Recent Advances in Reinforcement Learning for Traffic Signal Control: A Survey of Models and Evaluation

Hua Wei, Guanjie Zheng
College of Information Sciences and Technology
Penn State University
{hzw77,gjz5038}@ist.psu.edu

Vikash Gayah
Department of Civil Engineering
Penn State University
gayah@engr.psu.edu

Zhenhui Li
College of Information Sciences and Technology
Penn State University
jessieli@ist.psu.edu

ABSTRACT

Traffic signal control is an important and challenging real-world problem that has recently received a large amount of interest from both transportation and computer science communities. In this survey, we focus on investigating the recent advances in using reinforcement learning (RL) techniques to solve the traffic signal control problem. We classify the known approaches based on the RL techniques they use and provide a review of existing models with analysis on their advantages and disadvantages. Moreover, we give an overview of the simulation environments and experimental settings that have been developed to evaluate the traffic signal control methods. Finally, we explore future directions in the area of RL-based traffic signal control methods. We hope this survey could provide insights to researchers dealing with real-world applications in intelligent transportation systems.

1. INTRODUCTION

Traffic congestion is a growing problem that continues to plague urban areas with negative outcomes to both the traveling public and society as a whole. These negative outcomes will only grow over time as more people flock to urban areas. In 2014, traffic congestion costs Americans over $166 billion in lost productivity and wasted over 3.1 billion gallons of fuel [15]. Traffic congestion was also attributed to over 56 billion pounds of harmful CO_2 emissions in 2011 [54]. Mitigating congestion would have significant economic, environmental, and societal benefits. Signalized intersections are one of the most prevalent bottleneck types in urban environments, and thus traffic signal control plays a vital role in urban traffic management.

The typical approach that transportation researchers take is to cast traffic signal control as an optimization problem under certain assumptions about the traffic model, e.g., vehicles come in a uniform and constant rate [52]. Various assumptions have to be made in order to make the optimization problem tractable. These assumptions, however, usually deviate from the real world, where the traffic condition is affected by many factors such as driver's preference, interactions with vulnerable road users (e.g., pedestrians, cyclists, etc.), weather and road conditions. These factors can hardly be fully described in a traffic model. For a more comprehensive survey for the methods in transportation, we refer the interested readers to [47; 52; 40; 31; 17; 50; 64].

On the other hand, reinforcement learning methods can directly learn from the observed data without making unrealistic assumptions about the traffic model, by first taking actions to change the signal plans and then learning from the outcomes. In essence, an RL-based traffic signal control system observes the traffic condition first, then generates and executes different actions (i.e., traffic signal plans). It will then learn and adjust the strategies based on the feedback from the environment. However, in tradictional RL-based methods, the states in an environment are required discretized and low-dimensional, which is one of the major limitations of the traditional approaches.

Recent advances in RL, especially deep RL, offer the opportunity to efficiently work with high dimensional input data (like images), where the agent can learn a state abstraction and a policy approximation directly from its input states. A series of related studies using deep RL for traffic signal control have appeared in the past few years. This survey is to provide an overview on the recent RL-based traffic signal control approaches, including the state-of-the-art methods and their experimental settings for evaluation.

In this survey, we first introduce the formulation of traffic light control problems under RL, and then classify and discuss the current RL control methods from different aspects: agent formulation, policy learning approach, and coordination strategy when facing multiple intersections. In the third section, we review how current methods are evaluated, including simulators and experimental settings that affect the performance of these methods. We then discuss some future research directions. While [39; 71] provide surveys mainly on earlier studies before the popularity of deep RL, in this survey, we will mainly cover the recent deep RL methods. With the increasing interest on RL-based control mechanisms in intelligent transportation systems [24], such as autonomous driving [67] and road control [46; 68], we hope this survey could also provide insights on dealing with real-world challenges for other applications in intelligent transportation systems.

2. BACKGROUND

In this section, we first describe the reinforcement learning framework which constitutes the foundation of all the methods presented in this paper. We then provide background on conventional RL-based traffic signal control, including the problem of controlling a single intersection and multiple intersections.

2.1 Reinforcement learning

Usually a single agent RL problem is modeled as a Markov Decision Process represented by $\langle \mathcal{S}, \mathcal{A}, P, R, \gamma \rangle$, where their definitions are given as follows:

• Set of state representations $\mathcal{S}$: At time step t, the agent observes state $\mathbf{s}^t \in \mathcal{S}$.

• Set of action $\mathcal{A}$ and state transition function P: At time step t, the agent takes an action $\mathsf{a}^t \in \mathcal{A}$, which induces a transition in the environment according to the state transition function $P(\mathsf{s}^{t+1}|\mathsf{s}^t, \mathsf{a}^t) : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$

• Reward function R: At time step t, the agent obtains a reward r^t by a reward function: $R(\mathsf{s}^t, \mathsf{a}^t) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$

• Discount factor γ: The goal of an agent is to find a policy that maximizes the expected return, which is the discounted sum of rewards: $G^t := \sum_{i=0}^{\infty} \gamma^i \mathsf{r}^{t+i}$, where the discount factor $\gamma \in [0, 1]$ controls the importance of immediate rewards versus future rewards. Here, we only consider continuing agent-environment intersections which do not end with terminal states but goes on continually without limit.

Solving a reinforcement learning task means, roughly, finding an optimal policy π^* that maximizes expected return. While the agent only receives reward about its immediate, one-step performance, one way to find the optimal policy π^* is by following an optimal *action-value function* or *state-value function*. The action-value function (Q-function) of a policy π, $Q^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, is the expected return of a state-action pair $Q^\pi(\mathsf{s}, \mathsf{a}) = \mathbb{E}_\pi[G^t|\mathsf{s}^t = \mathsf{s}, \mathsf{a}^t = \mathsf{a}]$. The state-value function of a policy π, $V^\pi : \mathcal{S} \to \mathbb{R}$, is the expected return of a state $V^\pi(\mathsf{s}) = \mathbb{E}_\pi[G^t|\mathsf{s}^t = \mathsf{s}]$.

2.2 Problem setting

We now introduce the general setting of RL-based traffic signal control problem, in which the traffic signals are controlled by an RL agent or several RL agents. In *single traffic signal control problem*, the environment is the traffic conditions on the roads, and the agent controls the traffic signal. At each time step t, a description of the environment (e.g., signal phase, waiting time of cars, queue length of cars, and positions of cars) will be generated as the state s_t. The agent will predict the next action a^t to take that maximizes the expected return, where the action could be changing to a certain phase in the single intersection scenario. The action a^t will be executed in the environment, and a reward r^t will be generated, where the reward could be defined on traffic conditions of the intersection. Usually, in the decision process, an agent combines the exploitation of learned policy and exploration of a new policy.

In *multi-intersection traffic signal control problem*, there are N traffic signals in the environment, controlled by one or several agents. The goal of the agent(s) is to learn the optimal policies to optimize the traffic condition of the whole environment. At each timestep t, each agent i observes part of the environment as the observation o_i^t and make predictions on the next actions $\mathsf{a}^t = (\mathsf{a}_1^t, \ldots, \mathsf{a}_N^t)$ to take. The actions will be executed in the environment, and the reward r_i^t will be generated, where the reward could be defined on the level of individual intersections or a group of intersections within the environment. We refer readers interested in more detailed problem settings to [8].

3. RL-BASED TRAFFIC SIGNAL CONTROL

In this section, we introduce three major aspects investigated in recent RL-based traffic signal control literature: agent formulation, policy learning approach and coordination strategy.

3.1 Opportunities

In this subsection, we point out some high-level discussions about why RL and deep RL are appropriate for the traffic signal control problem.

Reinforcement learning methods learns from trail-and-error without making unrealistic assumptions on traffic model. The typical approach that conventional transportation methods take is to cast traffic signal control as an optimization problem under certain assumptions about the traffic model. For example, Webster's Formula method [28] is one of the widely-used method in field for a single intersection. Assuming the traffic flow is uniform during a certain period (i.e., past 5 or 10 minutes), it has a closed-form solution after optimization [52]. Other methods like Maxband [35] or SCATS [38] also make similar assumptions to make the optimization problem tractable. The key issue here is that these assumptions often deviate from the real world. The real-world traffic condition evolves in a complicated way, affected by many factors such as driver's preference, interactions with vulnerable road users (e.g., pedestrians, cyclists, etc.), weather and road conditions. These factors can hardly be fully described in a traffic model. On the other hand, reinforcement learning techniques can directly learn from the observed data without making unrealistic assumptions about the model. In essence, an RL system generates and executes different strategies (e.g., for traffic signal control) based on the current environment. It will then learn and adjust the strategy based on the feedback from the environment. This reveals the most significant difference between transportation approaches and our RL approaches: in traditional transportation research, the control model is static; in reinforcement learning, the control model is dynamically learned through trial-and-error in the real environment.

The combination of deep learning with reinforcement learning helps alleviate the "curse of dimensionality" problem. Traditionally, RL is concerned with the issue of the curse of dimensionality as the number of state-action pairs can grows exponentialy with the dimension of states and actions. Recent advances in deep learning helps the approximation of functions in RL like $Q(\mathsf{s}, \mathsf{a})$ or $V(\mathsf{s})$ by learning efficiently on a significantly smaller number of features instead of a large number of state-action pairs. This helps to improve scalability with reduced requirements on memory or storage capacity, as well as reduced learning time.

3.2 Agent formulation

A key question for RL is how to formulate the RL agent, i.e., the reward, state, and action definition. In this subsection, we focus on the advances in the reward, state, and action design in recent deep RL-based methods, and refer readers interested in more detailed definitions to [16; 39; 71].

3.2.1 Reward

The choice of reward reflects the learning objective of an RL agent. In the traffic signal control problem, although the ultimate objective is to minimize the travel time of all vehicles, travel time is hard to serve as a valid reward in RL. Because the travel time of a vehicle is affected by multiple actions from traffic signals and vehicle movements, the travel time as reward would be delayed and ineffective in indicating the goodness of the signals' action. Therefore, the existing literature often uses a surrogate reward that can be effectively measured after an action, considering factors like average queue length, average waiting time, average speed or throughput [55; 65]. The authors in [58] also take the frequency of signal changing and the number of emergency stops into reward. With different reward functions being proposed, researchers in [78; 62] find out that the weight on each factor in reward is tricky to set, and a minor difference in weight setting could lead to dramatically different results. Thus they set out to find a minimal set of factors, proving that using queue length as the reward for a single intersection scenario and using pressure, a variant of queue length in the multi-intersection scenario are equivalent to optimizing the global travel time.

3.2.2 State

At each time step, the agent receives some quantitative descriptions

of the environment as state to decide its action. Various kinds of elements have been proposed to describe the environment state, such as queue length, waiting time, speed and phase, etc. These elements can be defined on the lane level or road segment level, and then concatenated as a vector. In earlier work using RL for traffic signal control, people need to discretize the state space and use a simple tabular or linear model to approximate the state functions for efficiency [1; 7; 49]. However, the real-world state space is usually huge, which confines the traditional RL methods in terms of memory or performance. With advances in deep learning, deep RL methods are proposed to handle large state space as an effective function approximator. Recent studies propose to use images [9; 14; 18; 20; 22; 23; 32; 33; 42; 58; 65] to represent the state, where the position of vehicles are extracted as an image representation. With variant information used in state representation in different studies, [62; 78] shows that complex state definition and large state space do not necessarily lead to significant performance gain, and proposes to use simple state like lane-level queue length and phase to represent the environment state.

3.2.3 Action scheme

Now there are different types of action definitions for an RL agent in traffic signal control: *(1)* set current phase duration [4; 5], *(2)* set the ratio of the phase duration over pre-defined total cycle duration [1; 10], *(3)* change to the next phase in pre-defined cyclic phase sequence [39; 48; 58; 65], and *(4)* choose the phase to change to among a set of phases [2; 9; 12; 44; 42; 78]. The choice of action scheme is closely related to specific settings of traffic signals. For example, if the phase sequence is required to be cyclic, then the first three action schemes should be considered, while "choosing the phase to change to among a set of phases" can generate flexible phase sequences.

3.3 Policy learning

RL methods can be categorized in different ways. [3; 26] divide current RL methods to model-based methods and model-free methods. Model-based methods try to model the transition probability among states explicitly, while model-free methods directly estimate the reward for state-action pairs and choose the action based on this. In the context of traffic signal control, the state transition between states is primarily influenced by people's driving behaviors, which are diverse and hard to predict. Therefore, currently, most RL-based methods for traffic signal control are model-free methods. In this subsection, we take the categorization in [43]: value-based methods and policy-based methods.

3.3.1 Value-based methods

Value-based methods approximate the state-value function or state-action value function (i.e., how rewarding each state is or state-action pair is), and the policy is implicitly obtained from the learned value function. Most of the RL-based traffic signal control methods use DQN [41], where the model is parameterized by neural networks and takes the state representation as input [58; 30]. In DQN, discrete actions are required as the model directly outputs the action's value given a state, which is especially suitable for action schema (3) and (4) mentioned in Section *3.2.3*.

3.3.2 Policy-based methods

Policy-based methods directly update the policy parameters (e.g., a vector of probabilities to conduct actions under specific state) towards the direction to maximizing a predefined objective (e.g., average expected return). The advantage of policy-based methods is that it does not require the action to be discrete like DQN. Also, it can learn a stochastic policy and keep exploring potentially more rewarding actions. To stabilize the training process, the actor-critic framework is widely adopted. It utilizes the strengths of both value-based and policy-based methods, with an actor controls how the agent behaves (policy-based), and the critic measures how good the conducted action is (value-based). In the traffic signal control problem, [10] uses DDPG [34] to learn a deterministic policy which directly maps states to actions, while [4; 42; 69] learn a stochastic policy that maps states to action probability distribution, all of which have shown excellent performance in traffic signal control problems. To further improve convergence speed for RL agents, [51] proposed a time-dependent baseline to reduce the variance of policy gradient updates to specifically avoid traffic jams.

In the above-mentioned methods, including both value-based and policy-based methods, deep neural networks are used to approximate the value functions. Most of the literature use vanilla neural networks with their corresponding strengths. For example, Convolutional Neural Networks (CNN) are used since the state representation contains image representation [9; 20; 22; 23; 32; 33; 42; 58]; Recurrent Neural Networks (RNN) are used to capture the temporal dependency of historical states [60]. Special neural network structures are also proposed to incorporate prior knowledge about the states into the learning process [65; 77].

3.4 Coordination

Coordination could benefit signal control for multi-intersection scenarios. Since recent advances in RL improve the performance on isolated traffic signal control, efforts have been performed to design strategies that cooperate with multi-agent reinforcement learning (MARL) agents. Literature [13] categorizes MARL into two classes: Joint action learners and independent learners. Here we extend this categorization for the traffic signal control problem.

3.4.1 Joint action learners

A straightforward solution is to use a *single global agent* to control all the intersections [49]. It directly takes the state as input and learns to set the joint actions of all intersections at the same time. However, these methods can result in the curse of dimensionality, which encompasses the exponential growth of the state-action space in the number of state and action dimensions. *Joint action modeling* methods explicitly learns to model the joint action value of multiple agents $Q(o_1, \ldots, o_N, \mathbf{a})$. The joint action space grows with the increase in the number of agents to model. To alleviate this challenge, [58] factorizes the global Q-function as a linear combination of local subproblems, extending [66] using max-plus [27] algorithm: $\hat{Q}(o_1, \ldots, o_N, \mathbf{a}) = \Sigma_{i,j} Q_{i,j}(o_i, o_j, \mathsf{a}_i, \mathsf{a}_j)$, where i and j correspond to the index of neighboring agents. In other works, [74; 12; 57] regard the joint Q-value as a weighted sum of local Q-values, $\hat{Q}(o_1, \ldots, o_N, \mathbf{a}) = \Sigma_{i,j} w_{i,j} Q_{i,j}(o_i, o_j, \mathsf{a}_i, \mathsf{a}_j)$, where $w_{i,j}$ is the pre-defined weights. They attempt to ensure individual agents to consider other agents' learning process by adding a shaping term in the loss function of the individual agent's learning process and minimizing the difference between the weighted sum of individual Q-values and the global Q-value.

3.4.2 Independent learners

There is also a line of studies that use independent RL (IRL) agents to control the traffic signals, where each RL agent controls an intersection. Unlike joint action learning methods, each agent learns its control policy without knowing the reward signal of other agents.

IRL without communication methods treat each intersection individually, with each agent observing its own local environment and do not use explicit communication to resolve conflicts [39; 10; 78;

Table 1: Representative deep RL-based traffic signal control methods.

Citation	Method	Cooperation	Simulator	Road net (# signals)	Traffic flow*
[2]	Value-based	With communication	Matlab	Synthetic (5)	2,4
[5]	Policy-based	Without communication	Aimsun	Real (50)	5
[10]	Policy-based	Without communication	Aimsun	Real (43)	5
[11]	Value-based	Without communication	CityFlow	Real (2510)	5
[12]	Policy-based	Joint action	SUMO	Real (30)	4
[22]	Value-based	-	SUMO	Synthetic (1)	2
[30]	Value-based	-	Paramics	Synthetic (1)	4
[39]	Value-based	Without communication	SUMO	Synthetic (9)	2
[42]	Both studied	-	SUMO	Synthetic (1)	1
[44]	Value-based	With communication	SUMO	Synthetic (6)	2
[48]	Value-based	Without communication	AIM	Synthetic (4)	1
[49]	Both studied	Single global	GLD	Sythetic (5)	3
[51]	Policy-based	-	SUMO	Real (1)	5
[58]	Value-based	Joint action	SUMO	Synthetic (4)	2
[60]	Value-based	With communication	SUMO	Real (4)	5
[65]	Value-based	-	SUMO	Synthetic (1)	1,3,4,5
[62]	Value-based	Without communication	CityFlow	Real (16)	2,5
[63]	Value-based	With communication	CityFlow	Real (196)	2,5
[74]	Value-based	Joint action	SUMO	Synthetic (36)	1,2,3,4
[78]	Value-based	Without communication	CityFlow	Real (16)	3,5
[77]	Value-based	Without communication	CityFlow	Real (5)	4,5

* Traffic with arrival rate less than 500 vehicles/hour/lane is considered as light traffic in this survey, otherwise considered as heavy.
1. Synthetic light uniform; 2. Synthetic light dynamic; 3. Synthetic heavy uniform; 4. Synthetic heavy dynamic; 5. Real-world data

48; 36; 9; 23]. In some simple scenarios like arterial networks, this approach has performed well with the formation of several mini green waves. However, when the environment becomes complicated, the non-stationary impacts from neighboring agents will be brought into the environment, and the learning process usually cannot converge to stationary policies if there are no communication or coordination mechanisms among agents [45]. To deal with this challenge, the authors in [62] propose a specified reward that describes the demand for coordination between neighbors to achieve coordination.

IRL with communication methods enable agents to communicate between agents about their observations and behave as a group, rather than a collection of individuals in complex tasks where the environment is dynamic, and each agent has limited capabilities and visibility of the world [56]. Typical methods directly add neighbor's traffic condition [70] or past actions [21] into the observation of the ego agent, other than just using the local traffic condition of the ego agent. In this method, all the agents for different intersection share one learning model, which requires the consistent indexing of neighboring intersections. [44] attempts to remove this requirement by utilizing the road network structure with Graph Convolutional Network [53] to cooperate multi-hop nearby intersections. [44] models the influence of neighboring agents by the fixed adjacency matrix defined in Graph Convolutional Network, which indicates their assumption that the influences between neighbors is static. In other work, [63; 60] proposes to use Graph Attentional Networks [59] to learn the dynamic interactions between the hidden states of neighboring agents and the ego agent. It should be pointed out that there is a strong connection between methods employing max-plus [27] to learn joint action-learners and methods using Graph Convolutional Network to learn the communication, as both of them can be seen to learn the message passing on the graph, where the former kind of methods passing the reward and the later passing the state obervations.

4. EVALUATION

In this section, we will introduce some experimental settings that will influence the evaluation of traffic signal control strategies: evaluation metrics, simulation environment, road network setting, and traffic flow setting. A comparison of the settings that influence the evaluation are summarized in Table 1.

4.1 Evaluation metrics

The objective of traffic signal control is to facilitate safe and efficient movement of vehicles at the intersection. Safety is achieved by separating conflicting movements in time and is not considered in most related literature. Various measures have been proposed to quantify efficiency of the intersection from different perspectives, including the average travel time of all vehicles, the average number of stops that vehicles experience in the network, the average queue length in the road network, and the throughput of the road network. While the performance of the same method on queue length might differ with different definitions of a "waiting" state of a vehicle, travel time and throughput are widely adopted as evaluation metrics by recent literature.

4.2 Simulation environment

Since deploying and testing traffic signal control strategies in the real world involves high cost and intensive labor, simulation is a useful alternative before actual implementation. Simulations of traffic signal control often involve large, heterogeneous scenarios and vehicle-level information, thus most literature relies on microscopic simulation, in which movements of individual vehicles are represented through microscopic properties such as the position and velocity of each vehicle. Some representative open-source microscopic simulators are: *The Green Light District (GLD)*[1], *The Autonomous Intersection Management (AIM)*[2], *Simulation of Ur-*

[1]https://sourceforge.net/projects/stoplicht/
[2]http://www.cs.utexas.edu/~aim/

ban MObility (SUMO)[3], and *CityFlow* [73]. Other proprietary simulators like Paramics[4] and Aimsun[5] are also adopted in [30; 10; 5]. For a detailed comparison of the open-source simulators, please refer to [40].

4.3 Road network

Different road networks are explored in the current literature, including synthetic and real-world road network. At a coarse scale, a road network is a directed graph with nodes and edges representing intersections and roads, respectively. Specifically, a real-world road network can be more complicated than the synthetic network in the road properties (e.g., the number of lanes, speed limit of every lane), intersection structures and signal phases settings. Among all the road network properties, the number of traffic signals in the network largely influences the experiment results because the scale of explorations for RL agents to take increases with the scale of road network. Currently, most of the work still conducts experiments on relatively small road networks compared to the scale of a city, which could include thousands of traffic signals. Aslani *et al.* [5; 4] test their method in a real-world road network with 50 signals. In [63], a district with 196 signals is investigated. One of the most recent work [11] tests their methods on the real road network of Manhattan, New York, with 2510 traffic signals.

4.4 Traffic flow

Traffic flow demand in the simulation can influence the evaluation of control strategies. The simulator takes traffic demand data as input, with each vehicle described as (o, t, d), where o is the origin location, t is time, and d is the destination location. Locations o and d are both locations on the road network. Usually, the more dynamic and heavier the traffic demand is, the harder for an RL method to learn an optimal policy. This is because the dynamic traffic would require the RL agents learn in a non-stationary environment, and heavier traffic would require fast adaptation for RL policies. The vehicle behavior models, such as lane changing, speed changing and routing models, could also influence the traffic flow and further influence the evaluation of traffic signal control policies. But in existing literature, they are usually kept fixed during the learning process of traffic signal control methods.

5. CONCLUSION AND FUTURE WORK

In this survey, we present an overview of recent advances in reinforcement learning methods for traffic signal control, and provide an organization considering both the learning approach and evaluations of the research in this field. Here, we briefly discuss some directions for future research.

5.1 Benchmarking datasets and baselines

As discussed in Section 4.4, researchers use different road networks and traffic flow datasets, which could introduce large variances in final performance. Therefore, evaluating control policies using a standard setting could save the effort and assure a fair comparison and reproducability of RL methods [25]. An effort that could greatly facilitate research in this field is to create publicly available benchmark. Another concern for RL-based traffic signal control is that for this interdisciplinary research problem, existing literature of RL-based methods is often lack of comparison with typical methods from transportation area, like Webster's Formula [28] and MaxPressure [29].

[3]http://sumo.sourceforge.net
[4]https://www.paramics-online.com/
[5]https://www.aimsun.com

5.2 Learning efficiency

Existing RL methods for games usually require a massive number of update iterations and trial-and-errors for RL models to yield impressive results in simulated environments. These trial-and-error attempts will lead to real traffic jams in the traffic signal control problem. Therefore, how to learn efficiently is a critical question for the application of RL in traffic signal control. While there is some previous work using Meta-Learning [72] or imitation learning [69], there is still much to investigate on learning with limited data samples and efficient exploration in traffic signal control problem.

5.3 Safety issue

While RL methods learn from trial-and-error, the learning cost of RL could be critical or even fatal in the real world as the malfunction of traffic signals might lead to accidents. An open problem for RL-based traffic signal control problem is to find ways to adapt risk management to make RL agents acceptably safe in physical environments [19]. [37] directly integrates real-world constraints into the action selection process. If pedestrians are crossing the intersection, their method will not change the control actions, which can protect crossing pedestrians. However, more safety problems like handling collisions are still to be explored.

5.4 Transferring from simulation to reality

Most RL-based traffic signal control methods mainly conduct experiments in the simulator since the simulator can generate data in a cheaper and faster way than real experimentation. Discrepancies between simulation and reality confine the application of learned policies in the real world. While some work considers to learn an interpretable policy before applying to the real world [6] or to build a more realistic simulator [61; 75; 76] for direct transferring, there is still a challenge to transfer the control policies learned in simulation to reality.

6. ACKNOWLEDGMENTS

The work was supported in part by NSF awards #1652525 and #1618448, NSF Grant CMMI-1749200 and a seed grant through the Penn State Institute of CyberScience. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agencies.

7. REFERENCES

[1] M. Abdoos, N. Mozayani, and A. L. Bazzan. Hierarchical control of traffic signals using Q-learning with tile coding. *Applied intelligence*, 40(2):201–213, 2014.

[2] I. Arel, C. Liu, T. Urbanik, and A. Kohls. Reinforcement learning-based multi-agent system for network traffic signal control. *IET Intelligent Transport Systems*, 2010.

[3] K. Arulkumaran, M. P. Deisenroth, et al. A brief survey of deep reinforcement learning. *arXiv preprint*, 2017.

[4] M. Aslani, M. S. Mesgari, and M. Wiering. Adaptive traffic signal control with actor-critic methods in a real-world traffic network with different traffic disruption events. *TRB-C*, 2017.

[5] M. Aslani, S. Seipel, M. S. Mesgari, and M. Wiering. Traffic signal optimization through discrete and continuous reinforcement learning with robustness analysis in downtown tehran. *Advanced Engineering Informatics*, 2018.

[6] J. Ault, J. Hanna, and G. Sharon. Learning an interpretable traffic signal control policy. *arXiv preprint*, 2019.

[7] B. Bakker, S. Whiteson, L. Kester, and F. C. Groen. Traffic light control by multiagent reinforcement learning systems. In *Interactive Collaborative Information Systems*. 2010.

[8] A. L. Bazzan. Opportunities for multiagent systems and multiagent reinforcement learning in traffic control. *AAMAS*, 2009.

[9] J. A. Calvo and I. Dusparic. Heterogeneous multi-agent deep reinforcement learning for traffic lights control. In *AICS*, pages 2–13, 2018.

[10] N. Casas. Deep deterministic policy gradient for urban traffic light control. *arXiv preprint*, 2017.

[11] C. Chen, H. Wei, N. Xu, et al. Toward A Thousand Lights: Decentralized Deep Reinforcement Learning for Large-Scale Traffic Signal Control. In *AAAI*, 2020.

[12] T. Chu, J. Wang, L. Codecà, and Z. Li. Multi-agent deep reinforcement learning for large-scale traffic signal control. *arXiv preprint*, 2019.

[13] C. Claus and C. Boutilier. The dynamics of reinforcement learning in cooperative multiagent systems. *AAAI/IAAI*, 1998.

[14] M. Coşkun et al. Deep reinforcement learning for traffic light optimization. In *ICDMW*. IEEE, 2018.

[15] T. Economist. The cost of traffic jams. https://www.economist.com/blogs/economist-explains/2014/11/economist-explains-1, November 2014.

[16] S. El-Tantawy and B. Abdulhai. Comprehensive analysis of reinforcement learning methods and parameters for adaptive traffic signal control. Technical report, 2011.

[17] R. Florin and S. Olariu. A survey of vehicular communications for traffic signal optimization. *Vehicular Communications*, 2015.

[18] J. Gao, Y. Shen, J. Liu, M. Ito, and N. Shiratori. Adaptive traffic signal control: Deep reinforcement learning algorithm with experience replay and target network. *arXiv preprint arXiv:1705.02755*, 2017.

[19] J. Garcıa and F. Fernández. A comprehensive survey on safe reinforcement learning. *JMLR*, 2015.

[20] D. Garg, M. Chli, and G. Vogiatzis. Deep reinforcement learning for autonomous traffic light control. In *ICITE*, 2018.

[21] H. Ge, Y. Song, C. Wu, J. Ren, and G. Tan. Cooperative deep q-learning with q-value transfer for multi-intersection signal control. *IEEE Access*, 7:40797–40809, 2019.

[22] W. Genders and S. Razavi. Using a deep reinforcement learning agent for traffic signal control. *arXiv preprint*, 2016.

[23] Y. Gong, M. Abdel-Aty, Q. Cai, and M. S. Rahman. Decentralized network level adaptive signal control by multi-agent deep reinforcement learning. *Transportation Research Interdisciplinary Perspectives*, 1:100020, 2019.

[24] A. Haydari and Y. Yilmaz. Deep reinforcement learning for intelligent transportation systems: A survey. *arXiv:2005.00935*, 2020.

[25] P. Henderson, R. Islam, P. Bachman, J. Pineau, et al. Deep reinforcement learning that matters. In *AAAI*, 2018.

[26] L. P. Kaelbling, M. L. Littman, and A. W. Moore. Reinforcement learning: A survey. *Journal of artificial intelligence research*, 1996.

[27] J. R. Kok and N. Vlassis. Using the max-plus algorithm for multiagent decision making in coordination graphs. In *Robot Soccer World Cup*, pages 1–12. Springer, 2005.

[28] P. Koonce et al. Traffic signal timing manual. Technical report, United States. Federal Highway Administration, 2008.

[29] A. Kouvelas, J. Lioris, S. A. Fayazi, and P. Varaiya. Maximum Pressure Controller for Stabilizing Queues in Signalized Arterial Networks. *TRB*, 2014.

[30] L. Li, Y. Lv, and F.-Y. Wang. Traffic signal timing via deep reinforcement learning. *IEEE/CAA Journal of Automatica Sinica*, 2016.

[31] L. Li, D. Wen, and D. Yao. A survey of traffic control with vehicular communications. *IEEE TITS*, 2014.

[32] X. Liang, X. Du, G. Wang, and Z. Han. Deep reinforcement learning for traffic light control in vehicular networks. *arXiv preprint arXiv:1803.11115*, 2018.

[33] X. Liang, X. Du, G. Wang, and Z. Han. A deep reinforcement learning network for traffic light cycle control. *IEEE Transactions on Vehicular Technology*, 68(2):1243–1253, 2019.

[34] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint*, 2015.

[35] J. D. Little, M. D. Kelson, and N. H. Gartner. Maxband: A versatile program for setting signals on arteries and triangular networks. 1981.

[36] X.-Y. Liu, Z. Ding, S. Borst, and A. Walid. Deep reinforcement learning for intelligent transportation systems. *arXiv preprint arXiv:1812.00979*, 2018.

[37] Y. Liu, L. Liu, and W.-P. Chen. Intelligent traffic light control using distributed multi-agent q learning. In *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, pages 1–8. IEEE, 2017.

[38] P. Lowrie. Scats–a traffic responsive method of controlling urban traffic. roads and traffic authority, sydney. *New South Wales, Australia*, 1990.

[39] P. Mannion, J. Duggan, and E. Howley. An experimental review of reinforcement learning algorithms for adaptive traffic signal control. In *Autonomic Road Transport Support Systems*. 2016.

[40] F. J. Martinez, C. K. Toh, J.-C. Cano, C. T. Calafate, and P. Manzoni. A survey and comparative study of simulators for vehicular ad hoc networks (VANETs). *Wireless Communications and Mobile Computing*, 2011.

[41] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, et al. Human-level control through deep reinforcement learning. *Nature*, 2015.

[42] S. S. Mousavi et al. Traffic light control using deep policy-gradient and value-function-based reinforcement learning. *Intelligent Transport Systems*, 2017.

[43] O. Nachum, M. Norouzi, K. Xu, and D. Schuurmans. Bridging the gap between value and policy based reinforcement learning. In *NeurIPS*, 2017.

[44] T. Nishi, K. Otaki, K. Hayakawa, and T. Yoshimura. Traffic signal control based on reinforcement learning with graph convolutional neural nets. In *ITSC*. IEEE, 2018.

[45] A. Nowe, P. Vrancx, and Y. M. D. Hauwere. *Game Theory and Multi-agent Reinforcement Learning*. 2012.

[46] V. Pandey and S. D. Boyles. Multiagent reinforcement learning algorithm for distributed dynamic pricing of managed lanes. In *ITSC'18*. IEEE, 2018.

[47] M. Papageorgiou, C. Diakaki, V. Dinopoulou, A. Kotsialos, and Y. Wang. Review of road traffic control strategies. *Proceedings of the IEEE*, 2003.

[48] T. T. Pham, T. Brys, M. E. Taylor, T. Brys, et al. Learning coordinated traffic light control. In *AAMAS*, 2013.

[49] L. A. Prashanth and S. Bhatnagar. Reinforcement learning with average cost for adaptive control of traffic lights at intersections. *ITSC*, 2011.

[50] J. Rios-Torres and A. A. Malikopoulos. A survey on the coordination of connected and automated vehicles at intersections and merging at highway on-ramps. *IEEE TITS*, 2016.

[51] S. G. Rizzo, G. Vantini, and S. Chawla. Time critic policy gradient methods for traffic signal control in complex and congested scenarios. In *KDD*, 2019.

[52] R. P. Roess, E. S. Prassas, and W. R. McShane. *Traffic engineering*. Pearson, 2004.

[53] M. Schlichtkrull, T. N. Kipf, P. Bloem, R. Van Den Berg, I. Titov, and M. Welling. Modeling relational data with graph convolutional networks. In *European Semantic Web Conference*. Springer, 2018.

[54] D. Schrank, B. Eisele, T. Lomax, and J. Bak. 2015 urban mobility scorecard. 2015.

[55] M. Schutera, N. Goby, S. Smolarek, and M. Reischl. Distributed traffic light control at uncoupled intersections with real-world topology by deep reinforcement learning. *arXiv preprint arXiv:1811.11233*, 2018.

[56] S. Sukhbaatar, R. Fergus, et al. Learning multiagent communication with backpropagation. In *NeurIPS*, 2016.

[57] T. Tan, F. Bao, Y. Deng, A. Jin, Q. Dai, and J. Wang. Cooperative deep reinforcement learning for large-scale traffic grid signal control. *IEEE transactions on cybernetics*, 2019.

[58] E. van der Pol. Coordinated deep reinforcement learners for traffic light control. NeurIPS, 2016.

[59] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio. Graph attention networks. *ICLR*, 2018.

[60] Y. Wang, T. Xu, X. Niu, and other. STMARL: A Spatio-Temporal Multi-Agent Reinforcement Learning Approach for Traffic Light Control. *arXiv preprint*, 2019.

[61] H. Wei, C. Chen, C. Liu, G. Zheng, and Z. Li. Learning to simulate on sparse trajectory data. 2020.

[62] H. Wei, C. Chen, G. Zheng, K. Wu, V. Gayah, K. Xu, and Z. Li. PressLight: Learning Max Pressure Control to Coordinate Traffic Signals in Arterial Network. In *KDD*, 2019.

[63] H. Wei, N. Xu, H. Zhang, G. Zheng, et al. CoLight: Learning Network-level Cooperation for Traffic Signal Control. In *CIKM*, 2019.

[64] H. Wei, G. Zheng, V. Gayah, and Z. Li. A survey on traffic signal control methods. *arXiv:1904.08117*, 2019.

[65] H. Wei, G. Zheng, H. Yao, and Z. Li. IntelliLight: A Reinforcement Learning Approach for Intelligent Traffic Light Control. In *KDD*, 2018.

[66] M. Wiering. Multi-agent reinforcement learning for traffic light control. In *ICML*, 2000.

[67] C. Wu, A. Kreidieh, K. Parvate, E. Vinitsky, and A. M. Bayen. Flow: Architecture and benchmarking for reinforcement learning in traffic control. *arXiv preprint*, 2017.

[68] Y. Wu, H. Tan, and B. Ran. Differential variable speed limits control for freeway recurrent bottlenecks via deep reinforcement learning. *arXiv preprint arXiv:1810.10952*, 2018.

[69] Y. Xiong, G. Zheng, K. Xu, and Z. Li. Learning traffic signal control from demonstrations. In *CIKM*, 2019.

[70] M. Xu, J. Wu, L. Huang, R. Zhou, T. Wang, and D. Hu. Network-wide traffic signal control based on the discovery of critical nodes and deep reinforcement learning. *Journal of Intelligent Transportation Systems*, 24(1):1–10, 2020.

[71] K.-L. A. Yau, J. Qadir, H. L. Khoo, et al. A survey on reinforcement learning models and algorithms for traffic signal control. *ACM Computing Survey*, 2017.

[72] X. Zang, H. Yao, G. Zheng, N. Xu, K. Xu, and Z. Li. MetaLight: Value-based Meta-reinforcement Learning for Online Universal Traffic Signal Control. In *AAAI*, 2020.

[73] H. Zhang, S. Feng, C. Liu, et al. Cityflow: A multi-agent reinforcement learning environment for large scale city traffic scenario. In *The WebConf*, 2019.

[74] Z. Zhang, J. Yang, and H. Zha. Integrating independent and centralized multi-agent reinforcement learning for traffic signal network optimization. *arXiv preprint*, 2019.

[75] G. Zheng, C. Liu, H. Wei, C. Chen, and Z. Li. Rebuilding city-wide traffic origin destinationfrom road speed data. In *ICDE*, 2021.

[76] G. Zheng, H. Liu, and Z. Li. Learning to simulate vehicle trajectory from demonstrations. In *ICDE*, 2020.

[77] G. Zheng, Y. Xiong, X. Zang, J. Feng, H. Wei, et al. Learning Phase Competition for Traffic Signal Control. In *CIKM*, 2019.

[78] G. Zheng, X. Zang, N. Xu, H. Wei, Z. Yu, et al. Diagnosing Reinforcement Learning for Traffic Signal Control. *arXiv preprint*, 2019.