Traffic Signal Control with Deep Reinforcement learning

Tongyu Zhao[a], Peng Wang[b], Songjinag Li[c]

Institute of computer science and technology, Changchun University of Science and Technology

[a]zhaotongyu.emma@gmail.com, [b]wangpeng@cust.edu.cn, [c]lsj@cust.edu.cn

Keywords: Deep Reinforcement Learning, Traffic signal control, Partially Observable MDP, Vehicle network.

Abstract: To decrease the impact of Partially Observable MDP on deep reinforcement learning performance of intersection signal control, A deep reinforcement learning is proposed in this paper with utilizing the real-time GPS data as well as learning the control of the traffic lights in single intersection. We integrate deep reinforcement learning network (DRQN) with recurrent neural network (RNN) and apply deep network, experience pool and greedy strategy in deep reinforcement learning strategy. It solves the problem of overestimation of target Q value and insufficient long-term experience learning in the standard reinforcement learning of traffic signal control. The comparison of performance was made between the proposed method and standard Deep Q-Network (DQN) on the partial observation of traffic situations. The experimental results show that both DQN and DRQN methods can adjust their traffic signal timing control strategies according to the specific traffic conditions as well as calculating a lower average delay time of vehicle than that of fixed-time control. Besides, the simulation effect of DRQN learning method is better than that of DQN learning method in different probe vehicle proportion environment.

1. Introduction

The inefficient control of signals leads to many problems. The most serious problem is that the paralysis of the signal light control system leads to frequent traffic accidents [1]. Therefore, how to optimize the signal timing scheme and improve the operation efficiency of intersections has become an important key issue in the field of intelligent transportation. Because the complexity of traditional reinforcement learning increases exponentially and most successful reinforcement learning methods depend on the selection of artificial features, the quality of learning results mostly depends on the quality of feature selection [2] also increases the difficulty of control. The recent rapid development of Deep learning makes it possible to extract effective features directly from raw data. Some studies have proposed the application of deep reinforcement learning to solve traffic light control problems[3][4]. It is appropriate that modeling a partially observable MDP, namely POMDP, through a non-stationary. At least two factors make the traffic environment of a single intersection an object of local observation: (1)The traffic-flow pattern of intersections is unaware. (2) The vehicles at intersection are regarded as one part of entire traffic flow. The vehicles' data are captured from vehicle-mounted GPS devices(probe vehicle), it will upload real-time GPS data to remote data centers. However, it is still difficult for solving the problem in POMDP. Literature [5] proves that the state of a single intersection is partially observable Markov (POMDP). In order to further reduce the impact of POMDP characteristics on the performance of deep Q learning at intersections, this paper improves the structure of the deep neural network by introducing Deep Recurrent Q Network (DRQN). Experiments show that the performance of DRQN is better than that of DQN, and it is better than the traditional intersection control methods of timing control and Q learning control.

2. System Model

First, the traffic signal control problem is represented by a deep reinforcement learning problem with building a intersection model. The following four intersections, as shown in FIG. 1, are respectively two straight lanes, one left lane, and one right lane. Next, this section defines the state, action and reward functions in Agent learning respectively.

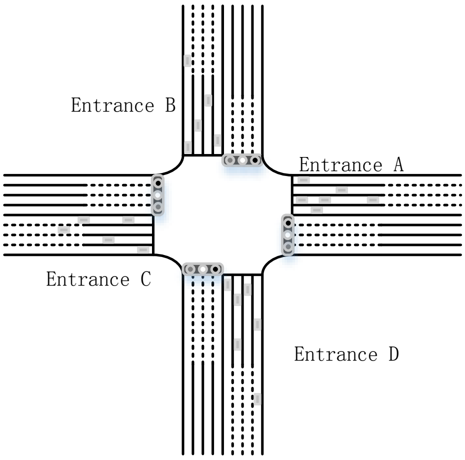
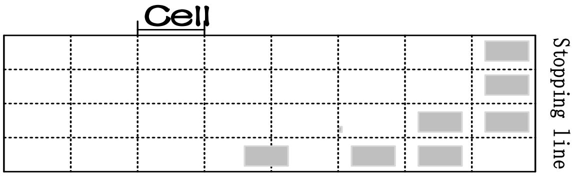


Fig. 1. The intersection model

2.1 State Space

Assuming that the vehicle at the current intersection is shown in FIG. 2 (a), the road is divided into several small sections called cells, and the length of each cell is C. The position and velocity matrix of vehicles at the current intersection are shown in Fig. 2 (b) (c), which shows the matrix representation of traffic network state (when a vehicle crosses two cells, a cell with a larger occupancy is selected). The state consists of two matrices: (1) binary matrix of vehicle position (Fig. 2b). (2) Vehicle speed matrix (Fig. 2c). Specifically, by creating a grid in which cells can be represented as a binary matrix, the lane is divided into several cells with a fixed length; An element in the corresponding matrix of each cell; Each element of matrix represents the number and speed of vehicles in the corresponding cell. The speed is expressed as the maximum allowable speed of the current vehicle.



(a) Entrance A Simulation Environment State

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 \end{bmatrix}$$

(b) Entrance A Position Matrix

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.2 & 0 \\ 0 & 0 & 0 & 0.8 & 0 & 0.4 & 0.1 & 0 \end{bmatrix}$$

(c) Entrance A Speed Matrix

Fig. 2. Environmental State Representation

2.2 Action Space

The traffic control model aims at find a optimization measurement. , and the accumulated return of the model is maximized with the training time. It must select an operation from the set of all available operations after agents observe the environmental state. In this paper, the possible action of Agent is traffic signal phase allocation (i.e. controlling the combination of traffic lights in each lane of the whole intersection). Possible actions are North-South Green (NSG), East-West Green (EWG), North-South Left-Green (NSLG), East-West Left-Green (EWLG). Formally, the set of all possible operations A is defined as A={NSG, EWG, NSLG, EWLG}.

2.3 Reward Definition

In terms of traffic signal control, several evaluation criteria are used to distinguish whether the control strategy improves traffic conditions, such as changes in vehicle queues, cumulative vehicle delays, and vehicle throughput. In this paper, we measure the total number of queued vehicles in front of the parking line at step t and define the reward as the change of cumulative vehicle delay between different behaviors. We use

$$r_t = d_t - d_{t+1} \tag{1}$$

as reward function, so that when the number of the halting vehicle decreases between time step t and t+1, the agent will receive a positive reward to encourage its decision.

2.4 Agent

The traffic signal control agent (Agent) is composed of the DRQN network is shown in FIG. 3. A combination of convolution neural network and cyclic neural network is used to represent Q function. The position matrix and velocity matrix are combined into a small two-channel image, and two convolution layers are used to capture available features. The first layer has four convolution filters with size 2 *2, and the second layer has eight convolution filters with size 2 *2. There is a 112-dimensional vector flattened at output layer. The next part of the DRQN network is a simple embedding layer, with a phase vector was encoded into 10-dimensional. Finally, 112-dimension and 10-dimension vectors are connected in series as an input of LSTM of 32 hidden units, and the continuous layer is the full-connection layer of 8 outputs and the full-connection layer of 2 outputs. The activation function uses ReLU, followed by the LSTM layer activated by ReLU. A more standard network structure has been established. The weight of deep neural networks can be updated through 2 methods. i.e. boot sequential update and boot random update[6]. Both updating methods can converge to similar performance. In this paper, the parameters of the deep neural network are adjusted by guiding random updating. In each training time step, a fixed length of experience sequence is randomly selected. Firstly, Q learning objectives are generated recursively by using the objective network based on the experience sequence. The classical back propagation algorithm is used for training. The gradient descent algorithm updates its internal parameters step by step. In general, the loss function of deep reinforcement learning is defined as:

$$\mathrm{L}(\theta) = \frac{1}{2} E_{s,a,r,s'}[(r_t + \gamma \max_{a'} Q'(s',a') - Q(s,a))^2] \tag{2}$$

Among them, $r_t + \gamma \max_{a'} Q'(s',a')$plays the role of learning goal in supervised learning. $Q'(s',a')$ is an estimate of the Q-value function of another neural network, i.e. the target network. The internal parameters of the target network can be updated from the main neural network after a certain time step. Use experience playback and target network [7] to stabilize the learning process. The update rule is defined as:

$$Q_{t+1}(S_t,a_t) = Q_t(S_t,a_t) + \alpha_t[r_t + \gamma max Q_{t+1}(S_{t+1},a_{t+1}) - Q_t(S_t,a_t)] \tag{3}$$

Among them, α represents the learning rate and $\alpha\epsilon[0,1]$The leaning rate determines the speed of updating the value function. γ is the discount factor $\gamma\epsilon[0,1]$, and the discount factor determines the importance of future rewards.

The prediction of Q target is not accurate at the beginning because the LSTM hidden state is zero at the beginning. Therefore, the second half of the predicted Q target is the only part we used as training target, and the error signals train the DRQN with time back propagation. The standard DQN is trained based on a random sampling experience and back propagation algorithm. Both networks are trained by Adam[8] algorithm. Agent will choose a random behavior to learn the optimal strategy with 1-ε probability for avoiding local extremum during the training process. During training, E decreases gradually with the number of training:

$$\varepsilon = \max(0.01, 1 - \frac{n}{N}) \tag{4}$$

Among them, n is the current number of training, N is the total number of training.

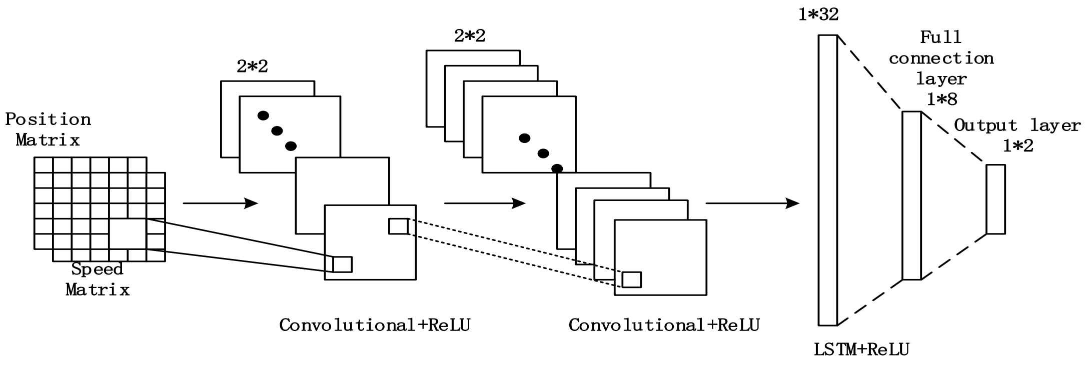


Fig. 3. The recurrent Q-network model

2.5 DRQN algorithm steps

Traffic signal algorithm control based on deep recurrent Q-learning is summarized as follows:

Algorithm: DRQN algorithm

```
Initialize DRQN network structure, and take as parameter θ
Initialize□ ε, γ, N
For epoch=1 to N do
    Initialize intersection state $s_0$ , action $a_0$
    For 1 to T do
        Choose action a according to ε-greedy policy
        Take action a, observe reward r and next state $s_{t+1}$
        If the size of memory m > M
            Delete the oldest memory in memory
        End if
        Store transition$(s_t, a_t, r_t, s_{t+1})$ in M
        Update $\theta$
        $s_t \rightarrow s_{t+1}$
    End for
End for
```

3. Simulation Results

The traffic was simulated by the simulation tool of urban mobility (SUMO) to verify the effectiveness of the proposed traffic control model. . Detailed simulation settings are as follows:

3.1 Simulation environment and parameter setting

Traffic roads and parameter settings:

Consider a four-lane intersection, each with four lanes, as shown in FIG. 1. The road length is 500 meters; the section L is 160 meters; the cell length C is 8 meters; the road speed limit is 19.444 m/s; the vehicle length is 5 meters; the minimum gap between vehicles is 2.5 meters. For traffic signal timing parameters, the green light phase interval g is 2 seconds; the yellow light interval y is 4 seconds. The minimum green light time t_{min} is 6 seconds and the maximum green light time t_{max} is 60 seconds. The traffic data of Huangshan Road and Tianzhi Road intersection in the Hefei demonstration area on October 15, 2018, are used. The representative time points are 08:00, 11:00 and 23:00. At this time, the flow corresponds to the higher-saturated flow, the near-saturated flow, and the low-saturated flow, respectively. The specific values are shown in Table I.

Table 1.Flow Value of Three Saturation Flows at Entrance

	Entrance A	***Entrance A***	***Entrance A***	***Entrance A***
higher-saturated flow	92	327	87	318
near-saturated flow	459	943	672	742
low-saturated flow	826	1798	729	1324

3.2 Simulation results and analysis

The reward value of the Agent depends on whether it can reduce the number of parked vehicles. The queue length and the average-waiting time are shortened by utilizing the strategy. Firstly, the convergence of DQN and DRQN methods is tested. Vehicles reach the intersection through the Poisson process, which is approximately binomial distribution. Assume that the probability of arrival of north-south (N-S) or north-south (S-N) direct vehicles is 1/4 (that is, one north-south (N-S) or north-south (S-N) direct vehicle will be produced every 4 seconds on average). As shown in FIG. 4 the performances of DQN and DRQN methods are superior to timing control as well as but the DRQN is significantly superior to the standard DQN method, which uses a sequence to update its internal parameters during each training, and performs exploratory guidance random update more effectively than DQN.

Real traffic environment full of randomness, in general, it is almost impossible that obtaining the state as real as the whole traffic environment, turning the environment into POMDP process, for example, only a small number of vehicles will upload their own position and speed to the data center during the process of real-time acquisition of vehicle data. The permeability [9] tend to be different when the probe vehicles detect the traffic environment in a particular scenario, even as the transport time of different changes this paper, the proportion of probe vehicles and the delay of data transmission are considered, and the permeability of detecting vehicles is changed by changing the proportion of observable vehicles. In the process of the simulation experiment, the position and velocity information of probe vehicles can decrease the value of the density vector and increase the sparsity of density vector and velocity vector correspondingly.

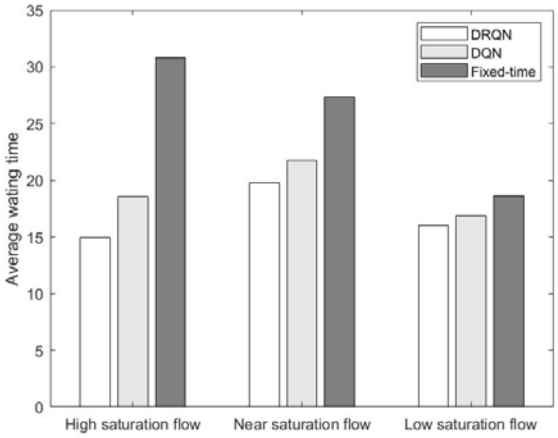


Fig. 4. The average waiting time of three traffic density network structures

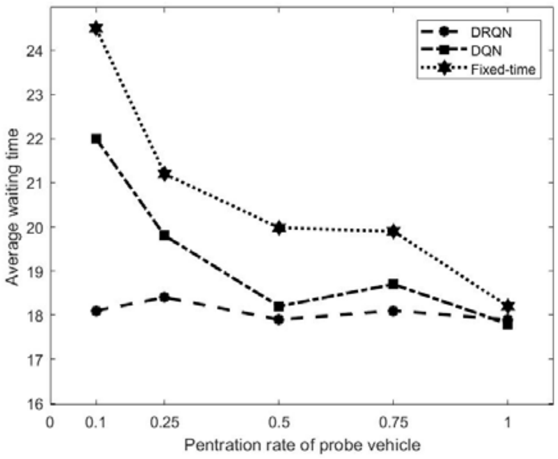


Fig. 5. Average waiting time of different permeability

FIG.5 shows the performance under different permeability. Testing and training are carried out in traffic environment with the same proportion of floating cars. Each data point expresses the average result in 20 test sets. The results show that the performance of DQN and DRQN is better than that of timing control. Adding deep loop network to DQN network can not only benefit from discrete representation of state, but also learn similar performance under different permeability. Combining the state of storage unit in LSTM with the input of current state, it can better identify the actual traffic environment state.

Can be seen from the FIG.6, under the environment of the permeability of 50% of the two kinds of Agent for training and testing, respectively in each time step input zero density and velocity vector will there is a 50% chance of ignoring the current vehicle information DRQN still keep good results show that the robustness DRQN Agent not only benefited from discrete status, but also benefit from the storage unit in the enrichment of historical information, the information under the different permeability may be using a loop similar to that of Q learning another advantage is that it can carry historical information data, even can't see the current state of environment, can also be accurately to make a decision.

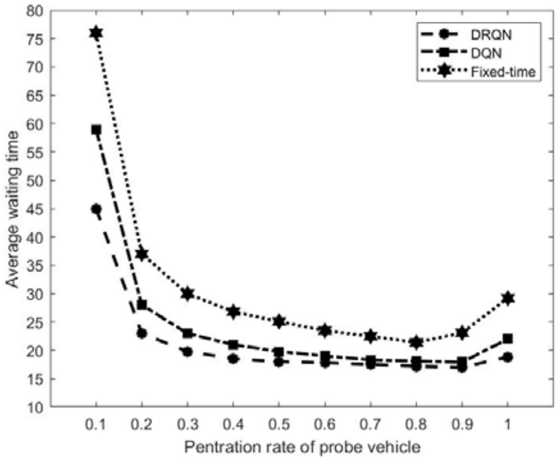


Fig. 6. Average waiting time of different permeability

4. Simulation Results

With the increasing complexity of urban traffic environment, hidden modes in traffic state are difficult to find. Deep reinforcement learning provides an effective method to mine hidden patterns from high-dimensional data and provides a solution for urban traffic control. We compare the performance of deep recurrent Q network with that of a standard Q network. Experiments show that the deep recurrent Q network learning can better approximate the strategy in the local observation environment. It is good for distinguishing the actual environmental status when integrating the historical status with current input status. In this paper, the urban road traffic signal optimization and other related issues are systematically studied, and some research results of theoretical and practical application value have been achieved. However, as the complexity of the urban traffic environment increases, the training time will be greatly increased. Many advanced deep reinforcement learning technologies will be considered in the future for decreasing the training time as well as improving the training effect, such as Dueling-DQN or Double- DQN.

Acknowledgments

This work was supported by the National Key Research and Development Program (Grant No.2017YFB1401800).

References

[1] S. S. Mousavi, M. Schukat, P. Corcoran, and E. Howley, “Traffic light control using deep policy-gradient and value-function based reinforcement learning,” arXiv preprint arXiv:1704.08883, April 2017.

[2] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Playing atari with deep reinforcement learning [C] //Proceedings of the NIPS Workshop on Deep Learning. Lake Tahoe: MIT Press, 2013.

[3] SHU Lingzhou, WU Jia, WANG Chen. Urban traffic signal control based on deep reinforement learning.[J]Journal of Computer Applications, 2019, 39(05):255-259.

[4] Wu C , Parvate K , Kheterpal N , et al. Framework for control and deep reinforcement learning in traffic[C]// IEEE International Conference on Intelligent Transportation Systems. IEEE, 2018.

[5] RITCHER S. Traffic light scheduling using policy-gradient reinforcement learning [C]//The International Conference on Automated Planning and Scheduling, ICAPS,2007.

[6] Hausknecht M, Stone P. Deep Recurrent Q-Learning for Partially Observable MDPs [J]. Computer Science, 2015.

[7] Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning [J]. Nature, 2015, 518(7540): 529.

[8] Kingma D P, Ba J. Adam: A Method for Stochastic Optimization [J] Computer Science, 2014.

[9] Krajzewicz D, Erdmann J, Behrisch M, et al. Recent Development and Applications of SUMO - Simulation of Urban Mobility [J]. International Journal on Advances in Systems & Measurements, 2012, 3&4(3 and 4):128-138.

[10] QIN Yan-yan, Wang Hao, Wang Wei, et al. Mixed Traffic Flow String Stability Analysis for Different CACC Penetration Ranges[J]. Journal of Transportation Systems Engineering and Information Technology, 2017(4).