Letter

Highway Lane Change Decision-Making via Attention-Based Deep Reinforcement Learning

Junjie Wang, Qichao Zhang, and Dongbin Zhao, *Fellow, IEEE*

Dear editor,

Deep reinforcement learning (DRL), combining the perception capability of deep learning (DL) and the decision-making capability of reinforcement learning (RL) [1], has been widely investigated for autonomous driving decision-making tasks. In this letter, we would like to discuss the impact of different types of state input on the performance of DRL-based lane change decision-making.

Note that the state representation is critical for the performance of DRL, especially for the autonomous driving task with multi-sensor data. Many previous works [2], [3] have targeted RL models with vector-based state representations, which lack the generalization ability for different road structures. On the one hand, road and lane line information is an important constraint on vehicle behaviors. Further research is needed on how these constraints can be better represented in DRL algorithms. On the other hand, for the case of many surrounding vehicles, it is necessary to find the interacting vehicles that have a more significant impact on the autonomous vehicle's decision to make a safe and effective decision behavior. Therefore, for the highway lane-changing task, we propose an appropriate state representation with dual inputs combining the local bird's-eye view (BEV) image with vector input and further implementing a combination of attention mechanisms and the DRL algorithm to enhance the performance of lane change decisions. Among the attention mechanisms, self-attention [4] is widely used. This letter employs different self-attention models for the BEV image and vector inputs to consider the key interacting vehicles with greater weights in the decision process. Note that the key interacting vehicles that have a considerable influence on self-driving cars' decision-making are identified with the self-attention mechanism. Fig. 1 gives some example BEV images and visualizes the results of feeding them into the trained attention module.

Related work: The application of deep reinforcement learning methods to lane-changing scenarios has been widely studied. Most existing works have used vectors [2] or grids [3] as forms of state representations, covering information such as surrounding vehicle positions and speeds that are critical for lane-changing decisions but do not explicitly consider spatial location relationships and interactions between vehicles. The attention mechanism can discover inter-dependencies among a variable number of inputs and is applicable to autonomous driving decision-making problems.

The self-attention mechanism [4]–[6] computes the response at a position in the sequence in a self-supervised manner. In [7], self-attention is extended to the more general class of non-local filtering operations that are applicable to image inputs. There are also works combining self-attention mechanisms with DRL methods for autonomous driving. In [8], an ego attention mechanism for vector inputs is developed to capture ego-to-vehicle dependencies. In [9], a multi-head attention model is introduced for trajectory prediction in autonomous driving. Unfortunately, there is no work to analyze the different types of state representations with the attention mechanism.

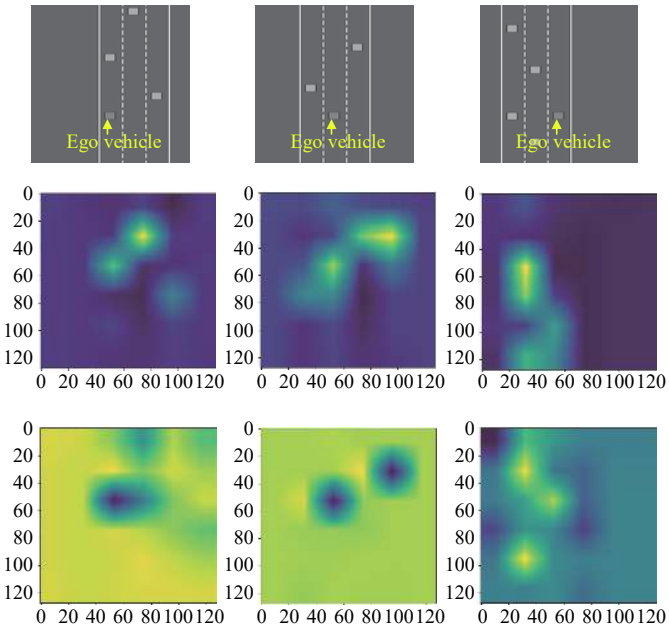


Fig. 1. Examples of image-based state representation and the visualization of the trained attention block with images as inputs. The top row shows the original image-based state; The middle row presents the output of softmax in the block, corresponding to the intermediate result of attention; The bottom row gives the output of the image state through the whole non-local block, representing the result of the entire attention module. We reshape the outputs to the same size as the inputs.

RL formulation for highway lane change decision-making: The process of lane change decision-making can be formulated as a Markov decision process (MDP). An MDP is a 5-tuple of the form $\langle \mathcal{S}, \mathcal{A}, P, R, \gamma \rangle$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ the action space, P the state transition model $P(s_{t+1} = s' | s_t = s, a_t = a)$ for each action, R the reward function $R(s_t = s, a_t = a) = \mathbb{E}(r_t | s_t = s, a_t = a)$, and $\gamma \in [0, 1]$ the discount factor. Also, s_t, a_t, and r_t are the state, action and reward at time t respectively. The transition model P and reward R are affected by the specific behavior a. The goal of RL is to learn an optimal policy $\pi^*(a|s)$ that maximizes the expected γ-discounted cumulative reward (also known as return) $J_\pi = \text{argmax}_\pi \mathbb{E}_\pi \left(\sum_{t=0}^{\infty} \gamma^t r_t \right)$. The state, action space, and reward function are defined as follows.

- Vector-based state input: For the highway lane change problem, the agent necessitates information about the ego and surrounding vehicles to make a decision. The related vector includes the location, heading, and velocity of vehicles. We use a 6-dimensional vector to characterize the information about vehicles:

$$s_t = (s_i)_{i \in [0,N]}, \quad s_i = \left[x_i \;\; y_i \;\; v_i^x \;\; v_i^y \;\; \cos\psi_i \;\; \sin\psi_i \right] \tag{1}$$

where N is the number of visible vehicles. The elements in s_i represent the vehicle's lateral location in the road, the longitudinal location, the lateral velocity of the vehicle, the longitudinal velocity, and the cos and sin values of the heading error between the vehicle and the lane orientation, respectively. The remaining states are filled with zeros when fewer than N vehicles are visible.

- Image-based state input: In addition to the direct vector-based state representation, an alternative is to formulate the state in image form. Although vector-based states inform the position of each vehicle in the road, it is not very intuitive to capture the relationship

Citation: J. J. Wang, Q. C. Zhang, and D. B. Zhao, "Highway lane change decision-making via attention-based deep reinforcement learning," *IEEE/CAA J. Autom. Sinica*, vol. 9, no. 3, pp. 567–569, Mar. 2022.

The authors are with the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, and also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: wangjunjie2017@ia.ac.cn; zhangqichao2014@ia.ac.cn; dongbin.zhao@ia.ac.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JAS.2021.1004395

between the road structure and vehicles. In contrast, the BEV image-based state representation can directly take the vehicles together with the road structure as the input and obtain all vehicles' location and heading information from a local BEV space. In addition, the combination of image-based states and convolutional neural networks (CNNs) can also help extract the position relationship between different vehicles. However, the image-based state does not easily represent the velocity information of the vehicles explicitly. Examples of BEV image-based state inputs are shown in the top row of Fig. 1. We represent the ego vehicle, the surrounding vehicles, and the road structure as a single-channel gray-scale image to characterize the position of the ego vehicle (the dark square in the image) and the surrounding vehicles (the light squares in the image) in the road. The position of the ego vehicle in the image is fixed.

● Action space: The discrete action space is set as the output of DRL agents for the lane change task, including both lateral and longitudinal commands of the ego vehicle, i.e., {no operation, change lanes to the left, change lanes to the right, acceleration, and deceleration}. At a certain time step, only one lateral or longitudinal action command will be given to the ego vehicle. Therefore, the agent needs to execute a series of actions coherently to produce a specific behavior. For example, when required to accelerate from the left to overtake the vehicle ahead, the agent needs to output the left lane change command in the current time step and then provide the acceleration actions in several subsequent time steps.

● Reward function: The design of the reward function requires a combination of safety and efficiency. To improve the safety of the policies learned by agents, we apply a penalty when a collision occurs, i.e., $r_c = 0$ (if no collision) or -1.0 (if collision happens) is given to the agent at each time step. In order to improve the efficiency, it is expected that ego speed should be fast, therefore, a reward $r_v = 0.2 \cdot (v_t - v_{\min}) / (v_{\max} - v_{\min})$ is offered at each time step, where v_t is the velocity of the ego vehicle at time step t, $v_{\max} = 30$ m/s, $v_{\min} = 20$ m/s. Thus, the total reward for each time step is $r = r_v + r_c$, which we clip it to $[0, 1]$.

Model-free DRL: To evaluate a particular policy π, the state value function $V^{\pi}(s)$ and state-action value function $Q^{\pi}(s,a)$ are formally defined as

$$V^{\pi}(s) = \mathbb{E}_{\pi}\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k} \middle| s_t = s\right] \tag{2}$$

$$Q^{\pi}(s,a) = \mathbb{E}_{\pi}\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k} \middle| s_t = s, a_t = a\right]. \tag{3}$$

For the continuous state space, we typically rely on function approximation techniques for the generalization over the input domain. In this letter, we utilize dueling double DQN (D3QN). DQN (deep Q-network) [10] incorporates Q-learning [11] with a deep neural network to fit the action-value function, denoted by $Q(s,a;\phi)$, where ϕ is the weights in the Q network. Dueling DQN [12] further divides the Q network into a value function part $V(s;\theta,\alpha)$ and an advantage function part $A(s,a;\theta,\beta)$, where θ is the part of the Q network common to V and A, and α and β are their respective parameters. Then, the final state-action value function can be re-expressed as $Q(s,a;\phi) = Q(s,a;\theta,\alpha,\beta) = V(s;\theta,\alpha) + A(s,a;\theta,\beta)$. The loss function of the neural network training can be defined as

$$L_i(\phi_i) = \mathbb{E}_{s\sim\pi}\left[\frac{1}{2}(y_i - Q(s,a;\phi_i))^2\right] \tag{4}$$

where $y_i = \mathbb{E}_{s'\sim E}\left[r + \gamma Q^T(s', \text{argmax}_{a'} Q(s',a';\phi_i); \phi^T{}_i) | s,a\right]$ is the target value of the i-th iteration in double DQN [13], and $Q^T\left(s,a;\phi^T\right)$, called the target network, is a copy of $Q(s,a;\phi)$. E is the environment. Note that the target network is updated less frequently than $Q(s,a;\phi)$ to improve the convergence of training. The above loss function for the gradient of the weights yields

$$\nabla_{\phi_i} L_i(\phi_i) = \mathbb{E}_{s'\sim E}\Big[\big(r + \gamma Q^T\left(s', \underset{a'}{\text{argmax}} Q(s',a';\phi_i); \phi_i^T\right) - Q(s,a;\phi_i)\big)\nabla_{\phi_i} Q(s,a;\phi_i)\Big]. \tag{5}$$

Stochastic gradient descent is usually employed to optimize the loss function rather than directly calculating the expectation value in (5).

Attention-based DRL framework: For different forms of state representations, we adopt different self-attention mechanisms to extract state features. The overall framework of the proposed attention-based D3QN framework combining the BEV image and vector states is depicted in Fig. 2.

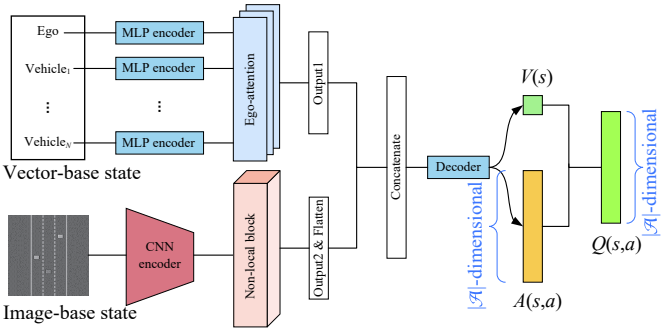


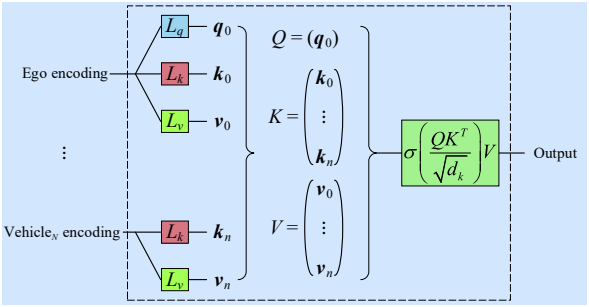
Fig. 2. The proposed attention-based DRL framework for highway lane change decision-making.

First, the dual inputs, including vector-based states and image-based states, are fed into their respective encoders: a multilayer perceptron (MLP) for vectors and a CNN for images. Next, the two encoder outputs are inputted into the corresponding attention modules, and the attention results of the two parts are concatenated together as a full feature. Subsequently, this feature goes through the dueling-structured MLP network to output the final Q value. Note that the overall architecture is jointly optimized by the D3QN algorithm.

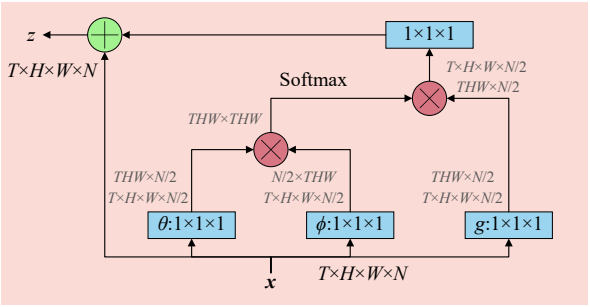
For the vector-based state input, the used attention model is the ego-attention [8]. This attention mechanism is a variant of the traditional social attention [14] mechanisms, in which only the ego state has query encoding. The architecture of an ego-attention head is represented in Fig. 3(a). This architecture can satisfy the requirements of variable sizes with permutation invariance, even when using a set of characteristic representations. It also naturally accounts for the interaction between the ego vehicle and surrounding vehicles.

For the image-based state input, the used attention model is the non-local block [7]. In some computer vision tasks, CNNs increase the receptive field of perception by stacking multiple convolutional modules. Convolution operators are all local operations in feature space. The way to capture a larger range of information in an image by repeated stacking has some shortcomings: inefficiency in capturing a large range of information, need for careful design of modules and gradients, and local operations are harder to implement when information needs to be passed between relatively distant locations. Compared with the traditional convolutional operation, the non-local block directly captures large range dependencies by computing the interaction between any two positions without limiting to adjacent points, which is equivalent to constructing a convolutional kernel as large as the size of the feature map and thus can sustain more information. In addition, the non-local block can be used as a component that can be easily combined with other network structures. The model structure is shown in the left panel in Fig. 3(b).

Experiments and analysis: For two different forms of state inputs, vector and image, we conduct different experiments to compare the effectiveness of the algorithms with the help of the open-source HighwayEnv environment (https://github.com/eleurent/highway-env). These experiments include vector input, vector input with ego-attention, image input, image input with non-local block, and dual



(a) Ego-attention [8]



(b) Non-local block [7]

Fig. 3. The utilized attention mechanisms for vector and image inputs.

Table 1. Test Results. The Data is Averaged Over 10 Runs

Input types	Average lane change times ↓	Average return ↑	Average step ↑	Lane change safety ↑
Vector	22.4	38.07	40.3	0.5
Vector with attention	12.8	45.35	48.8	**0.9**
Image	16.6	37.99	41.2	0.5
Image with attention	11.6	40.47	43.0	0.7
Vector with attention + Image with attention	**8.1**	**46.63**	**49.0**	**0.9**

input (vector input with ego-attention combined with image input with non-local block). Ten test episodes are conducted for each of the different experimental settings, with each test going through 50 time steps. The results are presented in Table 1. The average lane change times in the table indicate the average number of lane change actions taken by the ego vehicle in these ten test runs. The average return is the average cumulative reward of the feedback from the environment in 10 runs. The average step is the average of the time steps experienced in 10 runs (if the ego vehicle crashes during one test episode, the test will be terminated, then the test steps will be less than 50 time steps), and the lane change success rate indicates the rate of successful lane changes, i.e., without collision in 10 runs.

As can be seen from the results in Table 1, dual inputs combined with different attention mechanisms achieve the best results in all the evaluation metrics. For BEV images, the relationships between scene elements such as the lane with vehicles and the ego with vehicles can be captured. For vector inputs, more accurate spatially distant interactions between the ego and surrounding vehicles can be captured. The results prove that the dual input combining the vector with images is a better state representation for the decision-making task. In addition, comparing the odd rows of the table with the even rows, we can see that using the attention mechanism can further improve the performances of the original state input. This shows the effectiveness of the proposed method, and the attention mechanisms have a positive effect on different forms of state input.

In addition, we visualize the trained non-local block, and the results are given in Fig. 1. From the results, we can see that the regions with the presence of surrounding vehicles have higher weights, showing that the agent has greater attention to these regions. The interesting result is that, after the attention module, the agent only pays attention to vehicles close to and in front of itself and ignores the vehicles behind it (see the image in the bottom right-hand corner of Fig. 1). Coincidentally, the ego speed is usually higher than the surrounding vehicles, and the vehicles in the rear have very little effect on the ego vehicle.

Acknowledgments: This work was supported in part by the National Natural Science Foundation of China (NSFC) (62173325), and the Beijing Municipal Natural Science Foundation (L191002).

References

[1] D. Zhao, K. Shao, Y. Zhu, D. Li, Y. Chen, H. Wang, D. Liu, T. Zhou, and C. Wang, "Review of deep reinforcement learning and discussions on the development of computer Go," *Control Theory & Applications*, vol. 33, no. 6, pp. 701–717, 2016.

[2] D. Li, D. Zhao, and Q. Zhang, "Reinforcement learning based lane change decision-making with imaginary sampling," in *Proc. IEEE Symposium Series on Computational Intelligence (SSCI)*, 2019, pp. 16–21.

[3] J. Wang, Q. Zhang, D. Zhao, and Y. Chen, "Lane change decision-making through deep reinforcement learning with rule-based constraints," in *Proc. Int. Joint Conf. on Neural Networks (IJCNN)*, IEEE, 2019, pp. 1–6.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5998–6008.

[5] X. Li, Y. Liu, K. Wang, and F. Wang, "A recurrent attention and interaction model for pedestrian trajectory prediction," *IEEE/CAA Journal of Automatica Sinica*, vol. 7, no. 5, pp. 1361–1370, 2020.

[6] X. Zhao, Y. Chen, J. Guo, and D. Zhao, "A spatial-temporal attention model for human trajectory prediction," *IEEE/CAA Journal of Automatica Sinica*, vol. 7, no. 4, pp. 965–974, 2020.

[7] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7794–7803.

[8] E. Leurent and J. Mercat, "Social attention for autonomous decision-making in dense traffic," *arXiv preprint arXiv*: 1911.12250, 2019.

[9] K. Messaoud, N. Deo, M. M. Trivedi, and F. Nashashibi, "Trajectory prediction for autonomous driving based on multi-head attention with joint agent-map representation," in *Proc. IEEE Intelligent Vehicles Symposium (IV)*, 2021, pp. 165–170.

[10] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[11] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, USA: MIT Press, 2018.

[12] Z. Wang, T. Schaul, M. Hessel, H. Hasselt, M. Lanctot, and N. Freitas, "Dueling network architectures for deep reinforcement learning," in *Proc. Int. Conf. Machine Learning (ICML)*, PMLR, 2016, pp. 1995–2003.

[13] H. Van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double Q-learning," in *Proc. AAAI Conf. Artificial Intelligence (AAAI)*, vol. 30, no. 1, 2016.

[14] A. Vemula, K. Muelling, and J. Oh, "Social attention: Modeling attention in human crowds," in *Proc. IEEE international Conf. Robotics and Automation (ICRA)*, 2018, pp. 4601–4607.