Connected Autonomous Vehicle Platoon Control Through Multi-agent Deep Reinforcement Learning

Guangfei Xu[1,2], Bing Chen[3], Guangxian Li[4], and Xiangkun He[5](✉)

[1] Shandong University of Technology, Zibo 255000, China
gfxu@sdut.edu.cn
[2] Liaocheng Academy of Agricultural Sciences, Liaocheng 252000, China
[3] Bentron Information Technology Co. Ltd., Shenzhen 518000, China
gaae@esoar.com.cn
[4] Guangxi University, Nanning 530000, China
[5] School of Mechanical and Aerospace Engineering, Nanyang Technological University, Singapore 639798, Singapore

Abstract. The rise of the artificial intelligence (AI) brings golden opportunity to accelerate the development of the intelligent transportation system (ITS). The platoon control of connected autonomous vehicle (CAV) as the key technology exhibits superior for improving traffic system. However, there still exist some challenges in multi-objective platoon control and multi-agent interaction. Therefore, this paper proposed a connected autonomous vehicle latoon control approach with multi-agent deep reinforcement learning (MADRL). Finally, the results in stochastic mixed traffic flow based on SUMO (simulation of urban mobility) platform demonstrate that the proposed method is feasible, effective and advanced.

Keywords: Intelligent transportation system · Connected autonomous vehicle · Multi-objective platoon control · Multi-agent deep reinforcement learning

1 Introduction

In recent years, with the development of intelligent transportation system (ITS), people pay more attention to congestion, accident, fuel economy, et al. [1–3]. However, when vehicle flows running together, complex dynamic environment may make the running of vehicle flows hard to decide a target speed to deal with dynamic environment. It may be more difficult for vehicle flows satisfying all concerned objectives-high traffic efficient, energy, safe, and driving smoothness.

Single autonomous vehicle maybe relatively easy to obtain a proper speed to fulfil above mentioned aspects with artificial intelligent technology. It can be a important and effective way to solve multi-objectives problems in dynamic environment [4]. Deep learning (DL) as well as reinforcement learning (RL) are two main methods which are widespread adopted to make speed decision one after another [5]. Moreover, DL and RL make it easier to deal with the dynamic environment than other methods [6].

W. Xiang et al. (Eds.): BROADNETS 2021, LNICST 413, pp. 239–248, 2022.
https://doi.org/10.1007/978-3-030-93479-8_16

Lots of methods were proposed for single autonomous vehicle to obtain the proper speed. A rolling-horizon method can be effective to cope with complex trajectories [7]. However, proper speed which can be adequate to fulfill more objectives should be considered [8]. In the process of application of DL or RL, challenges may happen with policy prematurely converging to a local optimum. Therefore, research [9] considered PPO with entropy constraint to make the results better.

However, the learned speed may not suitable when put it into the convoy speed control. Namely, challenges also exist with how to determine a proper speed to make the whole convoy be high traffic efficient, safe and energy at the same time when facing with dynamic environment.

Therefore, the exploration of convoy speed control speed decision-making has become a hot spot. [10] designed different network to state a RL control method for CAVs to solve traffic congestion problem. And penetration rates are set with 2.5% which can be effective to have a better running flow. To improve the ability of RL control method, [11] setup four benchmarks to apply for different traffic problems.

Although multi-agents are concerned and applied in above researches, less attention is paid to multi-objectives emission of multi-agents in convoy speed control which make the convoy be put in a double squeeze.

As MADRL combines with both the advantage of deep neural network (DNN) and RL which can deal with large-scale dynamic information effectively when interacting with a dynamic environment.

Therefore, this paper proposes a connected autonomous vehicle platoon control through multi-agent DRL method. And the key contributions can be summarized as follows:

Firstly, a traffic control strategy is made using DRL with CAVs to deal with multi-objectives mission including high traffic efficient, safe and energy together on open road networks. It can balance various aspects for vehicle flow to achieve synthetically optimal state.

Secondly, it also be demonstrated that DRL can be adjusted to fulfil the requirement of convoy speed control. Namely, several traffic modes are formed which can be selected according to traffic situation.

The remainder of the article is organized as follows. Section 2 state the basic related knowledge of MADRL. Section 3 outlines the RL and multi-objective problem formulation for traffic efficient, safe and energy in open highway networks. Finally, Sect. 4 showed the simulation results of the proposed method.

2 Preliminaries

2.1 Markov Decision Processes

Markov Decision Processes (MDPs) is a description of transition from current state to next state. It is usually represented by a tuple: (S, A, R, P). Where, S means all the states of model including current state and next state, A is actions taken by model in the current state, R means the reward that the adopted A at the current state, P is the transition probability function [12].

2.2 Actor-Critic

Actor-Critic model is made up with the actor and critic model. The critic model updates through state-value function $V(s)$ and the action is evaluated by action-value function $Q(a|s)$. The actor model updates the critic model with the direction to make $V(s)$ higher [13].

2.3 Policy Gradient (PG)

PG method mainly considers the reward of the policy. The obtain of optimal policy is to use gradient descent [14]:

$$\overline{R} = E_t[\nabla_\theta \log \pi_\theta(a_t|s_t)\hat{A}_t], \tag{1}$$

where $\hat{A}_t$ is advantage function, π_θ is policy about parameter θ. And advantage function is written as follows:

$$\hat{A}_t = Q_{\pi_\theta}(s_t, a_t) - V_{\pi_\theta}(s_t), \tag{2}$$

To make the policy develop to the better way, a loss function is set as:

$$L(\theta) = E_t(\log \pi_\theta(a_t|s_t)A_t). \tag{3}$$

2.4 PPO

PPO is adequate to continuous state-action space. PPO usually has two developed forms: PPO-Penalty and PPO-Clip. The former is usually adopted for its simplified form. The purpose of PPO-Clip is to make the old and new policy similar when it is update [15].

PPO-Clip updates θ for the following equation:

$$\theta = \arg\max_{\theta} \underset{s,a \ \pi_{\theta_{\text{old}}}}{E} [L(s, a, \theta_{\text{old}}, \theta)]. \tag{4}$$

Let $r_t(\theta) = \frac{\pi_\theta(a|s)}{\pi_{\theta_{\text{old}}}(a|s)}$ donates the probability ratio for the current policy and the old policy. To obtain the objective, loss function $L(s, a, \theta_{\text{old}}, \theta)$ can be described as:

$$L(s, a, \theta_{\text{old}}, \theta) = \min(r_t(\theta)\hat{A}_t, clip(r_t(\theta), 1 - \varepsilon, 1 + \varepsilon)A_t), \tag{5}$$

where ε is a hyperparameter.

Then an advantage function is set to evaluate the update effectiveness:

$$L_{\hat{A}_t} = -\sum_{t=1}^{T} \left(\sum_{t' > t} \gamma^{t'-t}\tau_{t'} - V_{\pi_\theta}(s_t)\right)^2, \tag{6}$$

where γ represents discount factor. $\tau_{t'}$ means the reward at time *t'*.

3 Experimental Setup

3.1 Flow: Working Environment

The research of this paper is based on Flow [16]. Flow is open source which can be easy access and expand. Flow supports custom modules and permits the research of complex environments, agents, metrics, and algorithms. Flow is built upon SUMO (Simulation of Urban Mobility) [17] which is used to set vehicle and traffic model, Ray RLlib [18] which is used to execute reinforcement learning [19], and OpenAI gym [20] which is used to go on the MDP.

3.2 Problem Setup

This article is concerned with multi-objectives optimization for multi-agent in convoy speed control. Moreover, how to make the whole convoy be high traffic efficient, safe, energy and driving smoothness at the same time when there are only proportionate connected autonomous vehicles controlled by DRL and the other vehicles are human-driven vehicles in the convoy. And the human-driven vehicles are driven by the Intelligent Driver Model (IDM) which is set based on rules in SUMO.

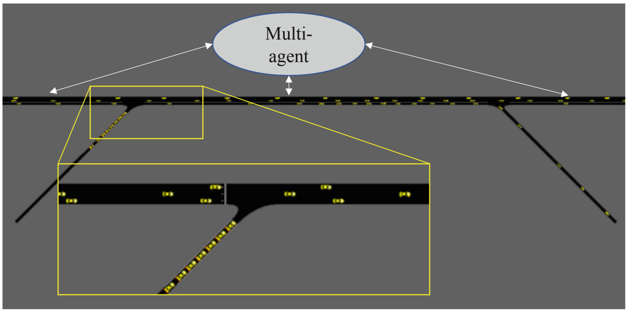

Fig. 1. Open highway network

3.3 Network Configuration

The setup of network can be seen in Fig. 1. It mainly includes a straight highway network and an on-ramp road to make the environment dynamic. The whole inflow rate and the on-ramp inflow rate are set as 4000 and 800 per hour, respectively. CAVs with a centralized controller are trained via MADRL to obtain multi-objectives. The length of main road, on ramp road and off ramp road are set as 1500 m, 250 m and 250 m, respectively. Meanwhile, lane number of them are set as 3, 1, 1, respectively.

3.4 Human-Driven Vehicles

The acceleration and deceleration of human-driven vehicles driven by IDM can be described as the following car following model:

$$f(l_n, v_{fn}, v_n) = a\left[1 - \left(\frac{v_n}{v_0}\right)^{\delta} - \left(\frac{l^*(v_n, \Delta v_n)}{l_n}\right)^2\right] \tag{7}$$

where Δv_n is relative velocity with the preceding vehicle, denoted by:

$$\Delta v_n = v_{fn} - v_n \tag{8}$$

where l^* is the desired headway of the vehicle which can be obtained by:

$$l^*(v_n, \Delta v_n) = s_0 + \max\left(0, v_n T + \frac{v_n \Delta v_n}{2\sqrt{ab}}\right) \tag{9}$$

where s_0, v_0, T, δ, a, b are calibrated parameters to model highway traffic [21].

3.5 Autonomous Vehicles

In the convoy speed control, the CAVs are added with a certain percentage to influence the whole vehicle flow in the network. The CAVs can be seen as multi-agents whose actions (acceleration or deceleration in this paper) are sampled from DRL strategy considering multi-objectives including traffic efficient, safe and energy. The total inflow rate of the network is set as 4000 per hour and the inflow rate of autonomous vehicles is set 20% of total inflow.

3.6 Observations and Actions

The observation space of the learning agent is decided by the multi-objectives which consists of speed, acceleration, fuel consumption, distance between the autonomous vehicle and other vehicles in front and rear, respectively. To improve the training speed and obtained a better training effectiveness, all the observation vectors were normalized [22].

The action space consists of acceleration n of each autonomous vehicles n. Considering the real situation of vehicles, the acceleration can not be infinite. Therefore, the acceleration is clipped into the range [−1, 1] in this paper.

3.7 Reward Designation

The speed control of CAVs should consider multi-objective tasks including traffic efficiency, fuel consumption, safety, driving smoothness at the same time. The designation of reward function can make the training result fulfill requirement.

(1) Traffic efficiency
Traffic efficiency is usually related to the speed of CAVs. And the speed of CAVs should not change sharply for the requirement of response time of all related CAVs. Therefore, a working efficiency reward function can be considered:

$$r_1 = \begin{cases} e^{-k_1 \cdot v_n} & v_{\min} \leq |v_n| \leq v_{\max} \\ -e^{-k_1 \cdot v_n} & v_n > v_{\max} \\ -e^{-k_1 \cdot (v_n + v_{limit})} & v_n < v_{\min} \end{cases}, \tag{10}$$

where $v_{\min}$ and $v_{\max}$ are the vehicle speed range respectively, v_{limit} is network speed limit, v_0 is the speed of CAVs, k_1 is a dynamically adjustable constant.

(2) Fuel consumption
Fuel consumption is considered in convoy control. The running convoy should be limited by the consumption as:

$$r_2 = e^{-k_2 \cdot Q_{cn}}, \tag{11}$$

where Q_{cn} is the fuel consumption of CAVs n, k_2 is a constant.

(3) Safety
When the CAVs are driving on road, Static or dynamic obstacles including surrounding vehicles, pedestrians, signal lights, et al. make it danger for CAVs. Therefore, a safety reward function can be set considering the distance between CAVs and others:

$$r_3 = -\frac{1}{\min(d_{fn}, d_{rn}) + 1}, \tag{12}$$

where d_f and d_r mean the distance between the CAVs and others in front and rear, respectively.

(4) Driving smoothness
Frequent acceleration and deceleration may make the convoy not smooth. Therefore, considering driving smoothness, the reward function can be set as:

$$r_4 = -1000 * |a_n| \tag{13}$$

where a_x and a_y mean the longitudinal and lateral acceleration, respectively.

(5) Multi-objectives
To make the training model be comprehensive in the above aspects, a multi-objective reward function can be obtained:

$$r = \frac{w_1}{\|\boldsymbol{w}\|_1} r_1 + \frac{w_2}{\|\boldsymbol{w}\|_1} r_2 + \frac{w_3}{\|\boldsymbol{w}\|_1} r_3 + \frac{w_4}{\|\boldsymbol{w}\|_1} r_4, \tag{14}$$

where w_i means the weights considering above four objectives, $\boldsymbol{w} = [\, w_1 \quad w_2 \quad w_3 \quad w_4 \,]$ is the weight vector.
The platoon speed matching the dynamic environment can be obtained by setting the proper value of the weight. The weight vector is set as: $\boldsymbol{w} = [\, 1 \quad 2 \quad 1 \quad 1 \,]$.

3.8 Neural Network Designation

In this paper, we consider a four-layer neural network structure to train the model. The neural network mainly has an input layer, two hidden layers, and an output layer. The hidden layers include 128 neurons. The output is acceleration of CAVs. The states of both set CAVs agents include 8 dimensions which are shown in Table 1.

Table 1. The neural network input variables

	Variables	Input meaning	Unites
1	v_n	Speed of autonomous vehicle n	m/s
2	v_{fn}	Velocity of other vehicle in front	m/s
3	v_{rn}	Velocity of other vehicle in behind	m/s
4	d_{fn}	Distance with front vehicle	m
5	d_{rn}	Distance with behind vehicle	m
6	n_i	Lane numbers	
7	a_n	Longitudinal acceleration of autonomous vehicle n	m/s^2
8	Q_{cn}	Fuel consumption of autonomous vehicle n	

4 Simulation

To verify the effectiveness of the proposed connected autonomous vehicle latoon control approach with MADRL, a training is carried out in SUMO.

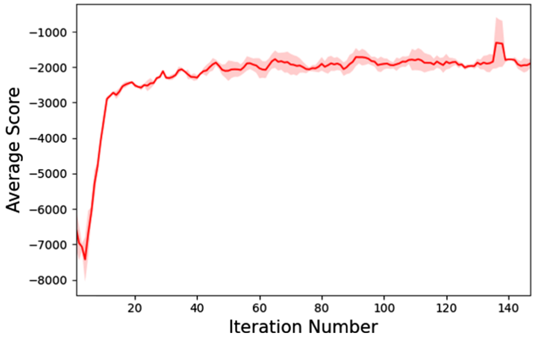

Fig. 2. Training results of connected autonomous vehicle latoon

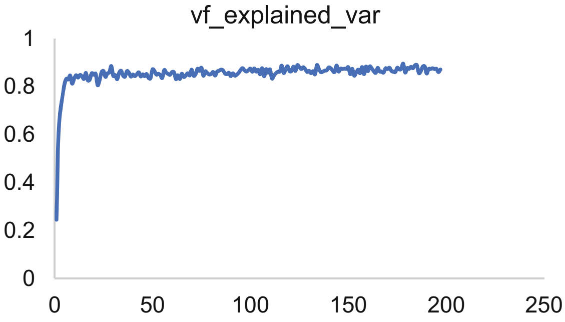

Fig. 3. vf_explained_var in the training

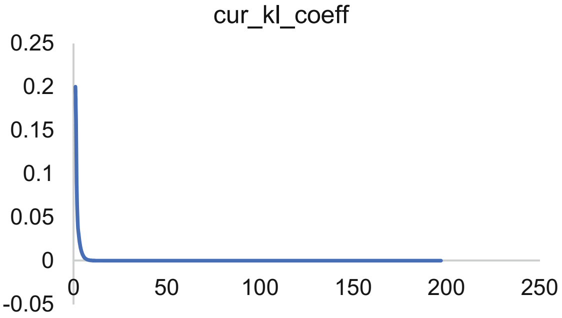

Fig. 4. cur_kl_coeff in the training

We can find that the training is rapidly converged to −2000 within 20 iterations from Fig. 2. And then the training is stable in the following iterations which means the whole designation of CAVs latoon control with MADRL is effective.

Figure 3 and Fig. 4 can reflect the training process inside the MADRL. vf_explained_var is the explained variation of those future rewards through the use of the value function. We want this to be higher if possible, and it tops out at 1; however, the results converge to 0.8 in the end which means it is effective to some extent. cur_kl_coeff is the difference between the old strategy and the new strategy at each time step. We want this to smoothly decrease as you train to indicate convergence. And it decreases to 0 in the end.

5 Conclusion

This paper presented a connected autonomous vehicle latoon control approach with multi-agent deep reinforcement learning (MADRL). In the designation of MADRL, multi-objectives are considered to achieve excellent comprehensive performance of latoon. The training results in stochastic mixed traffic flow based on SUMO platform represent that the proposed latoon control method is feasible, effective and advanced.

References

1. Liu, D., Wang, Y., Shen, Y.: Electric vehicle charging and discharging coordination on distribution network using multi-objective particle swarm optimization and fuzzy decision making. Energies **9**(3), 186 (2016)
2. Delgarm, N., Sajadi, B., Kowsary, F., et al.: Multi-objective optimization of the building energy performance: a simulation-based approach by means of particle swarm optimization (PSO). Appl. Energy **170**, 293–303 (2016)
3. Zhang, Y., Guo, L., Gao, B., Qu, T., Chen, H.: Deterministic promotion reinforcement learning applied to longitudinal velocity control for automated vehicles. IEEE Trans. Veh. Technol. **69**(1), 338–348 (2020). https://doi.org/10.1109/TVT.2019.2955959
4. Xu, G., et al.: Hierarchical speed control for autonomous electric vehicle through deep reinforcement learning and robust control. IET Control Theory Appl. 1–13 (2021). https://doi.org/10.1049/cth2.12211
5. Jardine, P.T.: A reinforcement learning approach to predictive control design: autonomous vehicle applications. Queen's University (Canada) (2018)
6. Hang, P., Lv, C., Huang, C., Cai, J., Hu, Z., Xing, Y.: An integrated framework of decision making and motion planning for autonomous vehicles considering social behaviors. Electr. Eng. Syst. Sci. 1–11 (2020)
7. Xu, J., Shu, H., Shao, Y.: Modeling of driver behavior on trajectory-speed decision making in minor traffic roadways with complex features. IEEE Trans. Intell. Transp. Syst. **20**(1), 41–53 (2019). https://doi.org/10.1109/TITS.2018.2800086
8. Liu, T., Wang, B., Cao, D., Tang, X., Yang, Y.: Integrated longitudinal speed decision-making and energy efficiency control for connected electrified vehicles. Electr. Eng. Syst. Sci. 1–11 (2020)
9. He, X., Fei, C., Liu, Y., Yang, K., Ji, X.: Multi-objective longitudinal decision-making for autonomous electric vehicle: a entropy-constrained reinforcement learning approach. In: 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC), Rhodes, Greece, pp. 1–6 (2020). https://doi.org/10.1109/ITSC45102.2020.9294736
10. Kreidieh, A.R., Wu, C., Bayen, A.M.: Dissipating stop-and-go waves in closed and open networks via deep reinforcement learning. In: 2018 21st International Conference on Intelligent Transportation Systems (ITSC), pp. 1475–1480. IEEE, November 2018
11. Vinitsky, E., et al.: Benchmarks for reinforcement learning in mixed-autonomy traffic. In: Conference on Robot Learning, pp. 399–409. PMLR, October 2018
12. Achiam, J., Held, D., Tamar, A., Abbeel, P.: Constrained policy optimization. In: International Conference on Machine Learning, pp. 22–31. PMLR, July 2017
13. Bhatnagar, S., Sutton, R.S., Ghavamzadeh, M., Lee, M.: Natural actor-critic algorithms. Automatica **45**(11), 2471–2482 (2009)
14. Cao, X.R.: A basic formula for online policy gradient algorithms. IEEE Trans. Autom. Control **50**(5), 696–699 (2005)
15. Schulman, J., Wolski, F., Dhariwal, P., et al.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017)
16. Wu, C., Kreidieh, A.R., Parvate, K., Vinitsky, E., Bayen, A.M.: Flow: a modular learning framework for mixed autonomy traffic. IEEE Trans. Robot. (2021)
17. Krajzewicz, D., Erdmann, J., Behrisch, M., Bieker, L.: Recent development and applications of sumo-simulation of urban mobility. Int. J. Adv. Syst. Meas. **5**(3&4) (2012)
18. Duan, Y., Chen, X., Houthooft, R., Schulman, J., Abbeel, P.: Benchmarking deep reinforcement learning for continuous control. CoRR, vol. abs/1604.06778 (2016). http://arxiv.org/abs/1604.06778

19. Liang, E., et al.: Ray RLlib: a composable and scalable reinforcement learning library. arXiv preprint arXiv:1712.09381 (2017)
20. Brockman, G., et al.: OpenAI Gym. arXiv preprint arXiv:1606.01540 (2016)
21. Treiber, M., Kesting, A.: Trajectory and floating-car data. In: Treiber, M., Kesting, A. (eds.) Traffic Flow Dynamics, pp. 7–12. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-32460-4_2.
22. Xu, G., et al.: Hierarchical speed control for autonomous electric vehicle through deep reinforcement learning and robust control. IET Control Theory Appl. (2021)