ISSN: 2664-9640

Research on Traffic Control Method Based on Deep Reinforcement Learning Network

Zhizhong Lv, Maojiang Yao [a]

College of Southwest Petroleum University, Sichuan 610500, China

[a]ymj19960413@qq.com

Abstract

In view of the current situation of frequent traffic congestion at urban intersections, a traffic control method based on deep reinforcement learning network is proposed in order to improve the utilization rate of urban traffic intersections resources, optimize vehicle traffic conditions at intersections, and improve the real-time and flexibility of traffic control at intersections. Firstly, the real-time video is analyzed and processed by machine vision to obtain traffic flow information such as queue length, traffic flow and speed of road vehicles. At the same time, the deep reinforcement learning network is trained and the agent is selected so that the agent can learn the traffic control method. Finally, based on deep reinforcement learning network and optimal agent, the intelligent traffic control method is simulated under the road network model built by VISSIM.

Keywords

Machine vision; Optimization of the traffic; Deep reinforcement learning; Intelligent traffic control; Joint traffic simulation.

1. Introduction

With the reform and opening up, China's economy has been rapid development, per capita GDP has been doubled. With the development of social productivity and the improvement of people's income, the social demand for all kinds of cars is becoming more and more large. According to statistics, China's new car sales volume in 2019 is 25.769 million, and as of December 31, 2019, the national car ownership has reached 260 million. Among them, Beijing, Chengdu and Chongqing rank the top three, and Chengdu ranks the second with 5.195 million vehicles. Taking an average of three people as calculation, Chengdu has about one car per household. A large number of vehicles put great pressure on urban road traffic, and the developed degree of road traffic also shows the potential of continued urban development. From our daily life, we can see that the traffic light time of each phase at most traffic intersections is fixed, and the traffic light system with fixed time is obviously contradictory to the real-time and changeable traffic environment. With the rapid development of computer technology, relevant scholars have begun to try to combine computer vision technology with traffic system, using computer vision to detect relevant traffic parameters, optimize the control of traffic lights at traffic intersections, and alleviate urban traffic congestion, environmental pollution and resource waste. But its still don't have a high level of intelligence and the intersection of generality is also very poor, so in this paper, combining with the rise in recent years and the rapid development of machine vision and the depth of the reinforcement learning technique to intelligent control of each intersection traffic lights, greatly improving the efficiency of roads and road traffic, with 5 g network will each intersection traffic information and traffic data and road intersection of each agent for data sharing, the agent in the use of local traffic flow information, also can use near the intersection traffic data such as the depth of the whole study, so as to improve the traffic efficiency of a region. So that each intersection can be

ISSN: 2664-9640

organically combined, unified coordination and command, so that the vehicle can pass quickly, reduce the occurrence of traffic jams.

Machine vision can obtain traffic flow information such as intersection traffic queue length, traffic flow and speed at ultra-high speed and ultra-precision. Deep reinforcement learning can learn effective control strategies from a large number of experimental data and form an optimal intelligent traffic control method.

The purpose of this paper is to use of machine vision in real time fast acquisition road intersection traffic flow information, combined with the depth of reinforcement learning network, through the study of the training of different traffic flow information, training the agent at the time of vehicles as per the minimum delay, the shortest queue length, maximum constantly interact training goal and the traffic environment, constantly explore the optimal control method. The main innovations of this paper are as follows:

1) Intelligent analysis and collection of traffic flow information by machine vision.

2) A variety of deep reinforcement learning networks are compared for training in the game environment, and agents are selected as the best ones.

3) Through the two software PyCharm and VisSim, simulation training was conducted for several times in different environments, and the control method of deep reinforcement learning for fast vehicle passing was highly reliable.

2. Machine vision

To sum up, machine vision is to use a variety of sensors instead of human eyes for detection and analysis. Machine vision system is to collect image information through the camera, the image signal collected to the image processing system, image pixel distribution, brightness, color and other information through a series of convolution, sampling, residual processing into digital signals, according to the processed digital signal characteristics of the image recognition and judgment.

Machine vision is a highly integrated technology, which includes image acquisition, image processing, engineering technology, mechanical control, electronic control, sensor technology, analog and digital video technology, computer software and hardware technology. A typical machine vision application system mainly includes image acquisition part, light source part, image processing part, electronic control part, intelligent decision part and mechanical body part.

2.1. Image preprocessing

The surveillance video collected by the surveillance camera is a continuous color image, that is, the image has three RGB color channel information, RGB three-channel image information will add a lot of computation and calculation steps to the image processing. In order to increase the processing and collection speed of real-time traffic flow information and reduce unnecessary operation process, the collected RGB three-channel images are firstly grayscale processed. Based on a large number of experiments and comparisons, the grayscale conversion formula of Formula (1) is adopted in this paper to process the images.

$$Gray = 0.301 * R + 0.586 * G + 0.113 * B \quad (1)$$

In the actual situation, due to weather, sound, light and other reasons, monitoring images often contain some noise, which will interfere with the extraction of traffic flow information. The method to eliminate noise interference is to filter the image information after gray processing by using filter.In this paper, the filter used in the filter pretreatment after grayscale image is Gaussian filter, Gaussian filter is a linear smoothing filter is known as the most useful filter. After Gaussian filtering, each pixel of the image is weighted average by itself and the value of other pixels in its neighborhood. The closer the weighted average coefficient is to itself, the

ISSN: 2664-9640

greater the coefficient is, and the farther away it is, the smaller the coefficient is, which can filter the image very effectively. Equation (2) is used for Gaussian filtering of the image:

$$G(x,y)=\frac{1}{2\pi\sigma^2}e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{2}$$

The result after image preprocessing is shown in Figure 1:

Fig. 1 Image preprocessing sample results

2.2. Intelligent Information Acquisition

Target detection has always been an important direction in the field of computer vision research and development. Target detection is an essential link in both graphic and text interaction and object tracking. Machine vision is the main direction of its development. Based on machine vision and the fast developing deep learning algorithm in recent years, intelligent collection of road traffic intersection information to better serve the smart city has become a very hot topic.

Intersection of intelligent information acquisition refers to the use of road cameras for real-time monitoring, traffic intersection by monitoring the video on the basis of image preprocessing through deep learning approach for real-time traffic flow information collection of video streams, traffic flow information mainly includes: lane detection, you lane, vehicle equipment, the vehicles quantity, intersection queue length, the pedestrian detection.

On the basis of image preprocessing, this paper adopts YOLO V3 model to intelligently collect traffic flow information. The construction of YOLO V3 network first requires the construction of Darknet-53 network, which includes 52 layers of convolution layer and 1 layer of output layer. Its main parameters include:

Inputs: Model variables

Index: The serial number of each convolution layer, which is convenient to load the pre-training weight according to the name

Weight: Weight for pre-training

Training: Is it a sign of training

Decay: The rate of decay when estimating moving average

Epsilon: Variance plus a very small number to prevent dividing by 0

ISSN: 2664-9640

Conv: the result after 52-layer convolution calculation. The size of the input image is 416×416×3, so the size of the output result is 13×13×1024

Route1: Returns the convolution calculation of the 26th layer with the size of 52×52×256 for later use

Route2: Returns the convolution calculation of the 43rd layer with the size of 26×26×512 for future use

Index: Convolutional layer count, easy to use when loading the pre-trained model

After the YOLO V3 network builds Darknet-53 network, it needs to set a prior box to decode the output of Darknet-53 network. YOLO V3 network needs to decode three feature layers. Shape of the three feature layers are (N,13,13,255), (N,26,26,255) and (N,52,52,255) respectively, and each image is convolved to the grid position of 13x13, 26x26 and 52x52.

The steps to set the prior box are as follows:

1. Reset the first feature layer (N,13,13,255) to the size [-1, 13,13, 3, 80 + 5], which represents the selected 169 center points and 3 priori boxes for each center point;

2. Separate X, Y, W and H channels represented by 5 of 80+5, 0 and 1 are the offsets of X and Y relative to the center point, 2 and 3 are the offsets of W and H, and 4 represents confidence;

3. Establish a 13×13 grid in advance to represent the center point of the grid after 13×13 image processing;

4. Use the calculation formula to calculate the position information of the actual center point of the image;

5. Multiply the confidence by 80 in 80+5 (the probability value corresponding to the target) to get the target detection value;

6. Reset the second feature layer (N,26,26,255) to size [-1, 26,26, 3, 80 + 5] and repeat 2 to 5 steps. Reset the third feature layer (N,52,52, 3, 80 + 5) to size [-1, 52,52, 3, 80 + 5] and repeat 2 to 5 steps.

The main parameters for setting the prior box include:

Feats: Feature mapping of YOLO output

Anchors: The location of the anchor object

Class_num: Number of categories

Input_shape: Enter the size of the image

Image_shape: Size of the image

Boxes: the position of the boxes of objects

BOXES_SCORES: The score of the object box, which is the product of confidence and category probability

After the YOLO V3 network set the prior box and decoded the output of Darknet-53 network, it needed to sort the score of target detection and screen the non-maximum suppression of target, and finally screen out the detection target. The main steps of score sequencing and target non-maximum inhibition screening include:

1. Take out the score of the corresponding target detection based on the prior box, and keep the score box and its score that are greater than the threshold value;

2. Based on the retained scoring box, the position information of its box is used to conduct non-maximum suppression, and the better scoring box is retained.

The main parameters of score sequencing and target non-maximum inhibition screening include:

Outputs: Preliminary outputs of the YOLO model

Image_shape: Size of the image

Max_boxes: Number of large boxes

ISSN: 2664-9640

Boxes: the position of the boxes of objects

Scores: The probability of determining the category of an object

Classes: the classes of objects

The trained network and traditional methods are used to detect the processed images, and the detection results are shown in Figure 2:

Fig. 2 Example of intelligent image acquisition results

Intelligent collection information is shown in Table 1:

Table 1 Traffic flow information collection table

lane	1	2	3	4	5
Nature of the driveway	Turn right	Go straight	Go straight	Go straight turn left	Turn left
Effective number of car	1	4	4	1	0
Queue length	1	2	2	1	0

3. Deep Reinforcement Learning Network Algorithm

3.1. DQN of Deep Reinforcement Learning Network

Compared with Q learning network, DQN network of deep reinforcement learning mainly has the following two improvements. A Fixed Q - the targets is to use the convolution neural network to approximate the value function, forecast Q estimation of the parameters of the neural network with the latest, and predict the parameters of the neural network using Q reality is long, long ago, experience value DQN one is to use two of the same structure but different parameters Q network support each other, makes the result more reliable;2 it is to use the Experience of playback Experience replay, generally the intensive study of observation data is ordered is step by step, the supervision of the general learning data were independent of each other, use Experience in DQN playback, which is a Memory to store the training of history data, each time you update the parameter extraction of part of the data from the Memory to one used to update the network value, in order to break the continuity between the data playback Experience randomly disrupted the correlation of training data, make the training more effective and accurate.

The algorithm flow of DQN network is mainly divided into the following steps:

1) Initialize the Memory component and give it an initial capacity N;

2) Initialize the Q learning network and randomly generate a weight ω;

3) Initialize the Fixed q-targets network and set the weights to ω-=ω;

ISSN: 2664-9640

4) Loop through the episode = 1,2,3... M;

5) Initialize initial state S1S1;

6) Loop through the Step = 1,2,3... T

3.2. A2C of Deep Reinforcement Learning Network

A2C(Advantage Actor-Critic) is a member of the deep reinforcement learning algorithm for Policy strategy, which divides two network Critic and Actor based on the Policy Gradient.Actor network refers to the probability distribution of input states and output actions, from which actions are selected as input to CRITIC network.CRITIC network refers to the q-value of input states and actions that predict the next state.In most other network models, the model outputs only one variable: either a policy or a value.However, A2C breaks this traditional rule. The neural network of A2C can output two variables.

A2C algorithm to optimize the PG algorithm, in order to solve the problem of Q value variance in PG algorithm is too large, set the average reward as a baseline unto a Q value idea, let the neural network output V (s) directly, instead of as a baseline, while the actual Q value (value calculated by the next state V) and the difference between the baseline V flagged as needing correction.The baseline of this correction is called CRITIC, and the corrected Q value is called Actor, thus forming the Advantage Actor-CRITIC algorithm.

The algorithm process of A2C network is mainly divided into the following steps:

1) Based on the PG algorithm, the algorithm sequence (S, A, R, S ', Done) is firstly obtained from the environmental training step;

2) Calculate the Q value of A2C network by using PG network formula, and calculate the reward of the total discount value;

3) Record the sequence in the network (s,a, reward,s',done);

4) Continue to explore the environment, and sample old data and new data in the environment;

5) Repeat steps 2-5 until the total amount of data reaches the expected set size, then stop sampling;

6) Start training, calculate Q value: val_ref= reward + net(s'), calculate policy and baseline: action_s,val_s = net(s).Net (S ') represents V(S '), that is, the V value of the next state;

7) Calculate the value loss function, which represents the mean square deviation of the baseline and the actual baseline (Q);

8) Calculate action probability and information entropy loss;

9) Calculated value-related loss = information entropy loss+ value loss;

10) Calculate the difference between the calculated baseline and the actual baseline.

11) Use the policy output by Advantage and neural network to calculate actual actions, and use cross entropy to calculate policy loss;

12) Calculate gradient descent for value-related loss and policy loss

3.3. DDPG of Deep Reinforcement Learning Network

DDPG(Deep Deterministic Policy Gradient) is also a member of Deep reinforcement learning algorithm of Policy strategies, which is a strategy learning method that integrates Deep learning neural network into DPG.DDPG can be summarized as two parts: Deep and Deterministic Policy Gradient. Deep and DQN are similar to have two sets of neural networks with the same structure but different updating frequencies. Deterministic Policy Gradient refers to a Deterministic Policy Gradient, which outputs an action value on a continuous action.DDPG is a model-free deterministic strategy gradient algorithm based on actor-critic. Artificial intelligence is to solve the multi-objective task of data preprocessing, multi-dimensional and sensitive input.DDPG absorbs and extracts the advantages of DQN algorithm, uses the different

ISSN: 2664-9640

strategy method, samples in the sample storage buffer (replay buffer) to minimize the correlation between samples, uses Q-network for training, and returns parameters regularly.

DDPG network has the following characteristics: Convolutional neural network is used to simulate the strategy function and Q function, and deep learning method is used to train, which proves the accuracy, high performance and convergence of nonlinear simulation function in reinforcement learning method;When the actor interacts with the environment, the transition data sequence generated is highly correlated in time. If these data sequences are directly used for training, the neural network will be overfitted and not easy to converge.The use of Target network and Online network makes the learning process more stable and the convergence more guaranteed.

The algorithm flow of DDPG network is mainly divided into the following steps:

1) random initialization of the Actor network mu | theta u (s) and the Critic network Q (s, a);

2) Initialize the Target network, copy Actor and CRITIC;

3) Initialize Replay Buffer R, which is used to disrupt the correlation between data and make the data independently distributed;

4) training Episode;

5) Initialize a random value of N, set to the exploration depth of the action;

6) Obtain the observed state S1;

7) Select folding action at according to the output of strategy network μ and exploration degree NT;

8) Execute the action at and get the reward RT and the next state ST +1;

9) Store and read Replay Buffer R, and store the sequence to R;

10) Then read the sequence in R in random batches for learning;

11) Update the CRITIC network structure and define YI;

12) Using root mean square error, the loss value of value function is directly used when parameter iteration is updated;

13) Update the Actor network structure and update the policy gradient of the network;

14) Soft updating method and τ delay parameter are used to update the Target network.

4. The application of deep reinforcement learning in traffic control

4.1. Combination of deep reinforcement learning and traffic control

The deep reinforcement learning model can be used to control traffic signals without directly calculating complex problems such as dynamic model and control phase change in signal control problems. Instead, the control strategy of signals can be continuously optimized by the continuous exploration of agent in learning and training under the correction of reward signals. Using deep reinforcement learning model to control traffic signals can greatly reduce the complexity of traffic control problems, and it is often highly adaptable to the traffic flow problems with strong real-time performance and large complexity. As shown in Figure 3, the framework combining deep reinforcement learning and traffic control is presented.

ISSN: 2664-9640

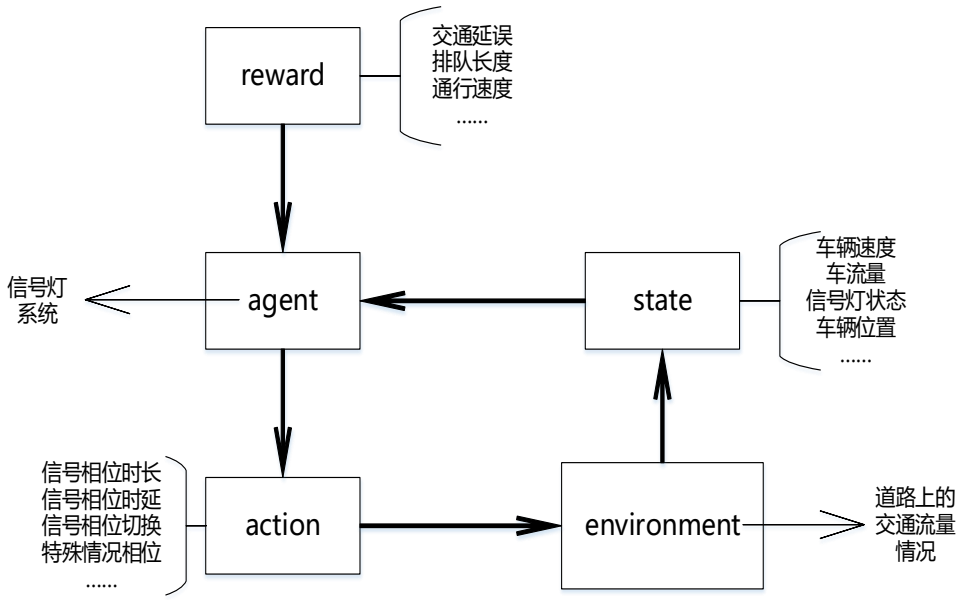


Fig. 3 Intelligent traffic control framework

State selection: The state selection adopted in this paper is to divide the road information into states according to time. One state represents the state of signal lights and average speed in the road unit within the current 30s.This kind of state selection is to replace all the information with the main information of the road, which is beneficial to reduce the data dimension and facilitate the agent to make decisions quickly and accurately.

Action selection: This paper considers the traffic control of Type 1, T and + intersections. Each Agent represents a single signal control lamp and is responsible for controlling the phase transition of a signal lamp. The phase setting in this paper is shown in Figure 4：

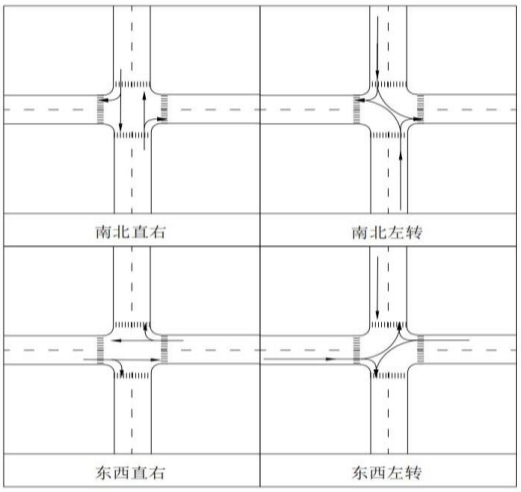


Fig. 4 Phase diagram

Reward value selection: Reward value is the main correction parameter of agent's training, which means that the agent chooses the action at time t and the environment rewards RT for the action feedback after the action at is executed. The reward value is calculated by formula (3) :

$$r_t = 0.3 * \mathrm{L}_t + 0.5 * V_t^{'} + 0.2 * S \tag{3}$$

Where, RT is the reward value; Lt is the queue length at the current time; Vt 'is the decreasing rate of the average velocity; S is the complexity of the road.

ISSN: 2664-9640

4.2. Deep reinforcement learning model trainingo

This section introduces the training process of deep reinforcement learning model in detail:

1) Set a model playback database Replay Buffer with a capacity of 50000, each of which can hold 50000 experimental data samples;

2) Set the correction parameter epsilon, which is used to determine the parameters and proportion of the explore/exploit paths. The initial parameter is set at 0.9, and then it automatically decreases by 0.005 every 64 time steps according to the epoch and continues to decay until the epsilon reaches the lower limit of 0.3.

3) Start to collect data samples (State, Action, next-state, Reward) to fill the Replay Buffer. Before the end of database collection and fill, follow the initial strategy, which adopts a simple timing control strategy;

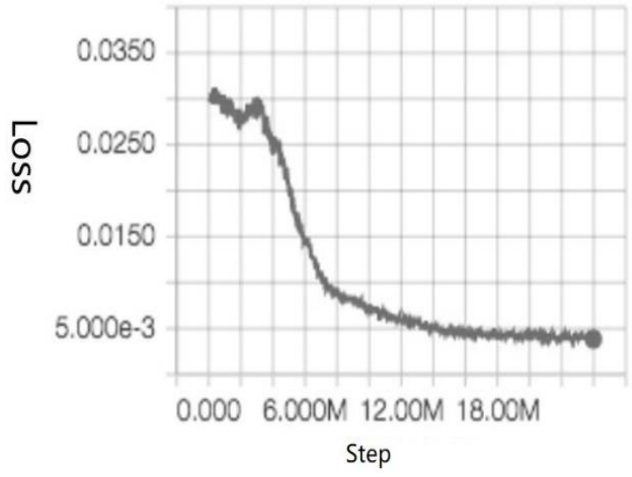
4) Then, after the collection of Replay buffer is completed, the model is trained by random sampling from the buffer, and the training goal is maximized by stochastic gradient descent;

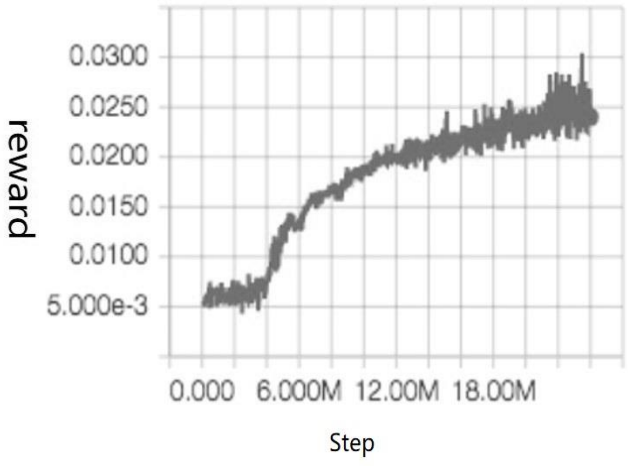
5) Loss loss is set as the difference between the reward value after the action feedback of sampled data and the road state change. When loss loss changes less in continuous time step training, the training is determined to be over.

4.3. Deep reinforcement learning model optimization

DQN, A2C and DDPG were used for learning and training at three intersections of type 1, type T and type +, and parameters such as loss decrease, average reward increase, change of queuing delay at intersections, and network Q expectation were compared respectively for three networks at four intersections.

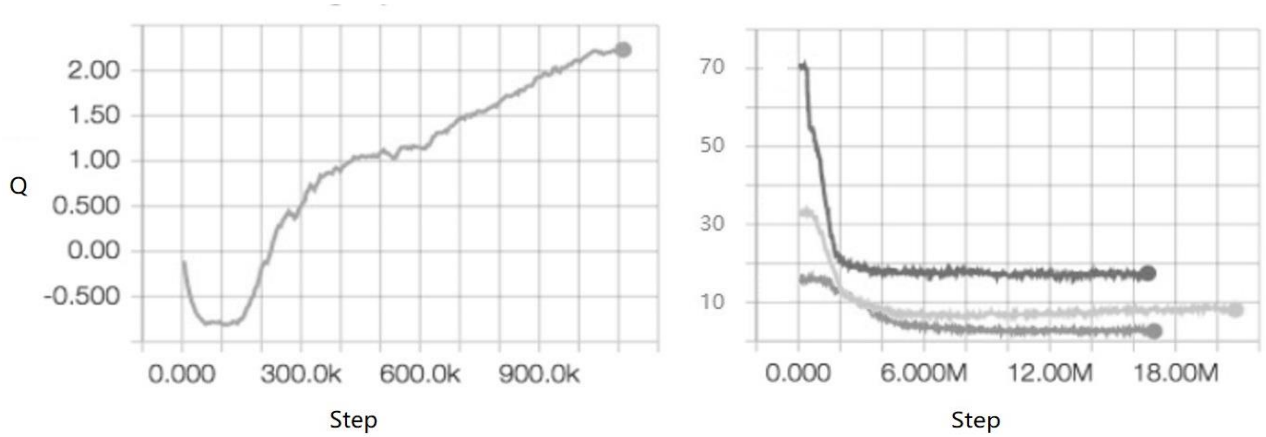
As shown in Fig. 5 (a), the overall trend of the optimized network in loss decline is obvious, with a fast rate of decline, and the final loss remains good.As shown in Fig. 5 (b), although the reward value for real-time and changeable traffic conditions fluctuated greatly in a short region, the overall upward trend was obvious and eventually tended to be stable.As shown in Fig. 5 (c), although the Q value expectation of the preferred network decreased in the trial and error process at the beginning, it also rose rapidly and tended to be stable in the continuous training.As shown in Fig. 5 (d), the queue length of the optimized network changes before and after the signal lamp phase conversion, and the optimization effect is obvious with the increase of training duration.





(a)Preferred network LOSS changes (b)Reward changes for the preferred network

ISSN: 2664-9640



(c)Preferred network Q value changes (d)Preferred network queue length variation

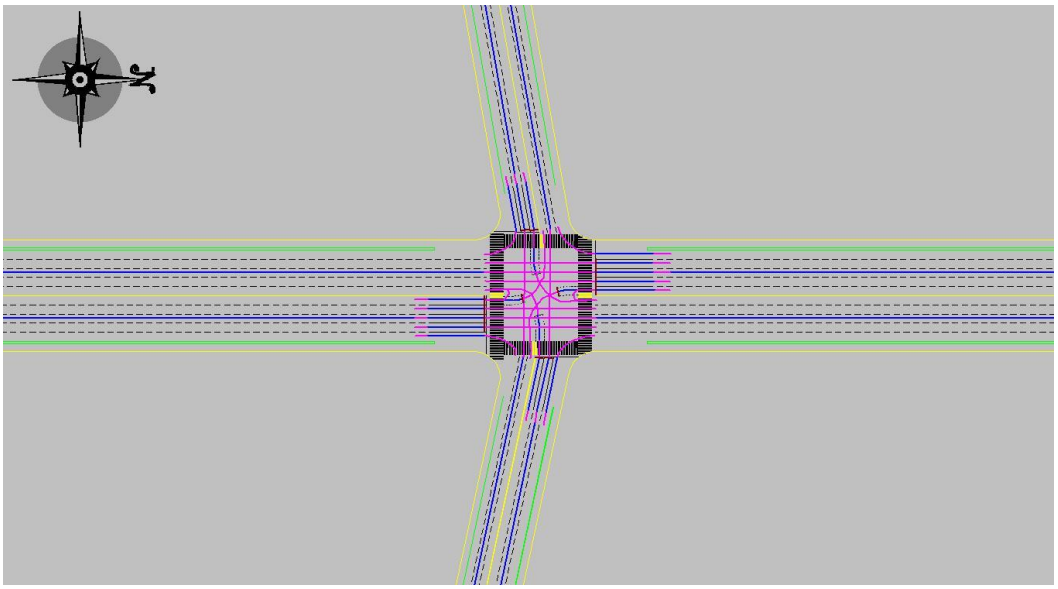
Fig. 5 Optimize network training results

5. VISSIM co-simulation

In order to further verify the effectiveness of the traffic control method based on machine vision and deep reinforcement learning proposed in this paper, the agent trained based on the preferred ascending reinforcement learning model builds a more complex traffic network on the VISSIM platform, and uses the agent to control the traffic network to check the optimization effect. VISSIM is a micro simulation software for urban traffic and public traffic operation, and is an effective tool for evaluating traffic engineering design and urban planning schemes.

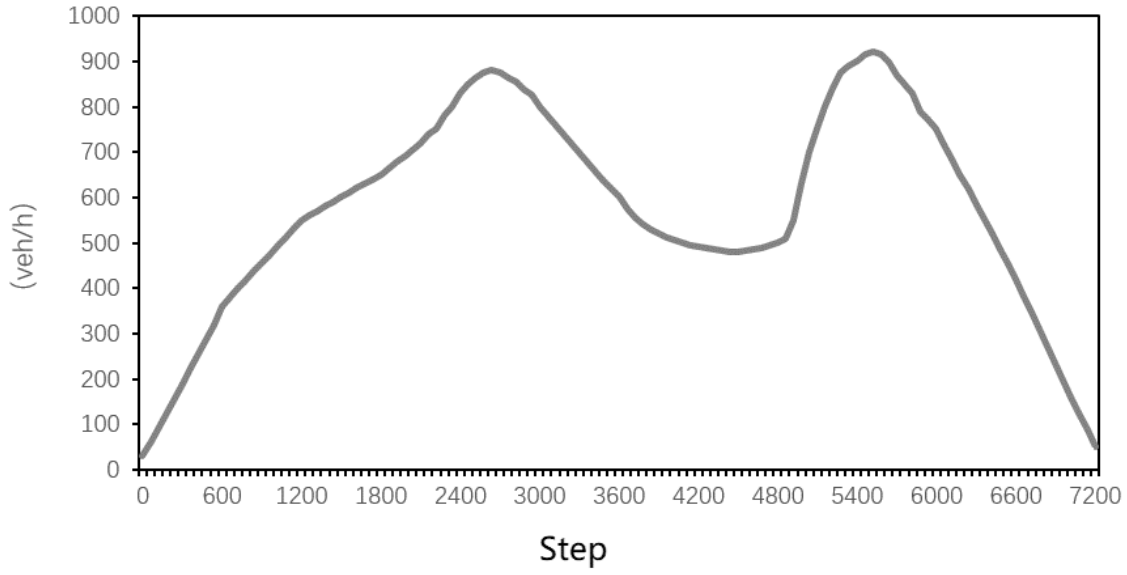
5.1. Experimental Simulation Construction

Based on the agent trained by the optimized deep reinforcement learning model, the road network model is built in VISSIM software, and the vehicle type, vehicle speed and intersection signal controller are set up.As shown in Fig. 6 (a), a complete road network model of intersections is established in the VISSIM traffic simulation software, and vehicle composition, expected speed, driving path, conflict zone, etc., are set respectively. As for the input of vehicle flow, this project set the input of vehicle flow based on the field research, as shown in Fig. 6 (b), to simulate the daily tidal distribution in the city. Road breakpoints are set at the intersections, and evaluation parameters such as queue length, traffic speed and traffic delay are carried out.



(a)VISSIM road network model construction

ISSN: 2664-9640



(b)Tidal flow input

Fig. 6 VISSIM Network Modeling and Traffic Input

5.2. Analysis of simulation results

According to the traffic flow input of tidal distribution, the established traffic network is simulated and analyzed. As shown in Fig. 7, the optimized Agent trained by deep reinforcement learning network has significantly better control over the intersection than the Fixed timing Fixed. In the case of Agent automatic matching, the queue length of intersections is significantly shorter than that of Fixed matching. On the overall trend, the optimization of the intersection by the agent is very obvious, which reduces about 28.6% compared with the queuing length of fixed timing.

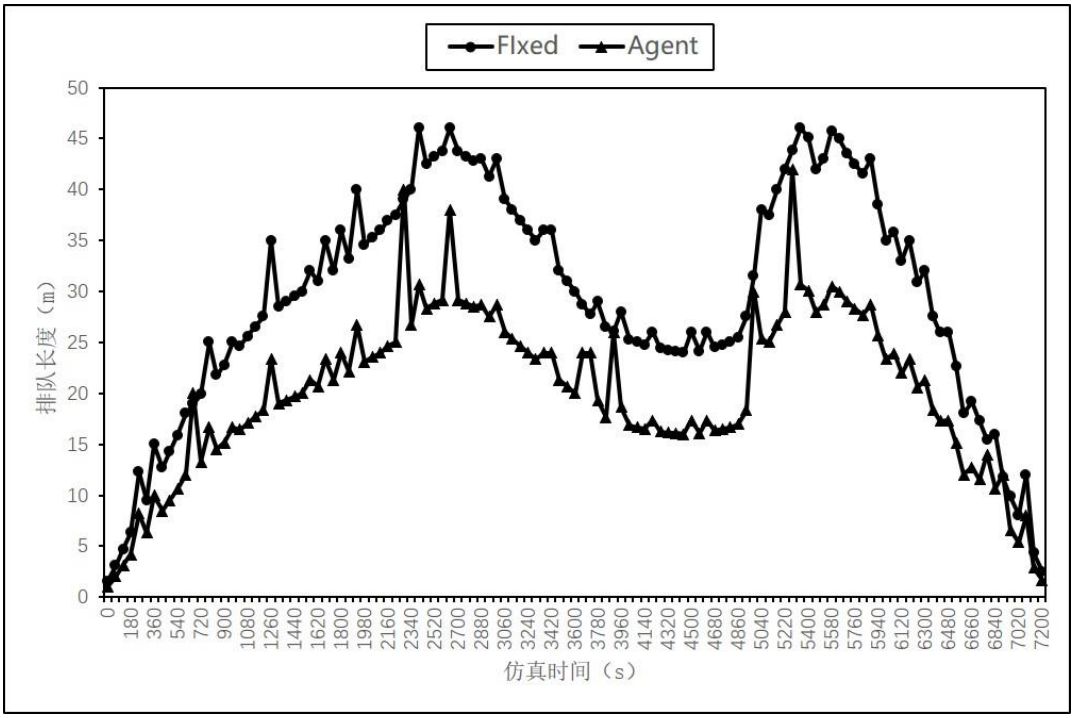


Fig. 7 Queuing length comparative analysis

The advantages and advancements of deep reinforcement learning in traffic control are further verified by micro-traffic simulation software VISSIM.

ISSN: 2664-9640

6. Conclusion

With the continuous development of cities, urban traffic problems will become increasingly severe. It is extremely urgent to extract the traffic flow information based on machine vision with ultra-high speed and ultra-precision, and then combine with deep reinforcement learning to optimize the control of urban traffic roads.In this paper, after optimizing deep reinforcement learning network to train Agent, the Agent is further verified and analyzed based on micro-traffic simulation software VISSIM. The results show that the traffic control method based on machine vision and deep reinforcement learning is highly effective in urban traffic road control.

References

[1]. Xu Haidong, Plum Alcohol. Analysis on The Operation of China's Automobile market in 2019 [J].Automobile Aspect,2020,(2): 56-59.
[2]. Zhao Li. Inventory of China's Auto Market in 2018 [J]. Automotive Horizon,2019,(2): 35-37.
[3]. Yao Guangzheng, LIU Xiaoming, Chen Yanyan, CUI Kaijun. Analysis and prediction of the limit value of urban car ownership,2020,(5): 110-119.
[4]. Zou Lei. A Brief Discussion on traffic flow detection Technology based on machine vision [J]. Jinxiu (the following ten-day issue),2020,(4).
[5]. Kong Fang fang, Song Beibei.Improved panoramic traffic monitoring target detection for YOLOv3 [J]. Computer Engineering and Applications,2020, Vol. 56 (8): 20-25.
[6]. Wu, Junta 1 (AUTHOR); Li, Huiyun 1 (AUTHOR). Deep Ensemble Reinforcement Learning with Multiple Deep Deterministic Policy Gradient Algorithm.[J].Mathematical Problems in Engineering,2020,: 1-13.
[7]. Xu Enchu, Zhu Hai-long, LIU Jingyu, Shi Ye-qiong, Yin Qi-tian.Urban intelligent Traffic control Method based on Asynchronous Intensive Learning [J]. Intelligent Computers and Applications, 2019, Vol. 9 (6): 164-167.