OPTIMIZING URBAN TRAFFIC FLOW: REINFORCEMENT LEARNING-BASED TRAFFIC LIGHT CONTROL

DEMETRE DZMANASHVILI

Thesis supervisor: ANAIS GARRELL ZULUETA (Department of Automatic Control)

Degree: Master Degree in Artificial Intelligence

Master's thesis

School of Engineering
Universitat Rovira i Virgili (URV)

Faculty of Mathematics
Universitat de Barcelona (UB)

Barcelona School of Informatics (FIB)
Universitat Politècnica de Catalunya (UPC) - BarcelonaTech

Abstract

Contents

1. Introduction

1.1 Overview

Traffic congestion is a persistent global issue, impeding daily commutes as a result of the ever-increasing urban population and transportation demands in cities worldwide [10][18]. One major contributor to this problem is the delay caused by red lights at intersections, where traffic signals typically operate on fixed-time schedules regardless of actual traffic conditions [14]. While such systems are effective in heavily congested areas, they often prove inefficient for low traffic density scenarios, resulting in unnecessary delays and fuel wastage [14].

Recent technological advancements have introduced the Adaptive Traffic Signal Control System, which utilizes sensors embedded in roads to synchronize traffic signals, thus responding to real-time traffic conditions [10]. However, this system's feasibility and cost-effectiveness have been questioned due to the need for embedded road infrastructure and power sources [10]. Additionally, optimizing traffic signal control to minimize delays while ensuring system stability remains a challenge [14].

This thesis aims to address these challenges by proposing a Traffic Control System based on reinforcement learning (RL), an artificial intelligence framework that learns optimal decision policies through continuous adaptation to real-time traffic scenarios. By moving away from fixed-time schedules and incorporating RL, we seek to develop an intelligent traffic control system that efficiently manages traffic flow, reduces environmental impact, such as air pollution and fuel wastage, and enhances road safety [14]. The research focuses on a 4-way intersection, analyzing incoming traffic density to optimize traffic signal control and improve overall transportation efficiency over time.

1.2 Motivation

My personal motivation for embarking on this thesis is deeply rooted in the persistent traffic problems that afflict my home country, Georgia. The congestion and inefficiency of traffic lights on some of the busiest streets in Georgia have long been a source of frustration for me and my fellow citizens. The resulting traffic jams not only waste valuable time but also contribute to environmental issues such as increased air pollution and fuel wastage. Moreover, the heightened risk of accidents in congested traffic conditions underscores the urgency of finding effective solutions.

Beyond my personal experiences, the global need for intelligent traffic control systems has never been more evident. Rapid urbanization and population growth have placed an ever-increasing burden on urban transportation infrastructure. As cities around the world grapple with the challenges posed by burgeoning traffic volumes, there is a pressing demand for innovative and adaptive solutions.

In this context, my motivation converges with a broader societal need for intelligent traffic light systems. These systems have the potential to revolutionize urban transportation by dynamically managing traffic flows, reducing congestion, and mitigating environmental concerns. By harnessing the power of reinforcement learning and artificial intelligence, I aim to contribute to the development of intelligent traffic control systems that can serve as a model for cities worldwide.

Through this research endeavor, I aspire to make a meaningful impact by fostering more efficient and sustainable urban transportation systems. By optimizing traffic light control, I seek not only to alleviate the traffic woes in my homeland but also to offer a scalable solution that addresses the global imperative for intelligent traffic management.

1.3 Objectives

The primary objectives of this thesis encompass a comprehensive investigation into optimizing urban traffic flow through a multi-faceted approach. These objectives are designed to address the complexities of traffic management, improve realism in simulations, and assess the performance of cutting-edge algorithms and baseline controllers. The key objectives are as follows:

1.3.1 Creation of Vake Map in SUMO

The first objective is to develop a complex and representative urban traffic simulation environment within the Simulation of Urban MObility (SUMO) framework. This entails the creation of a detailed Vake map, capturing the intricacies of traffic infrastructure, including road networks, intersections, and traffic lanes. The map should accurately reflect the real-world urban environment under investigation.

1.3.2 Generation of Realistic Traffic Patterns

To enhance the realism of the simulations, real-world traffic patterns are essential. This objective involves collecting real-time traffic data from the chosen location and meticulously recording the timing and behavior of traffic lights. The gathered data will then be integrated into the simulation environment to replicate actual traffic conditions.

1.3.3 Utilization of State-of-the-Art Algorithms

The core of this research lies in the exploration, implementation and adaptation of state-of-the-art traffic signal control algorithms. The following algorithms will be employed:

- **IDQN**: Implementing this deep reinforcement learning algorithm for traffic signal control, which has shown promise in optimizing signal timings.
- **IPPO**: Utilizing IPPO as another reinforcement learning algorithm to investigate its effectiveness in traffic management.
- **MPLIGHT**: Exploring MPLIGHT, a multi-phase control algorithm designed to adapt traffic signals dynamically.
- **FMA2C**: Investigating the potential of FMA2C for cooperative multi-agent traffic signal control.

1.3.4 Comparison with Baseline Controllers

To evaluate the performance of the selected state-of-the-art algorithms, this objective involves implementing and assessing the following baseline controllers:

- **Fixed Time Control**: A traditional control strategy with fixed signal timings that do not adapt to real-time traffic conditions.
- **Max-Pressure Control**: Implementing this controller, which focuses on minimizing congestion by prioritizing the most congested lanes at intersections.
- **Greedy Control**: Assessing the performance of a basic greedy controller that makes decisions based on immediate traffic conditions.

1.3.5 Comparative Analysis and Conclusions

Upon completing the simulations and experiments, the results from the various traffic signal control algorithms and baseline controllers will be rigorously analyzed and compared. The objective is to draw meaningful conclusions regarding the effectiveness of each approach in optimizing urban traffic flow. The research aims to provide insights into the potential for intelligent traffic management systems to alleviate congestion, improve efficiency, and reduce environmental impacts.

By achieving these objectives, this thesis seeks to contribute valuable knowledge to the field of urban traffic optimization and provide practical recommendations for enhancing traffic signal control systems in real-world urban settings.

2. Background

2.1 Reinforcement Learning

Reinforcement Learning (RL) is a paradigm in which an agent learns to make decisions by interacting with its environment. In the RL framework, the environment is often modeled as a Markov decision process (MDP), characterized by key components:

- S – the state space,
- A – the action space,
- $P(s_t, a, s_{t+1})$ – the transition function, mapping from state s_t and action a to the next state s_{t+1} with probabilities in the range [0, 1],
- $R(s, a)$ – the reward function, which assigns a real-valued reward to each state-action pair,
- γ – the discount factor, controlling the trade-off between immediate and future rewards.

The RL agent operates based on a policy π, which maps states to actions, i.e., $\pi : S \to A$. When the agent selects an action a_t in the current state s_t, it impacts the environment, leading to a new state s_{t+1} and an immediate reward r_t.

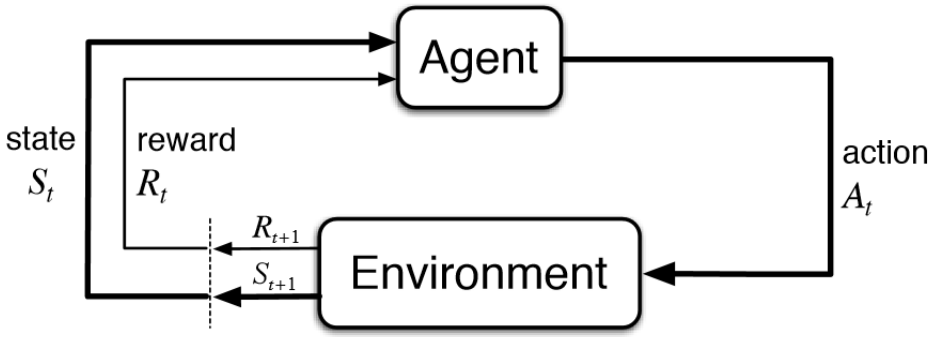

Figure 1: Reinforcement Learning Framework

The primary objective of the RL agent is to maximize the expected sum of discounted rewards, denoted as $J^\pi = \sum_{t=0}^{\infty} \gamma^t r_t$. The optimal policy, denoted as π^*, is the one that maximizes this objective.

There are various approaches for training a policy using RL:

- **Value-Based Approach:** This approach focuses on estimating the expected future utility from states (state value) or from action-state pairs (action value or q-value). The control policy is then directed towards actions or states that maximize the expected utility (J^π). A prominent example is the model-free deep Q-learning algorithm [13].

- **Policy-Gradient Approach:** In this approach, a policy is defined through a parameterized differential equation, and the parameters are updated incrementally following the policy gradient. These updates aim to achieve favorable outcomes as measured by the reward function. Estimations of state or action values are often used to define these favorable outcomes. This approach is commonly referred to as an actor-critic approach.

- **Actor-Critic Approach:** Actor-critic methods combine elements of both value-based and policy-gradient approaches. An actor (policy) learns to make decisions, while a critic (value function) evaluates these decisions. A state-of-the-art example of an actor-critic algorithm is the proximal policy optimization (PPO) algorithm [16].

These RL approaches provide a framework for training intelligent agents to make decisions in complex and dynamic environments, making them highly relevant to optimizing traffic signal control in urban settings.

2.2 Traffic Signal Control as an MDP

In the realm of traffic engineering, a signalized intersection represents a complex network of incoming and outgoing roads, each comprising one or more lanes. To efficiently manage traffic flow at such intersections, a set of phases, denoted as Φ, is defined. Each phase, $\varphi \in \Phi$, corresponds to a specific traffic movement through the intersection, as illustrated in Figure 2. It's crucial to note that two phases are considered conflicting if they cannot be simultaneously enabled due to intersecting traffic movements.

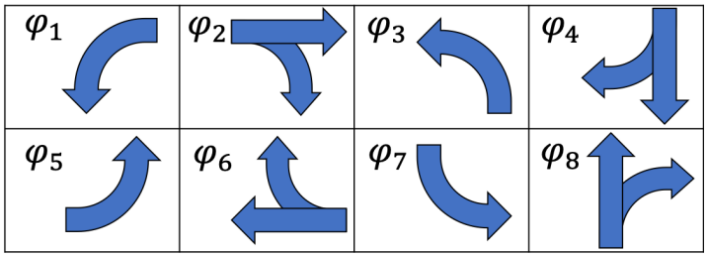

Figure 2: Example of Phases at a Signalized Intersection[3]

At each discrete time step, a signal controller is tasked with selecting a combination of non-conflicting phases to enable. The objective is to optimize a long-term objective function, which may vary depending on specific goals and constraints. In the context of Reinforcement Learning (RL)-based controllers, the signalized intersection environment is commonly modeled as a Markov Decision Process (MDP), with the following components:

- **State Space (S):** The state space encompasses the state of incoming traffic and the currently enabled phases. The definition of the state varies among studies, reflecting differing sensing capabilities. Some works assume state-of-the-art traffic sensing technologies, providing high-resolution data on incoming traffic, including information such as the number of approaching vehicles, accumulated waiting time, the number of stopped vehicles, and the average speed of approaching vehicles [4]. Others adopt less informative sensing capabilities, such as observing only the stopped queue length per lane [12] or solely the waiting time of the first vehicle in the queue [17].

- **Action Space (A):** In each time-step, the controller selects a set of non-conflicting phases to be assigned the right-of-passage (green light). If the chosen phases are different from the currently enabled ones, a mandatory yellow phase is enforced ny the system for a predefined duration. It's important to note that assigning yellow phases is not the part of the action space, it is a constraint imposed by the environment.

- **Transition Function (P):** The transition function describes the progression of traffic following the signal assignment. This progression can be defined within a simulated environment, as

commonly done in research [12], or based on real-world traffic progression in practical implementations.

- **Reward Function (R):** The reward function serves as a critical component in RL-based signal control. Different reward functions have been proposed in the literature. Commonly used reward functions include (minus) queue length summed over all incoming lanes [19], (minus) total delays imposed by the intersection [17], (minus) waiting time at the intersection [12], and (minus) traffic pressure [4]. These reward functions reflect various aspects of traffic performance and congestion alleviation.

The modeling of traffic signal control as an MDP provides a foundation for applying RL techniques to optimize signal operation, ultimately contributing to more efficient and adaptive traffic management strategies.

2.3 Evaluation environments for RL-based signal controllers

According to [3], previous research in the field of traffic signal control has often relied on custom-made scenarios tailored for evaluating specific Reinforcement Learning (RL) algorithms. For instance, Jinming and Feng (2020) utilized the well-established Simulation of Urban Mobility (SUMO) environment for their experiments. SUMO enjoys widespread acceptance within the transportation community and serves as a reasonable testbed choice for such studies. However, it's worth noting that Jinming and Feng's reported scenario, based on the real-world city of Monaco, was a modified version. This modified scenario included 18 synthetic traffic signals beyond the official "Monaco SUMO Traffic (MoST)" scenario and incorporated non-validated inflated traffic demands [6].

Another notable simulation testbed, CityFlow, was presented by Zhang et al.[20]. However, CityFlow has two primary limitations. Firstly, unlike SUMO, CityFlow lacks rigorous calibration and evaluation within the general transportation community. Although it claims to produce equivalent output as SUMO, this claim is primarily based on results from simplified grid network scenarios. Secondly, while CityFlow offers the Manhattan, New York network as a common benchmark scenario, the support for this scenario's representation of real-world city layouts and demands is limited.

Additionally, some relevant publications have conducted evaluations using the Autonomous Intersection Management (AIM) simulator. The primary drawback of the AIM simulator lies in its lack of traffic scenarios based on real-world cities. AIM typically generates simple grid networks with symmetric intersections. While one might draw parallels between such grid networks and the road layout in Manhattan, New York, a more in-depth analysis of traffic trends is needed to substantiate such claims and their relevance to the real world [15][7][12].

3. Related Work

3.1 Reinforced Signal Control (RESCO)

In this section, we review related work in the field of traffic signal control, with a focus on the Reinforced Signal Control (RESCO) toolkit, which serves as a baseline for our research.

The RESCO toolkit is a standard Reinforcement Learning (RL) traffic signal control testbed designed to achieve several key objectives:

1. Provide benchmark single and multi-agent signal control tasks based on well-established traffic scenarios.
2. Offer an OpenAI GYM interface within the testbed environment to facilitate the deployment of state-of-the-art RL algorithms.
3. Deliver a standardized implementation of state-of-the-art RL-based signal control algorithms.

RESCO is open-source and freely available under the GNU General Public License 3. It is built on top of SUMO-RL [1] and can be accessed on GitHub at `github.com/Pi-Star-Lab/RESCO`. The embedded traffic scenarios within RESCO have their own licensing, with Cologne-based scenarios under Creative Commons BY-NC-SA and Ingolstadt-based scenarios under the GNU General Public License 3.

State and Action Space

RESCO accommodates a wide range of sensing assumptions, including advanced sensing capabilities [6]. Users can select subsets of state features based on specific sensing assumptions. Features include information such as stopped vehicles' queue length, the number of approaching vehicles, total waiting time for stopped vehicles, and more, at the level of state, intersection, and lane. Additionally, users can define the effective sensing distance during initialization.

The action space in RESCO encompasses sets of non-conflicting phase combinations, following the methodology described in Section 2.2 of the RESCO documentation [6]. By default, actions are chosen for the next 10 seconds of simulation, with the first 3 seconds reserved for yellow signals, if necessary.

Reward Metrics

RESCO offers flexibility in terms of reward metrics. Users can designate any of the reward metrics defined in Section 2.2 of the RESCO documentation[6] or create custom weighted combinations of these metrics. When initializing a control task, users can pass a weight vector that assigns weights to different metrics in the reward function. These weights correspond to various aspects, such as system travel time, signal-induced delays, total waiting time at intersections, average queue length, and traffic pressure.

Benchmark Control Tasks

The signal control benchmark tasks in RESCO are based on two well-established SUMO scenarios: "TAPAS Cologne" and "InTAS" [15, 11]. These scenarios represent traffic within real-world cities, namely, Cologne and Ingolstadt in Germany. They include road network layouts and calibrated demands, making them suitable for comprehensive evaluation. RESCO defines three benchmark control tasks for each traffic scenario:

1. Controlling a single main intersection.
2. Coordinated control of multiple intersections along an arterial corridor.
3. Coordinated control of multiple intersections within a congested area (downtown).

Benchmark Algorithms

RESCO provides three baseline controllers and several RL-based controllers for comparative evaluation:

1. **Baseline Controllers**:
   (a) Fixed-time (Pre-timed) control, where phase combinations are enabled for fixed durations following predefined cycles, that was recorded physically from the real-world traffic signal controller.
   (b) Max-pressure control, which selects the phase combination with the maximum joint pressure. [4]
   (c) Greedy control, which chooses the phase combination with the maximum joint queue length and approaching vehicle count.[12]
2. **RL Controllers**:
   (a) IDQN (Independent DQN agents), employing convolutional layers for lane aggregation[2].
   (b) IPPO, which utilizes a deep neural network similar to IDQN[2].
   (c) MPLight, based on the FRAP open-source implementation, ChainerRL DQN[8], and pressure sensing[21].
   (d) Extended MPLight (MPLight*), an enhanced version of MPLight with additional sensing information.
   (e) FMA2C, built on top of the MA2C open-source implementation[5].

In each of the RL-based controllers, specific learning algorithms and hyperparameters are applied, allowing for a comprehensive evaluation of their performance [2, 4, 5, 12, 21].

In the case of IDQN, IPPO, and MPLight, the implementation of the learning algorithm is invoked directly from the ChainerRL [8] and the Preferred RL [9] libraries that is successor of ChainerRL, and customized to align with our specific map and requirements.

4. State of the Art

4.1 IDQN

Reinforcement Learning (RL) is a prominent area of machine learning where agents learn to make sequential decisions by interacting with an environment. DQN, short for Deep Q-Network, is a fundamental algorithm in RL that leverages deep neural networks to approximate optimal action-value functions.

4.1.1 Deep Q-Network (DQN)

DQN, proposed by Mnih et al. [13], is designed to address the challenges of learning Q-values in high-dimensional state spaces. It combines Q-learning, a well-established RL algorithm, with deep neural networks.

The Q-value, denoted as $Q(s, a)$, represents the expected cumulative reward when taking action a in state s. DQN approximates this Q-value using a deep neural network with parameters θ. The Q-network is trained to minimize the temporal difference (TD) error:

$$\delta = Q(s, a; \theta) - (r + \gamma \max_{a'} Q(s', a'; \theta^-))$$

Where:

δ - TD error

$Q(s, a; \theta)$ - Q-value predicted by the network

r - Immediate reward

γ - Discount factor

$Q(s', a'; \theta^-)$ - Target Q-value predicted by a target network with parameters θ^-

DQN employs experience replay and a target network to stabilize training. Experience replay stores past experiences in a replay buffer and samples mini-batches for training, breaking the temporal correlation in the data. The target network provides stable target Q-values for the TD error.

4.1.2 Independent Deep Q-Networks (IDQN)

IDQN is an extension of DQN tailored for multi-agent RL scenarios, where multiple agents operate independently to optimize their actions. Each agent in IDQN maintains its own Q-network and replay buffer.

The Q-value update rule in IDQN remains similar to DQN, but it is extended to accommodate multiple agents:

$$\delta = Q_i(s, a_i; \theta_i) - (r + \gamma \max_{a'} Q_i(s', a'; \theta^-))$$

Where:

δ - TD error for agent i

$Q_i(s, a_i; \theta_i)$ - Q-value predicted by agent $i's$ network

r - Immediate reward

γ - Discount factor

$Q_i(s', a'; \theta^-)$ - Target Q-value predicted by agent $i's$ target network

IDQN facilitates decentralized decision-making among multiple agents, making it suitable for scenarios involving cooperation or competition among agents.

To explore IDQN in more detail, the following paper[2] provide comprehensive insights into its theory and applications

4.2 IPPO

Proximal Policy Optimization (PPO) is a state-of-the-art reinforcement learning algorithm designed for optimizing parameterized policies in complex environments. IPPO, short for Independent Proximal Policy Optimization, is an extension of PPO tailored for multi-agent reinforcement learning scenarios, where multiple agents learn independently.

4.2.1 Proximal Policy Optimization (PPO)

Introduced by Schulman et al. [16], PPO addresses several challenges in policy optimization. It aims to maximize the expected cumulative reward while ensuring that policy updates are not too large, preventing catastrophic policy changes. PPO achieves this through the following objectives:

Objective Function

PPO optimizes a surrogate objective function that balances the trade-off between policy improvement and policy constraint. The objective function is given as:

$$\mathcal{L}(\theta) = \mathbb{E}\left[\min\left(r_t(\theta)\hat{A}_t, \text{clip}\left(r_t(\theta), 1-\epsilon, 1+\epsilon\right)\hat{A}_t\right)\right]$$

Where:

$\mathcal{L}(\theta)$ - Surrogate objective function

θ - Policy parameters

$r_t(\theta) = \dfrac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}$ - Importance ratio

$\hat{A}_t$ - Advantage estimate

ϵ - Clip parameter

PPO optimizes this objective function using stochastic gradient ascent.

Trust Region

PPO introduces a trust region constraint by clipping the surrogate objective. The clip function ensures that policy updates do not deviate significantly from the previous policy:

$$\text{clip}(x, a, b) = \begin{cases} x, & \text{if } x \in [a, b] \\ a, & \text{if } x < a \\ b, & \text{if } x > b \end{cases}$$

PPO efficiently balances policy updates to ensure stability and improved performance.

4.2.2 Independent Proximal Policy Optimization (IPPO)

IPPO extends the PPO algorithm for multi-agent RL scenarios, where multiple agents learn independently. Each agent in IPPO maintains its own policy and operates in the environment. IPPO's objective function for agent i remains similar to PPO:

$$\mathcal{L}_i(\theta_i) = \mathbb{E}\left[\min\left(r_t(\theta_i)\hat{A}_t^i, \text{clip}\left(r_t(\theta_i), 1-\epsilon, 1+\epsilon\right)\hat{A}_t^i\right)\right]$$

Where:

$\mathcal{L}_i(\theta_i)$ - Surrogate objective function for agent i

θ_i - Policy parameters for agent i

$r_t(\theta_i) = \dfrac{\pi_{\theta_i}(a_t^i|s_t)}{\pi_{\theta_{i_{\text{old}}}}(a_t^i|s_t)}$ - Importance ratio for agent i

$\hat{A}_t^i$ - Advantage estimate for agent i

IPPO facilitates decentralized learning among multiple agents, making it suitable for scenarios involving independent agents with their policies.

To explore IPPO in more detail, the following paper[2] provide comprehensive insights into its theory and applications

4.3 MPLight

MPLight[4] is a traffic light control system that utilizes the concept of pressure to coordinate multiple intersections efficiently. It operates by considering the pressure, which is the difference in queue lengths from incoming lanes of an intersection and the queue length on a downstream intersection's receiving lane. MPLight is designed to optimize traffic flow and reduce congestion in urban environments.

In MPLight, pressure serves as a critical metric for traffic signal coordination. It is calculated as the difference between the queue lengths of vehicles waiting to enter an intersection and the queue length on the downstream intersection's receiving lane. By considering pressure, MPLight aims to balance the traffic load across multiple intersections.

Chen et al. introduced MPLight as an approach to traffic light control that leverages reinforcement learning techniques. They utilized Deep Q-Networks (DQN) as the underlying framework for making traffic signal decisions. In this setup, a DQN agent is shared across all intersections.

In MPLight, pressure is not only used as a coordination metric but also as both the state and reward for the DQN agent. The state of the agent at a given time step includes information about the pressure values for all relevant intersections. The reward signal is derived from pressure differences and is used to guide the learning process of the DQN agent.

Chen et al.[4] reported significant improvements in traffic flow and travel times when implementing MPLight compared to existing methods. Specifically, MPLight achieved up to a 19.2% improvement in travel times over the next best compared method, PressLight.

4.4 FMA2C

FMA2C[5] is an advanced approach to traffic signal control that utilizes a hierarchical framework to optimize traffic flow in urban environments. It builds upon the prior work of MA2C (Multi-Agent Advantage Actor-Critic) by introducing managing agents to coordinate and oversee workers responsible for signal control at intersections.

4.4.1 Basic Concepts

Workers (Intersection-Level Agents)

In FMA2C, the core agents responsible for signal control at intersections are called workers. Each worker operates independently as an advantage actor-critic agent. The workers are tasked with making real-time decisions regarding traffic signal timings at their respective intersections.

Managing Agents (Region-Level Agents)

FMA2C introduces managing agents, which operate at a higher level of hierarchy compared to workers. Each managing agent is responsible for a specific region or area within the traffic network. These managing agents oversee multiple workers and have the responsibility of optimizing traffic flow within their assigned regions.

4.4.2 Hierarchical Reinforcement Learning

FMA2C leverages hierarchical reinforcement learning to improve traffic signal coordination. The hierarchy involves two levels: managing agents at the top level and workers at the lower level.

Managing Agent Training

Managing agents are trained to optimize traffic flow within their assigned regions. They receive high-level traffic-related goals and objectives, such as minimizing congestion or maximizing traffic throughput. The managing agents use these goals to make region-level decisions.

The training of managing agents can be formulated as a reinforcement learning problem, where the managing agent learns a policy π_m to maximize a region-specific objective function:

$$J_m(\pi_m) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R_t^m\right]$$

Where:

$J_m(\pi_m)$ - Expected cumulative reward for managing agent m
π_m - Policy of managing agent m
γ - Discount factor
R_t^m - Region-specific reward at time step t

Worker Training

Workers, on the other hand, are trained to incorporate the high-level goals set by their respective managing agents into their local decision-making process. This hierarchical training ensures that workers align their actions with the broader objectives of traffic flow optimization.

The training of workers also involves reinforcement learning, where each worker learns a policy π_w to maximize its intersection-specific objective function:

$$J_w(\pi_w) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R_t^w\right]$$

Where:

$J_w(\pi_w)$ - Expected cumulative reward for worker w
π_w - Policy of worker w
γ - Discount factor
R_t^w - Intersection-specific reward at time step t

4.4.3 Coordination Mechanisms

FMA2C employs various coordination mechanisms between managing agents and workers to ensure effective traffic signal control. These mechanisms may include communication of high-level goals, reward sharing, and coordination through a central mechanism.

4.4.4 Performance Improvement

FMA2C aims to improve traffic flow and reduce congestion by introducing a hierarchical framework that allows for coordinated decision-making at both the region and intersection levels. By aligning the actions of workers with the goals of managing agents, FMA2C seeks to optimize traffic signal timings efficiently.

5. Methodology

6. Experiments

7. Comparison

8. Conclusion

9. Future Work

References

[1] L. N. Alegre. Sumo-rl, 2019.

[2] J. Ault, J. Hanna, and G. Sharon. Learning an interpretable traffic signal control policy. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS 2020)*. International Foundation for Autonomous Agents and Multiagent Systems, May 2020.

[3] James Ault and Guni Sharon. Reinforcement learning benchmarks for traffic signal control. In *Proceedings of the Thirty-fifth Conference on Neural Information Processing Systems (NeurIPS 2021) Datasets and Benchmarks Track*, December 2021.

[4] Chacha Chen, Hongyu Wei, Nan Xu, Guanjie Zheng, Ming Yang, Yilin Xiong, Kewei Xu, and Zongzhang Li. Toward a thousand lights: Decentralized deep reinforcement learning for large-scale traffic signal control. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3414–3421, 2020.

[5] T. Chu, J. Wang, L. Codecà, and Z. Li. Multi-agent deep reinforcement learning for large-scale traffic signal control. *IEEE Transactions on Intelligent Transportation Systems*, 21(3):1086–1095, 2019.

[6] L. Codeca and J. Härri. Monaco sumo traffic (most) scenario: A 3d mobility scenario for cooperative its. In *SUMO 2018, SUMO User Conference, Simulating Autonomous and Intermodal Transport Systems, May 14-16, 2018, Berlin, Germany*, 2018.

[7] K. Dresner and P. Stone. A multiagent approach to autonomous intersection management. *Journal of artificial intelligence research*, 31:591–656, 2008.

[8] Yasuhiro Fujita, Prabhat Nagarajan, Toshiki Kataoka, and Takahiro Ishikawa. Chainerrl: A deep reinforcement learning library. *Journal of Machine Learning Research*, 22(77):1–14, 2021.

[9] Yasuhiro Fujita, Prabhat Nagarajan, Toshiki Kataoka, and Takahiro Ishikawa. Chainerrl: A deep reinforcement learning library. *Journal of Machine Learning Research*, 22(77):1–14, 2021.

[10] D. M. Levinson. Speed and delay on signalized arterials. *Journal of Transportation Engineering*, 124(3):258–263, 1998.

[11] S. C. Lobo, S. Neumeier, E. M. Fernandez, and C. Facchi. Intas-the ingolstadt traffic scenario for sumo, 2020.

[12] Jian Ma and Fan Wu. Feudal multi-agent deep reinforcement learning for traffic signal control. In *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 816–824, 2020.

[13] 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. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

[14] S. S. Mousavi, M. Schukat, and E. Howley. Traffic light control using deep policy-gradient and value-function-based reinforcement learning. *IET Intelligent Transport Systems*, 11(7):417–423, 2017.

[15] T. T. Pham, T. Brys, M. E. Taylor, T. Brys, M. M. Drugan, P. Bosman, M.-D. Cock, C. Lazar, L. Demarchi, and D. Steenhoff. Learning coordinated traffic light control. In *Proceedings of the Adaptive and Learning Agents workshop (at AAMAS-13)*, volume 10, pages 1196–1201, 2013.

[16] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[17] S. M. A. Shabestary and Baher Abdulhai. Deep learning vs. discrete reinforcement learning for adaptive traffic signal control. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 286–293, 2018.

[18] A. Tirachini. Estimation of travel time and the benefits of upgrading the fare payment technology in urban bus services. *Transportation Research Part C: Emerging Technologies*, 30:239–256, 2013.

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

[20] H. Zhang, S. Feng, C. Liu, Y. Ding, Y. Zhu, Z. Zhou, W. Zhang, Y. Yu, H. Jin, and Z. Li. Cityflow: A multi-agent reinforcement learning environment for large scale city traffic scenario. In *The World Wide Web Conference*, pages 3620–3624, 2019.

[21] G. Zheng, Y. Xiong, X. Zang, J. Feng, H. Wei, H. Zhang, Y. Li, K. Xu, and Z. Li. Learning phase competition for traffic signal control. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 1963–1972, 2019.