Reducing Societal Cost with Autonomous Vehicles: Multi-Agent Reinforcement Learning with Adaptive Reward

Chanin Eom[1], Dongsu Lee[1], Minhae Kwon[1,2,*]

[1]Department of Intelligent Semiconductors, [2]School of Electronic Engineering
Soongsil University

{eci0623, movementwater}@soongsil.ac.kr, *minhae@ssu.ac.kr

Abstract

Autonomous vehicles (AVs) have been introduced into conventional road environments. In this trend, AVs should provide driving strategies that can minimize the societal impact on the road system. To achieve this, we develop an autonomous driving strategy using multi-agent reinforcement learning (RL) within a multi-agent Markov decision process (MA-POMDP) framework. Furthermore, we employ an adaptive reward model designed to encourage agents to adjust their target speeds based on neighboring traffic conditions. Simulation results confirm that our proposed solution reduces societal costs by achieving lower travel times compared to conventional autonomous driving strategy.

Ⅰ. Introduction

The increasing attention to autonomous driving has led to mixed-autonomous traffic, where AVs and non-autonomous vehicles (NAVs) coexist. In this context, developing a driving strategy that minimizes societal impact is crucial. Deep RL offers a potential solution, as RL-based vehicles can perform well under complex road conditions. The core of RL is reward design, and most RL-based autonomous driving strategies employ a static reward model that encourages the agent to adhere to a fixed target speed [1]. However, an agent with such a static reward model can increase societal costs because it makes decisions with the priority of achieving its fixed goal. In this study, we aim to develop a driving strategy that reduces societal costs. To achieve this, we propose an adaptive reward model that allows the agent to dynamically adjust its target speed.

Ⅱ. Multi-agent Reinforcement Learning with Adaptive Reward Model

In this study, we take into account the mixed-autonomous traffic scenario. A set of N vehicles exists on the road, denoted by $C = \{C_{AV} \cup C_{NAV}\}$, where $C_{AV} = \{i | 1 \le i \le I\}$ is the set of AVs, and $C_{NAV} = \{i | I+1 \le i \le N\}$ is the set of NAVs.

Ⅱ.1. Multi-agent Partially Observable Markov Decision Process Design

In this work, the road consists of multiple AVs. It can be modeled by an MA-POMDP. The MA-POMDP is represented as a tuple $\langle I, S, A, O, A, R, \gamma \rangle$, where each element represents the agent $i \in I$, the state $s_t \in S$, the observation $o_{t,i} \in O$, the action $a_{t,i} \in A$, the reward function R, and the discount factor γ. The agent i can observe within a range of $2W$ distance across H lanes.

1) Observation: The observation $o_{t,i} \in O$ can be defined as follows.

$$o_{t,i} = \left[\Delta \boldsymbol{p}_{t,i}^T, \Delta \boldsymbol{v}_{t,i}^T, \boldsymbol{\rho}_{t,i}^T, \boldsymbol{\zeta}_{t,i}^T, v_{t,i}\right]^T,$$

where $\Delta \boldsymbol{p}_{t,i} = \left[\Delta p_{t,i,l_1}, \cdots, \Delta p_{t,i,H}, \Delta p_{t,i,f_1}, \cdots, \Delta p_{t,i,f_H}\right]^T$ is the relative distance to the leader/follower for each lane. The leader and follower refer to vehicles that have the minimum relative distance to the agents in their lane. $\Delta \boldsymbol{v}_{t,i} = \left[\Delta v_{t,i,l_1}, \cdots, \Delta v_{t,i,H}, \Delta v_{t,i,f_1}, \cdots, \Delta v_{t,i,f_H}\right]^T$ denotes the relative speed about leader/follower for each lane and $\boldsymbol{\rho}_{t,i} = \left[\rho_{t,i,1}, \cdots, \rho_{t,i,H}\right]^T$ represents the traffic density per lane. The persistence of lanes, denoted by $\boldsymbol{\zeta}_{t,i} = \left[\zeta_{t,i,1}, \cdots, \zeta_{t,i,H}\right]^T$ means the distance remaining until the lane connects or disconnects within the range of W. The absolute speed of the agent is defined as $v_{t,i}$.

2) Action: The agents perform two types of action $a_{t,i} = \{a_{t,i,acc}, a_{t,i,lc}\}$. The acceleration control action $a_{t,i,acc} \in [a_{min}, a_{max}]$ has a continuous space ranging from minimum value a_{min} to maximum value a_{max}. The lane change control $a_{t,i,lc} \in \{-1, 0, 1\}$ is defined in discrete action space. Herein, -1 and 1 mean lane changes to the right and left, respectively. The agent maintains the current lane when $a_{t,i,lc} = 0$.

3) Reward: The reward of the agent i at time t can be defined as follows.

$$R_{t,i} = \begin{cases} R_{t,crash}, & accident \\ \sum_{x=1}^{4} \eta_x R_{t,i,x}, & otherwise \end{cases} \quad (1)$$

In (1), $R_{t,i,x}$ represents the reward terms, with each term weighted by $\eta_x \ge 0$. If the agent collides with a neighboring vehicle, it incurs a maximum penalty $R_{t,crash}$. Otherwise, the agent receives a reward calculated as the linear combination of the weighted reward terms.

The first component $R_{t,i,1}$ evaluates how well the agent achieves the target speed v^*, without running beyond the speed limit v_{lim}.

$$R_{t,1} = \begin{cases} \frac{v_{t+1,i}}{v^*}, & v_{t+1,i} \le v^* \\ \frac{v_{lim} - v_{t+1,i}}{v_{lim} - v^*}, & v_{t+1,i} > v^* \end{cases} \quad (2)$$

In (2), when $v_{t+1,i} = v^*$, the agent gets maximum reward and incurs penalty when $v_{t+1,i} > v_{lim}$.

The second component $R_{t,2}$ is designed for successful lane change action. Therefore, the value of $R_{t,2}$ becomes zero when the agent does not change lanes ($|a_{t,i,lc} = 0|$).

$$R_{t,2} = |a_{t,i,lc}|(\Delta p_{t+1,i,\hat{l}} - \Delta p_{t,i,\hat{l}}), \quad (3)$$

where $\Delta p_{t,i,\hat{l}}$ represents a relative distance to the same lane leader. The agent gets a reward when $\Delta p_{t+1,i,\hat{l}} > \Delta p_{t,i,\hat{l}}$ because this lane change action ensures a greater

* Corresponding Author

driving range. Conversely, the agents receive a penalty when $\Delta p_{t+1,i,\hat{l}} - \Delta p_{t,i,\hat{l}} < 0$.

Both $R_{t,3}$, $R_{t,4}$ are related to safe driving. $R_{t,3}$ gives a penalty when the agent violates the safety distance to the same lane leader $\delta_{t,i,\hat{l}}$ as follows.

$$R_{t,3} = min\left[0,1 - \left(\frac{\delta_{t,i,\hat{l}}}{\Delta p_{t+1,\hat{l}}}\right)^2\right] \quad (4)$$

Based on (4), $R_{t,3}$ becomes negative if the relative distance to the same lane leader is less than the safety distance $\Delta p_{t+1,i,\hat{l}} < \delta_{t,i,\hat{l}}$. The safe distance to the same lane follower $\delta_{t,i,\hat{f}}$ is considered in $R_{t,4}$.

$$R_{t,4} = |a_{t,i,lc}| min\left[0,1 - \left(\frac{\delta_{t,i,\hat{f}}}{\Delta p_{t+1,\hat{f}}}\right)^2\right] \quad (5)$$

This component only works when the agent changes lane ($|a_{t,i,lc} = 0|$). It is because maintaining a safe distance is a duty for followers.

Ⅱ.2. Adaptive Reward Model

In this subsection, we provide a proposed adaptive reward model. We consider reward functions adaptively adjusting reward standards by road conditions.

Definition 1 (Adaptive target speed): The agent with the adaptive reward model dynamically sets the target speed $v^*_{t,i} = v^\dagger_{t,i}$ in the first reward term $R_{t,1}$ as the speed of the fastest leader, i.e.,

$$v^\dagger_{t,i} = clip(v_{t,i} + max(\Delta v_{t,i,l}), v^*_{min}, v^*_{max}). \quad (6)$$

In (6), $\Delta v_{t,i,l} = [\Delta v_{t,i,l_1}, \cdots, \Delta v_{t,i,l_H}]$ means the relative speed vector of leaders. v^*_{min} and v^*_{max} represent the lower and upper bounds of the adaptive target speed, respectively. Based on this definition, the first reward term in (2) can be modified as follows.

$$R_{t,1} = \begin{cases} \frac{v_{t+1,i}}{v^\dagger_{t+1,i}}, & v_{t+1,i} \le v^\dagger_{t+1,i} \\ \frac{v_{lim} - v_{t+1,i}}{v_{lim} - v^\dagger_{t+1,i}}, & v_{t+1,i} > v^\dagger_{t+1,i} \end{cases} \quad (7)$$

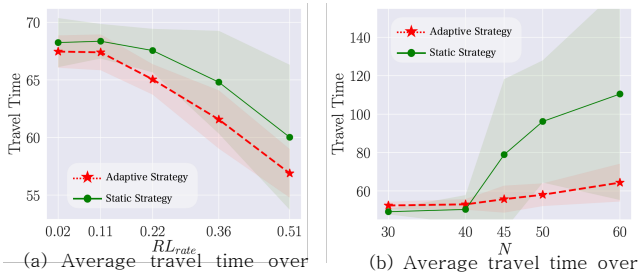
Ⅲ. Simulation Results

In this study, we set up the road environment and perform simulation using the FLOW framework [2]. The total number of vehicles is $N = 45$, including 10 AVs. The speed limit of the road is set to $v_{lim} = 31.389m/s$. AVs have an acceleration range with a maximum acceleration of $a_{max} = 5.4m/s^2$ and a minimum acceleration of $a_{min} = 5.4m/s^2$. Each episode lasts for a total of $3000ts$, where $1ts$ is equivalent to $0.1\ second$. The minimum and maximum target speed of agents are $v^*_{min} = 5m/s$ and $v^*_{max} = 22m/s$. To train agents, we use a deep deterministic policy gradient algorithm [3].

Ⅲ.1. Baselines and Evaluation Methods

To evaluate the societal impact of autonomous driving, we examine the following strategies.

- **Adaptive strategy**: This is the proposed RL-based strategy. The agent is trained with the adaptive reward model in (7).
- **Static strategy**: The agent with this strategy is trained through the static reward model in (2). The target speed is set to $v^* = 13.686m/s$.



(a) Average travel time over autonomous vehicle ratio RL_{rate}

(b) Average travel time over traffic volume N

Figure 1. Travel time comparison of adaptive and static

Ⅲ.2 . Travel Time Comparison

Figure 1 illustrates the comparison of average travel time over the AV ratio RL_{rate} and traffic volume N. In this figure, the solid line represents the average speed of all vehicles on the road, and the shaded area indicates one standard deviation based on 10 random seeds.

Figure 1(a) displays the travel time across various RL_{rate} when the number of vehicles in the road $N = 45$. The results demonstrate that the proposed solution consistently achieves faster travel times with narrower variance compared to the static approach. This supports the effectiveness of the proposed adaptive reward model. Figure 1(b) shows the travel time under heavy traffic conditions when $RL_{rate} = 1$. In this figure, the travel time for both RL-based strategies increases as the traffic volume N increases. Notably, the increase rate for the adaptive strategy is significantly lower than that for the static strategy. This result confirms that the proposed solution can reduce societal costs under heavy traffic conditions.

Ⅳ. Conclusion

In this study, we propose an autonomous driving strategy aimed at reducing the societal costs of road systems. To achieve this, we introduce an adaptive reward model, enabling the agent to adapt to the surrounding traffic flow quickly. Simulation results demonstrate that our proposed solution can effectively reduce travel time compared to the static strategy.

ACKNOWLEDGMENT

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2023-00278812)

REFERENCE

[1] L. Xiong, et al., "Integrated decision making and planning based on feasible region construction for autonomous vehicles considering prediction uncertainty", *IEEE Transactions on Intelligent Vehicles*, vol. 4, no. 11, pp. 4515-4523, 2023.

[2] C. Wu, et al., "FLOW: A modular. learning framework for mixed autonomy traffic," *IEEE Transactions on Robotics*, vol. 38, no. 2, pp. 1270-1286, 2021.

[3] T. Lillicrap, et al., "Continuous control with deep reinforcement learning," *ICLR*, 2016.