Reinforcement Learning Based Velocity Control for Autonomous Driving with Multi-Objectives: Safety, Efficiency, and Comfort

Rune Yu
Graduate Student
School of Transportation Engineering
Tongji University
4800 Cao'an Road, Jiading District
Shanghai, 201804, China
Phone: +86-15021901395
Email: yurune1111@outlook.com
(Corresponding author)

Hangfei Lin, PhD
Professor
School of Transportation Engineering
Tongji University
4800 Cao'an Road, Jiading District
Shanghai, 201804, China
Email: linhangfei@126.com

Meixin Zhu
Graduate Student
School of Transportation Engineering
Tongji University
4800 Cao'an Road, Jiading District
Shanghai, 201804, China
Email: 1151208@tongji.edu.cn

November 2017
Presentation at the TRB Annual Meeting

1 INTRODUCTION

A driver model is critical for velocity control systems. In general, driver models related to car following have been established with two approaches: rule-based and supervised learning (*1*). Rule-based approach mainly refers to traditional car-following models, such as the Gaxis-Herman-Rothery model (*2*) and the intelligent driver model (*3*). Supervised learning approach relies on data typically provided through human demonstration in order to approximate the relationship between car-following state and acceleration.

These two approaches all intend to emulate human drivers' car-following behavior. However, imitating human driving behaviors may not be the best solution in autonomous driving. Firstly, human drivers may not drive in an optimal way. Secondly, users may not want their autonomous vehicles driving in a way like them (*4*). Thirdly, driving should be optimized with respect to safety, efficiency, comfort, and so on, rather than imitating human drivers.

To resolve the problem, we propose a deep reinforcement learning (RL) based on car-following model for autonomous velocity control. This model does not try to emulate human drivers, rather, it directly optimizes driving safety, efficiency, and comfort, by learning from trial and interaction with a simulation environment.

Specifically, the deep deterministic policy gradient (DDPG) algorithm (*5*) that performs well in continuous control field was utilized to learn an actor network together with a critic network. The actor is responsible for policy generation: outputting following vehicle accelerations based on speed, relative speed and spacing. The critic is responsible for policy improvement: update the actor's policy parameters in the direction of performance improvement.

To evaluate the proposed model, real-world driving data collected in the NGSIM project (*6*) were used to train the model. And car-following behavior simulated by the DDPG model were compared with that observed in the empirical NGSIM data, to demonstrate the model's ability to follow a leading vehicle safely, efficiently, and comfortably.

2 METHODOLOGY

2.1 Safety

Safety should be the most important element of autonomous car following. Time to collision (TTC) was used to emphasize safety. A 7-second safety limit corresponding to the 10 percentiles of TTC distribution in NGSIM data was chosen. Then the TTC feature was constructed as:

$$F_{TTC} = \log(TTC/7) \qquad (1)$$

2.2 Efficiency

Time headway is defined as the elapsed time between the arrival of the lead vehicle (LV) and the following vehicle (FV) at a designated point. Keeping a short headway within the safety bounds can improve traffic flow efficiency because short headways correspond to large roadway capacities (*7*).

A headway feature was constructed as the probability density value of the estimated headway lognormal distribution:

$$F_{headway} = f_{\text{lognormal}}(headway|\mu{=}0.4226, \sigma = 0.4365)$$

According to this headway feature, headways around 1.3 seconds correspond to large headway feature values (about 0.65); while headways being too long or too short correspond to low feature values. In this way, efficient headways are encouraged while unsafe or too long headways are discouraged.

2.3 Comfort

Jerk, defined as the change rate of acceleration, was used to measure driving comfort because it has a strong influence on comfort of the passengers (*8*). A jerk feature was constructed as:

$$F_{jerk} = \frac{jerk^2}{3600}$$

The squared jerk was divided by a base value (3600) to scale the feature into the range of [0 1]. The base value was determined by the following intuition:

1) The sample interval of the data is 0.1s;
2) The acceleration is bounded between -3 to 3 m/s^2;
3) Therefore the largest jerk value is $\frac{3-(-3)}{0.1} = 60\ m/s^3$, if squared we get 3600.

2.4 Reward function

In RL, the reward function, *r*(*s*, *a*), is used as a training signal to encourage or discourage behaviors in the context of a desired task. The reward provides a scalar value reflecting the desirability of a particular state transition that is observed by performing action *a* starting in the initial state *s* and resulting in a successor state s'.

For the task of autonomous car following, a reward function was established based on a linear combination of the feature constructed:

$$r = -w_1 F_{TTC} + w_2 F_{headway} - w_3 F_{jerk} \qquad (2)$$

where w_1, w_2, w_3 are coefficients of the features, all set as 1 in the current study.

2.5 Network Architecture

Two separate neural networks were used to represent the actor and critic respectively. At time step *t*, the actor network takes a state $s_t = (v_n(t), \Delta v_{n-1,n}(t), \Delta S_{n-1,n}(t))$ as input and outputs a continuous action: the FV acceleration $a_n(t)$; the critic network takes as input a state s_t and action a_t and outputs a scalar *Q*-value $Q(s_t, a_t)$.

As shown in Figure 1, both the actor and critic networks have three layers: an input layer taking the input signals to the whole neural network, an output layer generating the output signal, and a hidden layer containing 30 neurons between the former two layers. We had tested even deeper neural networks, but experiment results showed that considering neural networks deeper than one hidden layers was unnecessary for our problem, which has only three or four input variables.

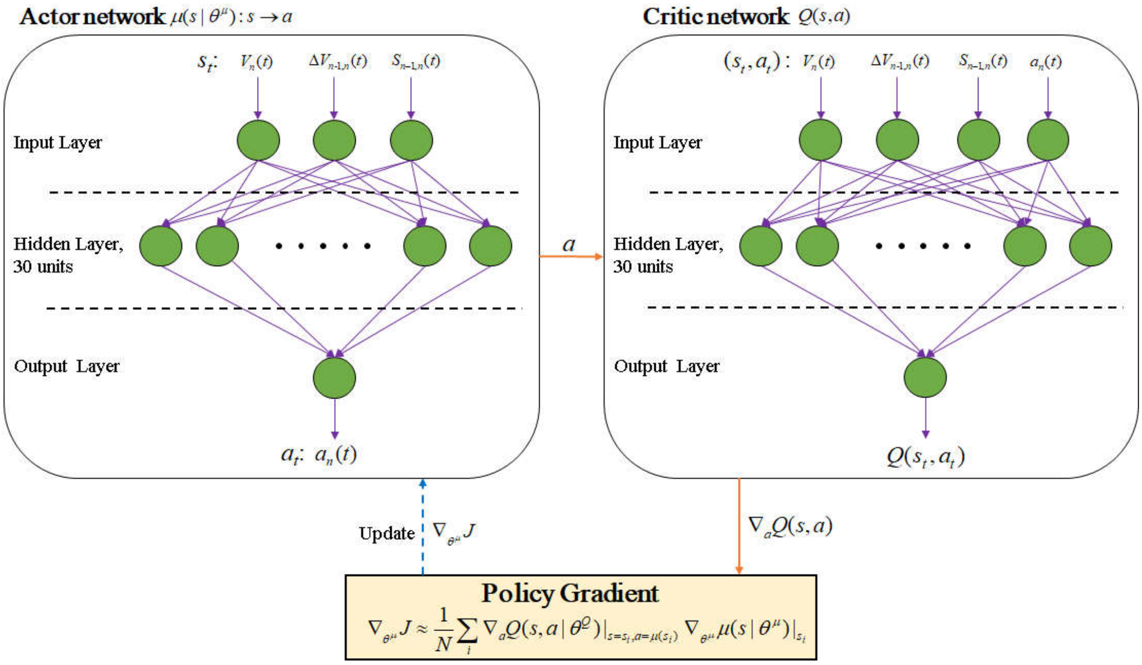
Actor network $\mu(s \mid \theta^{\mu}): s \rightarrow a$
s_t: $V_n(t)$ $\Delta V_{n-1,n}(t)$ $S_{n-1,n}(t)$
Input Layer
Hidden Layer, 30 units
Output Layer
a_t: $a_n(t)$
a
Critic network $Q(s,a)$
(s_t, a_t): $V_n(t)$ $\Delta V_{n-1,n}(t)$ $S_{n-1,n}(t)$ $a_n(t)$
Input Layer
Hidden Layer, 30 units
Output Layer
$Q(s_t, a_t)$
Update $\nabla_{\theta^{\mu}} J$
$\nabla_a Q(s,a)$
Policy Gradient
$\nabla_{\theta^{\mu}} J \approx \frac{1}{N}\sum_i \nabla_a Q(s,a \mid \theta^{Q})|_{s=s_i, a=\mu(s_i)} \nabla_{\theta^{\mu}} \mu(s \mid \theta^{\mu})|_{s_i}$

FIGURE 1 Architecture of the actor and critic networks.

3 RESULTS

3.1 Safe Driving

Driving safety is evaluated based on minimum TTC during a car-following event. Figure 2 shows the cumulative distributions of minimum TTC for NGSIM empirical data and DDPG simulation. Nearly 35% of NGSIM minimum TTCs were lower than 5s, while only about 8% of DDPG minimum TTCs were lower than 5s. It means car-following behavior generated by DDPG model is much safer than drivers' behavior recorded in NGSIM data.

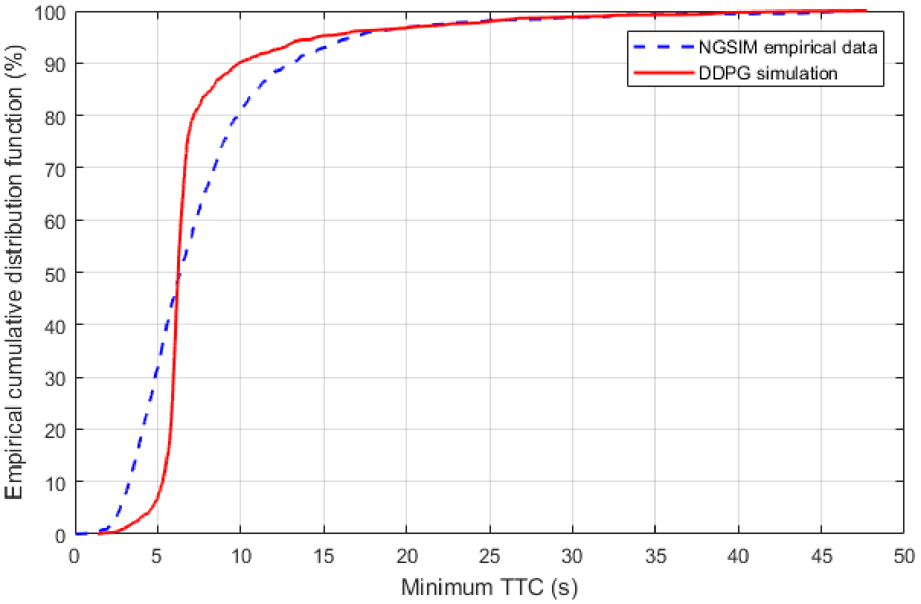

FIGURE 2 Cumulative distribution of minimum TTC during car following.

3.2 Efficient Driving

Time headway during car-following process was used to evaluate driving efficiency. Time headway was calculated at every time step of a car-following event, and the distribution of these time headways is shown in Figure 3.

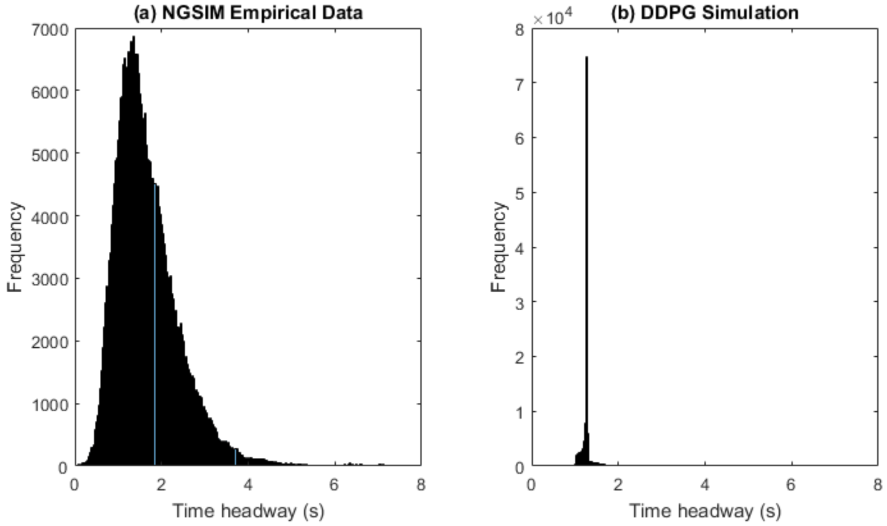

FIGURE 3 Histograms of time headway during car following for (a) NGSIM empirical data and (b) DDPG simulation.

As can been seen, the DDPG model produced car-following trajectories that always maintained a time headway in the range of 1s to 2s. While the NGSIM data had a much wider range of time headway distribution (0s to 6s). This included some dangerous headways that were less than 1s, and also some inefficient headways that were larger than 3s. Therefore, drawing to the conclusion that the DDPG model has the ability to follow the leading vehicle with an efficient and safe time headway.

3.3 Comfortable Driving

Driving comfort was evaluated based on jerk values during car following. Similar to time headway, it was calculated for every time step of a car-following event. Figure 4 presents the histograms of jerk values during car following.

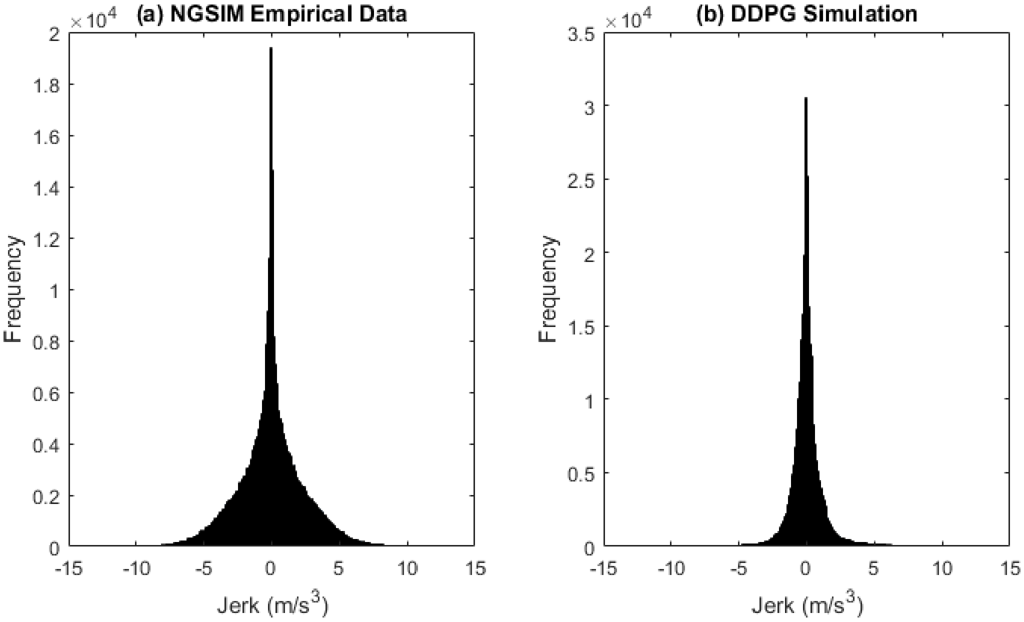

FIGURE 4 Histograms of jerk during car following for (a) NGSIM empirical data and (b) DDPG simulation.

It is obvious that the DDPG model produced trajectories with lower values of jerk. Firstly, the DDPG trajectories had a narrow jerk distribution range (–5 to 5 m/s^3) than NGSIM data (–5 to 5 m/s^3). Second, jerk values were centered more closely to zero in DDPG simulation trajectories than in NGSIM data. As smaller absolute values of jerk correspond to more comfortable driving, it can be concluded that the DDPG model can control vehicle velocity in a more comfortable way than human drivers in the NGSIM data.

To summarize, the DDPG model demonstrated the capability of safe, efficient, and comfortable driving in that it 1) had small percentages of dangerous minimum TTC values that is less than 5 seconds; 2) could maintain efficient and safe headways within the range of 1s to 2s; and 3) followed the leading vehicle comfortably with smooth acceleration.

4 CONCLUSION

The model used for velocity control during car-following was proposed based on deep RL. The model uses deep deterministic policy gradient (DDPG) algorithm to learn from trials and interaction, with a reward function signaling the RL agent performed in terms of driving safety, efficiency, and comfort. Results show that the proposed DDPG car-following model demonstrated a better capability of safe, efficient, and comfortable driving compared to human drivers in the real world. The results indicate that reinforcement learning methods could contribute to the development autonomous driving systems.

REFERENCES

1. Kuefler, A., Morton, J., Wheeler, T., and Kochenderfer, M. Imitating Driver Behavior with Generative Adversarial Networks. arXiv preprint arXiv:1701.06699, 2017.
2. Gazis, D.C., R. Herman, and R. W. Rothery. Nonlinear Follow-the Leader Models of Traffic Flow. *Operations Research,* Vol. 9, 1961, pp. 545–567.
3. Treiber, M., A. Hennecke, and D. Helbing. Congested Traffic States in Empirical Observations and Microscopic Simulations. *Physical Review E,* Vol. 62, 2000, pp. 1805–1824.
4. Basu, C., Yang, Q., Hungerman, D., Singhal, M., and Dragan, A. D. Do You Want Your Autonomous Car to Drive Like You? ACM/IEEE International Conference, 2017, pp.417–425.
5. Lillicrap, T. P., J. J. Hunt, A. Pritzel, N. Heess, T. Erez, and, Y. Tassa. Continuous Control with Deep Reinforcement Learning. *Computer Science*, Vol. 8, No. 6, 2015, pp. 187.
6. NGSIM, 2005. Next Generation SIMulation. US Department of Transportation, Federal Highway Administration http://ops.fhwa.dot.gov/trafficanalysistools/ngsim.htm (last accessed 12. 11.16.).
7. Zhu, M., X. S. Wang, A, P, Tarko. Calibrating Car-Following Models on Urban Expressways for Chinese Drivers Using Naturalistic Driving Data. Transportation Research Board 96th Annual Meeting, Washington D.C., USA, 2017. pp. 1–12.
8. Jacobson, I. D., L. G., Richards, and A. R. Kuhlthau. Models of Human Comfort in Vehicle Environments. *Human Factors in Transport Research*, Vol. 2, 1980, pp. 24–32.