Back to Basics: Deep Reinforcement Learning in Traffic Signal Control

Sierk Kanis
sierkkanis@hotmail.com
University of Amsterdam
Amsterdam, The Netherlands

Laurens Samson
Daan Bloembergen
l.samson@amsterdam.nl
d.bloembergen@amsterdam.nl
CTO, City of Amsterdam
Amsterdam, The Netherlands

Tim Bakker
t.b.bakker@uva.nl
University of Amsterdam
Amsterdam, The Netherlands

ABSTRACT

In this paper we revisit some of the fundamental premises for a reinforcement learning (RL) approach to self-learning traffic lights. We propose RLight, a combination of choices that offers robust performance and good generalization to unseen traffic flows. In particular, our main contributions are threefold: our lightweight and cluster-aware state representation leads to improved performance; we reformulate the Markov Decision Process (MDP) such that it skips redundant timesteps of yellow light, speeding up learning by 30%; and we investigate the action space and provide insight into the difference in performance between acyclic and cyclic phase transitions. Additionally, we provide insights into the generalisation of the methods to unseen traffic. Evaluations using the real-world Hangzhou traffic dataset show that RLight outperforms state-of-the-art rule-based and deep reinforcement learning algorithms, demonstrating the potential of RL-based methods to improve urban traffic flows.

KEYWORDS

Reinforcement Learning, Deep Reinforcement Learning, Q-learning, Adaptive Traffic Signal Control

1 INTRODUCTION

Adaptive Traffic Signal Control (ATSC) aims to facilitate smooth and safe traffic flow, thereby reducing travel times and cutting CO_2 emissions[24]. While conventional methods rely on handcrafted rules based upon specific traffic scenarios, the recent increase in traffic data and enormous advances in optimization techniques suggest that more opportunities are at hand [20].

In recent years, combining deep learning with reinforcement learning (RL) has led to strong performance in many complex environments, often achieving super-human performance [8, 12, 14]. These methods have proven to be able to handle dynamic environments, starting with a blank slate and learning directly from feedback by trial and error without relying on simplistic data assumptions. It is natural to wonder whether deep RL could be just as beneficial for ATSC.

However, ATSC presents a challenge from an RL perspective. While in some RL applications the environment straightforwardly translates to a Markov Decision Process (MDP), ATSC has no principled source of raw data and rewards, while also lacking a fixed set of actions or action-rate. This means that success is highly reliant on the quality of the MDP representation.

Models that aim to equip the agent with all possible information are unnecessarily complex which consequently hurts performance [25]. On the contrary, Light-Intellight (LIT) [25] aims to propose a minimal set of features based upon a uniform traffic distribution and shows good performance. However, traffic distributions in dense urban areas are often non-uniform and consist of clusters, see Figure 1. No prior work seems to specifically design their state representation to fit clustered traffic data.

We propose Reinforcement-Learning-Light (RLight)[1], a simple yet concise RL framework that exploits inductive biases relevant to the ATSC problem. We build upon the advantages of LIT using a simple shaping reward, yet expand their state representation to fit a more clustered data distribution. We represent approaching and waiting vehicles separately and add the mean speed and distance of the approaching cluster to the state as an approximation, thereby exploiting the specific structure of clusters to keep the state representation compact and digestible. By using only structured data, our method can exploit the current sensors in the ground and GPS data, therefore contributing to the transition of adopting a reinforcement learning approach in the real world.

Additionally, to our knowledge, no prior work has acknowledged the ambiguity in modelling yellow light, while it does impact the performance. So far the ATSC problem has mainly been formulated as a standard MDP, whereas an underlying inductive bias could be built in to skip redundant timesteps during yellow light, thus cleaning up the reward signal by ignoring irrelevant state-action pairs. We show that this reduces learning time by approximately one third.

Furthermore, prior work does not agree on whether it is safe to have an acyclic action space or not, whereas we find that limiting the agent to cyclic control impacts the performance significantly. Just as a human traffic warden would do, we think it is more natural to be able to freely choose the optimal phase at every timestep. We investigate this difference to provide more ground for utilizing acyclic control.

We apply our approach to real data of five intersections in the city of Hangzhou using the CityFlow simulator [23]. Since this dataset is fixed and thus, in RL terms, the environment is deterministic, we split up the data into train, validation and test sets to provide insights into the generalisation of our method. To test our approach against a rule-based method, we propose an extension of the Self-Organizing Traffic Lights (SOTL) method [2] that works in multi-phase settings as well as in the originally implemented two-phase setting. Empirically, we show that our proposed method

[1] https://github.com/Amsterdam-Internships/Self-Learning-Traffic-Lights

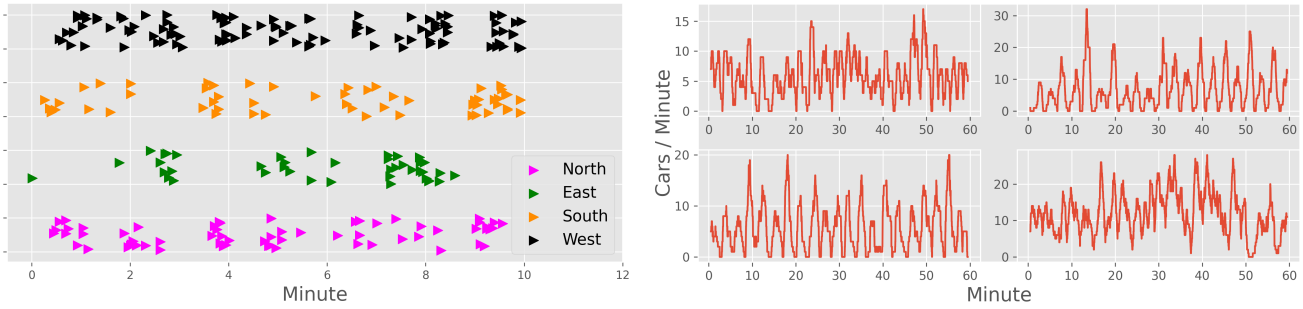

Figure 1: The left panel shows data from the first ten minutes of inflow on intersection Hangzhou 2, with every triangle indicating a vehicle. The right panel shows the rate of vehicles per minute. Notice how traffic often appears in clusters.

RLight consistently outperforms state-of-the-art methods on all five intersections.

In short, our contributions are the following:

- We propose RLight, a simple yet concise RL framework that exploits inductive biases relevant to the ATSC problem:
  - Our lightweight and cluster-aware state representation outperforms state-of-the-art methods on the Hangzhou dataset.
  - We reformulate the MDP such that it skips redundant timesteps of yellow light, speeding up learning.
  - We investigate the action space and provide insight in the difference in performance between acyclic and cyclic phase transitions.
- In addition, we extend the rule-based algorithm SOTL to work in an acyclic fashion for multi-phase intersections.

Our single agent is able to successfully control varying traffic densities, enabling it to generalize to unexpected real-world situations such as a football event, a nearby accident or, let's say, a world-wide pandemic.

2 BACKGROUND

In this section we elaborate on the basics of traffic signal control, we give an overview of reinforcement- and deep reinforcement learning, and we go over the basics of the Self-Organising Traffic Lights (SOTL) algorithm.

2.1 Traffic Signal Control

We address the traffic signal control problem using a set of incoming lanes $\boldsymbol{\nu}$ of cardinality J, a set of outgoing lanes and a set of phases $\boldsymbol{\varphi}$ of cardinality I. Each phase is a combination of green and red lights and the set of phases consists of all possible combinations in which conflicting directions do not have green light simultaneously. In particular, this means every phase consists of two green lights, while the other lights are red. A fixed yellow period of five seconds is enforced after every switch of lights. For simplicity, we assume that separating conflicting directions and forcing yellow light periods account for safety and we do not consider safety in more detail.

2.2 Reinforcement Learning

Reinforcement learning is a computational approach characterized by its learning from direct interaction with the environment without requiring supervision or a complete model of the environment [15].

Markov Decision Process. We consider a Markov Decision Process (MDP), which is a formalization of sequential decision making where actions not only influence immediate rewards, but also subsequent states and through those future rewards. In an MDP, the agent and environment interact at each of a sequence of discrete timesteps, $t = 0, 1, ..., T$. At every timestep t, the agent receives a representation of the environment state s_t and selects an action a_t, after which the environment transitions to the next state s_{t+1} and the agent receives a reward r_{t+1}. The goal of reinforcement learning is to learn a policy, i.e. a mapping from perceived states to actions in those states $\pi : S \to \Delta_A$ that maximizes the expected return, which is defined as:

$$Q_\pi(s, a) = \mathbb{E}_\pi \left[\sum_{t \geq 0} \gamma^t r_t \middle| S_0 = s, A_0 = a \right]$$

for each initial state-action pair $(s, a) \in S \times A$, where $\gamma \in [0, 1]$ is the time-discount factor that prioritises short-term rewards over long-term rewards [15].

State. At each time step, the agent receives a quantitative representation of the environment, i.e. a state representation S_t. Ideally, the state representation fully describes the environment, such that the agent is given all information necessary to perform the right action. However, an overly complex state representation toughens distilling the useful information, while on the other hand, a simple state representation may result in different states appearing to be identical, making it impossible to learn the appropriate behaviour. In some reinforcement learning applications the environment straightforwardly translates to a state representation, e.g. pixels on a screen. In the traffic signal control setting, however, no principled source of raw data exists, which means success is highly reliant on the quality of the hand-crafted state representation.

Reward. On each timestep the environment sends the RL-agent a reward signal R_t indicating the quality of its chosen action. This allows the agent to estimate a value function, i.e. the total amount

of reward the agent can expect to accumulate over the future from each state. Subsequently, actions are taken based on this estimated value function, seeking to reach states with higher values.

Again, there is no clear-cut reward signal in the traffic signal control setting, like points in Tetris or winning a game of chess. In the case of very sparse reward signals like average travel time, the addition of intermediate shaping rewards can be helpful to steer the agent towards the goal [13]. However, rewards must be provided in such a way that in maximizing them the agent will also achieve the goals one want to be accomplished.

Actions. Each timestep t the agent selects an action A_t from the action space $A = \{1, ..., K\}$, i.e. the set of possible actions. Ideally, the action space allows the agent maximum freedom in the environment, such that all options are available necessary to perform the right action. However, a too big action space imposes the agent with the task to choose the right action from a bigger set, possibly harming performance. In traffic signal control there is no fixed set of available options in the environment, e.g. the legal moves in the game of chess. This means that performance is reliant on the choice of the action space.

2.3 Deep Reinforcement Learning

Deep reinforcement learning can be understood as consisting of three fundamental components: a function approximator, a learning algorithm and a mechanism for generating training data. We parameterize an approximate value function $Q(s, a; \theta_i)$ using a deep neural network architecture in which θ_i are the parameters of the network at iteration i. Samples of experiences generated by our traffic simulator $(s, a, r, s') \sim U(D)$ are drawn uniformly at random from the memory of saved transitions and are used for learning by updating the local neural network at iteration i by the following loss function:

$$L_i(\theta_i) = \mathbb{E}_{(s,a,r,s')\sim U(D)} \left[r + \gamma \max_{a'} Q(s', a'; \theta_i^-) - Q(s, a; \theta_i) \right]$$

in which θ_i are the parameters of the local neural network at iteration i and θ_i^- are the parameters of the target network at iteration i [12]. The target network is added to stabilize learning and its parameters θ_i^- are being soft-updated with factor τ to slowly catch-up with the local network [10].

2.4 Self-Organizing Traffic Lights

Next to deep reinforcement learning, we examine the rule-based Self-Organizing Traffic Lights (SOTL) algorithm, in order to compare our RLight agent to a rule-based approach. SOTL is a distributed adaptive traffic light system with no communication between intersections. The method is considered self-organizing, because the global performance is only dependent on the local rules of each intersection; each intersection is unaware of the state of other intersections [2]. Still, it is able to achieve global coordination of traffic because of the probabilistic formation of vehicle platoons, i.e. convoys of vehicles [5].

Each traffic phase $\boldsymbol{\varphi}_i$ keeps a count κ_i, with i from 1 to I, in which I is the number of phases. At every time-step, κ_i adds the number of vehicles on the currently red lanes of $\boldsymbol{\varphi}_i$, independently of whether a vehicle is moving or waiting. When κ_i, representing the integral of vehicles on the lanes of $\boldsymbol{\varphi}_i$ over time, reaches a threshold θ, the current phase switches to phase $\boldsymbol{\varphi}_i$ and κ_i is reset to zero. To prevent fast switching of lights, additional constraints are tuned per intersection [5]. The original implementation by Gershenson only supports two phases, in Section 5 we extend the method to work in multi-phase settings too.

3 RELATED WORK

The ATSC problem has traditionally been approached using rule-based methods that use manually tuned parameters, as well as machine learning techniques [24]. Although approaches based on RL go back more than two decades [22], the advent of Deep RL has led to a new boom in research in this direction [7]. Many RL approaches discuss various ways in which to define the underlying Markov Decision Process (MDP) in terms of its state, action and reward representation [20].

State. The introduction of deep learning has shown that training directly from raw inputs can lead to better representations than handcrafted features [12], but the ATSC setting does not have a principled source of raw data. As an attempt to fully describe the traffic situation recent studies have used grids [9] or images [17], which led to state representations with thousands of dimensions. This abundance of information however does not necessarily lead to a gain in performance [20].

Reward. A principal goal of ATSC is to reduce the average travel time in the system. However, this metric is hard to estimate outside a simulator and additionally leads to highly delayed signals. Recent studies define the reward function as an ad-hoc weighted linear combination of several direct traffic measures such as queue length, waiting time, speed, and throughput [11, 17, 19]. Such multi-component reward exhibits two drawbacks. First, there is no guarantee that the desired goal is optimized. Second, small changes in the component weights could lead to drastically different results [25]. Unfortunately, there is no principled approach to tune weights within a RL reward function [20].

The work closest to ours is Light-Intellight (LIT) [25]. Their approach uses the cumulative queue length as a proxy for average travel time. The authors show that queue length is proportional to travel time and therefore optimizing queue length amounts to optimizing travel time. With this reward, Zheng et al. claim that only the number of vehicles and the traffic signal phase are needed to fully describe the system, under the assumption of uniform traffic inflow. Under this assumption their method worked well, but as shown in Figure 1, urban areas with dense intersections cause traffic to be highly clustered. In this work, we adopt the reward function of Zheng et al. but challenge their assumption of uniformity by expanding our state representation to adapt to more fragmented traffic distribution.

Action. The LIT approach, among others, additionally assumes that a predetermined phase order aligns best with drivers' expectations and avoids safety issues [9, 25]. Yet, there are urban traffic control systems that do utilize acyclic phases (e.g. Amsterdam). Since limiting the agent to a fixed cycle restricts it to learn the

optimal policy [4], we investigate the reduction of travel time when allowing acyclic phase transitions.

Coordination. In recent years, studies have scaled up their approach to large multi-agent systems, proposing multiple forms of coordination and communication between the agents [1, 7, 11]. Van der Pol and Oliehoek [17] have extended their single agent DQN solution by making use of transfer planning and the max-plus coordination algorithm while Wei et al. [18] have introduced a graph attentional network to facilitate cooperation. However, under certain circumstances, explicit coordination between intersections may not be a necessity [6]. Traffic lights themselves already structure traffic flow in clusters of vehicles. When intersections follow each other rapidly, clusters likely persist until the next intersection. The large empty areas that appear between clusters could then be used by crossing clusters; if agents can learn to anticipate this approaching traffic, the system can naturally become self-organizing without the need for explicit coordination [2]. Empirically, Zheng et al. [25] have shown that LIT outperforms the multi-agent approach of Van der Pol and Oliehoek [17] without adding explicit coordination, which motivates us to investigate the quality of the basic assumptions of single-agent control further.

4 RLight AGENT DESIGN

Our RLight agent is based on the standard reinforcement learning framework [15]. In the following, we describe how we formulate ATSC as a Markov Decision Process (MDP) and discuss the state and action representation, as well as the reward function. We consider the scenario in which the agent controls a single intersection with J incoming lanes, and I traffic light phases.

4.1 Markov Decision Process

To our knowledge, no prior work has investigated the difference in modelling the action rate in ATSC problems. Assuming a timestep rate of one simulation second per transition and a fixed period of yellow light enforced by the environment, we consider two options.

MDP. The agent selects an action at every simulation second and we accept that actions during yellow light are being ignored by the environment. This means that the agent would have to learn that actions in such states do not have any impact on the final reward, and therefore, every action is equally good. In this way, training effort is put into learning irrelevant state-action pairs, while it can also be prone to making the reward signal noisier. Consider the scenario when the agent has not fully learned to distinguish yellow and non-yellow states, in which case it appears to the agent as if it gets rewarded for its action, yet in a yellow state it would get this reward anyway.

SMDP. The agent only chooses actions when its decisions actually have an impact on the environment, meaning it is inactive during yellow light. This can be seen as a Semi-Markov Decision Process (SMDP) [16]. Assume that the yellow period has a fixed length of ψ timesteps. Whenever the agent decides to switch to another phase (and thus the yellow light period starts) at timestep τ, in state s_τ with action a_τ, it receives the cumulative discounted rewards during the yellow period added to the discounted immediate reward after the yellow period:

$$r_\tau(s_\tau, a_\tau) = \sum_{t=1}^{\psi+1} \gamma^t r_{\tau+t}(s_{\tau+t})$$

The advantage of this way of modelling is that irrelevant state-action pairs do not have to be learned, saving training time and simplifying optimization. Note that, since the queue length of each timestep is still implicitly perceived by the agent, the proportionality of queue length and average travel time which we discus in Section 4.4 still holds.

4.2 State representation

At each time step, the agent receives a quantitative representation of the environment, i.e. a state representation. Our goal is to choose a representation that is easily digestible, yet concise enough to contain the information necessary to select the appropriate actions.

As a starting point, we consider only the number of vehicles and the current phase as used in LIT [25]. The current phase is necessary for the agent to know whether lights will switch by taking an action and the waiting and approaching vehicles are represented jointly based upon the idea that, if the inflow of traffic is constant, the agent can estimate these instead of perceiving them directly.

This leads to:

$$\boldsymbol{s_t} = [\boldsymbol{w}_t^\intercal + \boldsymbol{a}_t^\intercal, \boldsymbol{p}_t^\intercal],$$

in which $\boldsymbol{w} + \boldsymbol{a} \in \mathbb{R}^J$ is a transposed vector representing the total number of vehicles (waiting, $\boldsymbol{w}$, plus approaching, $\boldsymbol{a}$) on each lane and $\boldsymbol{p}$ is the phase represented as a one-hot vector of size I, where I and J are the number of phases and lanes, respectively. While LIT uses phase-gate [21] to condition the action-value function on the current state, we simply encode the phase as a one-hot vector and append it to the state.

However, not distinguishing approaching and waiting vehicles may result in different states appearing to be identical, making it impossible to learn the appropriate behaviour. In particular, this state representation cannot distinguish patterns in the traffic flow, and thus also cannot be used for implicit coordination between traffic lights.

We extend this approach by explicitly separating the waiting and approaching vehicles and thereby fitting the more fragmented distribution of urban traffic. Additionally, we concatenate the average speed and distance of the approaching traffic to improve traffic anticipation, i.e. switch lights earlier if a cluster moves faster or closer.

This results in the following state representation:

$$\boldsymbol{s}_t = [\boldsymbol{w}_t^\intercal, \boldsymbol{a}_t^\intercal, \boldsymbol{d}_t^\intercal, \boldsymbol{s}_t^\intercal, \boldsymbol{p}_t^\intercal]$$

in which $\boldsymbol{w}$ represents the number of waiting vehicles, $\boldsymbol{a}$ the number of approaching vehicles, $\boldsymbol{d}$ the average distance of approaching vehicles, $\boldsymbol{s}$ the average speed of approaching vehicles, and $\boldsymbol{p}$ the phase represented as a one-hot vector as before. All values are normalized by dividing by their maximum observed value. We use the average speed and distance of the cluster as an approximation to keep the state representation compact, under the assumption that vehicles behave like steady convoys. In an eight-approach, twelve-phase setting ($J = 8, I = 12$) this results in a state-space of dimension $4 \times 8 + 12 = 44$. Note that even under uniform traffic

scenarios our expansion will be minimally as descriptive as LIT's approach.

4.3 Action Space

Each timestep t the agent selects an action a_t from the available set of actions $A = \{1, ..., K\}$. Our goal is to allow the agent maximum freedom, such that all necessary possibilities are available to select the right action. We investigate the following two options.

Cyclic. We use a predetermined phase sequence in which the agent can either keep the current phase or switch to the next phase. Specifically, the neural network outputs a value to switch or to keep.

Acyclic. We use an acylic phase sequence in which the agent can freely choose which phase to switch to next, allowing more flexible control. In this case, the network outputs as many values as there are phases.

4.4 Reward Function

At every timestep the agent receives a numerical reward r_t defined by the reward function. Our goal is to define a reward function to minimize average travel time which for simplicity we assume to be the main objective.

Since average travel time can only be computed at the end of a simulation trajectory, using this as the reward signal leads to extremely sparse and delayed rewards. In addition, the average travel time is generally not readily available outside a simulator. Following Zheng et al. [25] we use the total queue length at the intersection as a shaping reward:

$$r_t(s_t) = -\sum_{j=1}^{J} w_j(s_t)$$

in which w_j is the queue length at the lane j and J is the number of lanes. This shaping reward is proven to be approximately proportional to the average travel time, when neglecting speed changes for simplicity [25].

5 SELF-ORGANIZING TRAFFIC LIGHTS 2.0

In this section, we explain how we generalize the rule-based Self-Organizing Traffic Lights (SOTL) algorithm to multi-phase settings.

The original method was only implemented for two-phase settings, which means it was implemented to switch to the next phase instead of explicitly switching to the phase with the most waiting time. In a two-phase setting, this amounts to the same policy, because in a two-phase setting cyclic and acyclic control is ambiguous. In a multi-phase setting, however, a phase should not only switch but also decide which phase to switch to. Also, in a multi-phase setting, multiple phases can get above the threshold simultaneously. Therefore we augment the algorithm by letting it choose the phase corresponding to the maximum κ instead of the next phase in the sequence, such that the phase with the most cumulative waiting time is chosen.

The original implementation brings along another issue when adapting it to a multi-phase intersection: Resetting κ to zero is not anymore sufficiently informing the system that vehicles have passed through green, as a green light is part of multiple phases.

Algorithm 1: SOTL generalized to multi-phase settings

```
initialize κ and ρ to 0
for t ← 1 to T do
    for j ← 1 to J do
        ρ_j += vehicles_jt
    for i ← 1 to I do
        κ_i = Σ_{j=1}^{J} ρ_j · 𝟙_{φ_i}(v_j)
    if φ_green > φ_min then
        if not 0 < vehicles_{φ_green} < μ then
            if max κ > θ then
                action = arg max_i κ
                ρ^i = 0
```

Consider κ_i and κ_j $(i \neq j)$, which share one lane. If phase $\boldsymbol{\varphi}_i$ is set to green, vehicles on the shared lane will pass through green, yet only κ_i is reset, while κ_j does not get updated. Consequently, the passed vehicles are still present in the integral of κ_j, which results in phase $\boldsymbol{\varphi}_j$ being chosen next, although no vehicles are waiting anymore. Since SOTL uses counters to represent the integrals, there is no way to remove only the waiting time of one lane. This means that to keep functioning solely with counters (such that only one induction loop per lane is necessary) an additional parameter is called for.

Let us introduce $\boldsymbol{\rho} \in \mathbb{N}^J$, where $\boldsymbol{\rho} = (\rho_1, ..., \rho_j)$ is a set of counters which represent the integrals of vehicles per lane over time and J is the number of approaching lanes. Now, when lane v_j gets green light, independently which phase turned this light into green, ρ_j corresponding to lane v_j is reset to zero.

Then, instead of counting κ_i directly, κ_i is calculated by summing its corresponding ρ-values:

$$\kappa_i = \sum_{j=1}^{J} \rho_j \cdot \mathbb{1}_{\boldsymbol{\varphi}_i}(v_j)$$

where $\mathbb{1}_{\boldsymbol{\varphi}_i}(v_j)$ is a function that indicates whether lane v_j has green light in phase $\boldsymbol{\varphi}_i$. Note that κ_i remains the cumulative waiting time of phase $\boldsymbol{\varphi}_i$, like in the original SOTL method.

Now, whenever a lane v_j gets green light, all the corresponding κ-values are updated accordingly, because they are all dependent on the same ρ_j. So, by introducing an additional parameter, passed cars will be removed from all corresponding integrals, and consequently, true to the original SOTL method, the phase with the most waiting time is chosen.

The full algorithm is presented in Algorithm 1, in which ϕ_{green} is the current duration of the phase, ϕ_{min} is the minimal duration a phase must be, $vehicles_{\phi_{green}}$ is the number of vehicles within a hand-tuned distance from the green lights, μ is a tunable parameter indicating the number of vehicles that is needed to split up a cluster and $\boldsymbol{\rho}^i$ is the set of ρ values corresponding to phase $\boldsymbol{\varphi}_i$. Note that in the original two-phase, four-approach setting SOTL and SOTL-2.0 are identical.

6 EXPERIMENT SETUP

In this section, we discuss the experimental setup to evaluate our method. We describe the dataset, model architecture and baseline methods.

6.1 Data

We perform experiments on five simulated intersections in Hangzhou, China. The data are based upon real-life traffic recordings and contain two hours of vehicle trajectories per intersection [18]. We simulate the traffic in CityFlow[2] [23] by feeding the route and spawn time of each vehicle and let our RL agent control the traffic lights. Each green light is followed by five seconds of yellow light. Each intersection has eight approaches, four going straight and four turning left. We use all possible phases as actions, i.e. non-conflicting green light configurations. Note that our method can be applied to intersections of any size.

Usually, RL environments are either deterministic or contain some randomness that makes every run different. However, when simulating vehicles based upon a fixed dataset, every run is very similar, while in reality traffic is always different. Therefore, we split the data into training, validation and test sets to avoid overfitting to a specific hour of traffic flow data, similar to supervised learning. Besides that, using test and validation sets allows us to compare the generalisation of the different methods.

In order to increase generalization to unseen traffic scenarios and simultaneously make maximum use of the available data, we train our model on four intersections and use the two different hours of the remaining fifth intersection for validating and testing. We repeat this for all five intersections.

6.2 Model Architecture

To test our framework we use the Deep Q-learning algorithm as proposed by Mnih et al. [12] with soft-updates [10]. Our neural network consists of two fully-connected hidden layers of 64 nodes with ReLu activation followed by a fully-connected linear layer with a single output for each phase. We use the hyperparameters as proposed by Mnih et al. [12], except for using a minibatch size of 512, a learning rate of $1e^{-3}$ and a replay memory size of 360k. We optimize our network with the Adam algorithm and use an ϵ-greedy behaviour policy. We use the same network architecture and hyperparameters across all datasets, showing that our approach is robust to a variety of intersections.

6.3 Baselines

In addition to the earlier described RL method LIT [25], we also report scores for a set of rule-based methods to give an intuition about the complexity of the intersection. The method labeled fixed is a cyclic phase order with 20 seconds of green time per phase and the method labeled random is a policy that selects actions uniformly at random. SOTL-1.0 is an implementation[3] of the rule-based Self-Organizing Traffic Lights (SOTL) algorithm [2]. Note that, although Zheng et al. use this implementation throughout their studies, it resembles more the cut-off method designed by Fouladvand et al. than it does resemble the original SOTL method as introduced by Gershenson. SOTL-2.0 is our augmented version of the original SOTL method and picks the phase with the most waiting time in a multi-phase setting as well as in a two-phase setting, as described in section 5.

[2]https://cityflow-project.github.io/

[3]https://traffic-signal-control.github.io/code.html

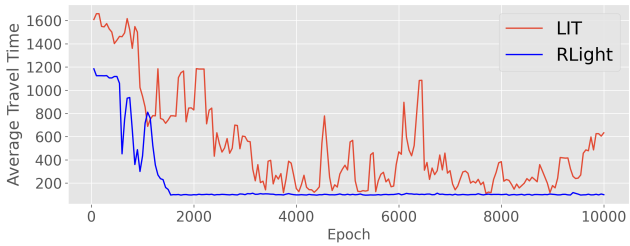

Figure 2: This plot shows how the average travel time per 50 epochs evolves during training on Hangzhou 1 of RLight (blue) compared to LIT (orange). Every 50 epochs correspond to 45000 weight updates or approximately 12 minutes of training time.

7 RESULTS

In this section, we discuss how the performance of our learning algorithm is influenced by choices regarding the state and action representation as well as the nature of the MDP formalisation. For simplicity, we use average travel time as an evaluation metric, which we compute every 50 epochs on the validation set during training. We save the model corresponding to the best score on the validation set and use that model on the test set to compute the final evaluation score. We only perform one run for each experiment, since the low variance between runs, due to the deterministic traffic scenarios used for data generation, does not impact the relative performance of the methods.

State representation. We compare our results with some of the best-performing methods from the ATSC literature, both rule-based and RL-based [2, 25]. The lower four rows of Table 1 show the scores on the validation/test set for each part of our state representation compared to the state representation of LIT. Note how we consistently outperform LIT on all five intersections.

LIT learns policies that are on par with SOTL-2.0 on the validation set but fails to generalize to the test set on intersections 1 and 5. As shown in Figure 2, the validation performance of LIT is quite unstable during training, which goes some way towards explaining its poor test time performance. This figure also shows that our streamlining of the framework and consequently of the optimization process has paid out, resulting in very stable learning.

From our different state representations, $[\boldsymbol{w}^\intercal, \boldsymbol{a}^\intercal, \boldsymbol{d}^\intercal]$ achieves the best results on four of the five intersections. The addition of the average speed $\boldsymbol{s}$ of the cluster of approaching vehicles leads to slightly worse performance, which might indicate that it does not contribute enough to compensate for the expansion in state space.

Additionally, we show that SOTL-2.0 consistently achieves at least 30% lower travel times than SOTL-1.0, with an average reduction of 38%. Still, our RLight agent using state representation $[\boldsymbol{w}^\intercal, \boldsymbol{a}^\intercal, \boldsymbol{d}^\intercal, \boldsymbol{p}_t^\intercal]$ outperforms SOTL-2.0 by a substantial margin of

Table 1: The upper table shows the average travel time in seconds for various rule-based methods on the validation/test set. The lower table reports results of each component in our state representation on the validation/test set, while being trained on the other four intersections. $\boldsymbol{w}$ represents the number of waiting vehicles, $\boldsymbol{a}$ the number of approaching vehicles, $\boldsymbol{d}$ the average distance of approaching vehicles and $\boldsymbol{s}$ the average speed of approaching vehicles. Additionally, every state contains the current phase as described in Section 4.2. The experiments are performed with the Semi-Markov Decision Process model.

	Hangzhou 1	Hangzhou 2	Hangzhou 3	Hangzhou 4	Hangzhou 5
Random	981 / 1086	613 / 520	649 / 844	830 / 1110	967 / 1064
Fixed time	440 / 632	283 / 213	252 / 278	319 / 599	539 / 613
SOTL-1.0	221 / 335	160 / 165	117 / 132	174 / 290	215 / 325
SOTL-2.0	120 / 234	77 / 78	81 / 87	96 / 159	115 / 193
$[\boldsymbol{w}^\top + \boldsymbol{a}^\top]$ (LIT)	112 / 422	79 / 73	96 / 111	100 / 148	96 / 842
$[\boldsymbol{w}^\top, \boldsymbol{a}^\top]$	103 / **138**	71 / 74	85 / 92	88 / 104	99 / 116
$[\boldsymbol{w}^\top, \boldsymbol{a}^\top, \boldsymbol{d}^\top]$	**95** / 151	**66** / **65**	**78** / **83**	**82** / **100**	**85** / **107**
$[\boldsymbol{w}^\top, \boldsymbol{a}^\top, \boldsymbol{d}^\top, \boldsymbol{s}^\top]$	**95** / 142	67 / **65**	80 / 87	**82** / **100**	**85** / 144

at least 34% on all five intersections despite incorporating almost no prior knowledge about traffic flows.

Note that intersections 1, 4 and 5 are more crowded than the other intersections, which increases the impact of each action and therefore requires more precise control. We show that in these cases SOTL and LIT do not generalize well from the validation set to the test set, while our method excels especially in these cases, achieving a striking average of 75% lower travel times than LIT.

Finally, we show that the results on the validation and test set vary significantly from each other, suggesting the methods are prone to overfitting to a specific hour of traffic flow at the intersection. This confirms the decision to split the data into train, validation and test sets, which is uncommon in reinforcement learning. This is probably because we generate experiences based upon real data, which only allows for a finite amount of experiences, whereas many reinforcement learning applications allow for the generation of an infinite stream of non-deterministic data.

Action space. We compare the effect of using a cyclic versus acylic phase order in Table 2. The freedom the agent gets in choosing a phase independently of the last phase improves performance consistently on all intersections. Moreover, selecting a wrong action in cyclic control has a longer effect on the queues since each action determines the next actions, becoming more sensitive to mistakes. Additionally, we show that also when adopting a cyclic phase order, our state representation outperforms LIT's, even when LIT's state representation was specifically aimed towards cyclic phase orders.

Markov Decision Process. In Table 3 we compare modelling the process as a Markov Decision Process (MDP) versus a Semi-Markov Decision Process (SMDP). Both methods result in similar scores, but the SMDP requires 37% less training time than the MDP, because the replay memory is filled only with valuable timesteps, thereby needing fewer gradient updates relative to the MDP.

Reward function. Figure 3 visually shows the proportionality of our reward function and the average travel time, which confirms that negative cumulative queue length is a good shaping reward for average travel time.

Table 2: The average travel time in seconds of a cyclic and acyclic action space using our state representation $[\boldsymbol{w}^\top, \boldsymbol{a}^\top, \boldsymbol{d}^\top, \boldsymbol{p}_t^\top]$ relative to the state representation of LIT. Note that our state representation also outperforms LIT's when adopting a cyclic phase order.

	H1	H2	H3	H4	H5
Cyclic LIT	422	88	119	470	184
Cyclic RLight	248	78	93	142	142
Acyclic RLight	**151**	**65**	**83**	**100**	**107**

Table 3: The difference in average travel time in seconds of modelling the problem as a Markov Decision Process (MDP) versus a Semi Markov Decision Process (SMDP) using our state representation $[\boldsymbol{w}^\top, \boldsymbol{a}^\top, \boldsymbol{d}^\top, \boldsymbol{p}_t^\top]$. The results are similar but the SMDP requires significantly less training time.

	H1	H2	H3	H4	H5
MDP	134	68	83	97	101
SMDP	151	65	83	100	107

Generalisation. To investigate the generalization we input artificial states with zero to five vehicles per direction into a trained Q network in a simple four-approach two-phase intersection. We depict the corresponding Q values in Figure 4 to show the decision boundary of choosing one over the other phase. The consistent linearity between the Q values indicates that the model has learned to generalize well to unseen states, which suggests that there exists an underlying function over the traffic densities, independent of a specific intersection. Consequently, we hypothesize that the more diverse the training samples are, the more this underlying function gets explored and learned, resulting in good understanding of where the decision boundary lies. This suggests that training on multiple intersections and augmenting the data could increase

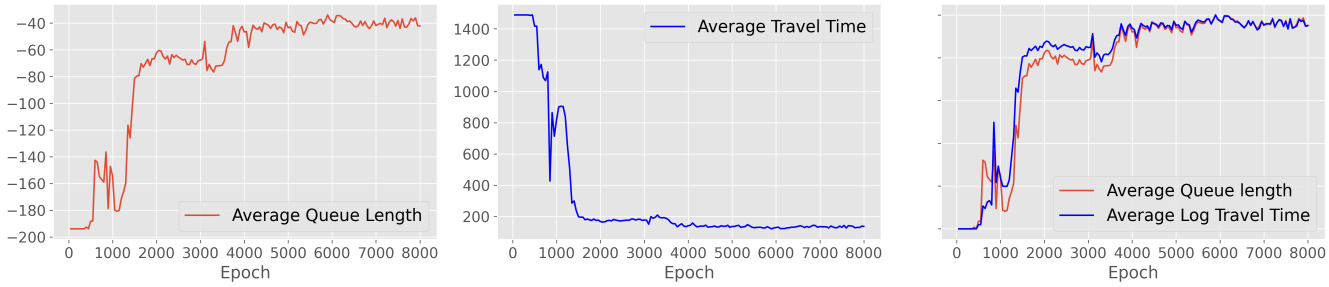

Figure 3: The left and middle plots show the average negative queue length and average travel time respectively during training. The right plot shows the proportionality between these two quantities by depicting the log travel time upside-down.

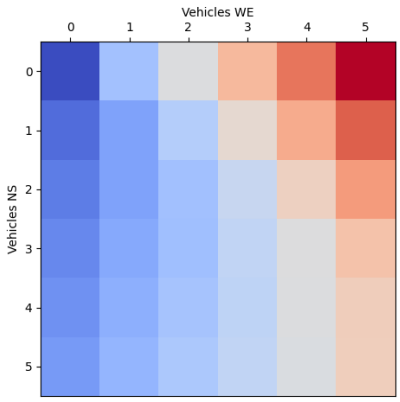

Figure 4: This figure shows the difference between the two Q-values corresponding to switching to WE or keeping phase NS in a simple four-approach two-phase intersection. Warmer colors indicate a switch, cooler colors indicate keeping the phase and grey values indicating indifference. This shows how the decision boundary is fairly symmetrical in the diagonal, indicating an underlying function over the traffic densities. Note how the decision boundary is shifted to the right due to the costs of switching (the addition of yellow light).

generalisation, which consequently could increase performance in the unpredictable traffic scenarios of the real-world.

8 DISCUSSION

While our results look promising, we have only tested our approach on data from one source with only two hours of traffic data per intersection. In addition, the layouts of the intersections are identical, although we have designed our method to be applicable to any form of intersection.

Our results are also dependent on the realism of the simulation. Firstly, all vehicles are simulated with the maximum speed, which unrealistically makes it an undescriptive factor in our RL environment. Secondly, vehicles are unable to collide in the simulator, which should be built in to provide a balanced view on safety.

Our state representation can be extended to incorporate bicycle lanes, which we reckon to be fairly straightforward due to the compact nature of our state representation. Furthermore, in extremely dense urban environments, augmenting the state representation by including the amount of free space on the outgoing lanes could become useful. This is because in these cases not every vehicle might be able to pass through green and accordingly, the agent should change its decisions.

The reward function could potentially be augmented to prioritize certain vehicles (e.g. trucks, public transport, bicycles) by assigning each of them a different weight. Also, each vehicle could be weighted by its waiting time in order to decrease the variance of the travel time at the expense of the average travel time.

Interesting follow-up research would be to investigate whether our algorithm performs strongly in a multi-agent scenario as well, where several connected intersections are controlled by individual agents. We hypothesize that the natural formation of clusters allows for implicit coordination between intersections, due to our agent's ability to deal with clustered traffic.

9 CONCLUSION

In this work, we have opted to overcome the lack of a straightforward reinforcement learning framework in traffic signal control by investigating choices regarding the fundamental premises of a deep reinforcement learning agent. Concretely, we have sought to find ways to exploit relevant inductive biases regarding the state, reward, action and MDP formulation. We have shown that we can speed up learning by reformulating the MDP as an SMDP by discarding yellow phase time from the learning process, that acyclic phase transitions consistently yield better performance than cyclic phase transitions, and that distinguishing approaching and waiting vehicles is necessary to effectively deal with non-uniform traffic flows. Additionally, we have shown that splitting up the data into train-validation-test sets is appropriate in traffic signal control and that our generalisation of the Self-Organising Traffic Lights leads to a strong baseline in multi-phase intersections.

Our empirical evaluations on real-world datasets have shown that our proposed RLight method produces stable learning and outperforms the chosen baseline methods by a substantial margin. These results demonstrate the potential of reinforcement learning methods to improve urban traffic flows, thereby reducing travel time and cutting CO2 emissions.

REFERENCES

[1] Chacha Chen, Hua Wei, Nan Xu, Guanjie Zheng, Ming Yang, Yuanhao Xiong, Kai Xu, and Zhenhui Li. 2020. Toward A Thousand Lights: Decentralized Deep Reinforcement Learning for Large-Scale Traffic Signal Control. *Proc. of the AAAI Conference on Artificial Intelligence* 34, 04 (2020), 3414–3421. https://doi.org/10.1609/aaai.v34i04.5744

[2] Seung Bae Cools, Carlos Gershenson, and Bart D'Hooghe. 2008. Self-Organizing Traffic Lights: A Realistic Simulation. *Advanced Information and Knowledge Processing* (2008), 41–50. https://doi.org/10.1007/978-1-84628-982-8_3 arXiv:0610040 [nlin]

[3] M Ebrahim Fouladvand, Zeinab Sadjadi, and M Reza Shaebani. 2004. Optimized traffic flow at a single intersection: traffic responsive signalization. *Journal of Physics A: Mathematical and General* 37, 3 (2004), 561.

[4] Wade Genders and Saiedeh Razavi. 2019. An open-source framework for adaptive traffic signal control. *arXiv* X, X (2019), 1–11. arXiv:1909.00395

[5] Carlos Gershenson. 2004. Self-Organizing Traffic Lights. (2004). arXiv:0411066 [nlin] http://arxiv.org/abs/nlin/0411066

[6] Jan-Torben Girault, Vikash V Gayah, Ilgin Guler, and Monica Menendez. 2016. Exploratory analysis of signal coordination impacts on macroscopic fundamental diagram. *Transportation Research Record* 2560, 1 (2016), 36–46.

[7] Ammar Haydari and Yasin Yilmaz. 2020. Deep reinforcement learning for intelligent transportation systems: A survey. *IEEE Transactions on Intelligent Transportation Systems* (2020).

[8] Pablo Hernandez-Leal, Bilal Kartal, and Matthew E Taylor. 2019. A survey and critique of multiagent deep reinforcement learning. *Autonomous Agents and Multi-Agent Systems* 33, 6 (2019), 750–797.

[9] Xiaoyuan Liang, Xunsheng Du, Guiling Wang, and Zhu Han. 2019. A deep reinforcement learning network for traffic light cycle control. *IEEE Transactions on Vehicular Technology* 68, 2 (2019), 1243–1253.

[10] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. 2015. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971* (2015).

[11] Patrick Mannion, Jim Duggan, and Enda Howley. 2016. An experimental review of reinforcement learning algorithms for adaptive traffic signal control. *Autonomic road transport support systems* (2016), 47–66.

[12] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. 2015. Human-level control through deep reinforcement learning. *Nature* 518, 7540 (2015), 529–533.

[13] Andrew Y Ng, Daishi Harada, and Stuart Russell. 1999. Policy invariance under reward transformations: Theory and application to reward shaping. In *Icml*, Vol. 99. 278–287.

[14] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. 2018. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science* 362, 6419 (2018), 1140–1144.

[15] Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction* (2nd ed.). MIT press.

[16] Richard S Sutton, Doina Precup, and Satinder Singh. 1999. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence* 112, 1-2 (1999), 181–211.

[17] Elise Van der Pol and Frans A Oliehoek. 2016. Coordinated deep reinforcement learners for traffic light control. In *NIPS'16 Workshop on Learning, Inference and Control of Multi-Agent Systems*.

[18] Hua Wei, Nan Xu, Huichu Zhang, Guanjie Zheng, Xinshi Zang, Chacha Chen, Weinan Zhang, Yanmin Zhu, Kai Xu, and Zhenhui Li. 2019. CoLight: Learning network-level cooperation for traffic signal control. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 1913–1922.

[19] Hua Wei, Huaxiu Yao, Guanjie Zheng, and Zhenhui Li. 2018. IntelliLight: A reinforcement learning approach for intelligent traffic light control. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2018), 2496–2505. https://doi.org/10.1145/3219819.3220096

[20] Hua Wei, Guanjie Zheng, Vikash Gayah, and Zhenhui Li. 2019. A survey on traffic signal control methods. *arXiv preprint arXiv:1904.08117* (2019).

[21] Hua Wei, Guanjie Zheng, Huaxiu Yao, and Zhenhui Li. 2018. Intellilight: A reinforcement learning approach for intelligent traffic light control. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2496–2505.

[22] Marco A Wiering. 2000. Multi-agent reinforcement learning for traffic light control. In *Machine Learning: Proceedings of the Seventeenth International Conference (ICML'2000)*. 1151–1158.

[23] Huichu Zhang, Siyuan Feng, Chang Liu, Yaoyao Ding, Yichen Zhu, Zihan Zhou, Weinan Zhang, Yong Yu, Haiming Jin, and Zhenhui Li. 2019. CityFlow: A multi-agent reinforcement learning environment for large scale city traffic scenario. In *The World Wide Web Conference*. 3620–3624.

[24] Dongbin Zhao, Yujie Dai, and Zhen Zhang. 2011. Computational intelligence in urban traffic signal control: A survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 42, 4 (2011), 485–494.

[25] Guanjie Zheng, Xinshi Zang, Nan Xu, Hua Wei, Zhengyao Yu, Vikash Gayah, Kai Xu, and Zhenhui Li. 2019. Diagnosing reinforcement learning for traffic signal control. *arXiv preprint arXiv:1905.04716* (2019).