



Article

Multi-Agent Reinforcement Learning for Highway Platooning

Máté Kolat  and Tamás Bécsi * 

Department of Control for Transportation and Vehicle Systems, Budapest University of Technology and Economics, H-1111 Budapest, Hungary; mate.kolat@edu.bme.hu

* Correspondence: becsi.tamas@kjk.bme.hu

Abstract: The advent of autonomous vehicles has opened new horizons for transportation efficiency and safety. Platooning, a strategy where vehicles travel closely together in a synchronized manner, holds promise for reducing traffic congestion, lowering fuel consumption, and enhancing overall road safety. This article explores the application of Multi-Agent Reinforcement Learning (MARL) combined with Proximal Policy Optimization (PPO) to optimize autonomous vehicle platooning. We delve into the world of MARL, which empowers vehicles to communicate and collaborate, enabling real-time decision making in complex traffic scenarios. PPO, a cutting-edge reinforcement learning algorithm, ensures stable and efficient training for platooning agents. The synergy between MARL and PPO enables the development of intelligent platooning strategies that adapt dynamically to changing traffic conditions, minimize inter-vehicle gaps, and maximize road capacity. In addition to these insights, this article introduces a cooperative approach to Multi-Agent Reinforcement Learning (MARL), leveraging Proximal Policy Optimization (PPO) to further optimize autonomous vehicle platooning. This cooperative framework enhances the adaptability and efficiency of platooning strategies, marking a significant advancement in the pursuit of intelligent and responsive autonomous vehicle systems.

Keywords: deep learning; reinforcement learning; platooning; road traffic control; multi-agent systems



Citation: Kolat, M.; Bécsi, T. Multi-Agent Reinforcement Learning for Highway Platooning. *Electronics* **2023**, *12*, 4963. <https://doi.org/10.3390/electronics12244963>

Academic Editor: Jose Eugenio Naranjo

Received: 3 November 2023

Revised: 30 November 2023

Accepted: 8 December 2023

Published: 11 December 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The platooning of vehicles represents a significant advancement within the automotive sector, with the primary objective of enhancing safety, fuel efficiency, travel time, and overall performance. The utilization of autonomous vehicles operating closely in computer-guided platoons promises advantages such as fuel savings, expanded highway capacity, and enhanced passenger comfort. The integration of automation into road traffic holds the potential to address crucial challenges, including accidents, traffic congestion, environmental pollution, and energy consumption, offering valuable solutions to these pressing issues [1]. The concept of automated highways has a longstanding history, dating back to the 1939 World's Fair when General Motors showcased a functional model. Significant research efforts have been dedicated to the development of automated highway systems (AHS), culminating in the National Automated Highway System Consortium (NAHSC) successfully demonstrating the technical feasibility of AHS in San Diego in August 1997. Early steps in AHS research date back to the 1960s [2], with operational testing of prototype equipment as early as 1970. Dr. Shladover has extensively reviewed the progress made in AHS research [3]. Since those initial efforts, the field has witnessed considerable advancement. In 1995, a research report provided a comprehensive account of the aerodynamic performance of platoons, highlighting a notable reduction of approximately 55% in the drag coefficient, which accounts for both vehicle size and velocity, in two-, three-, and four-vehicle platoons. This reduction in drag coefficient translates to lower fuel consumption, as reported to the California Partners for Advanced Transit and Highways (PATH) [4]. Platooning involves the creation of groups or "linked" vehicles that operate collectively on the Automated Highway System, effectively acting as a single

unit [5–8]. These vehicles maintain an extremely close headway or spacing between them, often just a few meters. They are interconnected through headway control mechanisms like radar-based systems. Within the platoon, the lead vehicle, also known as the leader, continually shares information about the AHS conditions and any planned maneuvers with the other vehicles, referred to as followers. In vehicle platooning, a group or platoon of vehicles operates with precise automated control over both longitudinal and lateral movement. These vehicles maintain a consistent and fixed gap between them, even at high highway speeds. This reduced spacing contributes to an increased capacity on the highway. Safety is substantially improved through automation and the coordinated actions of the vehicles. The close proximity of vehicles within the platoon ensures that even rapid accelerations and decelerations have minimal impact on passenger comfort [4], which was conclusively demonstrated during the platooning scenario presented by the PATH program [9].

In summary, one of the primary advantages of platooning is improved fuel efficiency. Vehicles in a platoon experience reduced air resistance, resulting in fuel savings for each vehicle, particularly in long-haul transportation, where fuel costs constitute a substantial portion of operational expenses. This helps optimize traffic flow by reducing the space between vehicles, leading to smoother traffic patterns, less congestion, and improved overall traffic management, particularly on highways and major roadways. Platooning can increase the capacity of highways by allowing vehicles to travel more closely together. This space optimization can lead to better infrastructure utilization, potentially reducing the need for costly expansions. Lastly, platooning's automated systems and communication technologies contribute to increased road safety. Vehicles within a platoon can respond quickly to the lead vehicle's speed or direction changes, reducing the likelihood of accidents.

1.1. Related Work

Extensive literature is available covering various aspects of traffic optimization. In [10], the authors significantly reduce vehicle congestion and pedestrian congestion utilizing multi-agent traffic light control, while an energy management strategy for HEVs is introduced in [11]. In [12], the paper introduces a deep reinforcement learning-based medium access control (MAC) protocol, MCT-DLMA, for optimizing multi-channel transmission in heterogeneous wireless networks, demonstrating superior throughput compared to conventional schemes and enhanced robustness in dynamic environments. In [13], the author presents a novel rate-diverse encoder and decoder for a two-user Gaussian multiple access channel (GMAC); the encoder and decoder are beneficial for traffic optimization as they enable users to transmit with the same codeword length but different coding rates, accommodating diverse user channel conditions. Moreover, there is numerous literature considering using adaptive Traffic Signal Control (TSC) for traffic optimization, such as [14–17]. In [18], the authors propose a method to optimize trajectory and traffic lights to achieve significant reductions in CO₂ emissions and delays. There is also a substantial body of literature in the realm of platooning. In [19], the author explores the effect of the coordinated formation of vehicle platooning, assesses the effectiveness of an autonomous mobility-on-demand system (AMoD), and employs an agent-based model (ABM) for analysis. In [20], the authors investigate the possibility of flexible platooning, utilizing different merging methods. The research of [21] presents an economic model predictive control (EMPC) approach to solve the problem of the vehicle platoon subject to nonlinear dynamics and safety constraints with a focus on fuel efficiency. In [22], ecological cooperative adaptive cruise control (Eco-CACC) is implemented on the ego battery electric vehicle (BEV) of a two-vehicle platoon by using a Nonlinear Model Predictive Control (NMPC) framework concentrating on CO₂ emission reduction. Moreover, there is literature regarding the investigation of vehicle safety under unreliable vehicle communication [23]. The article [24] investigates the advantage of sliding-mode control theory in forming and maintaining stable platooning, while [25] investigates the protection of platooning from possible cyber attacks. Furthermore, reinforcement learning is also used as a preference

in the case of research on platooning. In [26], the authors utilize a particular version of DDPG called the platoon-sharing deep deterministic policy gradient algorithm (PSDDPG), while [27] uses DDPG for mixed mixed-autonomy vehicle platoons. Furthermore, from the examined topic's perspective, many studies in the field of platooning can be distinguished. Aki et al. discuss the need for improved braking systems in platooning based on the results of a driving simulator study [28]; Jones et al. analyze the human factor challenges associated with CACC and suggest research questions that need to be further investigated [29]; Ogitsu et al. analyze equipment failures during platooning and describe strategies to deal with failures based on their severity [30]; and in [31], the authors discuss the value of various information for safely controlling platooning.

1.2. Contribution

Few research papers address the platooning problem within the reinforcement learning (RL) framework using Multi-Agent Reinforcement Learning, and most existing studies involve only a limited number of participants. This paper presents an innovative approach to highway platooning, focusing on enabling adaptive responses to varying velocity conditions and encompassing different platooning speeds. This adaptation is achieved through the introduction of a novel reward mechanism. The research employs a reward strategy in which the following distance is dynamically adjusted according to the platooning's velocity. This adjustment is determined by using the following time as a metric rather than relying solely on a fixed distance. The research paper demonstrates the suitability of this innovative reward method within a multi-agent environment through the utilization of Proximal Policy Optimization. It effectively organizes a platooning scenario characterized by minimal velocity variance among participating vehicles, showcasing its ability to manage fluctuations in velocity conditions adequately. It is important to highlight that we have focused on building a strong foundation rather than directly comparing our method with other solutions. The goal has been to set the stage for future work, providing a starting point that others can use as a reference for more detailed studies. The paper is structured as follows: Section 2 presents the literature background of the utilized methods. Section 3 introduces the implemented traffic environment and the used RL properties. In Section 4, the results are shown and compared to other methods. Ultimately, Section 5 summarizes the research and makes further development suggestions.

2. Methodology

This study presents a highway platooning method, applying Multi-Agent Reinforcement Learning with Proximal Policy Optimization algorithm. Section 2.1 introduces the literature background of reinforcement learning. Section 2.2 presents the basis of MARL, while Section 2.4 shows the motivation of PPO.

2.1. Reinforcement Learning

While reinforcement learning has seen notable successes in the past, as evidenced by the works of [32–35], these earlier approaches encountered various limitations. These limitations encompassed issues related to dimensions, memory, computational demands, and sample complexity, as highlighted by [36]. The emergence of deep neural networks introduced innovative solutions to tackle these challenges.

Deep learning, a groundbreaking advancement in machine learning, has significantly enhanced multiple domains, including object detection, language translation, and speech recognition. Its application in reinforcement learning gave birth to a novel field known as deep reinforcement learning (DRL). In a landmark development in 2015, a DRL algorithm was devised to achieve superhuman-level performance in playing Atari 2600 games solely based on gameplay images, as demonstrated by Mnih et al. (2015). This marked the first instance of a neural network being trained on raw, high-dimensional observations—unstructured data—solely guided by a reward signal. Subsequently, there were

notable achievements, including an RL algorithm defeating the human world champion in AlphaGo, as reported by [37].

Since these breakthroughs, RL has found applications in various domains, spanning from robotics [38] to video games [39] and navigation systems [40]. A fundamental focus within machine learning (ML) involves sequential decision making. This area revolves around determining, through experiential learning, the sequence of actions necessary to navigate an uncertain environment and achieve predefined objectives. The central principle of RL is depicted in Figure 1.

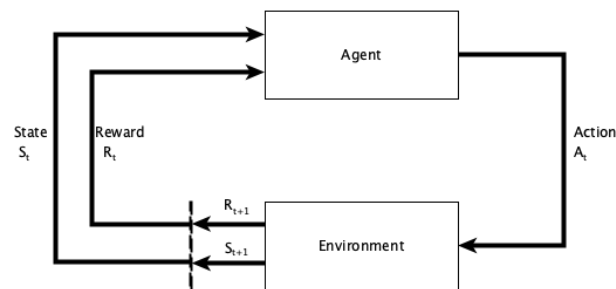


Figure 1. The reinforcement learning basics.

Figure 1 illustrates the core concept of RL, where an agent, acting as a decision-maker, observes its environment and takes actions based on the current state. Each action is assessed for its quality, and the agent is subject to penalties or rewards based on the desirability of its actions. RL is formally modeled as a Markov decision process (MDP), which can be defined in terms of the tuple $\langle S, A, R, P \rangle$:

- S represents the set of observable states.
- A corresponds to the set of possible actions available to the agent.
- R signifies the set of rewards obtained based on the quality of the chosen actions.
- P denotes the policy responsible for determining which action to take in a given state.

An effectively designed reward strategy in RL plays a pivotal role in influencing the agent's neural network (NN) performance as it strives to achieve the desired behavior. This form of learning aims to fine-tune a network's responses, enabling it to generalize its actions effectively in unfamiliar environments that were not part of its training dataset.

2.2. Multi-Agent Reinforcement Learning

Multi-Agent Reinforcement Learning is a subfield of reinforcement learning in artificial intelligence that focuses on training multiple agents to make decisions within an environment. Traditional reinforcement learning typically involves a single agent learning to interact with its environment to maximize its reward signal. MARL has multiple agents, each making decisions in the same environment and often taking actions with unique goals. In certain intricate scenarios, an individual agent alone may not suffice to address a specific problem. In such instances, a Multi-Agent Reinforcement Learning system can be employed, wherein multiple interacting agents operate within a shared environment. Nonetheless, the effectiveness of MARL systems hinges greatly upon the coordination among these agents. The MARL technique can be categorized in various ways based on its coordination strategy. In the work by Claus and Boutilier [41], the authors differentiate between two primary classes: independent learners (ILs) and joint action learners (JALs). For ILs, each agent treats the environment as stationary and learns its policy independently without considering the actions of other agents. In contrast, JALs consider the actions of other agents by integrating reinforcement learning with equilibrium learning methods. ILs are straightforward but may not perform well in scenarios requiring coordination. At the same time, JALs aim to improve the overall performance of a team of agents by considering the impact of their actions on the collective outcome.

2.3. Stable Baseline3

Reinforcement learning (RL) has received considerable attention in recent years because it enables machines to learn and make decisions through interaction with their environment. Among the various tools and libraries available for implementing RL algorithms, Stable Baselines3 (SB3) stands out as a robust and easy-to-use library developed by OpenAI. Stable Baselines3 is an evolution of the previous Stable Baselines and aims to provide a stable and reliable set of implementations for various RL algorithms. The library is based on OpenAI Gym, a toolkit for developing and comparing RL algorithms, allowing one to experiment with different environments and tasks easily. SB3 includes implementations of several state-of-the-art RL algorithms, making it a versatile choice for researchers and practitioners. Supported algorithms include Proximal Policy Optimization (PPO), Deep Deterministic Policy Gradients (DDPG), and more. As the successor to Stable Baselines, SB3 maintains compatibility with OpenAI Gym environments. This means that users can seamlessly integrate SB3 into a variety of existing environments, making it easy to experiment with different tasks. SB3 is designed with a modular architecture that allows users to easily modify and experiment with different components of the algorithm. This flexibility is particularly beneficial for researchers who want to customize their algorithms for specific use cases.

2.4. PPO

Reinforcement learning has made significant progress in recent years, and one notable contribution to this progress is the Proximal Policy Optimization algorithm. PPO is widely popular due to its good empirical performance and ease of implementation. PPO is in the field of model-free RL, particularly in the field of policy gradient methods. Policy gradient methods operate by calculating an estimate of the policy gradient and incorporating it into a stochastic gradient ascent algorithm. The commonly employed gradient estimator takes the following form:

$$\hat{g} = \hat{\mathbb{E}}_t \left[\nabla_{\theta} \log \pi_{\theta}(a_t | s_t) \hat{A}_t^i \right] \quad (1)$$

where π_{θ} represents a stochastic policy and \hat{A}_t is an estimator for the advantage function at timestep t . The expectation $\hat{\mathbb{E}}_t$ denotes the empirical average over a finite batch of samples in an algorithm that alternates between sampling and optimization. Implementations utilizing automatic differentiation software construct an objective function whose gradient corresponds to the policy gradient estimator. The estimator \hat{g} is derived by differentiating the objective:

$$L_{PG}(\theta) = \hat{\mathbb{E}}_t \left[\theta \log \pi_{\theta}(a_t | s_t) \hat{A}_t^i \right] \quad (2)$$

While the idea of performing multiple optimization steps on this loss L_{PG} using the same trajectory is appealing, it lacks a solid justification. Empirically, such an approach often results in excessively large policy updates. PPO is characterized by its focus on optimizing policies while ensuring stability and efficient learning. This is achieved by enforcing “proximal” constraints on policy updates and preventing drastic changes that might destabilize learning. The core idea of PPO is to iteratively update the policy by optimizing a surrogate objective function that includes a clipping mechanism.

$$L_{PPO}^{CLIP}(\theta) = \hat{\mathbb{E}}_t \left[\min \left(\frac{\pi_{\theta}(a_t | s_t)}{\pi_{\theta_{old}}(a_t | s_t)} \hat{A}_t, \text{clip} \left(1 - \epsilon, 1 + \epsilon, \frac{\pi_{\theta}(a_t | s_t)}{\pi_{\theta_{old}}(a_t | s_t)} \right) \hat{A}_t \right) \right] \quad (3)$$

where:

- $\pi_{\theta}(a_t | s_t)$ is the probability of taking action a_t given state s_t under policy π_{θ} ;
- $\pi_{\theta_{old}}(a_t | s_t)$ is the probability under the old policy;
- \hat{A}_t is an estimator of the advantage function at timestep t ;
- $\text{clip}(a, b, x)$ clips x to the interval $[a, b]$;
- ϵ is a hyperparameter controlling the size of policy updates.

This mechanism constrains policy updates to a range that does not differ significantly from the previous policy. This way, PPO prevents policy updates that could lead to catastrophic performance drops and ensures smoother convergence. PPO has several advantages and is recommended for various RL applications. It has impressive sampling efficiency and can learn proficient policies with fewer interactions with the environment. Additionally, PPO provides discrete and continuous action spaces, making it versatile for various problems. Furthermore, PPO includes regularizing entropy, encouraging exploration, and preventing premature convergence to suboptimal policies. This feature is especially useful in scenarios where exploration is essential to discover optimal strategies. In recent years, PPO has been applied to many real-world problems, including robotics, autonomous agents, and gaming agents, proving its adaptability and robustness in various fields.

3. Environment

Creating the right setting is crucial in reinforcement learning because it shapes the training data based on the agent's choices. In this study, we have set up a customized environment. Nine vehicles are driving one after the other on a highway lane. The first vehicle periodically changes its speed randomly or maintains its current speed every 5 s. In response to these speed changes, the other vehicles adjust their acceleration. This adjustment, determined by the training agent, is carefully tuned to maintain a safe and steady following distance. The custom environment is managed through PettingZoo, an extension of Gymnasium specifically designed for multi-agent scenarios.

To make the training process more varied and help the model adapt effectively, we begin each training episode with the vehicles randomly positioned. This randomness introduces a range of traffic situations during training, leading to a robust and flexible outcome.

3.1. State Representation

State representation in reinforcement learning is pivotal in shaping an agent's understanding of its environment. The choice of state representation profoundly impacts an agent's decision-making process, and in the context of platooning, this choice becomes particularly critical.

In this study, the state space corresponds to the dynamics and positions of the vehicles within the platoon. Specifically, the state representation encompasses data regarding the relative distance, global velocity, and global acceleration of the ego vehicle—the vehicle under the agent's control—and the two vehicles behind and the two vehicles in front of it within the platoon. Additionally, the state space includes each vehicle's *inLane* position as a Boolean value. Figures 2 and 3 depict the state space representations under different ego vehicle configurations. Specifically, Figure 2 illustrates the composition of the state space when the third vehicle is the ego. In contrast, Figure 3 showcases the state space when the fourth vehicle assumes the ego role. The ego vehicle is consistently positioned as the third vehicle within the state space representation to maintain consistency across decision-making processes. In cases where there is a deficiency of two vehicles either behind or in front of the ego vehicle, such as in the scenario of the last vehicle, the *inLane* attribute is assigned a value of 0, and the state values are uniformly set to 0.

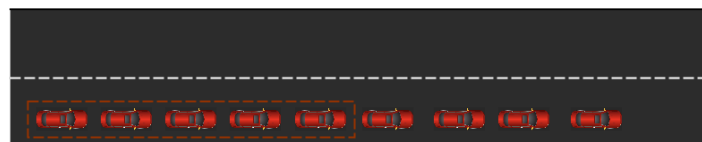


Figure 2. State space representation corresponding to the third vehicle.

This comprehensive state representation grants the RL agent a holistic view of the platooning environment, allowing it to discern not only the behavior of its immediate followers but also anticipate the actions of the vehicles ahead. Considering these neighbor-

ing vehicles' relative dynamics and positions, the agent can make informed decisions to optimize platoon stability and ensure safe and efficient driving.

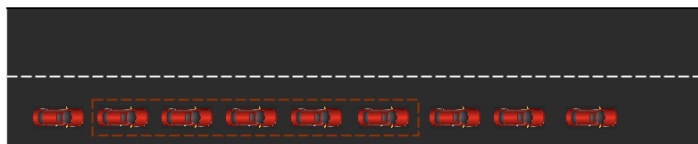


Figure 3. State space representation corresponding to the fourth vehicle.

Furthermore, within this state representation framework, including ego data, which comprises the ego vehicle's velocity and acceleration, provides the agent with a self-awareness crucial for cooperative platooning. This self-awareness enables the agent to factor its actions into the decision-making process, ensuring that it harmoniously integrates with the platoon while adhering to its operational constraints. The above-mentioned characteristics align with the principles of Cooperative Multi-Agent Reinforcement Learning (Cooperative MARL). In Cooperative MARL, agents collaborate to achieve shared objectives, and state representation plays a central role in facilitating effective cooperation. The state representation supports a shared goal of optimizing platoon stability and ensuring safe and efficient driving. Cooperative MARL is designed for scenarios where agents work collectively towards a common objective. The state representation captures the interdependence among vehicles within the platoon. Actions of one vehicle impact others, reinforcing the need for cooperation—a core principle of Cooperative MARL. The inclusion of ego data, representing velocity and acceleration, fosters self-awareness. In Cooperative MARL, self-aware agents contribute meaningfully to the collective decision-making process.

3.2. Action Space

In this dynamic environment, the action space for each vehicle, except the leader, is defined as a continuous range between -1 and 1 , representing acceleration and deceleration. Notably, the rate of acceleration or deceleration is directly proportional to the chosen action value. Furthermore, vehicles are constrained by a minimum speed of 10 m/s, ensuring they cannot come to a complete stop, while their maximum speed is capped at 36.1 m/s. These actions are executed at intervals of 0.1 s, contributing to the precise and coordinated movement of the vehicles within the system.

3.3. Reward

Rewards in reinforcement learning constitute pivotal feedback signals, serving as evaluative metrics for agent performance. These numerical indicators encapsulate the desirability of actions taken within an environment, thereby facilitating the optimization of decision-making processes. The acquisition of rewards engenders adaptive learning, where agents strive to obtain maximize cumulative rewards, emblematic of the quintessential RL paradigm. The rewarding strategy takes center stage in our specific context of optimizing the following time within a dynamic highway platooning system. Our primary objective is to maintain a target following time of 3 s between vehicles. The essence of our rewarding strategy lies in a well-defined threshold of 0.5 s. This threshold serves as a critical boundary, indicating the permissible deviation from the target following time of 3 s. When the actual following time falls within this 0.5 s threshold, signifying that the vehicles are maintaining close proximity to the desired 3 s gap, a maximum reward of 1 is granted. This reward serves as positive reinforcement, encouraging the platoon to maintain an optimal following distance, which is vital for minimizing collision risks and promoting efficient traffic flow. However, if the deviation from the target following time exceeds the 0.5 s threshold, indicating a substantial departure from the desired spacing, a penalty is applied to the reward. The following equation determines the penalty:

$$penalty = max_penalty(-1) \times min(1.0, deviation / target_time) \quad (4)$$

Here, *max_penalty* represents the maximum penalty value set to -1 , signifying the most severe penalty possible. The *min* function ensures the penalty is scaled proportionally to the deviation from the target time. If the deviation equals or surpasses the target time, the penalty reaches its maximum value of -1 , acting as a strong deterrent against such behavior.

Conversely, if the deviation exceeds the target time, the penalty decreases linearly, encouraging vehicles to return to the desired 3 s following time. We have established a bounded reward space, with a minimum reward of -1 and a maximum reward of 1, ensuring that the rewarding strategy remains within defined limits.

3.4. Training

As previously discussed, the carefully selected diverse dataset is crucial for enhancing the network's ability to generalize and yield superior performance. In this study, we ensure the diversity of our training data through the random generation of traffic in each episode. We employ a competitive Multi-Agent Reinforcement Learning technique characterized by immediate rewards. Under this approach, each agent prioritizes its own acceleration decisions and receives rewards based solely on its actions following each step. A training episode concludes either when it reaches the predefined time step limit or when collisions occur. During the training process, the neural network receives the state representation of each vehicle and, in response, provides action recommendations for every agent.

4. Results

Considering road transportation, safety cannot be emphasized enough. This study introduced a highway-based platooning method, where vehicles travel at high speed, highlighting the importance of the adequately selected following distance and time. However, the following distance highly depends on the speed chosen by vehicles participating in road traffic. Evaluating the results with the utilization of the following time metric is necessary. Therefore, the results are evaluated based on the deviation of the vehicle's velocity from the target following time. The proposed approach underwent testing through 1000 individual testing episodes, each featuring unique randomly generated traffic flows to simulate real-world traffic scenarios. This testing protocol was specifically designed to assess the performance and effectiveness of the control strategy developed during the training phase.

4.1. Training Process

Because hyperparameters are the primary limiting factors in all training processes, a random grid search was carried out to identify the optimal hyperparameters of the algorithm. The key parameters can be found in Table 1.

Table 1. The training parameters.

Parameter	Value
Learning rate (α)	0.00005
Discount factor (γ)	0.97
Num. of ep. after params. are upd. (ξ)	20
Num. of hidden layers	4
Num. of neurons	128,256,256,128
Hidden layer activation function	RELU
Layers	Dense
Optimizer	Adam
Kernel initializer	Xavier normal

4.2. Performance Evaluation

As mentioned before, the evaluation measure is the deviation from the target following time of the participants. Figure 4. presents the agents' performance considering the above-mentioned deviation during the 1000 evaluation episodes.

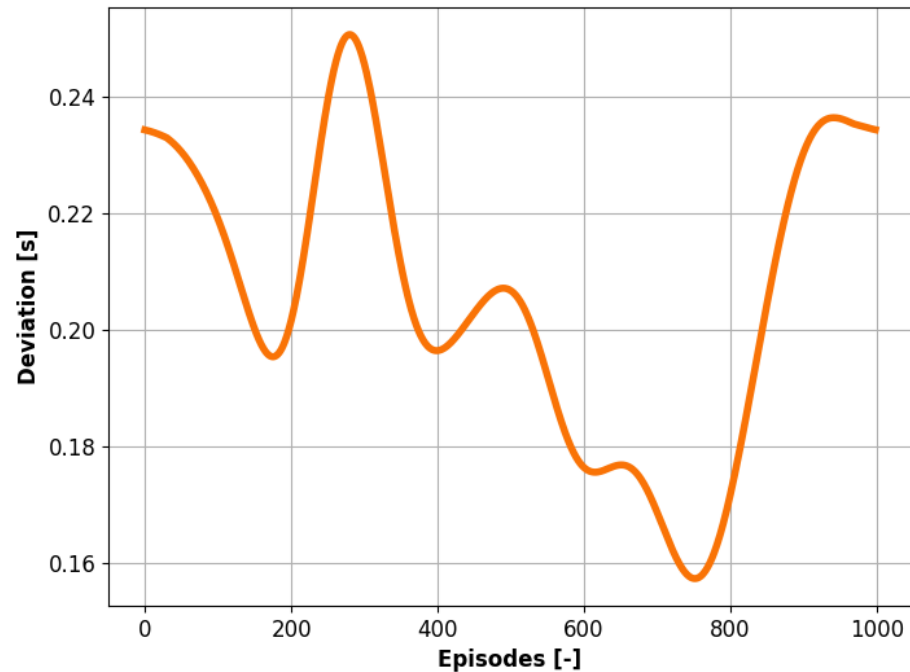


Figure 4. The average deviation from the target following time of the participants

In Figure 4, the absolute deviation of the following distance is depicted in seconds (s) for each episode. The figure illustrates that the participants' average absolute deviation from the target following time per episode consistently falls well below the specified threshold of 0.5 s and ranges between 0.15 and 0.26. Therefore, the figure validates the performance of the agents of the utilized system since the deviation from the standard and suggested 3 s following distance is relatively small. As the leading vehicle executes various dynamic maneuvers, such as acceleration, speed maintenance, and deceleration, there will always be some inherent reaction time, as the state space does not include the leading vehicle's exact next-time step maneuver. Consequently, in this scenario, the objective was not to reach a deviation time of zero but rather to minimize it.

Figure 5 shows the the test result for one specific episode.

Here, the focus is on a specific test episode having 1600 time steps and plotting the velocity of the leading vehicle and the average velocity of the following vehicles during the whole episode. It can be observed that the leading vehicle (yellow line) performs various acceleration and deceleration maneuvers during the episode. The green dotted line denotes the average velocity of the following vehicles. The data show that the following vehicle responds effectively to the acceleration and deceleration actions of the vehicle in front, as shown in the figure, maintaining an appropriate speed range for the lead vehicle throughout the episode. These variations in speed reflect the dynamics of real-world traffic situations, where the lead vehicle often responds to external factors such as road conditions or the actions of other vehicles. Finally, once the speed of the vehicle in front stabilizes after some fluctuations, the vehicle behind will follow suit.

The trend in Figure 6 indicates that the learning algorithm effectively optimizes the policy to make decisions that, on average, result in positive rewards. In RL, achieving convergence to a desired reward level is a key objective, and in this case, a reward above 0.9 suggests that the agent is successfully navigating its environment and making advantageous decisions.

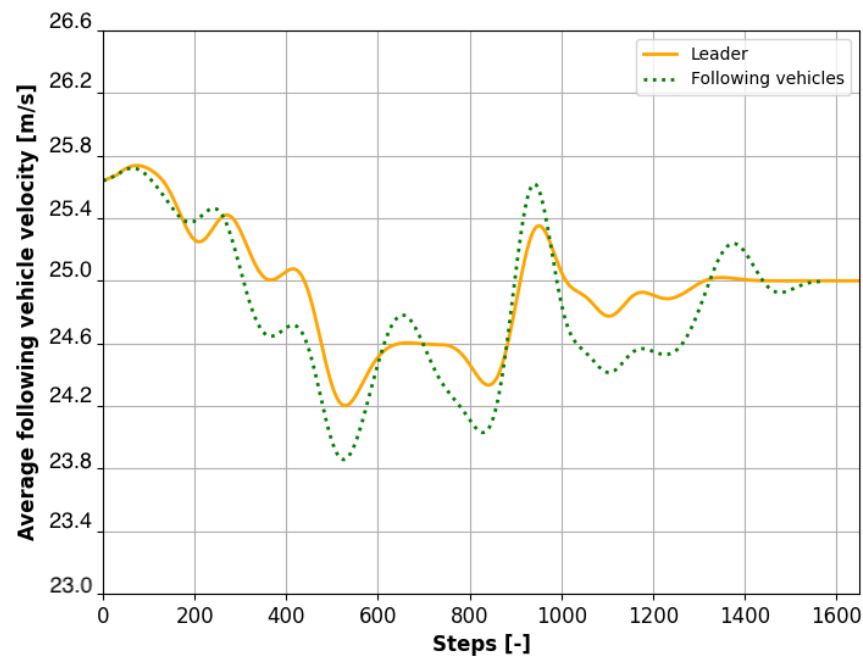


Figure 5. The average velocity of the participants during an episode.

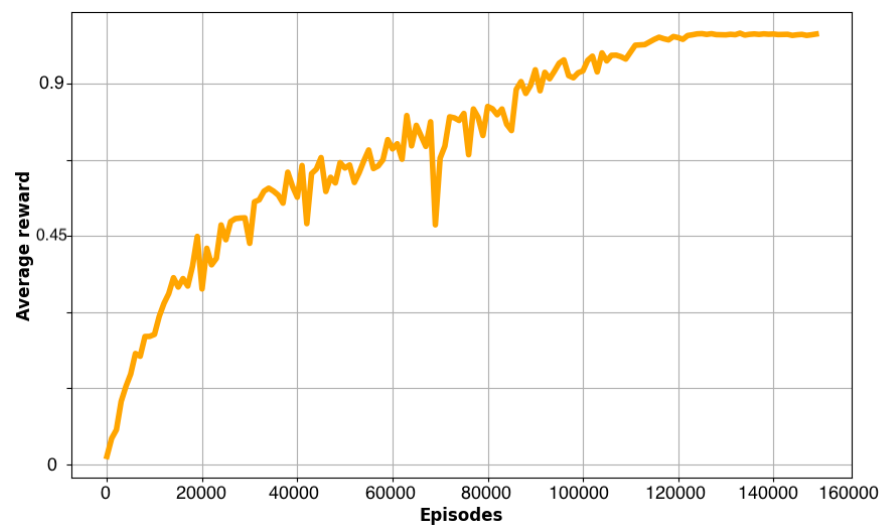


Figure 6. The average reward per episode during training.

5. Conclusions

The advent of self-driving cars offers an opportunity to improve the efficiency and safety of transportation significantly. Platooning, characterized by the synchronized movement of vehicles over short distances, offers an opportunity to alleviate traffic congestion, reduce fuel consumption, and improve overall road safety. The paper presents a promising potential Multi-Agent Reinforcement Learning-based platooning solution utilizing Proximal Policy Optimization. The research results show that the proposed method can establish strong stability within the platooning system. Furthermore, using the introduced reward mechanism, this approach can be effectively adapted to different speed conditions in the system. This study uses a reward strategy in which the following distance is dynamically adjusted based on speed, as the distance is flexibly determined based on following time metrics rather than relying on a fixed distance. This approach results in slight variations in follow-up time within a platoon. However, it is essential to note that various events can occur in road traffic, such as merging into a platoon, which were not investigated in this

study. Therefore, the future focus of this research will be on addressing merging scenarios. Additionally, improved system performance can be achieved by including the leading vehicle decision in the state space. In presenting the outcomes of the study, it is essential to highlight that we have focused on building a solid foundation rather than directly comparing it with other solutions. The goal has been to set the stage for future work. In our next project, we aim to tackle the complexities of merging dynamics in the context of platooning. While our current paper keeps things simple and focuses on the basics of platooning, our upcoming work will dive into the more complex challenges of merging.

Author Contributions: Conceptualization, T.B.; Methodology, M.K. and T.B.; Software, M.K.; Writing—original draft, M.K.; Writing—review & editing, T.B.; Visualization, M.K.; Supervision, T.B.; Funding acquisition, T.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the European Union within the framework of the National Laboratory for Autonomous Systems (RRF-2.3.1-21-2022-00002). Project no. TKP2021-NVA-02 has been implemented with the support provided by the Ministry of Culture and Innovation of Hungary from the National Research, Development and Innovation Fund, financed under the TKP2021-NVA funding scheme. T.B. was supported by BO/00233/21/6, János Bolyai Research Scholarship of the Hungarian Academy of Sciences.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

MARL	Multi-Agent Reinforcement Learning
PPO	Proximal Policy Optimization
AHS	Automated Highway Systems
NAHSC	National Automated Highway System Consortium
PATH	Partners for Advanced Transit and Highways
AMoD	Autonomous Mobility on Demand system
ABM	Agent-based Model
EMPC	Economic Model Predictive Control
TSC	Traffic Signal Control
Eco-CACC	Ecological Cooperative Adaptive Cruise Control
BEV	Battery Electric Vehicle
NMPC	Nonlinear Model Predictive Control
DDPG	Deep Deterministic Policy Gradient
PSDDPG	Platoon Sharing Deep Deterministic Policy Gradient
RL	Reinforcement Learning
ML	Machine Learning
DRL	Deep Reinforcement Learning
MDP	Markov Decision Process
NN	Neural Network
ILs	Independent Learners
JALs	Joint Action Learners

References

1. Krizsik, N.; Sipos, T. Social Perception of Autonomous Vehicles. *Period. Polytech. Transp. Eng.* **2023**, *51*, 133–139. [[CrossRef](#)]
2. Hanson, M.E. *Project METRAN: An Integrated, Evolutionary Transportation System for Urban Areas*; Number 8; MIT Press: Cambridge, MA, USA, 1966.
3. Shladover, S.E. Review of the state of development of advanced vehicle control systems (AVCS). *Vehicle Syst. Dyn.* **1995**, *24*, 551–595. [[CrossRef](#)]
4. Levedahl, A.; Morales, F.; Mouzakitis, G. *Platooning Dynamics and Control on an Intelligent Vehicular Transport System*; CSOIS, Utah State University: Logan, UT, USA, 2010; pp. 1–7.
5. Bergenheim, C.; Huang, Q.; Benmimoun, A.; Robinson, T. Challenges of platooning on public motorways. In Proceedings of the 17th World Congress on Intelligent Transport Systems, Busan, Republic of Korea, 25–29 October 2010; pp. 1–12.

6. Dávila, A.; Nombela, M. Sartre: Safe road trains for the environment. In Proceedings of the Conference on Personal Rapid Transit PRT@ LHR, London, UK, 21–23 September 2010; Volume 3, pp. 2–3.
7. Robinson, T.; Chan, E.; Coelingh, E. Operating platoons on public motorways: An introduction to the sartre platooning programme. In Proceedings of the 17th World Congress on Intelligent Transport Systems, Busan, Republic of Korea, 25–29 October 2010; Volume 1, p. 12.
8. Little, J.D.; Kelson, M.D.; Gartner, N.H. MAXBAND: A versatile program for setting signals on arteries and triangular networks. In Proceedings of the 60th Annual Meeting of the Transportation Research Board, Washington, DC, USA, 12–16 January 1981.
9. Shladover, S.E.; Desoer, C.A.; Hedrick, J.K.; Tomizuka, M.; Walrand, J.; Zhang, W.B.; McMahon, D.H.; Peng, H.; Sheikholeslam, S.; McKeown, N. Automated vehicle control developments in the PATH program. *IEEE Trans. Veh. Technol.* **1991**, *40*, 114–130. [[CrossRef](#)]
10. Wu, T.; Zhou, P.; Liu, K.; Yuan, Y.; Wang, X.; Huang, H.; Wu, D.O. Multi-Agent Deep Reinforcement Learning for Urban Traffic Light Control in Vehicular Networks. *IEEE Trans. Veh. Technol.* **2020**, *69*, 8243–8256. [[CrossRef](#)]
11. Zhang, H.; Peng, J.; Dong, H.; Ding, F.; Tan, H. Integrated velocity optimization and energy management strategy for hybrid electric vehicle platoon: A multi-agent reinforcement learning approach. *IEEE Trans. Transp. Electrif.* **2023**, *1*. [[CrossRef](#)]
12. Zhang, X.; Chen, P.; Yu, G.; Wang, S. Deep Reinforcement Learning Heterogeneous Channels for Poisson Multiple Access. *Mathematics* **2023**, *11*, 992. [[CrossRef](#)]
13. Chen, P.; Shi, L.; Fang, Y.; Lau, F.C.M.; Cheng, J. Rate-Diverse Multiple Access Over Gaussian Channels. *IEEE Trans. Wirel. Commun.* **2023**, *22*, 5399–5413. [[CrossRef](#)]
14. Egea, A.C.; Howell, S.; Knutins, M.; Connaughton, C. Assessment of reward functions for reinforcement learning traffic signal control under real-world limitations. In Proceedings of the 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Toronto, ON, Canada, 11–14 October 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 965–972.
15. Cools, S.B.; Gershenson, C.; D’Hooghe, B. Self-organizing traffic lights: A realistic simulation. In *Advances in Applied Self-Organizing Systems*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 45–55.
16. Gershenson, C. Self-organizing traffic lights. *arXiv* **2004**, arXiv:nlin/0411066.
17. Shabestary, S.M.A.; Abdulhai, B. Deep learning vs. discrete reinforcement learning for adaptive traffic signal control. In Proceedings of the 2018 21st International Conference on Intelligent Transportation Systems (ITSC), Maui, HI, USA, 4–7 November 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 286–293.
18. Feng, Y.; Yu, C.; Liu, H.X. Spatiotemporal intersection control in a connected and automated vehicle environment. *Transp. Res. Part Emerg. Technol.* **2018**, *89*, 364–383. [[CrossRef](#)]
19. Wang, S.; de Almeida Correia, G.H.; Lin, H.X. Effects of coordinated formation of vehicle platooning in a fleet of shared automated vehicles: An agent-based model. *Transp. Res. Procedia* **2020**, *47*, 377–384. [[CrossRef](#)]
20. Maiti, S.; Winter, S.; Kulik, L.; Sarkar, S. The Impact of Flexible Platoon Formation Operations. *IEEE Trans. Intell. Veh.* **2020**, *5*, 229–239. [[CrossRef](#)]
21. He, D.; Qiu, T.; Luo, R. Fuel efficiency-oriented platooning control of connected nonlinear vehicles: A distributed economic MPC approach. *Asian J. Control* **2020**, *22*, 1628–1638. [[CrossRef](#)]
22. Lopes, D.R.; Evangelou, S.A. Energy savings from an Eco-Cooperative Adaptive Cruise Control: A BEV platoon investigation. In Proceedings of the 2019 18th European Control Conference (ECC), Naples, Italy, 25–28 June 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 4160–4167.
23. Li, Z.; Hu, B.; Li, M.; Luo, G. String Stability Analysis for Vehicle Platooning Under Unreliable Communication Links With Event-Triggered Strategy. *IEEE Trans. Veh. Technol.* **2019**, *68*, 2152–2164. [[CrossRef](#)]
24. Peng, B.; Yu, D.; Zhou, H.; Xiao, X.; Fang, Y. A Platoon Control Strategy for Autonomous Vehicles Based on Sliding-Mode Control Theory. *IEEE Access* **2020**, *8*, 81776–81788. [[CrossRef](#)]
25. Basiri, M.H.; Pirani, M.; Azad, N.L.; Fischmeister, S. Security of Vehicle Platooning: A Game-Theoretic Approach. *IEEE Access* **2019**, *7*, 185565–185579. [[CrossRef](#)]
26. Lu, S.; Cai, Y.; Chen, L.; Wang, H.; Sun, X.; Jia, Y. A sharing deep reinforcement learning method for efficient vehicle platooning control. *IET Intell. Transp. Syst.* **2022**, *16*, 1697–1709. [[CrossRef](#)]
27. Chu, T.; Kalabić, U. Model-based deep reinforcement learning for CACC in mixed-autonomy vehicle platoon. In Proceedings of the 2019 IEEE 58th Conference on Decision and Control (CDC), Nice, France, 11–13 December 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 4079–4084.
28. Aki, M.; Zheng, R.; Nakano, K.; Yamabe, S.; Lee, S.Y.; Suda, Y.; Suzuki, Y.; Ishizaka, H. Evaluation of safety of automatic platoon-driving with improved brake system. In Proceedings of the 19th Intelligent Transport Systems World Congress, ITS 2012, Vienna, Austria, 22–26 October 2012.
29. Jones, S. *Cooperative Adaptive Cruise Control: Human Factors Analysis*; Technical Report; Office of Safety Research and Development, Federal Highway Administration: Washington, DC, USA, 2013.
30. Ogitsu, T.; Fukuda, R.; Chiang, W.P.; Omae, M.; Kato, S.; Hashimoto, N.; Aoki, K.; Tsugawa, S. Decision process for handling operations against device failures in heavy duty trucks in a platoon. In Proceedings of the 9th FORMS/FORAMAT 2012 Symposium on Formal Methods for Automation and Safety in Railway and Automotive Systems, Braunschweig, Germany, 12–13 December 2012.

31. Xu, L.; Wang, L.Y.; Yin, G.; Zhang, H. Communication information structures and contents for enhanced safety of highway vehicle platoons. *IEEE Trans. Veh. Technol.* **2014**, *63*, 4206–4220. [[CrossRef](#)]
32. Kohl, N.; Stone, P. Policy gradient reinforcement learning for fast quadrupedal locomotion. In Proceedings of the IEEE International Conference on Robotics and Automation, 2004, Proceedings, ICRA'04, New Orleans, LA, USA, 26 April–1 May 2004; IEEE: Piscataway, NJ, USA, 2004; Volume 3, pp. 2619–2624.
33. Ng, A.Y.; Coates, A.; Diel, M.; Ganapathi, V.; Schulte, J.; Tse, B.; Berger, E.; Liang, E. Autonomous inverted helicopter flight via reinforcement learning. In *Experimental Robotics IX*; Springer: Berlin/Heidelberg, Germany, 2006; pp. 363–372.
34. Singh, S.; Litman, D.; Kearns, M.; Walker, M. Optimizing dialogue management with reinforcement learning: Experiments with the NJFun system. *J. Artif. Intell. Res.* **2002**, *16*, 105–133. [[CrossRef](#)]
35. Tesauro, G. Temporal difference learning and TD-Gammon. *Commun. ACM* **1995**, *38*, 58–68. [[CrossRef](#)]
36. Strehl, A.L.; Li, L.; Wiewiora, E.; Langford, J.; Littman, M.L. PAC model-free reinforcement learning. In Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, USA, 25–29 June 2006; pp. 881–888.
37. Silver, D.; Huang, A.; Maddison, C.J.; Guez, A.; Sifre, L.; Van Den Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M.; et al. Mastering the game of Go with deep neural networks and tree search. *Nature* **2016**, *529*, 484–489. [[CrossRef](#)]
38. Levine, S.; Finn, C.; Darrell, T.; Abbeel, P. End-to-end training of deep visuomotor policies. *J. Mach. Learn. Res.* **2016**, *17*, 1334–1373.
39. Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A.A.; Veness, J.; Bellemare, M.G.; Graves, A.; Riedmiller, M.; Fidjeland, A.K.; Ostrovski, G.; et al. Human-level control through deep reinforcement learning. *Nature* **2015**, *518*, 529–533. [[CrossRef](#)] [[PubMed](#)]
40. Zhu, Y.; Mottaghi, R.; Kolve, E.; Lim, J.J.; Gupta, A.; Fei-Fei, L.; Farhadi, A. Target-driven visual navigation in indoor scenes using deep reinforcement learning. In Proceedings of the 2017 IEEE International Conference on Robotics and Automation (ICRA), Singapore, 29 May–3 June 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 3357–3364.
41. Claus, C.; Boutilier, C. The dynamics of reinforcement learning in cooperative multiagent systems. In Proceedings of the AAAI '98/IAAI '98: Proceedings of the Fifteenth National/Tenth Conference on Artificial Intelligence/Innovative Applications of Artificial Intelligence, Madison, WI, USA, 26–30 July 1998; pp. 746–752.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.