DL24: Eco Driving Strategies for Autonomous Vehicles based on Q-learning and SUMO

Shuo Zhang, Heng Quan, Zerun Liu

[1]Civil and Urban Engineering Department
sz4580@nyu.edu, hq322@nyu.edu, zl3280@nyu.edu

Abstract

Transportation sector in the US contributes 29% to the green house gas emission (GHG) in which 77% is due to land transportation. Leverage autonomous vehicles to reduce GHG levels of vehicles when approaching and leaving a signalized intersection. Our goal of this project is **Reducing fuel consumption** at intersection while minimizing stoppage. Reinforcement Learning (RL) enhances autonomous driving by enabling vehicles to make complex decisions in dynamic environments and continuously adapt their driving strategies for safety and efficiency. It leverages simulations to rapidly learn and refine behaviors, handling everything from daily traffic variations to rare critical situations. Model-free **Reinforcement learning** for multi-agent control is selected to solve this problem. Moreover, SUMO (Simulation of Urban MObility) is essential for training autonomous vehicles (AVs) with reinforcement learning (RL), providing a risk-free, controlled environment for developing and testing driving algorithms. Finally, we proposed a Q-learning based framework to optimize AVs fuel control than traditional car following model.

The public repo is here:

https://github.com/shuo9898-zs/NYU-24Spring-DL-course-project

Introduction

The transportation sector, primarily fueled by gasoline and diesel, remains a major contributor to global greenhouse gas emissions. These fuels, derived from petroleum, release carbon dioxide—a key greenhouse gas—when burned, significantly impacting environmental health. Efficient fuel management in vehicles is therefore crucial in mitigating the environmental footprint of land transportation.

To address these challenges, the Simulation of Urban MObility (SUMO) is employed as a vital tool in our project. SUMO is an open-source, highly portable, and versatile simulation suite designed to model large-scale road networks. It allows for the realistic simulation of vehicular traffic, including the detailed behaviors of each vehicle under various traffic conditions and scenarios. This simulation environment is particularly useful for testing and validating autonomous vehicle technologies before their implementation in real-world scenarios, ensuring that the systems are both efficient and safe.

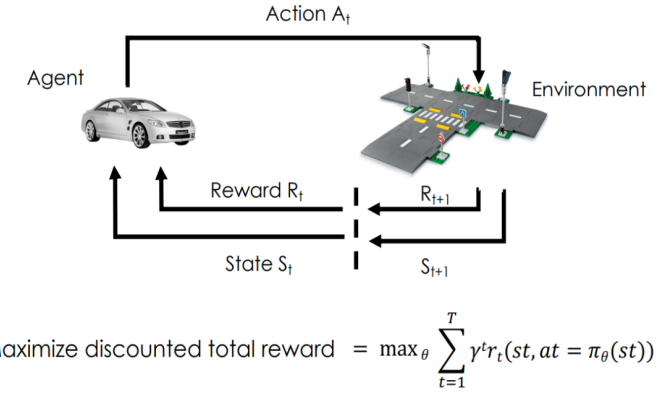

Figure 1: POMDP based Reinforcement Learning framework

Reinforcement Learning (RL) is integral to our strategy for minimizing fuel consumption at intersections by enabling autonomous vehicles (AVs) to make informed, real-time decisions that lead to more efficient driving patterns. As a branch of machine learning, RL operates on the principle of action and reward: agents (in this case, vehicles) interact with a dynamic environment (like traffic systems) and learn from the consequences of their actions through rewards or penalties. This feedback loop allows RL to refine the decisions of AVs continuously.

In the specific context of autonomous driving, model-free RL, which does not require a model of the environment and learns purely from interaction data, is particularly useful. This approach allows AVs to learn optimal driving strategies directly from raw experience, without the need for preconceived models of traffic behavior. Such strategies include learning when and how aggressively to accelerate or decelerate to avoid frequent stopping, thereby conserving fuel. Over time, through trial and error, RL helps AVs develop a nuanced understanding of how to navigate intersections in the most efficient way possible, optimizing fuel usage and reducing emissions effectively. This makes model-free RL a robust tool for enhancing fuel efficiency in the ever-evolving scenarios of urban traffic.

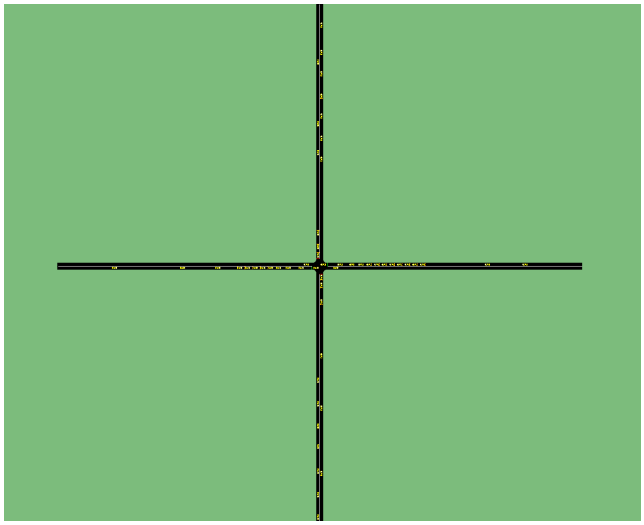

Figure 2: SUMO UI visualization & vehicle control

In summary, we contributes:

- Build a framework by SUMO and model-free RL for emission control
- Create a Signalized intersection environment through SUMO
- self-design states, observations, actions and reward function
- achieve better performance than traditional car-following model, Intelligent driver model (IDM)

Related works

In this section we want to introduce the development of RL, SUMO and our reference. Shoham et.al(Mannor and Shamma 2007) explore the applicability of multi-agent learning in engineering contexts, addressing specific challenges and contrasting the assumptions and concerns of engineering applications with those prevalent in economic game theory. And the field of Cooperative AI (Dafoe et al. 2020)seeks to enhance the ability of both artificial and human agents to collaborate effectively, addressing global and daily cooperative challenges through innovative research in multi-agent systems, game theory, and other interdisciplinary areas, with the aim of fostering cooperation and improving human welfare. The concept of an autocurriculum(Leibo et al. 2019) in multi-agent systems suggests that the intrinsic dynamics of competition and cooperation naturally generate a progression of social tasks, fostering continual innovation and increasing complexity that demand escalating adaptability from the agents involved. The Dyna(Sutton 1991) AI architecture uniquely combines learning, planning, and reactive execution, using learning to refine plans and update world models, enabling direct action from perception, which highlights its versatility and invites comparisons to other systems. SUMO (Simulation of Urban MObility)(Hay 2005) is an open-source traffic simulation software that models large-scale road networks, enabling the realistic simulation of traffic flow, vehicle behavior, and transportation systems for research and planning purposes. This report introduces EcoLight(Agand and Chen 2023), a reinforcement learning-based reward shaping scheme designed to optimize traffic signal control, significantly reducing CO2 emissions and improving travel times, with a comparative performance analysis of algorithms like Q-Learning, DQN, SARSA, and A2C across various traffic scenarios. (Jayawardana and Wu 2022) contributes the code on GitHub, which is our reference.

Methodology and Result analysis

Environment design

In SUMO, we internalize two-way lane at a signalized intersection. The total system travel time will continue 3600 seconds (not the RL training time), figure 2 shows how our framework implement. All vehicles are pure Autonomous driving vehicles, instead of Connected autonomous driving or human-driven vehicles. A partially observable Markov decision process (POMDP) is utilized for our decision making. POMDP requires clear definition, including observation, states, actions and most important reward function. For observations, we have four candidates: Ego-vehicle velocity, Headway, Ego-vehicle position and Lead vehicle position. Headway is the distance between ego-vehicle and previous vehicle. Initially, we only consider two states: Ego-vehicle velocity and Headway. Ego-vehicle velocity is discredited divided into 6 groups (index) from 0-12 m/s. And Headway is discredited divided into 5 groups (index) from 0-100 m. Actions are important, because actions are directly linked to the reward and loss function. Also, the number of actions will influence the quality of control. Theoretically, the more actions we involved in this RL system, the more precise control we can achieve. 7 actions are -3, -2, -1, 0, 1, 2, 3, which are accelerations.

Reward function is always an issue for RL system design, due to Complexity of Objectives, Sparse Rewards, Scaling and Normalization, Temporal Credit Assignment, Non-stationarity and other issues. In addressing the problem of high greenhouse gas (GHG) emissions associated with frequent acceleration and deceleration, we have devised a reward function aimed at promoting speed stability within vehicular traffic. The formulated reward incentivizes vehicles to maintain a constant speed and minimize instances of coming to a complete stop. Additionally, a stability reward has been integrated, which encourages vehicles to sustain a stable velocity. In situations where changes in speed are necessary, this component of the reward function favors gradual variations over abrupt alterations, such as a transition from 0 to 12 units of speed, thereby enhancing both fuel efficiency and emissions profiles. Reward function is shown in equation (1).

$$Reward = movement\ reward + stability\ reward \quad (1)$$

Q-Learning

Q-learning is a model-free reinforcement learning algorithm designed to optimize action-selection policies in fi-

nite Markov decision processes by learning an action-value function, which determines the expected utility of actions in various states. The process involves initializing a Q-table where rows represent states and columns represent actions. The agent then interacts with its environment by selecting actions based on the Q-table, typically using an ϵ-greedy policy to balance exploration and exploitation, observes the outcomes, and updates the Q-values based on the rewards received using the Bellman equation.

The Bellman equation for Q-learning is expressed as:

$$Q(s,a) \leftarrow Q(s,a) + \alpha \left[r + \gamma \max_{a'} Q(s',a') - Q(s,a) \right] \tag{2}$$

In Partially Observable Markov Decision Processes (POMDPs) where the environment's states are not fully observable, traditional Q-learning struggles due to incomplete state information, compelling agents to operate based on a belief state—a probability distribution over potential states refined through Bayesian updating or particle filtering. Adapting Q-learning to these complexities requires significant modifications such as integrating belief states into the learning algorithm and redesigning the loss function to account for the expanded state space and computational demands, which are critical for addressing the inherent uncertainties and ensuring practical applicability through advanced approximation methods.

$$Loss = \frac{1}{N} \sum_{i=1}^{N} \left(r_i + \gamma \max_{a'} Q(s_i', a'; \theta^-) - Q(s_i, a_i; \theta) \right)^2 \tag{3}$$

Where:

- r_i is the reward received after taking action a_i in state s_i,
- γ is the discount factor,
- $Q(s_i', a'; \theta^-)$ is the maximum predicted Q-value for the next state s_i', computed by the target network with parameters θ^-,
- $Q(s_i, a_i; \theta)$ is the predicted Q-value from the network for action a_i in state s_i, with current network parameters θ,
- N is the number of samples in the batch.

Experiment and Results

The microscopic traffic simulator SUMO was utilized to provide data on speed, location, and headways of vehicles, with actions implemented in SUMO through the TraCI API. The training process involved 100 to 300 episodes, with each episode consisting of 10 steps per vehicle to update the Q-table.

The simulation environment consisted of a single intersection with only through-traffic and standard passenger cars. The VT-CPFM fuel consumption model was applied to assess fuel consumption. A fixed-time traffic signal control cycle was used, with uniform vehicle arrivals to ensure consistency across simulations.

The implementation of the Q-learning algorithm in this setup demonstrated a notable improvement in the efficiency

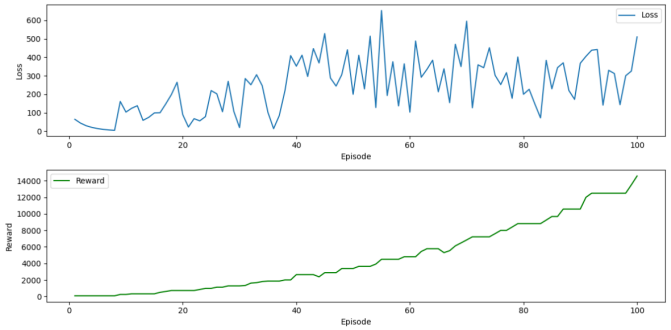

Figure 3: 100 Epoch Training

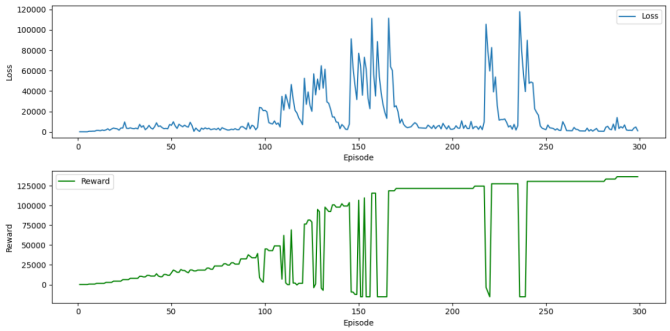

Figure 4: 300 Epoch Training

of AVs at the intersection. The AVs successfully learned to optimize their actions to minimize delays and improve traffic flow. Fuel consumption metrics indicated a reduction in fuel use, reflecting the effectiveness of the VT-CPFM model in conjunction with the optimized driving behaviors. Overall, the combination of Q-learning, SUMO simulations, and the fixed-time signal control proved to be an effective strategy for enhancing AV performance at intersections.

Figure 3 and 4 show our training result. We applied an on-policy training. But due to the limit of computation resource, SUMO has a high requirement for CPU. We stop by 300 epochs. The loss curve exhibits fluctuations, with peaks corresponding to increases in traffic flow. The reward function curve shows a steady increase. Both the loss and reward are cumulative. Figure 5 demonstrates the greater capacity of our method to efficiently regulate the flow of vehicles at the intersection compared to the IDM car-following model applied in SUMO. More vehicles are collected together passing the intersection with few accelerations. As a greater number of vehicles collect at the intersection, enduring minimal accelerations.

Future work and Conclusion

In this project, we leveraged the Q-learning algorithm within the SUMO (Simulation of Urban MObility) framework to reduce fuel consumption and minimize stoppage for autonomous vehicles (AVs) at signalized intersections. Given the significant contribution of the transportation sector to

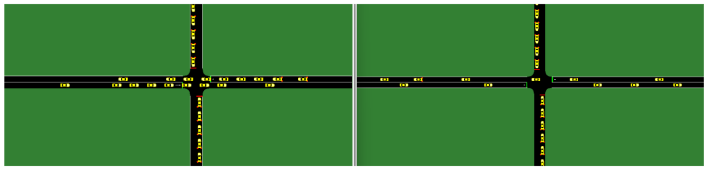

Figure 5: Comparison between our Q-Learning eco-driving(left) method and SUMO based IDM(right)

greenhouse gas emissions, our approach aimed to optimize fuel efficiency through intelligent AV decision-making. The Q-learning framework enabled AVs to learn and refine driving strategies in a simulated environment, effectively handling various traffic conditions. The experimental setup, featuring a single intersection with through-traffic, demonstrated that AVs could minimize delays and improve traffic flow, ultimately reducing fuel consumption. This study highlights the potential of reinforcement learning, particularly model-free RL, in optimizing AV performance and contributing to more efficient and environmentally friendly transportation systems.

During our training process, we refined the sets of states and actions, which resulted in improved performance. However, the current states and observations are quite limited, primarily consisting of speed and headway. Given the complexity introduced by intersections, we have learned from the literature that incorporating traffic signal changes into the state representation could further enhance the training outcomes. Additionally, conducting more epochs and adjusting various parameters could provide further improvements. Therefore, future work should focus on expanding the state space to include traffic signal information and systematically exploring a broader range of parameters through extended training epochs to optimize the model's performance.

References

Agand, I. A., Pedram; and Chen, M. 2023. Deep Reinforcement Learning-based Intelligent Traffic Signal Controls with Optimized CO2 emissions.

Dafoe, A.; Hughes, E.; Bachrach, Y.; Collins, T.; McKee, K. R.; Leibo, J. Z.; Larson, K.; and Graepel, T. 2020. Open problems in cooperative ai. *arXiv preprint arXiv:2012.08630*.

Hay, R. T. 2005. SUMO: a history of modification. *Molecular cell*, 18(1): 1–12.

Jayawardana, V.; and Wu, C. 2022. Learning eco-driving strategies at signalized intersections. In *2022 European Control Conference (ECC)*, 383–390. IEEE.

Leibo, J. Z.; Hughes, E.; Lanctot, M.; and Graepel, T. 2019. Autocurricula and the emergence of innovation from social interaction: A manifesto for multi-agent intelligence research. *arXiv preprint arXiv:1903.00742*.

Mannor, S.; and Shamma, J. S. 2007. Multi-agent learning for engineers. *Artificial Intelligence*, 171(7): 417–422.

Shi, J.; Qiao, F.; Li, Q.; Yu, L.; and Hu, Y. 2018. Application and evaluation of the reinforcement learning approach to eco-driving at intersections under infrastructure-to-vehicle communications. *Transportation Research Record*, 2672(25): 89–98.

Sutton, R. S. 1991. Dyna, an integrated architecture for learning, planning, and reacting. *ACM Sigart Bulletin*, 2(4): 160–163.

Zhang, J.; Tang, T.-Q.; Yan, Y.; and Qu, X. 2021. Eco-driving control for connected and automated electric vehicles at signalized intersections with wireless charging. *Applied Energy*, 282: 116215.