Alexandria Engineering Journal 106 (2024) 381–391

Contents lists available at ScienceDirect

Alexandria Engineering Journal

journal homepage: www.elsevier.com/locate/aej

Reinforcement learning based adaptive control method for traffic lights in intelligent transportation

Zhongyi Huang

Geography and Environment Department, London School of Economics and Political Science, London, England WC2A2AE, United Kingdom

ARTICLE INFO

Keywords:
Machine learning
Transportation engineering
Intelligent transportation system
Q-learning algorithm
DQN control theory

ABSTRACT

Addressing the requirements and challenges of traffic light control, a reinforcement learning based adaptive optimal control model for traffic lights in intelligent transportation systems is proposed. In the model design, we combined Markov decision process, Q-learning algorithm, and Deep Q-Learning Network (DQN) control theory to establish a comprehensive signal light Adaptive Optimal Control of Signal Lights in Intelligent Transportation Systems (AOCITL) control model. Through simulation experiments on the model and the application of actual road scene data, we have verified the superiority of the model in improving traffic system efficiency and reducing traffic pressure. The experimental results show that compared with traditional fixed cycle signal light control, the adaptive optimal control model based on reinforcement learning can significantly improve the traffic efficiency of roads, reduce the incidence of traffic accidents, and enhance the overall operational effectiveness of urban transportation systems. The proposed method is possible to further optimize the model algorithm, expand its application scope, and promote the development and practical application of intelligent transportation systems.

1. Introduction

With the acceleration of urbanization and the continuous growth of car ownership, traffic congestion has become a common problem faced by major cities around the world [1]. Traditional traffic signal control methods, such as fixed duration control and simple adaptive control, are no longer suitable for increasingly complex traffic environments. For this reason, researchers are constantly exploring new technological means to improve traffic management efficiency. Among them, the adaptive optimal control model for traffic lights in intelligent transportation systems based on reinforcement learning has received widespread attention [2].

As an important part of urban traffic system, traffic lights' control strategies directly affect the efficiency and safety of road traffic [3]. Traditional traffic signal control methods are often based on fixed time or simple adaptive logic, which cannot sense the change of traffic environment in real time and make corresponding adjustments. This not only easily leads to increased traffic congestion, but also may increase the risk of traffic accidents. In recent years, with the development of sensor technology, big data analysis and artificial intelligence technology, new possibilities have been brought to traffic signal control [4]. Through real-time monitoring of road traffic flow, speed, pedestrian density and other information, we can more accurately grasp the change of traffic environment, and provide a more scientific basis for traffic signal control.

Reinforcement learning, as a machine learning paradigm based on environmental interaction, aims to learn an optimal strategy that maximizes cumulative rewards in a given environment through continuous trial and error. This learning method has shown great potential for the application of reinforcement learning in traffic signal control. Reinforcement learning can perceive real-time changes in the traffic environment and adjust the control strategy of traffic lights based on these changes [5]. Through continuous interaction and trial and error with the environment, reinforcement learning algorithms can learn an optimal traffic light control strategy, thereby maximizing traffic efficiency and reducing traffic congestion. Reinforcement learning can handle complex traffic scenarios. In actual traffic environments, there are often multiple influencing factors, such as road type, vehicle type, number of pedestrians, etc [6]. These factors interact with each other, making the traffic scene exceptionally complex. And reinforcement learning algorithms have strong learning capabilities, which can handle such complex scenes and learn an optimal traffic light control strategy [7].

Although reinforcement learning has shown great potential in traffic signal control, current research is still in its early stages. The existing research mainly focuses on traffic signal control at single intersections, while there is relatively little research on traffic network signal control

E-mail address: z.huang55@lse.ac.uk.

https://doi.org/10.1016/j.aej.2024.07.046
Received 2 March 2024; Received in revised form 27 June 2024; Accepted 9 July 2024
Available online 15 July 2024

at multiple intersections and regions. In addition, existing research also has some limitations, such as high model complexity and long training time. This study aims to explore an adaptive optimal control model for traffic lights in intelligent transportation systems based on reinforcement learning. By constructing an efficient reinforcement learning algorithm framework, adaptive control of traffic lights in multi intersection and multi area traffic networks can be achieved. This study can not only improve the efficiency of urban traffic and reduce traffic congestion, but also provide new ideas and methods for urban traffic management. Research contributions include:

1) This paper proposes a signal control strategy based on multi-agent cooperation, in which each intersection is regarded as an independent agent. This model can realize unified scheduling and optimization of multiple intersection signal lights in the region, so as to avoid the spread and transfer of traffic congestion.

2) Through continuous learning and optimization, the model can gradually adapt to different traffic scenarios and needs to achieve more efficient and intelligent signal control.

The first chapter describes the background and main purpose of the study. Chapter 2 reviews the latest relevant research to provide the basis for the follow-up research. The third chapter combines Markov decision to complete the traffic state decomposition, uses extended Kalman filter to construct the backbone network of intelligent traffic signal system and complete the feature fusion, and finally establishes the traffic light adaptive optimal control AOCITL model. Chapter 4 Test and analysis. The fifth chapter is the discussion of the research results and the suggestions for the traffic signal system. The last chapter is the summary of the full text, the limitations of the study and the possible direction of follow-up research.

2. Literature review

With the increasingly serious problem of urban traffic congestion, the development of intelligent transportation systems (ITS) has become the key to alleviating traffic pressure and improving road efficiency. Among them, traffic lights, as the core element of traffic control, have become a hot research topic in terms of their adaptive optimal control. In recent years, reinforcement learning (RL) technology has shown great potential in the field of traffic signal control due to its unique learning and optimization capabilities. Li, et al. [8] have applied a multi-objective multi-agent framework to traffic signal control at a single intersection. They treat each signal phase as an intelligent agent and use reinforcement learning algorithms to enable these agents to learn the optimal signal light control strategy. Their research not only considers the degree of traffic congestion, but also fuel economy, optimizing these goals by setting appropriate reward functions. Haydari, et al. [9] proposed a traffic signal control model based on deep reinforcement learning. They use deep neural networks to approximate complex traffic environments and train agents through interaction with the environment. Their model can perceive traffic conditions in real-time and adjust the control strategy of traffic lights based on these conditions. Through experiments in actual traffic scenarios, they demonstrated that the model can effectively improve traffic efficiency and reduce congestion. Baumgart, et al. [10]. have studied multi intersection traffic signal control based on reinforcement learning. They proposed a distributed learning architecture where each intersection has an independent agent to learn its optimal control strategy. These intelligent agents share information through communication to achieve coordinated control of the entire transportation network. Their research suggests that reinforcement learning can enable traffic signal controllers at multiple intersections to work together to improve the performance of the entire network. Ge, et al. [11]. have combined reinforcement learning with fuzzy logic to propose a new traffic signal control method. They use fuzzy logic to handle the uncertainty of traffic conditions and use reinforcement learning to optimize the parameters of fuzzy controllers. This method can demonstrate good adaptability in different traffic scenarios and effectively adjust the control strategy of traffic lights in real-time traffic environments.

Drawing on the research achievements of the above scholars and combining with the application of reinforcement learning in intelligent transportation systems, we can construct a more efficient and intelligent traffic signal adaptive optimal control model, effectively alleviating traffic congestion, improving transportation system efficiency, and bringing important technological innovation and improvement to urban traffic management.

3. Intelligent traffic signal light control system and reinforcement learning

3.1. Basic concepts of traffic signal control systems

Traffic signal control system is an important part of urban traffic management [12,13]. It guides and manages traffic flow by controlling changes in traffic signals [14,15], improves road traffic efficiency, reduces traffic congestion, and improves traffic safety [16,17]. The traffic light adaptive optimal control model of intelligent transportation system based on reinforcement learning is a new research direction [18,19]. The core idea is to use reinforcement learning algorithm to dynamically adjust the traffic light control strategy according to real-time traffic conditions and environmental changes, and realize the adaptive optimal control of traffic lights.

The traffic signal control system is a crucial component of urban traffic management, and the reinforcement learning based intelligent traffic system signal adaptive optimal control model represents an innovative direction in traffic signal control technology, which will bring more efficient and intelligent solutions to urban traffic management.

3.2. Markov decision process

The Markov Decision Process (MDP) is a mathematical framework used to describe decision problems [20,21], especially in environments involving randomness and uncertainty. When studying the adaptive optimal control model of signal lights in intelligent transportation systems [22], Markov decision process can provide effective tools and methods for system modeling.

Reinforcement learning refers to the process in which agents interact with the surrounding environment to achieve self-learning and optimize goals without the need for predetermined data [23]. Agents that make decisions or control the environment through reinforcement learning are called agents [24]. Its learning process is carried out through the state of the interactive environment (State) $s \in S$, where the agent executes an action (Action) $a \in$A, and the environment executes a reward (Reward) r. Among them, S_A is the set of environmental states. A_t time t, the intelligent agent selects the action at corresponding to state s_t, and the reward obtained by taking action at in the environment will be obtained at time $t+1$, represented by r_t.

The environmental state of reinforcement learning problems satisfies the Markov requirement, which means that the action A_t is taken in the current state St and transferred to the next state S_{t+1} without considering the previous state S_{t-1}, S_1. The state S_t captures all associated state information in history. If the state S_{t-1}, S_1 is discarded, but the transition probability from the current state S to the next state S_{t+1} is still obtained. The transition probability is:

$$P = [S_{t+1}|S_t] = P[S_{s+1}|S_1, S_2, S_3, S_t] \tag{1}$$

Markov processes belong to the category of stochastic processes, also known as Markov chains. It describes the stochastic process of spatial state transition from one state to another, represented by a binary M=<S, P>, where S is a finite set of states and P is the probability of state transition. All state transition probabilities are combined to form a state transition matrix [25],

$$P_{ss'} = P[S_{t+1} = s' | S_t = s] \tag{2}$$

The transition relationship between states with rewards is described using Markov Reward Process (MRP). When each state transitions, different reward values are given, which is the process of obtaining Markov rewards. The reward process of Markov includes quadruples<S, P, R, γ>, the process is shown in Fig. 1.

In the intelligent transportation system, the control of signal light is an important link that affects the smooth and efficiency of traffic flow. However, due to changes in the number of vehicles on the road, speed, driving direction and other factors, the traditional timing control method of signal lights may not fully adapt to real-time traffic conditions, resulting in traffic congestion and low efficiency [26].

$$R_{S,A} = E(R_t | S_t = s, A_t = a) \tag{3}$$

It refers to the expected reward obtained when the state at time t is transmitted to various possible states within time t+1. By judging the results obtained after completing action a∈A, it guides the direction of intelligent agent learning.

$$G_t = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \ldots \tag{4}$$

The reward attenuation coefficient in the formula is γ∈ [0,1] refers to the sum of all rewards r obtained from the current t-sampling to the termination time with decay. If γ= If 0, it means that the harvest will only be determined by the current immediate reward, and it is not related to the delayed reward in the future; If γ= 1. It indicates that there is no difference in the impact of the current two types of rewards on the overall harvest. Specifically, they are subsequent delayed rewards and immediate rewards.

3.3. Traffic state decomposition

We introduce the principle of "traffic state decomposition", that is, traffic inference should maintain rotation, inversion and other symmetries, so that the model can learn from the data more effectively and adapt to different intersection structures more easily. We divide traffic reasoning into four steps: traffic state decomposition, traffic information representation, traffic reasoning and traffic state synthesis. By decomposing traffic state, action and future information and input it into the reasoning model based on lane group, the principle of invariance can be guaranteed [27]. The formula is as follows:

$$\left(s^{(1)}, s^{(2)}, \ldots, s^{(n)}\right) = Decomp(s) \tag{5}$$

$$\left(a^{(1)}, a^{(2)}, \ldots, a^{(n)}\right) = Decomp(a) \tag{6}$$

$$\left(f^{(1)}, f^{(2)}, \ldots, f^{(n)}\right) = Decomp(f) \tag{7}$$

Where n represents a group of lanes in the same direction and phase. Encode the collected lane group status, actions, and future information separately, embed them, and obtain their representation vectors. Its function is to convert the raw data into representation vectors that can be recognized by neural networks. The process of q_i traffic state decomposition is as follows.

Specifically, we will represent the collected lane group status, vehicle or pedestrian actions, and future information as $s^{(i)}$, $a^{(i)}$, $f^{(i)}$, respectively. Take $s^{(i)}$, $a^{(i)}$, $f^{(i)}$ as inputs and pass them into three independent K1 layer fully connected neural networks to generate corresponding representation vectors e^s, e^a, e^f.

$$h_1^s = ReLU\left(W_1^s e^s + b_1^s\right) \tag{8}$$

Traffic flow refers to the number of vehicles passing through a certain section of road, which can describe the current traffic state based on the density and speed of vehicles. The size and distribution of traffic flow directly affect the degree of road congestion and traffic efficiency, which is of great significance for optimizing traffic signal control. The speed of the vehicle reflects the operation of the vehicle on the road. A fast vehicle usually indicates smooth traffic, while a slow vehicle may indicate traffic congestion [28]. Monitoring speed changes can help traffic managers adjust signal control in time and optimize traffic flow.

$$h_k^s = ReLU\left(W_k^s h_{k-1}^s + b_k^s\right), k \in \left[2, K_1\right] \tag{9}$$

$$e^s = h_{K_1}^s \tag{10}$$

The status of traffic signals includes the duration and phase settings of traffic lights at each intersection, and different signal control schemes will have different impacts on the operation of the traffic system. Optimizing the status of traffic lights requires consideration of factors such as vehicle flow, speed, and queue length to achieve smooth and efficient operation of traffic flow.

3.4. Extended Kalman filter combined with a single intersection traffic signal adaptive method

Extended Kalman filter is a filter used to handle nonlinear systems, combined with adaptive control of traffic lights in intelligent transportation systems, which can better estimate and predict traffic states such as vehicle flow, speed, and queue length, thereby achieving more optimized traffic light control. The traffic signal timing system is

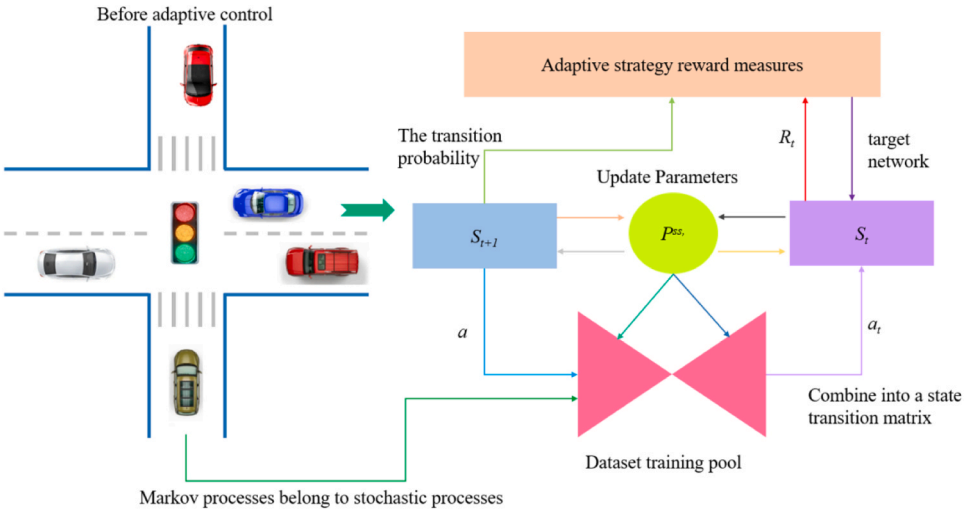

Fig. 1. Reward process of Markov.

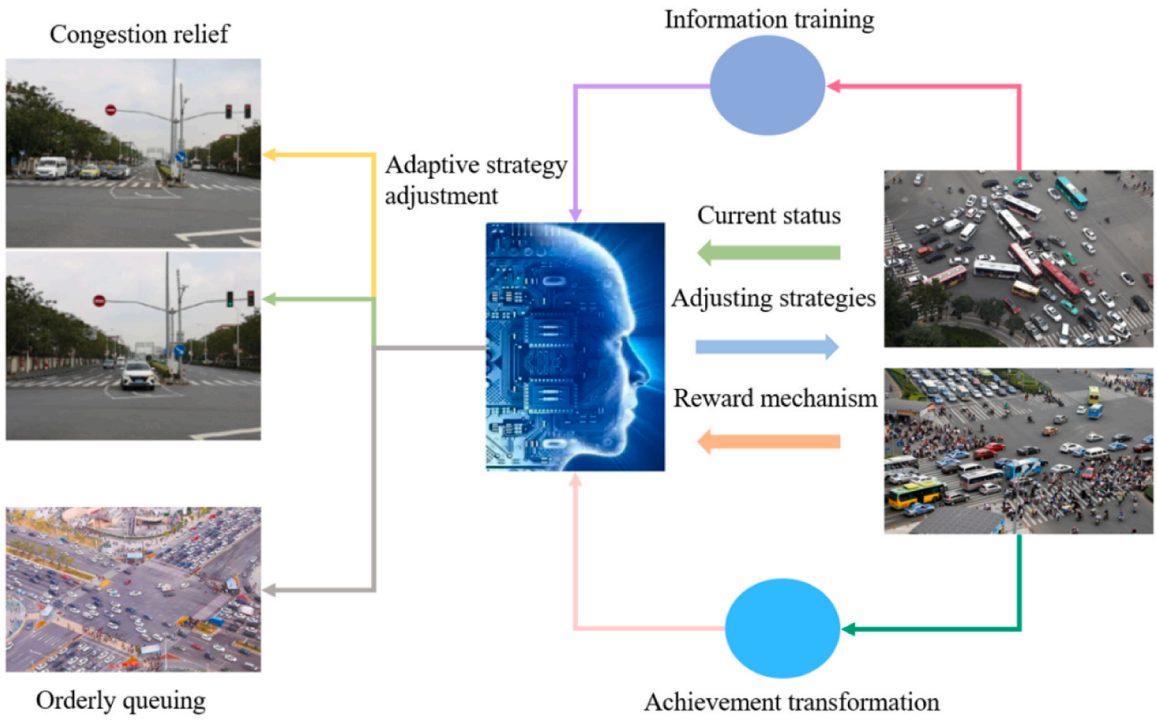

Fig. 2. Traffic State Decomposition.

abstracted as an intelligent agent, the traffic environment at intersections is abstracted as a controlled object, the timing scheme of traffic signals is abstracted as actions, and the changes in the cumulative waiting time of vehicles are abstracted as rewards. Firstly, the traffic signal timing system selects a timing scheme based on the vehicle information provided by the traffic environment; After implementing the timing plan, the traffic environment will provide feedback on the vehicle information and accumulated waiting time to the traffic signal timing system, and this process will be repeated continuously; Finally, the traffic signal timing system continuously updates with the goal of reducing the cumulative waiting time of vehicles until the optimal timing scheme is obtained Among them, s_t is the state at time t, at is the action at time t, and rt is the reward at time t; D_t is the playback memory unit at time t. In the traffic signal timing system, a main measure of vehicle traffic efficiency is the waiting time of the vehicle, so the reward is defined as the difference in cumulative waiting time between adjacent cycles [29].

$$r_t = W_{t+1} - W_t \tag{11}$$

Among them, r_t is the reward at time t, and W_t represents the cumulative waiting time of the vehicle at time t. The position matrix and speed matrix of vehicles at a single intersection are shown in Fig. 3.

$Q(s_t, a_t)$ is the estimated network value function for taking action at a certain moment in state s_t, and the environment will provide feedback on the corresponding reward $r(S_t, a_t)$ based on the action executed by the agent. The update formula for Q-learning is as follows:

$$Q(s_t, a_t) = Q(s_t, a_t) + \alpha\left[r(s_t, a_t) + \gamma \max_{a_{t+1}} Q(s_t, a_{t+1}) - Q(s_t, a_t)\right] \tag{12}$$

The parameters of the target network are not iteratively updated, but are periodically copied from the main network [30]. Therefore, the structures of the two networks are the same, but the parameters are different.

$$Loss(\theta_t) = \left[r(s_t, a_t) + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1}|\widetilde{\theta}) - Q(s_t, a_t|\theta_t)\right]^2 \tag{13}$$

In the formula, Z_t is the value function of the target network at time t, $\widetilde{\theta}$ is the fixed weight of the target network, $Q(s_t, a_t|\theta_t)$ is the value function of the estimated network at time t, and θ_t is the uncertainty parameter of the estimated network at time t. In the adaptive method for single intersection traffic signals, extended Kalman filtering can combine traffic state data and sensor information to effectively estimate the current traffic state and predict future traffic state changes. By utilizing real-time data and information in intelligent transportation systems, extended Kalman filtering can dynamically adjust the control

Position matrix

1	0	1	0	0	1
0	1	0	1	1	1
0	0	1	1	0	0

Speed matrix

0	0	0	0.3	0.5	0.8
0	0.2	5	1	1.2	0.3
0	0	0.2	0.2	0.3	0.6

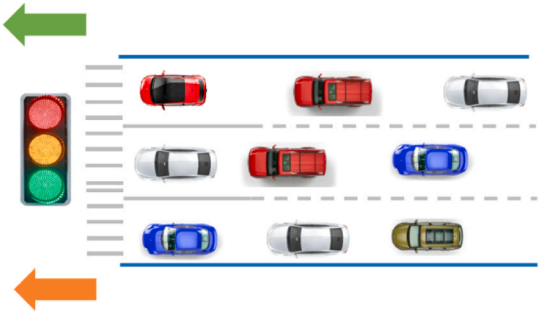

Fig. 3. Position matrix and speed matrix of vehicles at a single intersection.

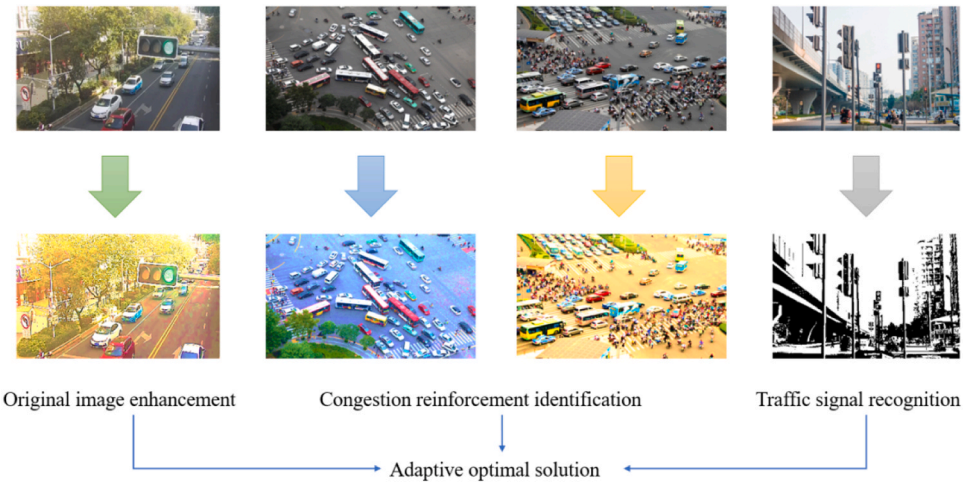

Fig. 4. Schematic diagram of adaptive optimal solution operation.

parameters of traffic lights to adapt to changes in traffic flow and minimize traffic congestion.

3.5. Backbone network and feature fusion structure

In the process of studying the adaptive optimal control model for intelligent transportation system traffic lights based on reinforcement learning, the backbone network and feature fusion structure are two key technical components that can help improve the performance and stability of the traffic light control model. In intelligent transportation systems, backbone networks are typically used to process the input and output of traffic status data. The backbone network can be a deep neural network responsible for learning complex traffic state features, such as vehicle flow, vehicle speed, and queue length. Through the backbone network, the model can extract representative features from the raw data, providing effective input for signal light control decision-making. In the transportation system, there may be complex correlations and mutual influences between different traffic state characteristics. The function of feature fusion structure is to fuse and integrate features from different sources, in order to improve the model's understanding and representation ability of traffic status. Through the feature fusion structure, the model can more comprehensively consider the dependency relationships between different features, making signal light control decisions more accurate and robust.

Feature learning and representation backbone networks can learn complex traffic state features through deep learning, thereby better describing the operation of transportation systems. The feature fusion structure can integrate multiple feature information from the backbone network, improving the model's representation ability of traffic status. Decision optimization integrates the features extracted from the backbone network and the feature fusion structure, allowing the model to better understand the complexity of traffic conditions and provide more accurate inputs for signal control decisions. This helps to optimize the timing and phase settings of traffic lights and achieve adaptive control of intelligent transportation systems. The signal light control model combining backbone network and feature fusion structure can improve the overall performance of intelligent transportation systems, effectively reduce traffic congestion, optimize vehicle traffic efficiency, improve road utilization, and achieve intelligent and efficient operation of transportation systems.

3.6. Traffic signal control algorithm based on DQN

Traffic signal control is the key link to ensure efficient and safe operation of road network. With the development of deep reinforcement learning, Deep Q network (DQN) algorithm has become an important method to solve traffic signal control problems. By combining the feature extraction capability of deep learning with the decision-making capability of reinforcement learning, DQN algorithm enables traffic lights to achieve adaptive optimal control, thus improving the efficiency and safety of the traffic system.

Using formula 10 as the definition of the state action value function, and based on this, updating the Q value can be achieved through the formula [31]. The new Q value is equal to the old Q value plus the learning rate multiplied by the difference, where the difference is the actual Q value minus the estimated Q value.

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \cdot \delta_t \tag{14}$$

$$\delta_t = R_{t+1} + \gamma \cdot \max_a Q\left(s_{t+1}, a\right) - Q\left(s_t, a_t\right) \tag{15}$$

Among them a is the learning rate. In traditional Q-learning algorithms, tables are usually used to store the Q-values corresponding to each state action, and the Q-value table is queried to determine the action taken by the agent in the next control step. In traffic signal control problems, due to the large traffic state space, it is not realistic to continue using tables to store Q values and select actions by querying the Q value table.

DQN algorithm is a deep reinforcement learning algorithm based on value function, which approximates the state-action value function (Q function) through neural network, so as to realize the evaluation and selection of actions. In the traffic signal control problem, DQN algorithm can take the real-time state of the traffic intersection (such as the number of vehicles, traveling speed, traffic flow, etc.) as the input, calculate the Q value of each action (such as changing the state of the signal light) through the neural network, and then select the best action to execute. Based on these states and actions, the DQN algorithm can learn an optimal control strategy, making the adjustment of traffic signals more intelligent and adaptive.

$$L(\theta) = [Q_t(s, a) - Q_e(s, a; u)]^2 \tag{16}$$

In the formula, Q_t is the Q-reality, Q_e is the Q-estimation, s is the current state, a is the neural network's output action based on state information, and u is the weight parameter of eval.net. Update using random gradient descent method. The weight parameters of target.net are not updated every iteration, but are directly copied to target.net after completing several time steps of the update.

Using state information as the input of a convolutional neural network, the Q-value of the action taken in that state is processed by the convolutional neural network, and this value is used to update and

iterate. Using the velocity matrix and position matrix corresponding to each entrance direction at the intersection as inputs to the neural network, feature information is extracted using two convolutional layers. The first layer consists of 16 4 * 4 filters with a step size of 2, and the second layer consists of 32 2 * 2 filters with a step size of 1. Input the data processed by these two convolutional layers and the intersection phase information into two fully connected layers. The fully connected layer connects all features and outputs a series of Q-values corresponding to various actions that the intelligent agent may take in the current input state.

Through DQN algorithm, traffic lights can adjust signal timing adaptively according to real-time traffic conditions, achieve reasonable allocation of green time, reduce vehicle waiting time, and improve road traffic efficiency. At the same time, the DQN algorithm can also learn according to historical data and real-time traffic conditions, and constantly optimize the control strategy to adapt to the needs of different time periods and different traffic conditions.

3.7. AOCITL model for adaptive optimal control of traffic lights

The study established an AOCITL (Adaptive Optimal Control of Signal Lights in Intelligent Transportation Systems) model based on reinforcement learning. It can adjust the control strategy of signal lights based on real-time traffic conditions and environmental changes through learning and optimization, achieving adaptive optimal control of signal lights. In this article, we conducted experiments on bidirectional six lane intersections. The total number of lanes entering the intersection is 12, so we define the state as the number of vehicles currently parked on each lane. This is an example of an intersection. To calculate the number of stops on the 12 lanes entering the intersection, we normalize the vector according to the following formula to accelerate the training speed.

$$x = (x - x_{\min})/(x_{\max} - x_{\min}) \tag{17}$$

x_{min} represents the minimum value in the data, and x_{max} represents the maximum value in the data.

When the intelligent agent obtains the current system state, it needs to perform an operation to set the phase of the next cycle. In this model, we define the action space as {a1, a2, a3, …, a9} by selecting the duration of each phase in the next cycle. Actions a1 to a4 indicate an increase of 5 seconds in the corresponding traffic light stage; Action a5 indicates that the duration of the corresponding traffic light phase remains unchanged; Actions a6 to a9 indicate a reduction of 5 seconds in the corresponding traffic light stage. For example, when action a1 is selected, the duration of the north-south straight phase green light increases by 5 seconds. When action a2 is selected, the duration of the north-south left turn phase green light increases by 5 seconds. When action a5 is selected, it means that no changes will be made to the signal light. When action a6 is selected, the duration of the north-south straight phase green light decreases by 5 seconds. When action a7 is selected, the duration of the north-south left turn phase green light decreases by 5 seconds, and so on. With this small phase change interval, the duration of the phase in the current state should change steadily. In this model, we use quadruples<t1, t2, t3, t4>to represent the duration of the four phases in the signal lamp cycle. In addition, we set the maximum legal duration of a phase to 60 seconds and the minimum legal duration to 0 seconds.

Set the delay time for vehicles entering each lane at the intersection to d, the sum of the waiting queue lengths for all vehicles entering the lane to q, the waiting time for all vehicles entering the lane to w, the state switching of vehicle phases to p, the emergency braking stop of vehicles at the intersection to e, and the number of vehicles choosing to leave after executing actions to n.

$$R_t = k_1 d + k_2 q + k_3 w + k_4 p + k_5 e + k_6 n \tag{18}$$

The observed environmental state S_t is mapped to the Q-Value value associated with the current action, and a deep neural network is constructed for systematic training. The input of this signal control system is the IDR (Environmental State Vector) vector with a time step of t, while the output of the deep neural network is the Q-Value value from the action of state s_t.

$$n_{k,t}^{in} = -IDR_{k,t} \tag{19}$$

Among them, *IDRk* represents the nth input element of the neural network when the time step is t, and $IDR_{k,t}$ is the kth element of the vector IDR when the time step is t. The input here in this traffic signal control system is the five dimensional state vector mentioned in the previous establishment of the traffic signal control model, that is, the system state S=($Qn1$,$Qn2$,$Qn3$,$Qn4$….Q_{ni};P_n), which is input into the neural network for training. This way, regardless of the state, it can be included in the vector without omission.

The Q-Value update uses the following update formula:

$$Q(s_t, a_t) = r_{t+1} + \gamma E[max_A Q'(s_{t+1}, a_{t+1})] \tag{20}$$

The reward r_t+1 of the traffic signal control system is an immediate reward obtained only after s_t takes the action, while $Q(s_t a_t)$ is the Q-Value value obtained after s_{t+1} takes the relevant action, that is, the state of the next system after taking the action, discount factor γ Compared with immediate rewards, future rewards will follow the progression of time step t, and the magnitude of punishment will become smaller and even have little impact. The data and perceptron types are shown in Table 3 below.

4. Experimental results

4.1. Data source and experimental parameter setting

The research data comes from the simulated environment provided by the City Brain Challenge competition. Based on the data set generated by SUMO (Simulation of Urban MObility) open source traffic simulation software, in order to verify the applicability and effect of the model in the actual traffic environment, we cooperated with the local traffic management department and selected typical traffic intersections for field testing. In the field testing process, we used a variety of sensors and data acquisition devices, such as cameras, radar, GPS, etc., to collect key data such as traffic flow, vehicle speed, and driving trajectory in real time. The field test data set not only contains real-time traffic information, but also covers the data of different time periods, different weather conditions and different traffic conditions, providing us with a wealth of actual traffic scene data. By comparing the simulation data set with the field test data set, we can more accurately evaluate the performance and applicability of the model, and provide strong data support for the optimization and improvement of the model. The test parameter Settings are shown in Table 4 below.

Table 1
Increment of Traffic Signal Control System.

Increment of transportation system (increasing)	Constant variation	Indicate
Signal lights	Including signal states such as red light, yellow light, green light, etc	Different states correspond to different traffic rules
traffic flow	The number of vehicles or pedestrians passing through a certain road or intersection per unit time	Road traffic efficiency
control strategy	Adjusting the rhythm and duration of signal light changes	Optimal traffic efficiency
adaptive control	Control strategy for autonomously adjusting signal lights	To achieve optimal traffic efficiency

Table 2
Backbone Network and Traffic Feature Fusion Calculation.

1 :	Input: The set of traffic environment states S_A, at time t, the intelligent agent selects the corresponding action a_t state s_t, lane group state $s^{(i)}$, vehicle or pedestrian action $a^{(i)}$, and future information $f^{(i)}$.
2 :	Take action A_t in the current state S_t and transfer to the next state S_{t+1}
3 :	Capture all associated state information in history
4 :	$P=[S_{(t+1)} \mid S_t]=P[S_{(s+1)} \mid S_1,S_2,S_3,S_t]$
5 :	**for all** i =1 to t **do**
6 :	All state transition probabilities are combined to form a state transition matrix
7 :	$P_{ss'}=P[S_{(t+1)}=s' \mid S_t=s]$
8 :	By judging the results obtained after completing the action
9 :	Guiding the direction of intelligent agent learning
10 :	**for** t 1: T
11 :	$G_t=r_t+\gamma r_{t+1}+\gamma^2 r_{t+2}+\ldots$
12 :	**if** γ=0
13 :	The harvest will only be determined through current immediate rewards
14 :	**else**
15 :	There is no difference in the impact of the current two types of rewards on the overall harvest
16 :	**end for**
17 :	**end for**

Table 3
Types of data and perceptron.

Type of perceptron for data and perception	
Data type	Sensor type
flow	Road sensors, cameras, etc
Signal light status	Signal light control system
environment	temperature sensor
Transportation History	Deep learning sensing system

Table 4
Experimental parameter Settings.

Parameter setting name	Specific setting	Remark
Learning rate	0.05	Adjustable
Activation function	Relu	-
Loss function	Paddle	-
Cnov_1	3*3256	-
Cnov_2	3*3256	-
Cnov_3	3*3256	-

4.2. Basic single intersection reward value

In an environment controlled by traffic lights, the reward value is usually related to the smoothness and degree of delay of traffic flow. By designing reward values reasonably, the system can continuously try and adjust the control scheme of traffic lights during the learning process, in order to achieve the optimal traffic flow control effect. The setting of basic single intersection reward values needs to consider multiple factors, such as vehicle speed, delay time, queue length, etc. By quantifying and weighting these factors, an effective feedback mechanism can be provided for reinforcement learning algorithms, guiding the system to continuously learn and optimize traffic light control strategies. When designing basic single intersection reward values, it is also necessary to consider the dynamics and complexity of the transportation system. The changes in traffic flow during different time periods and under different weather conditions will have an impact on the setting of reward values. Therefore, it is necessary to establish a flexible reward value calculation framework to adapt to the signal light control requirements under various traffic conditions. The results are shown in Fig. 5.

When discussing the change of the reward value of different algorithms in single-intersection traffic signal control with the increase of step size, we can observe several remarkable features. First of all, when the step size of Q-learning algorithm gradually increased from 200 steps

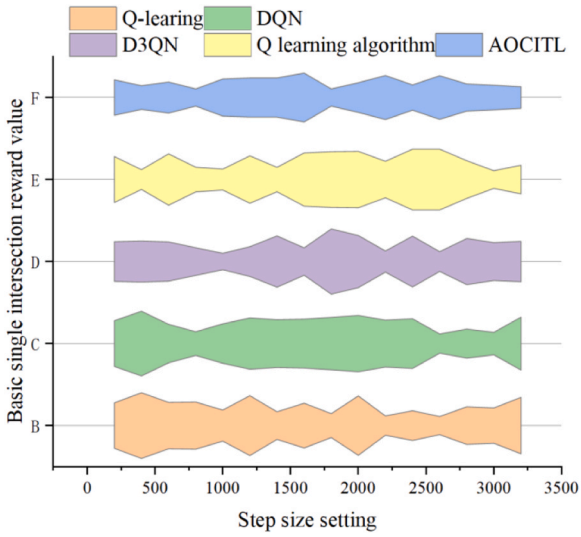

Fig. 5. Partial Display of Reward Values for Basic Single Intersection.

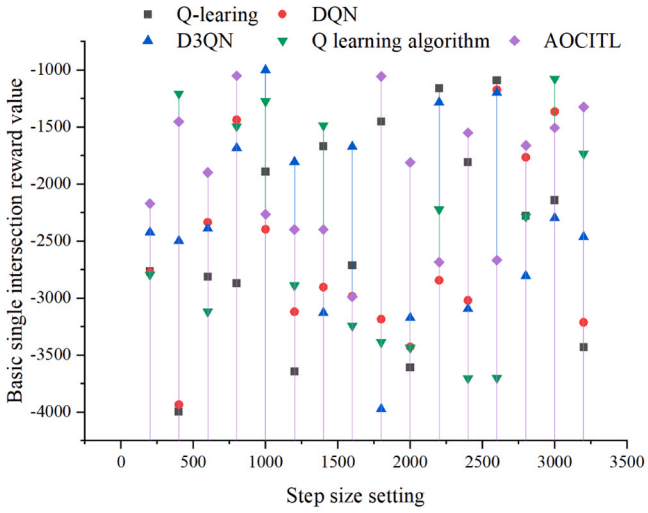

Fig. 6. Basic Single Intersection Reward Value Vertical Content Display.

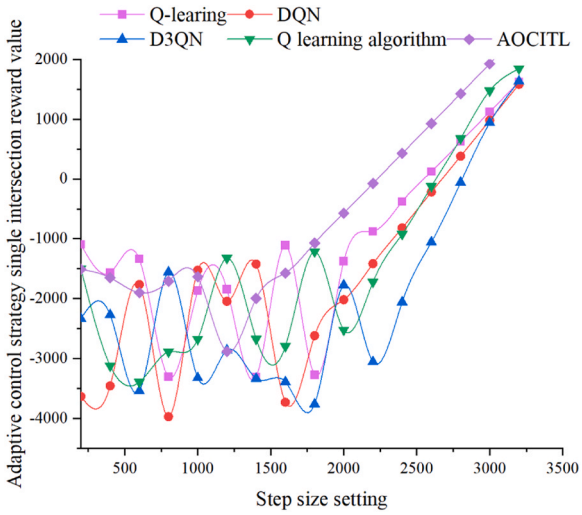

Fig. 7. Adaptive Control Strategy Single Intersection Reward Values.

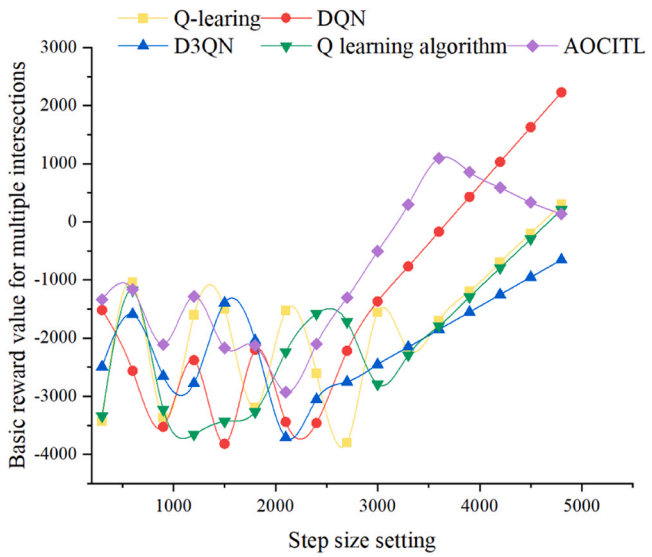

Fig. 8. Basic reward values for multiple intersections.

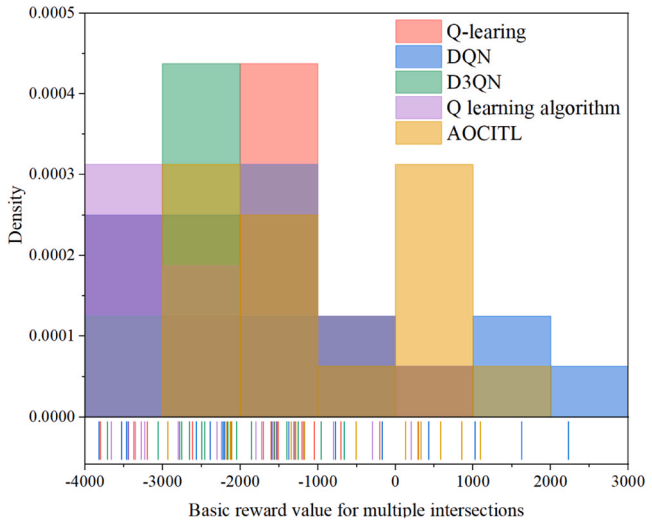

Fig. 9. Results of opening to traffic at multiple intersections.

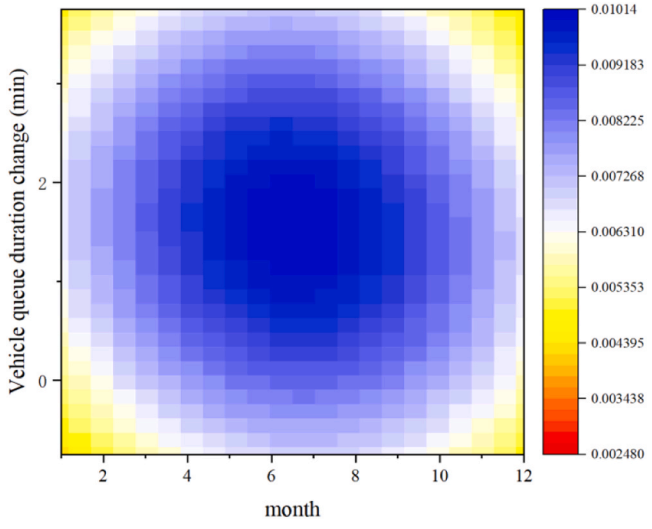

Fig. 10. Changes in vehicle queue duration after adaptive control.

to 3000 steps, its reward value experienced large fluctuations, showing a downward trend at the beginning, but then rebounded slightly. The DQN algorithm also shows a large fluctuation in the similar step size range, first decreasing and then gradually increasing. However, in the process of D3QN algorithm, its reward value fluctuates more violently, and the overall trend is more inclined to decline. Compared with the above algorithms, when the step size of Q learning algorithm increases, the reward value also shows a large fluctuation, but in general, it shows a trend of continuous decline. However, when the AOCITL algorithm increased the step size, the reward value fluctuated relatively smoothly, with only slight changes, and showed a relatively stable reward value at 3000 steps. From the above analysis, it can be seen that different algorithms have different performances in single-intersection traffic signal control, and AOCITL algorithm has better performance in the stability of reward value, which is more in line with the expected control effect. Of course, in practical applications, we also need to choose the appropriate algorithm according to the specific application scenarios and needs to achieve the best control effect.

4.3. Adaptive control strategy single intersection reward value

In the reinforcement learning based adaptive optimal control model for intelligent transportation system signal lights, the adaptive control strategy is crucial. The design of reward values for single intersections is one of the key considerations. In this scenario, the setting of reward values needs to comprehensively consider factors such as traffic flow, vehicle delay time, energy consumption, and vehicle emissions. For traffic flow, the reward value can be set to reduce vehicle waiting time or maximize vehicle traffic efficiency. By using reinforcement learning algorithms to learn the optimal control strategy, traffic lights can be dynamically adjusted according to real-time traffic conditions, effectively reducing traffic congestion and improving road traffic efficiency.

By observing the above data, it can be seen that there are certain fluctuations and differences in the reward value performance of different algorithms in adaptive traffic light control strategies. When the step size is 200, the Q-learning algorithm and DQN algorithm perform relatively well, while the AOCITL algorithm performs weakly. As the step size increases, some algorithms exhibit significant fluctuations in their reward values, such as the D3QN algorithm and AOCITL algorithm. At certain step sizes, such as 1600 and 2200, the D3QN algorithm performs relatively well. At a step size of 2800, the reward value of the Q-learning algorithm significantly decreases, while the reward value of the AOCITL algorithm also changes significantly. The performance of different algorithms in adaptive traffic light control strategies will be affected by step size settings, and it is necessary to choose appropriate algorithms and adjust parameters according to specific situations to achieve better traffic control effects and reward values.

4.4. Basic reward value for multiple intersections

The basic reward value can be determined based on various factors, such as the average speed of the vehicle, the waiting time of the vehicle, the traffic congestion index, etc. By comprehensive evaluation of these factors, the basic reward value of each signal control period can be calculated, which is used as the feedback signal of reinforcement learning algorithm and guides the system to adjust the control strategy of the signal. Through the data collection and analysis of the actual traffic scene, the calculation method and influencing factors of the basic reward value of multiple intersections are determined. By optimizing the reinforcement learning model, the signal control system can dynamically adjust the time interval and phase of the signal light according to the real-time traffic situation, so as to achieve the optimal control of traffic flow.

This paper discusses the influence of Q-learning, DQN, D3QN and AOCITL algorithms on the control reward value of multi-intersection traffic lights under the asynchronous length setting. Q-learning

algorithm is a table-based reinforcement learning algorithm that directly stores and updates the Q value of each state-action pair. In multi-junction traffic light control, the reward value fluctuates greatly with the increase of step length. In the process of gradually increasing the step size from a smaller value, the reward value shows a clear downward trend in the step size range of 300–1800. This may be because the large step size causes the algorithm to over-trust the new information, resulting in unstable updates, affecting the overall performance. DQN algorithm combines the feature extraction ability of deep learning with the decision making ability of Q-learning. In multi-junction traffic light control, the fluctuation range of reward value of DQN algorithm is smaller than that of Q-learning algorithm. This may be because the DQN algorithm approximates the Q function through a neural network, which is better able to handle high-dimensional state Spaces and reduce fluctuations due to table updates.

D3QN algorithm is an improvement of DQN algorithm, aiming to solve the overestimation problem in DQN algorithm. However, in multi-intersection traffic light control, the reward value of D3QN algorithm fluctuates sharply, and the overall trend is not obvious. This may be because the D3QN algorithm not only reduces overestimation, but also introduces some randomness, resulting in large fluctuations in the performance of the algorithm under non-synchronous length. Under some steps, the reward value of D3QN algorithm is relatively stable, but the overall level is average. AOCITL algorithm is a combination of adaptive online control and reinforcement learning algorithm. In multi-intersection traffic light control, the AOCITL algorithm has a large fluctuation range of reward value under non-synchronous length, and the overall level is relatively low. This may be because the AOCITL algorithm requires a larger step size to quickly update its strategy when adapting to different traffic conditions. However, excessive step size can also lead to instability of the algorithm. At some steps, the reward value of AOCITL algorithm increases, but the overall performance is mediocre. However, the AOCITL algorithm has a wider applicability because it can adapt its step size to traffic conditions.

4.5. Changes in vehicle queue duration after adaptive control at complex intersections

In traditional traffic signal control systems, fixed timing control is often difficult to adapt to changes in traffic flow, leading to traffic congestion and an increase in vehicle queuing time. With the help of reinforcement learning algorithms, traffic signals can dynamically adjust their timing and cycle based on real-time data, enabling the transportation system to better adapt to current traffic demands. Through continuous learning and optimization, traffic lights can make the best decisions in a short period of time, thereby reducing vehicle queuing time and improving road traffic efficiency. The adaptive optimal control model for traffic signals in intelligent transportation systems based on reinforcement learning can effectively optimize the control strategy of traffic signals and reduce the queuing time of vehicles. The results are shown in the following figure.

In the initial first month, the D3QN algorithm achieved remarkable results in reducing the queuing time of vehicles, successfully reducing the queuing time by 3 minutes, while other algorithms were mediocre. In the second month, the DQN algorithm stood out, and its optimization effect reduced the queue time by 2 minutes, which was slightly inferior to other algorithms. Over the next few months, the performance of the various algorithms has fluctuated. In the third month, Q-learning algorithm and AOCITL algorithm both showed good performance, reducing the queue time by 3 minutes. In the fourth month, the DQN algorithm and AOCITL algorithm further improved the effect, reducing the queue time by 4 minutes. In the fifth month, the AOCITL algorithm achieved the best results in reducing the queue time, successfully reducing the queue time by 4 minutes, while other algorithms failed to achieve such an effect. In the next six months, Q-learning and D3QN also performed well, reducing the queue time by 2 minutes. In the second half of the year, AOCITL algorithm continued to show its superiority. In both the seventh and eighth months, it managed to reduce the queue time by four minutes, and in the ninth month, this achievement was maintained. By the tenth month, DQN algorithm and Q learning algorithm also achieved good results, reducing the queue time by 3 minutes. At the eleventh month, the D3QN algorithm regained the lead, reducing the queue time by 2 minutes. In the last month, the 12th month, both the DQN algorithm and the D3QN algorithm showed good performance, reducing the queue time by 2 minutes. Combining the above data, we can find that AOCITL algorithm has the best performance in reducing vehicle queuing time, followed by DQN algorithm. The two algorithms have achieved significant optimization effects in different months, and have made positive contributions to easing traffic congestion.

5. Discussion

In the field of intelligent transportation Systems (ITS), the adaptive optimal control of signal lights has been a hot research topic. With the rapid development of reinforcement learning (RL) technology, more and more researchers begin to explore the application of RL in traffic signal control to achieve more efficient and intelligent traffic management. The adaptive optimal control model of intelligent transportation system signal light based on reinforcement learning has achieved remarkable results in both simulation and field tests. However, while exploring this model in depth, we also noted important factors such as the limitations of other relevant studies, adaptability to different flow conditions, and potential integration challenges.

In the field of traffic signal control, although the traditional fixed time control method is simple and easy, it can not adapt to the dynamic change of traffic flow. However, rule-based adaptive control methods, such as fuzzy control and genetic algorithm, can adjust signal timing according to traffic flow to a certain extent, but their decision-making process often relies on predefined rules and thresholds, which lacks flexibility and adaptability.

In contrast, control methods based on reinforcement learning are able to learn from historical data and real-time traffic conditions and continuously optimize control strategies. However, the existing research on traffic signal control based on reinforcement learning still has some limitations. For example, some studies only focus on the signal control of a single intersection, but ignore the collaborative optimization of multiple intersections. Other studies are limited by computational resources and the difficulty of obtaining data, and cannot be applied in actual large-scale transportation networks.

The model presented in this study shows good adaptability under different flow conditions. Through simulation experiments and field tests, we verify the control effect of the model in different scenarios such as peak hours, off-peak hours and sudden traffic incidents. The experimental results show that the model can dynamically adjust the signal timing according to the real-time change of traffic flow, effectively alleviate traffic congestion and improve road traffic efficiency. However, we also note that under certain extreme traffic conditions, such as extreme congestion or extreme sparsity, the performance of the model may suffer somewhat. Therefore, in future studies, we will further explore how to improve the adaptability and robustness of the model under extreme flow conditions.

Integrating reinforcement learning-based traffic signal control models into existing intelligent transportation systems faces some potential challenges. First of all, the existing intelligent transportation systems often adopt a variety of technologies and methods, such as video surveillance, traffic flow detection, road state assessment, etc., how to effectively integrate reinforcement learning models with other technologies to achieve information sharing and collaborative work is a problem that needs to be solved. Because the traffic system is a complex network structure, it involves the connection of multiple intersections and roads. It is a challenging task to realize the cooperative optimization and global optimal control of the entire traffic network while ensuring

the optimal control of a single intersection.

The adaptive optimal control model of intelligent transportation system signal light based on reinforcement learning has achieved remarkable results in both simulation and field tests. However, we also note important factors such as the limitations of other relevant studies, adaptability to different flow conditions, and potential integration challenges. In future studies, we will further explore these issues in depth to promote the continued development and application of intelligent transportation systems.

6. Conclusion

The aim of this study is to explore a reinforcement learning based adaptive optimal control model for traffic signals in intelligent transportation systems, in order to improve the efficiency and intelligence level of urban traffic signal systems. The adaptive optimal control model for traffic signal lights in intelligent transportation systems based on reinforcement learning has shown significant advantages in improving the efficiency of traffic signal light systems. The traditional fixed time sequence signal control method often cannot adapt to changes in traffic flow, which can easily lead to traffic congestion and energy waste. The reinforcement learning based model proposed in this study can dynamically adjust traffic flow data in real-time, making the control of traffic lights more flexible and efficient, effectively alleviating traffic congestion and improving road traffic efficiency.

This study also found that the adaptive optimal control model for traffic lights in intelligent transportation systems based on reinforcement learning has good adaptability and generalization ability. Through experimental verification of models under different traffic scenarios and road conditions, we found that the model can achieve good control effects in different environments, demonstrating strong adaptability and generalization ability. This lays a solid foundation for the application of the model in actual transportation systems.

The research model may be limited by computational resources and time when dealing with large-scale transportation systems, and further optimization of algorithms and improvement of computational efficiency are needed; In addition, the robustness and security of the model also need to be strengthened to cope with various unexpected situations and malicious attacks.

In summary, the adaptive optimal control model for traffic lights in intelligent transportation systems based on reinforcement learning has significant advantages and application potential, and is of great significance in improving transportation system efficiency, reducing traffic congestion, and saving energy. However, in order to further promote the application of this model in actual transportation systems, we still need to conduct further in-depth research and improvement, continuously improve the theoretical basis and practical application effects of the model.

CRediT authorship contribution statement

Zhongyi Huang: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of Competing Interest

The authors declared that they have no conflicts of interest to this work.

References

[1] K. Bálint, T. Tamás, B. Tamás, Deep reinforcement learning based approach for traffic signal control, Transp. Res. Procedia 62 (2022) 278–285.
[2] H. Wang, W. Li, Z. Zhao, Z. Wang, M. Li, D. Li, Intelligent Distribution of Fresh Agricultural Products in Smart City], IEEE Trans. Ind. Inform. 18 (2) (2021 ,) 1220–1230.
[3] X. Wang, B. Abdulhai, S. Sanner, A critical review of traffic signal control and a novel unified view of reinforcement learning and model predictive control approaches for adaptive traffic signal control, Handb. Artif. Intell. Transp. (2023) 482–532.
[4] V.G. Stepanyants, A.Y. Romanov, A survey of integrated simulation environments for connected automated vehicles: requirements, tools, and architecture, IEEE Intell. Transp. Syst. Mag. (2023).
[5] L. Koch, T. Brinkmann, M. Wegener, et al., Adaptive traffic light control with deep reinforcement learning: an evaluation of traffic flow and energy consumption, IEEE Trans. Intell. Transp. Syst. (2023).
[6] S.M.A. Shabestary, B. Abdulhai, Adaptive traffic signal control with deep reinforcement learning and high dimensional sensory inputs: case study and comprehensive sensitivity analyses, IEEE Trans. Intell. Transp. Syst. 23 (11) (2022) 20021–20035.
[7] R. Kumar, N.V.K. Sharma, V.K. Chaurasiya, Adaptive traffic light control using deep reinforcement learning technique, Multimed. Tools Appl. (2023) 1–22.
[8] D. Li, J. Wu, M. Xu, et al., Adaptive traffic signal control model on intersections based on deep reinforcement learning, J. Adv. Transp. 2020 (2020) 1–14.
[9] A. Haydari, Y. Yılmaz, Deep reinforcement learning for intelligent transportation systems: a survey, IEEE Trans. Intell. Transp. Syst. 23 (1) (2020) 11–32.
[10] Baumgart U., Burger M. Optimal Control of Traffic Flow Based on Reinforcement Learning[C]//International Conference on Vehicle Technology and Intelligent Transport Systems. Cham: Springer International Publishing, 2021: 313-329.
[11] Ge Z. Reinforcement learning-based signal control strategies to improve travel efficiency at urban intersection[C]//2020 International Conference on Urban Engineering and Management Science (ICUEMS). IEEE, 2020: 347-351.
[12] L. Kuang, J. Zheng, K. Li, et al., Intelligent traffic signal control based on reinforcement learning with state reduction for smart cities, ACM Trans. Internet Technol. (TOIT) 21 (4) (2021) 1–24.
[13] R. Zhang, A. Ishikawa, W. Wang, et al., Using reinforcement learning with partial vehicle detection for intelligent traffic signal control, IEEE Trans. Intell. Transp. Syst. 22 (1) (2020) 404–415.
[14] H. Wang, Y. Yuan, X.T. Yang, et al., Deep Q learning-based traffic signal control algorithms: Model development and evaluation with field data, J. Intell. Transp. Syst. 27 (3) (2023) 314–334.
[15] D. Ma, B. Zhou, X. Song, et al., A deep reinforcement learning approach to traffic signal control with temporal traffic pattern mining, IEEE Trans. Intell. Transp. Syst. 23 (8) (2021) 11789–11800.
[16] Du Y., ShangGuan W., Rong D., et al. RA-TSC: Learning adaptive traffic signal control strategy via deep reinforcement learning[C]//2019 ieee intelligent transportation systems conference (itsc). IEEE, 2019: 3275-3280.
[17] N. Kumar, S. Mittal, V. Garg, et al., Deep reinforcement learning-based traffic light scheduling framework for sdn-enabled smart transportation system, IEEE Trans. Intell. Transp. Syst. 23 (3) (2021) 2411–2421.
[18] M.A. Khamis, W. Gomaa, Adaptive multi-objective reinforcement learning with hybrid exploration for traffic signal control based on cooperative multi-agent framework, Eng. Appl. Artif. Intell. 29 (2014) 134–151.
[19] Miletić M., Kušić K., Gregurić M., et al. State complexity reduction in reinforcement learning based adaptive traffic signal control[C]//2020 International Symposium ELMAR. IEEE, 2020: 61-66.
[20] J. Hurtado-Gomez, J.D. Romo, R. Salazar-Cabrera, et al., Traffic signal control system based on intelligent transportation system and reinforcement learning, Electronics 10 (19) (2021) 2363.
[21] S. El-Tantawy, B. Abdulhai, H. Abdelgawad, Design of reinforcement learning parameters for seamless application of adaptive traffic signal control, J. Intell. Transp. Syst. 18 (3) (2014) 227–245.
[22] Chen P., Zhu Z., Lu G. An adaptive control method for arterial signal coordination based on deep reinforcement learning[C]//2019 IEEE Intelligent Transportation Systems Conference (ITSC). IEEE, 2019: 3553-3558.
[23] Z.E. Liu, Q. Zhou, Y. Li, et al., Safe deep reinforcement learning-based constrained optimal control scheme for HEV energy management, IEEE Trans. Transp. Electrification (2023).
[24] C.H. Wan, M.C. Hwang, Adaptive traffic signal control methods based on deep reinforcement learning, Intell. Transp. Syst. Everyone'S. Mobil. (2019) 195–209.
[25] J.S. Yang, S.J. Park, A neural network approach for adaptive control: application to traffic signal control, J. Intell. Fuzzy Syst. 2 (2) (1994) 115–123.
[26] M. Yazdani, M. Sarvi, S.A. Bagloee, et al., Intelligent vehicle pedestrian light (IVPL): a deep reinforcement learning approach for traffic signal control, J.]. Transp. Res. Part C: Emerg. Technol. 149 (2023) 103991.
[27] A.A. Agafonov, A.S. Yumaganov, V.V. Myasnikov, Adaptive traffic signal control based on neural network prediction of weighted traffic flow, Optoelectron., Instrum. Data Process. 58 (5) (2022) 503–513.
[28] Liu X.Y., Zhu M., Borst S., et al. Deep Reinforcement Learning for Traffic Light Control in Intelligent Transportation Systems[J]. arxiv preprint arxiv:2302.03669, 2023.

[29] A. Boukerche, D. Zhong, P. Sun, A novel reinforcement learning-based cooperative traffic signal system through max-pressure control, IEEE Trans. Veh. Technol. 71 (2) (2021) 1187–1198.
[30] X. Ma, A multi-objective agent-based control approach with application in intelligent traffic signal system, IEEE Trans. Intell. Transp. Syst. 20 (10) (2019) 3900–3912.
[31] M. Wang, L. Wu, J. Li, et al., Traffic signal control with reinforcement learning based on region-aware cooperative strategy, IEEE Trans. Intell. Transp. Syst. 23 (7) (2021) 6774–6785.