MATLight: Traffic Signal Coordinated Control Algorithm Based on Heterogeneous-Agent Mirror Learning With Transformer

Extended Abstract

Haipeng Zhang
Guangxi University of Science and Technology
LiuZhou, China
221068705@stdmail.gxust.edu.cn

Zhiwen Wang*
Guangxi University of Science and Technology
LiuZhou, China
wzw69@126.com

Na Li
Guangxi University of Science and Technology
LiuZhou, China
512164081@qq.com

ABSTRACT

In order to better handle the issue of real-time multi-intersection traffic signal coordinated control, we expect that multi-agent decision-making can benefit from the advantages of large sequence models. In this paper, we propose a method for multi-intersection traffic signal coordinated control based on heterogeneous-agent mirror learning and Transformer to sequential multi-agent cooperative decision. First, multi-intersection traffic signal control is modeled as a sequential problem based on the heterogeneous-agent mirror learning framework. We convert real-time multi-intersection traffic signal control into a multi-agent sequential decision-making process. It completely capitalizes on the surprising connection between the multi-agent reinforcement learning decision process and sequential model prediction. And it provides strong theoretical guarantees. Then the Transformer sequence model is used to cleverly implement the sequential update scheme to learn the optimal traffic signal coordination control strategy online with a new training paradigm. The proposed method has theoretical policy promotion and convergence, alleviates the credit assignment problem in the process of multi-intersection traffic signal coordinated control, reduces the complexity of the joint policy optimization, and improves the learning efficiency of few-shot samples. We used LibSignal, a unified framework for traffic signal control tasks, for comparison testing. According to experimental results, our method can significantly improve the efficiency and performance of few-shot online learning, outperform the baseline methods in both network-level and arterial coordination, and simplify the complexity of algorithm implementation.

KEYWORDS

Sequential decision; Traffic signal coordinated control; Multi-agent reinforcement learning; Transformer; Few-shot; Online learning

ACM Reference Format:
Haipeng Zhang, Zhiwen Wang, and Na Li. 2024. MATLight: Traffic Signal Coordinated Control Algorithm Based on Heterogeneous-Agent Mirror Learning With Transformer: Extended Abstract. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), Auckland, New Zealand, May 6 – 10, 2024*, IFAAMAS, 3 pages.

1 INTRODUCTION

The current traffic signal control (TSC) scheme in multi-agent reinforcement learning (MARL) has glaring faults. First, the scheme must accurately attribute the agent's specific acts to the environment's overall rewards. But they lack both direct advice on policy optimization and clear insights into it. Therefore, it is difficult to guarantee that the joint policy and the policy of a single agent are both improved. Furthermore, the current update scheme could cause training instability and non-convergence because of the high variance of their estimations. Nowadays researchers have proposed various generic solutions[4, 7, 11, 12] to accelerate model learning on new data. But there are no theoretical assurances of policy monotonic improvement or convergence in the majority of MARL methods. It is essential to create new mechanisms to reduce traffic congestion and improve transportation effectiveness.

In this work, we propose a coordinated control algorithm for sequential multi-intersection traffic signal named MATLight. Our algorithm ensures that the sequential modeling multi-intersection TSC model, regardless of the number of intersections, is an effective online learning few-shot learner on novel tasks and hence has a more excellent learning capability. The experimental results show that our solution can greatly enhance performance while reducing the complexity of the TSC algorithm's implementation. The following is a summary of the contributions made by our work:

(1) We innovatively formulate traffic signal coordinated control as a multi-agent sequence decision process based on Heterogeneous-Agent Mirror Learning (HAML) [3] framework. It can be online learning.
(2) We design and build an efficient online few-shot TSC model combined with transportation theory.
(3) Our method combines the Transformer[6] architecture with the proximal policy optimization (PPO) algorithm[5]. All our comparison experiments are performed on LibSignal to obtain uniform and fair comparison results. And our method performs better than baselines.

2 METHOD

According to the sequential property of the framework, we directly treat multiple intersections as a sequence of agents, and the Transformer fits well with the sequential update scheme of

*Corresponding Author

Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), N. Alechina, V. Dignum, M. Dastani, J.S. Sichman (eds.), May 6 – 10, 2024, Auckland, New Zealand. © 2024 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

the framework.We combine the basic principles of the PPO algorithm and implement TSC by using Transformer-based multi-agent trust-region learning derived from the HAML framework. Through the continuous training of the Transformer, agents can learn the optimal coordinated control policy of traffic signals at multiple intersections. Moreover, HAML provides us with theoretical guarantees for policy improvement and convergence, and Transformer can perform parallel computing to speed up the training process.

- *Observation*: The agent's observation is chosen to be the length of the incoming lane's queue [9].
- *Action*: The agent chooses one of the eight non-conflicting candidate phases s at time step t.
- *Reward*: The reward $r^i = -P_i$ of agent i is defined as the negative value of the pressure [8] at the intersection.

HAPPO is derived from HAML when the HADF $\mathfrak{D}^{i,\nu}$ is $\mathbb{E}_{\boldsymbol{a}^{i_{1:m}} \sim \boldsymbol{\pi}_{\text{old}}^{i_{1:m}}}$ [$\text{ReLU}\left(\left[\text{r}\left(\hat{\pi}^{i_m}\right) - \text{clip}\left(\text{r}\left(\hat{\pi}^{i_m}\right), 1 \pm \epsilon\right)\right] A_{\boldsymbol{\pi}_{\text{old}}}^{i_{1:m}}\left(s, \boldsymbol{a}^{i_{1:m}}\right)\right)\Big]$, neighborhood operator is the policy-space of every agent i, and sampling distribution $\beta_{\boldsymbol{\pi}}$ is equal to drift distribution $\nu_{\boldsymbol{\pi},\hat{\pi}}$.

Several researches[1, 2, 13] have proved the great potential of Transformer in MAS. The MAT architecture[10] implements the sequential update scheme of HAML and conforms to its basic principles. The mapping between the multi-agent input observation sequence $\left(o^{i_1}, ..., o^{i_n}\right)$ and the multi-agent output action sequence $\left(a^{i_1}, ..., a^{i_n}\right)$ can be considered a sequence modeling task similar to machine translation. So we can execute cooperative multi-agent decision-making tasks with a large SM like the Transformer. The action a^{i_m} is dependent on the decisions of all former agents $\boldsymbol{a}^{i_{1:m-1}}$. The MATLight comprises two components: a decoder that autonomously generates actions for each agent and an encoder that produces representations of the joint observations. Each agent's actions are produced in an auto-regressive manner.

The Transformer's Encoder. The output of the observation as $\left(\hat{o}^{i_1}, ..., \hat{o}^{i_n}\right)$ by encoder not only encodes the information of the agents $(i_1, ..., i_n)$ but also the high-level interrelationships that reflect the agents' interactions. During the training phase, the encoder is utilized to approximate the value function with the goal of minimizing the Bellman error by Eq. (1).

$$L(\phi) = \frac{1}{Tn} \sum_{m=1}^{n} \sum_{t=0}^{T-1} [R(\boldsymbol{o}_t, \boldsymbol{a}_t) + \gamma V_{\bar{\phi}}(\hat{o}_{t+1}^{i_m}) - V_{\phi}(\hat{o}_t^{i_m})]^2. \tag{1}$$

The Transformer's Decoder. The output of the last decoding block is a joint action sequence representation $\left\{\hat{\boldsymbol{a}}^{i_{0:j-1}}\right\}_{j=1}^{m}$. It is input into an MLP that outputs the probability distribution of the i_m action, namely the policy $\pi_\theta^{i_m}\left(a^{i_m} \mid \hat{\boldsymbol{o}}^{i_{1:n}}, \hat{\boldsymbol{a}}^{i_{1:m-1}}\right)$. To train the decoder, we use Eq. (2) to minimize the following objective of the proximal policy optimization clip algorithm (PPO-clip).

$$\begin{aligned} L(\theta) &= -\frac{1}{Tn} \sum_{m=1}^{n} \sum_{t=0}^{T-1} \min(\text{r}_t^{i_m}(\theta)\hat{A}_t, \text{clip}(\text{r}_t^{i_m}(\theta), 1 \pm \epsilon)\hat{A}_t) \\ \text{where } \text{r}_t^{i_m}(\theta) &= \frac{\pi_\theta^{i_m}(a_t^{i_m} \mid \hat{\boldsymbol{o}}_t^{i_{1:n}}, \hat{\boldsymbol{a}}_t^{i_{1:m-1}})}{\pi_{\theta_{\text{old}}}^{i_m}(a_t^{i_m} \mid \hat{\boldsymbol{o}}_t^{i_{1:n}}, \hat{\boldsymbol{a}}_t^{i_{1:m-1}})}. \end{aligned} \tag{2}$$

3 EXPERIMENTS

Coordinated control of multi-intersection traffic signal can be regarded as a sequential decision process. MATLight is compared against baseline methods on several publicly available datasets. To test the stability, the experiment of each RL algorithm on each dataset is repeated 5 times. The findings of the experiment indicate that our algorithm works better on these publicly available datasets than baseline methods. We set the memory capacity of the replay buffer for PressLight, CoLight, and MPLight to 360. It is the same number of decisions made by the simulation in an hour. MATLight has only one full 60-minute trajectory for each policy update. This makes RL models learn in a few-shot space. In this scenario, MATLight's sample efficiency is higher than that of CoLight, MPLight, and PressLight.

We utilize average traveling time (ATT) to evaluate the performance of various TSC models. The total time (in seconds) it takes for all vehicles to arrive and exit the region is averaged to calculate the average traveling time. It is the most common performance metric for TSC problems in the traffic domain. The results of comparing the traditional and RL methods for the ATT are shown in Table 1. According to the experimental results, our model outperforms the comparative baseline methods. Moreover, the convergence speed of MATLight is also better than the baseline methods.

Table 1: Evaluations of different methods

Models	Average Traveling Time(s)			
	Hangzhou$_{4\times4}$	**Jinan$_{3\times4}$**	6 × 1 Flat	6 × 1 Peak
FixedTime	575.56	425.31	149.36	162.51
MaxPressure	365.06	328.96	99.41	99.66
SOTL	354.13	333.58	117.22	120.40
CoLight	343.54	682.42	98.20	98.73
MPLight	334.54	311.21	96.41	98.05
PressLight	436.81	430.16	110.92	106.66
MATLight	**322.08**	**283.75**	**93.59**	**97.05**

4 CONCLUSION

In this paper, we formulate the coordinated control of multi-intersection traffic signal as a sequential decision-making process, design an algorithm based on heterogeneous-agent mirror learning and Transformer, that is, MATLight, and cleverly implement the sequential multi-agent decision-making scheme by using Transformer, generating a set of the best possible actions for the agent sequence. Judging from the comparative experimental results on Hangzhou, Jinan, and arterial datasets, our method shows good performance and generalization ability in few-shot online learning, which is suitable for practical application. The TSC task can be expressed as an online sequential multi-agent cooperative decision process, which brings a new research direction.

ACKNOWLEDGMENTS

This work was funded by the National Natural Science Foundation of China, grant number 61962007, 62266009; Guangxi Key Laboratory of Big Data in Finance and Economics, Grant No.FEDOP2022A06.

REFERENCES

[1] Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. 2021. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems* 34 (2021), 15084–15097.

[2] Siyi Hu, Fengda Zhu, Xiaojun Chang, and Xiaodan Liang. 2021. UPDET: Universal multi-agent reinforcement learning via policy decoupling with transformers. In *Proceedings of the 9th International Conference on Learning Representations (ICLR 2021)*. 1–15.

[3] Jakub Grudzien Kuba, Xidong Feng, Shiyao Ding, Hao Dong, Jun Wang, and Yaodong Yang. 2022. Heterogeneous-agent mirror learning: A continuum of solutions to cooperative marl. *arXiv preprint arXiv:2208.01682* (2022).

[4] Afshin Oroojlooy, Mohammadreza Nazari, Davood Hajinezhad, and Jorge Silva. 2020. Attendlight: Universal attention-based reinforcement learning model for traffic signal control. *Advances in Neural Information Processing Systems* 33 (2020), 4079–4090.

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

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[7] Min Wang, Libing Wu, Man Li, Dan Wu, Xiaochuan Shi, and Chao Ma. 2022. Meta-learning based spatial-temporal graph attention network for traffic signal control. *Knowledge-Based Systems* 250 (2022), 109166.

[8] Hua Wei, Chacha Chen, Guanjie Zheng, Kan Wu, Vikash Gayah, Kai Xu, and Zhenhui Li. 2019. Presslight: Learning max pressure control to coordinate traffic signals in arterial network. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 1290–1298.

[9] 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.

[10] Muning Wen, Jakub Kuba, Runji Lin, Weinan Zhang, Ying Wen, Jun Wang, and Yaodong Yang. 2022. Multi-agent reinforcement learning is a sequence modeling problem. *Advances in Neural Information Processing Systems* 35 (2022), 16509–16521.

[11] Xinshi Zang, Huaxiu Yao, Guanjie Zheng, Nan Xu, Kai Xu, and Zhenhui Li. 2020. Metalight: Value-based meta-reinforcement learning for traffic signal control. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 1153–1160.

[12] Huichu Zhang, Chang Liu, Weinan Zhang, Guanjie Zheng, and Yong Yu. 2020. Generalight: Improving environment generalization of traffic signal control via meta reinforcement learning. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 1783–1792.

[13] Qinqing Zheng, Amy Zhang, and Aditya Grover. 2022. Online decision transformer. In *International Conference on Machine Learning*. 27042–27059.