERALL OBSERVATION & CONCLUSION:**

The three reinforcements learning algorithms DQN, PPO and A2C across driving environments. We observed that each algorithm has different and strengths and limitations depending on the task complexity and environment characteristics.

The DQN algorithm performed consistently well in environments that required stable and fast decision making, such as straight forward highway situations. It performed well when focusing on immediate actions and clearly defined rewards, but it struggled a little bit in environments that needed long term decision making.

The PPO algorithm showed stability and efficiency across more complex tasks, such as intersections and roundabouts. PPO’s performance showed a fine-tuning learning rates and discount factors, allowing it to effectively balance exploration and exploitation. The flexible policy updates provided a clear advantage in environments where long term planning is important

At last, the A2C model offered faster initial learning compared to DQN, though it was more sensitive to hyperparameter settings and occasionally exhibited instability in complex situations. This model required more careful parameter tuning and reward shaping to maintain stable improvements over training episodes.

Overall, through these experiments, we learned the importance of matching the RL algorithm to the specific characteristics of the driving environment. A short-term reactive algorithm like DQN can excel in simple task whereas policy gradient methods like PPO and A2C are more suitable to handle complex and dynamic decision-making situations.

**CHALLENGES FACED WHILE WORKING:**

This project has been a valuable learning experience, as it provides practical insights into how reinforcement learning algorithms like DQN, PPO and A2C behave under different hyperparameter settings. While working on this project we gained a deeper understanding of how crucial parameters such as learning rates and discount factors significantly affect the model’s learning process and overall effectiveness. Managing and interpreting large amounts of training data TensorBoard visualization were somewhat overwhelming at first but became clear after detailed analysis. Lastly, we could incorporate additional advanced RL algorithms such as SAC or hybrid methods in future which enrich our understanding of which approaches work best under different environmental conditions. Overall, this project was challenging yet rewarding and great enhancement for our practical understanding of reinforcement learning in autonomous driving situations.