东南大学

Feudal Multi-Agent Deep Reinforcement Learning for Traffic Signal Control

AAMAS-2020

汇报人/朱晓璇

时间/2021.04.22

目 录/CONTENT

Part. 1

研究背景

- 背景介绍
- 研究瓶颈
- 本文工作

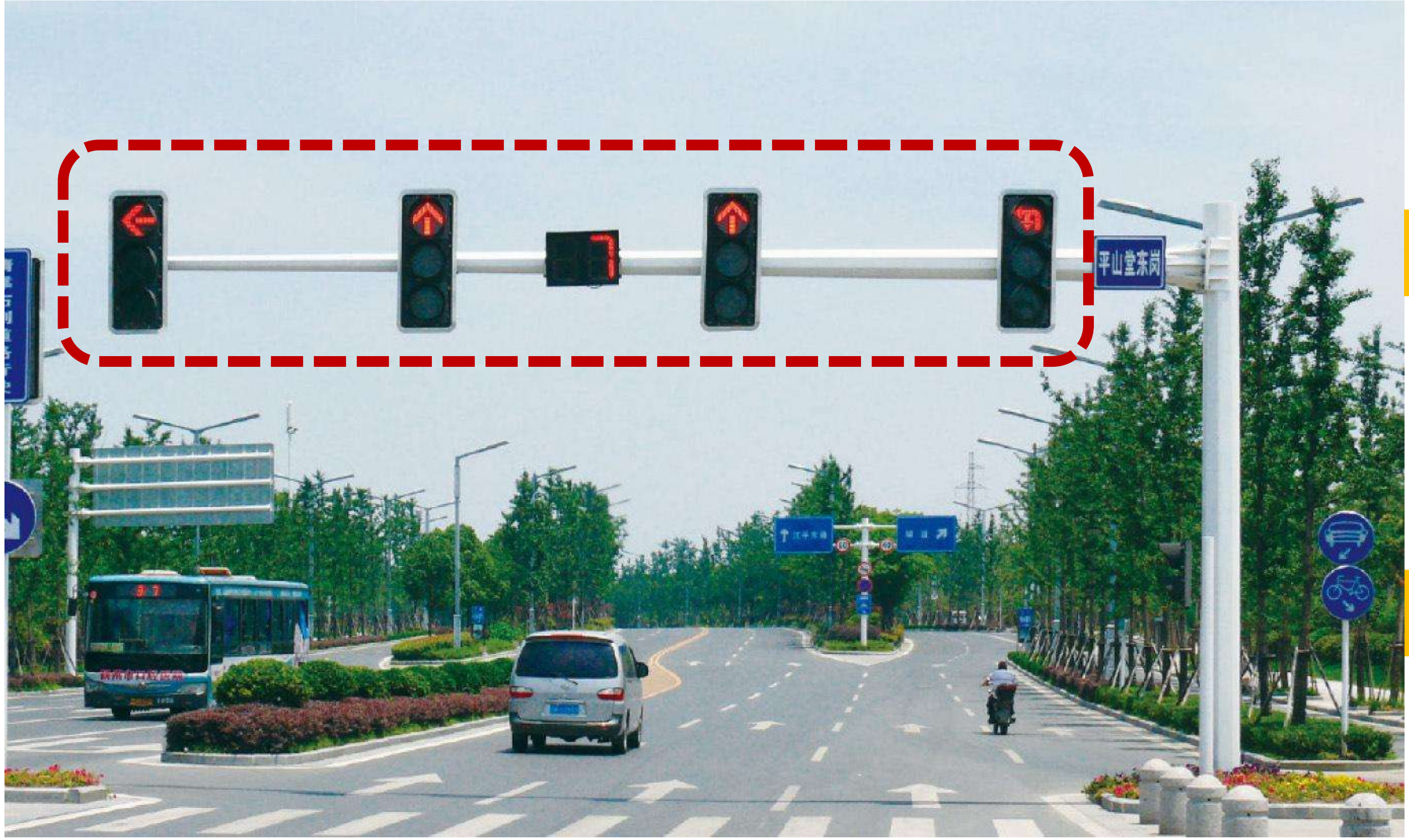
背景介绍

Introduction

堵！堵！堵！

背景介绍

Introduction



相位顺序怎么安排？

每个相位持续多久？

研究瓶颈

challenge

Centralized RL

多个agent观测本地状态，形成全局状态对信号灯进行调控；

- 较高的延迟。收集网络中的所有环境检测值形成一个全局状态在实践中会导致延迟
- 可扩展问题。信号灯交叉口数量的增加，agent之间的联合行动空间维度出现指数级增长，导致训练不收敛，无法扩展

Decentralized MARL

每个交叉口由单个agent控制，具有局部观测值，学习自己的策略，多个agent进行协作；

- 独立学习：难以达到全局最优
- 集中优化：可扩展问题，需要在巨大的联合行动空间上实现最大化

MA2C

每个agent只独立学习自己的策略，但是引入：

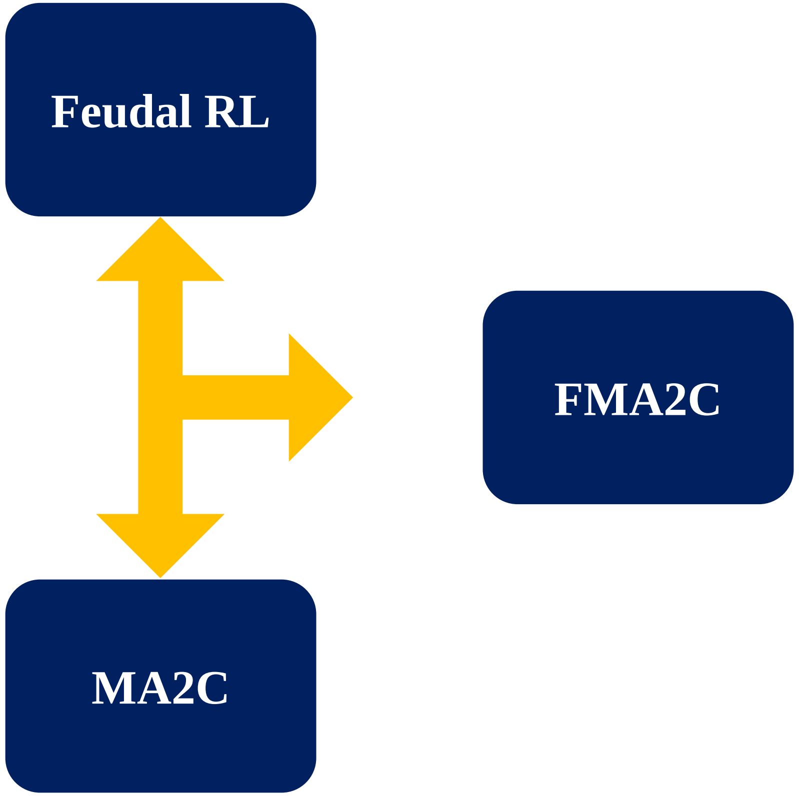
a） 状态中包含相邻agent的observation和fingerprint

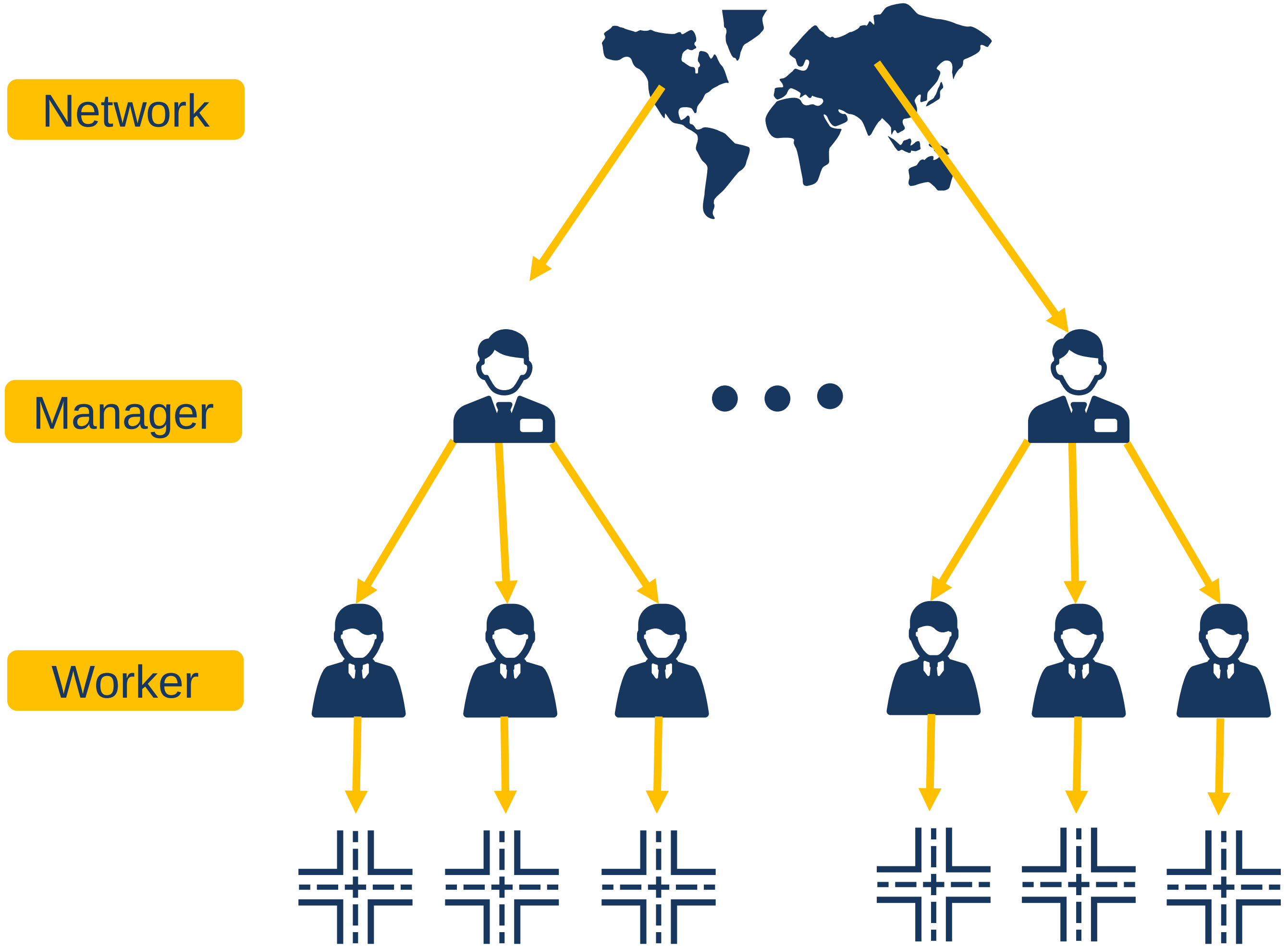
b） 引入空间折扣因子，缩小了相邻agent的观察和奖励信号，使每个agent更专注于改善附近的交通

不足：缺乏全局协调，容易陷入局部最优

本文工作

Contribution of This Paper





本文工作

Contribution of This Paper

Manager

1. 与其他manager协作
2. 为自己的worker制定目标

Worker

1. 满足manager的目标
2. 同时满足自己的局部目标

模型设计

Part. 2

- 分层结构的交通网络
- 部分可见马尔可夫决策
- 策略学习

分层结构的交通网络

Traffic Network with Hierarchical Structure

$\mathcal{N}_k$： manager k 的邻居

$\mathcal{U}_k = \mathcal{N}_k \cup \{k\}$： manager k与其邻居的集合

- **交通路网** $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$

  $v_i \in \mathcal{V}$： 交叉口

  $e = (v_i, v_j) \in \mathcal{E}$： 连接两个交叉口的道路

  $\mathcal{N}_i$ ： agent i 的邻居集合
  $\mathcal{U}_i = \mathcal{N}_i \cup \{i\}$： agent i 与其邻居的集合

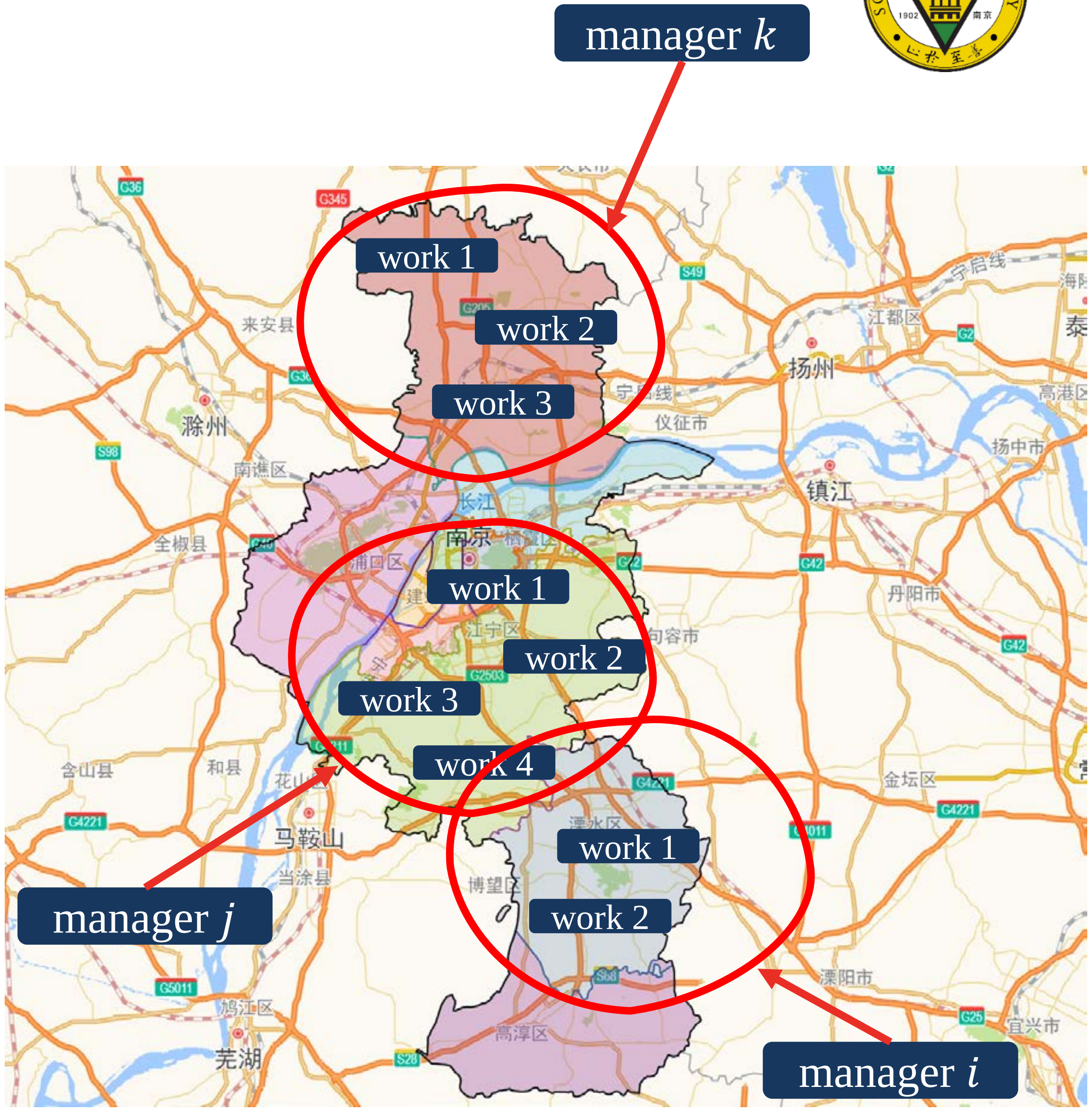
  $d(i,j)$： 两个agent之间的距离

- **Disjoint sub-networks**

  $\mathcal{G} = \{\mathcal{V}_1 \dots \mathcal{V}_m\}$， $\forall\, \mathcal{V}_i, \mathcal{V}_j$ ， $\mathcal{V}_i \cap \mathcal{V}_j = \emptyset$， $\cup_{k=1}^{m} \mathcal{V}_k = \mathcal{G}$

  $\forall i, j \in \mathcal{V}_k$存在一条连接 i 和 j 的路径

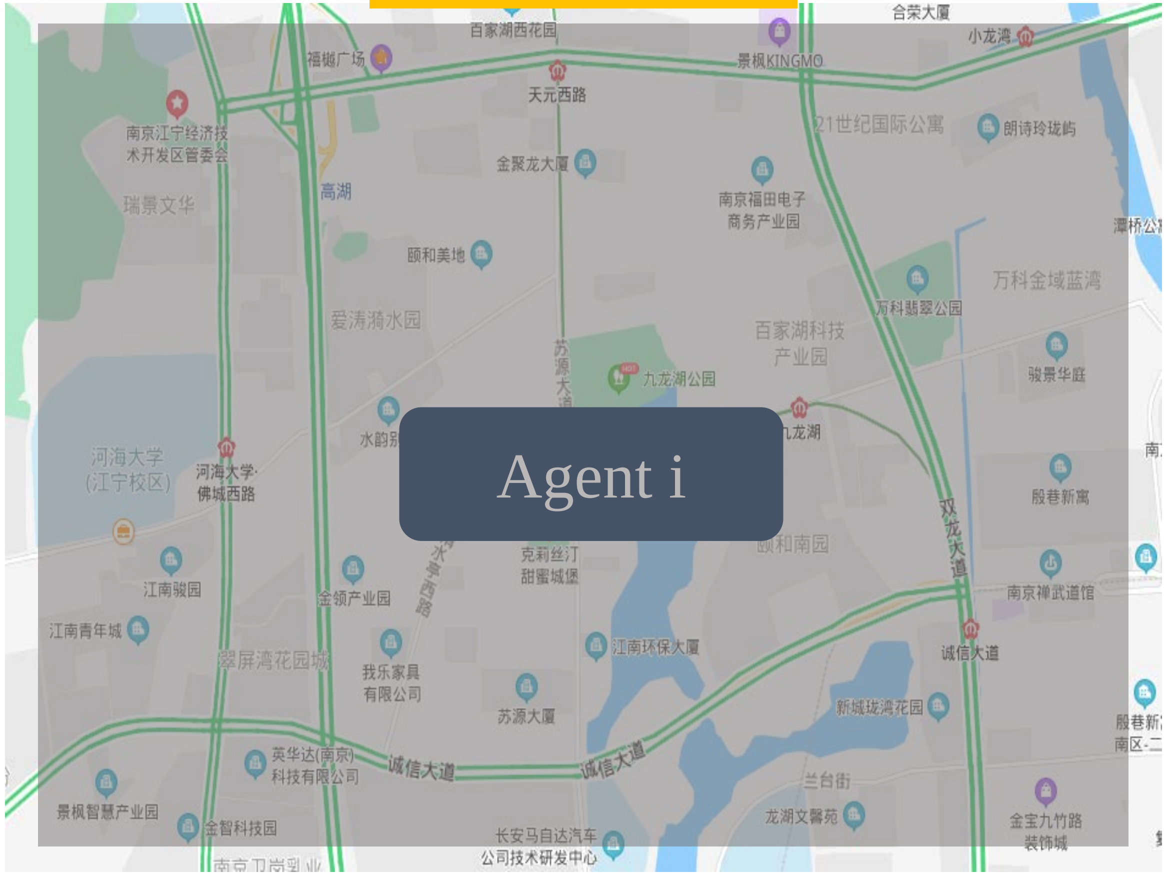
  Region： $\mathcal{V}_k \subseteq \mathcal{G}$



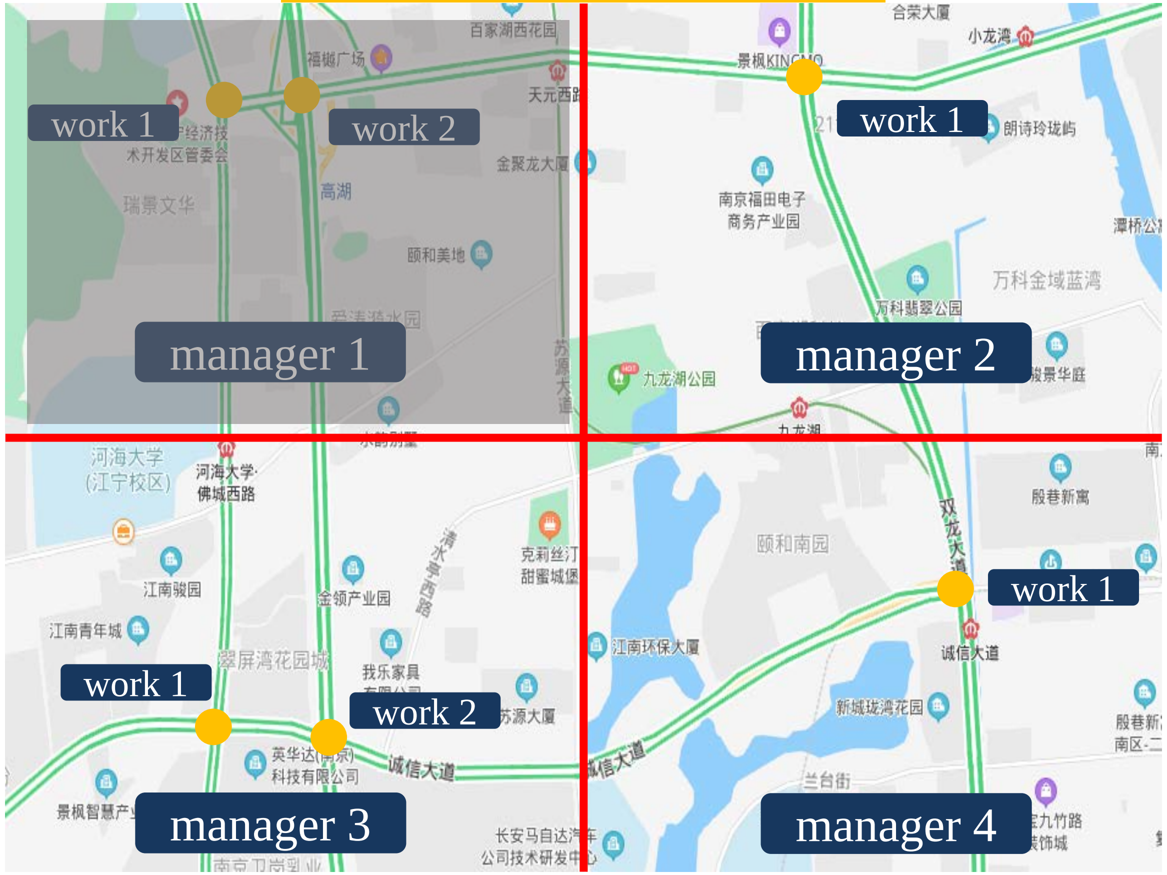
部分可见马尔可夫决策

Partial Observable Markov Game

马尔可夫决策



部分可见马尔可夫决策



部分可见马尔可夫决策

Partial Observable Markov Game

- 马尔可夫决策过程

$$\langle S, A, P, R \rangle$$

- Manager

$$\mathcal{M}^M = \langle S^M, \{O_k^M\}, \{A_k^M\}, P^M, R^M \rangle$$

- Worker

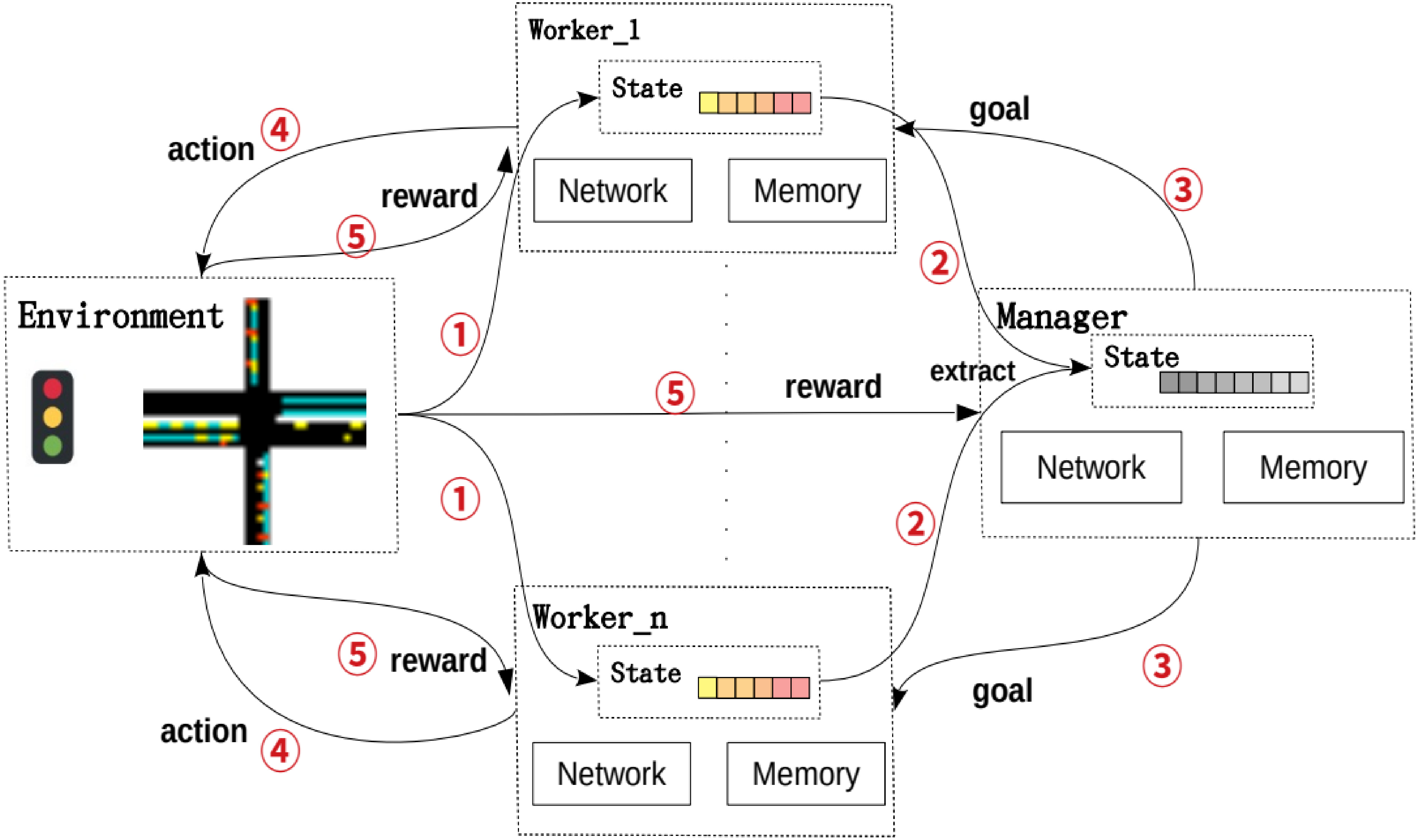
抽象　　目标

$$\mathcal{M}^W = \langle S^W, \{O_i^W\}, \{A_i^W\}, P^W, R^W \rangle$$

$$\hat{r}_{t,i}^W = r_{t,i}^W + \boxed{\sigma\left(o_{t,i}^w, a_{t,k}^M\right)}$$

部分可见马尔可夫决策

Partial Observable Markov Game



策略学习

Learning Policies For Managers and Workers

怎样增加全局协作信息？

- manager、worker加入其neighbor的observation来扩大自己的observation，获取更多信息；
- manager、worker加入其neighbor上一轮的策略来更新自身策略；
- worker加入manager分发的goal（action）进行策略更新；
- manager所得即时奖励引入其neighbor的奖励；
- worker所得即时奖励引入其neighbor及manager的奖励；
- 全局状态由自身状态与邻居状态生成；

$$\pi_{t,k}^{M} = \pi_{\theta_k^-}\left(\cdot \middle| o_{t,\mathcal{U}_k}^{M}, \pi_{t-1,\mathcal{N}_k}^{M}\right)$$

$$\pi_{t,k}^{W} = \pi_{\theta_i^-}\left(\cdot \middle| o_{t,\mathcal{U}_i^-}^{W}, \pi_{t-1,\mathcal{N}_i^-}^{W}\right)$$

$$\mu_{t,i}^{W} = \mu_{\theta_i^-}\left(\cdot \middle| o_{t,\mathcal{U}_i^-}^{W}, \mu_{t-1,\mathcal{N}_i^-}^{W}, a_{t,\mathcal{U}_k}^{M}\right)$$

$$\tilde{r}_{t,i}^{M} = r_{t,k}^{M} + \sum_{j\in\mathcal{N}_k} \alpha \cdot r_{t,j}^{M}$$

$$\tilde{r}_{t,i}^{W} = \textstyle\sum_{d=0}^{D_i}\left(\sum_{j\in\mathcal{V}_k|d(i,j)=d} \alpha^{d} \cdot \hat{r}_{t,j}^{W}\right) + \tilde{r}_{t,k}^{M}$$

$$\tilde{s}_{t,\mathcal{U}_k}^{M} = \left[o_{t,k}^{M}\right] \cup \alpha\left[o_{t,j}^{M}\right]_{j\in\mathcal{N}_k}$$

$$\tilde{s}_{t,\mathcal{U}_i}^{W} = \left[o_{t,i}^{W}\right] \cup \alpha\left[o_{t,j}^{W}\right]_{j\in\mathcal{N}_i^-}$$

策略学习

Learning Policies For Managers and Workers

NN的输出

• 累积奖励

$$\tilde{R}^M_{t,k} = \hat{R}^M_{t,k} + \gamma^{T-t} V^M_{\omega_k^-}\left(\tilde{s}^M_{T,\mathcal{U}_k}, \pi^M_{T-1,\mathcal{N}_k} \middle| \pi^M_{\theta_{-k}^-}\right)$$

$$\tilde{R}^W_{t,i} = \hat{R}^W_{t,i} + \gamma^{T-t} V^W_{\omega_i^-}\left(\tilde{s}^W_{T,\mathcal{U}_i}, \pi^W_{T-1,\mathcal{N}_i} \middle| \pi^W_{\theta_{-i}^-}\right)$$

Advantage function

$$\tilde{A}^M_{t,k} = \tilde{R}^M_{t,k} - V^M_{\omega_k^-}\left(\tilde{s}^M_{t,\mathcal{U}_k}, \pi^M_{t-1,\mathcal{N}_k}\right)$$

$$\tilde{A}^W_{t,i} = \tilde{R}^W_{t,i} - V^W_{\omega_i^-}\left(\tilde{s}^W_{t,\mathcal{U}_i}, \pi^W_{t-1,\mathcal{N}_i}\right)$$

• Critic loss function

$$\mathcal{L}(\omega_k^M) = \frac{1}{2|B|}\sum_{t\in B^M}\left(\tilde{R}^M_{t,k} - V^M_{\omega_k}\left(\tilde{s}^M_{t,\mathcal{U}_k}, \pi^M_{t-1,\mathcal{N}_k}\right)\right)^2$$

$$\mathcal{L}(\omega_i^W) = \frac{1}{2|B|}\sum_{t\in B^W}\left(\tilde{R}^W_{t,i} - V^W_{\omega_i}\left(\tilde{s}^W_{t,\mathcal{U}_i}, \pi^W_{t-1,\mathcal{N}_i}\right)\right)^2$$

Regularization term

• Actor loss function

$$\mathcal{L}(\theta_k^M) = -\frac{1}{|B|}\sum\nolimits_{t\in B^M}\left(\log\pi^M_{\theta_k}\left(a^M_{t,k} \middle| \tilde{s}^M_{t,\mathcal{U}_k}, \pi^M_{t-1,\mathcal{N}_k}\right)\tilde{A}^M_{t,k} - \beta\sum\nolimits_{a_k\in A_k^M}\pi^M_{\theta_k}\log\pi^M_{\theta_k}\left(a_k \middle| \tilde{s}^M_{t,\mathcal{U}_k}, \pi^M_{t-1,\mathcal{N}_k}\right)\right)$$

$$\mathcal{L}(\theta_i^W) = -\frac{1}{|B|}\sum\nolimits_{t\in B^W}\left(\log\pi^W_{\theta_i}\left(a^W_{t,i} \middle| \tilde{s}^W_{t,\mathcal{U}_i}, \pi^W_{t-1,\mathcal{N}_i}\right)\tilde{A}^W_{t,i} - \beta\sum\nolimits_{a_i\in A_i^W}\pi^W_{\theta_i}\log\pi^W_{\theta_i}\left(a_i \middle| \tilde{s}^W_{t,\mathcal{U}_i}, \pi^W_{t-1,\mathcal{N}_i}\right)\right)$$

Part. 3

实验设计

- 模型定义
- 合成网络下实验
- 真实网络下实验

模型定义

Model Setting

- Observation

Manager： $o_{t,k}^{M} = \{Nwave_t[l], Ewave_t[l], Swave_t[l], Wwave_t[l]\}_{l \in L_k}$

Worker： $o_{t,i}^{W} = \{wave_t[l], wait_t[l]\}_{l \in L_i}$

- Action

Manager： possible traffic flow， four combinations of north-south and east-west traffic flows.

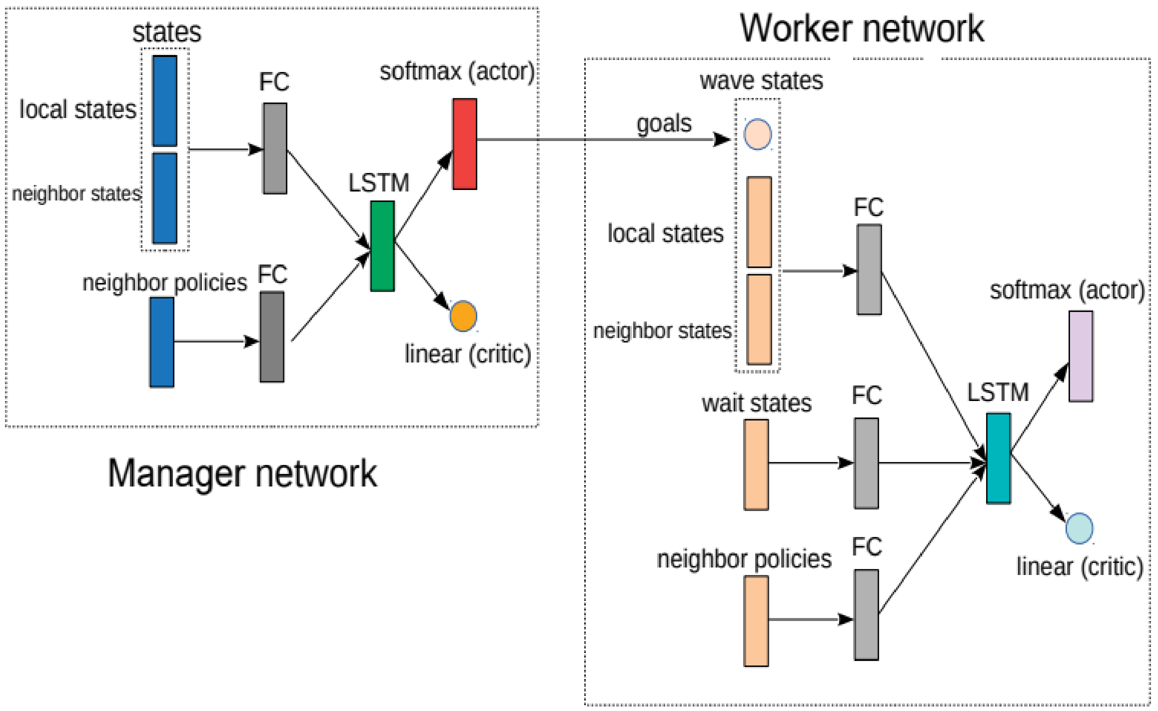
Worker： possible phrase， east-west straight and right-turn phase, east-west left-turn and right-turn phase, and three straight, right-turn and left-turn phases for east, west, and north-south

- Reward

Manager： $r_{t,k}^{M} = \sum_{l \in L_k}\left(arrival_{t+\Delta t}[l] + \sum_{i \in \mathcal{V}_k} liquid_{t+\Delta t}[l]\right)$

Worker： $r_{t,i}^{W} = -\sum_{l \in L_i}(wave_{t+\Delta t}[l] + a \cdot wait_{t+\Delta t}[l])$

- NN structure

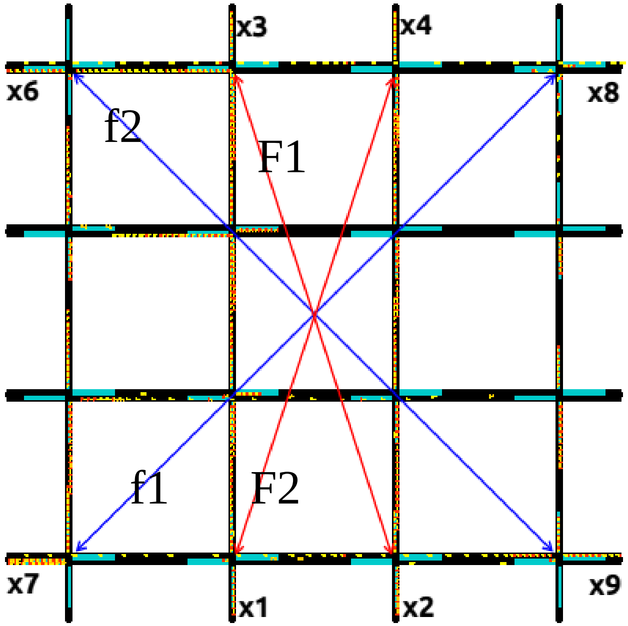


对比算法

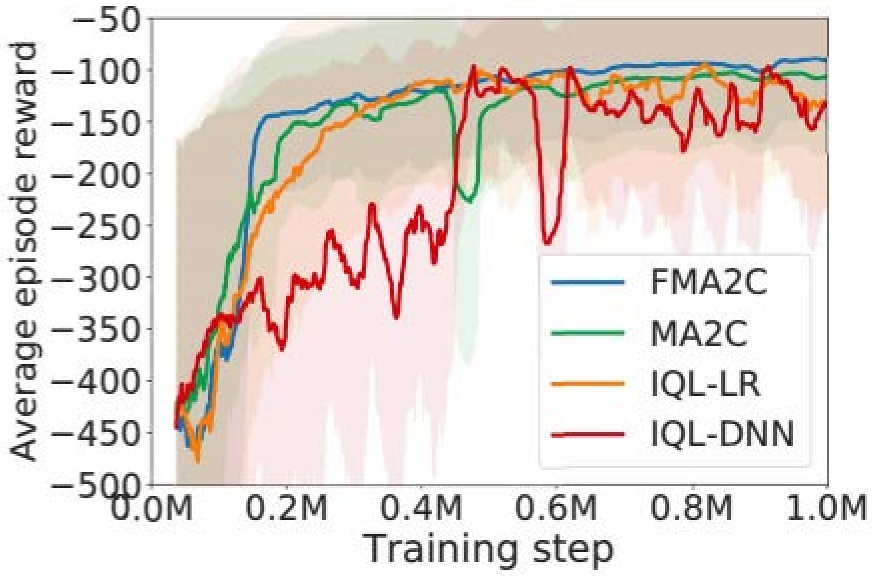
- FMA2C：层次结构的多智能体强化学习
- MA2C：目前在多智能体研究中领先的RL方法，缺乏全局协作
- IQL-DNN：带有DNN的独立Q学习
- IQL-LR：带线性回归的独立Q学习
- Greedy：代理选择贪婪行为

合成网络下的实验

Synthetic Traffic Grid



(a) 4×4 synthetic traffic grid



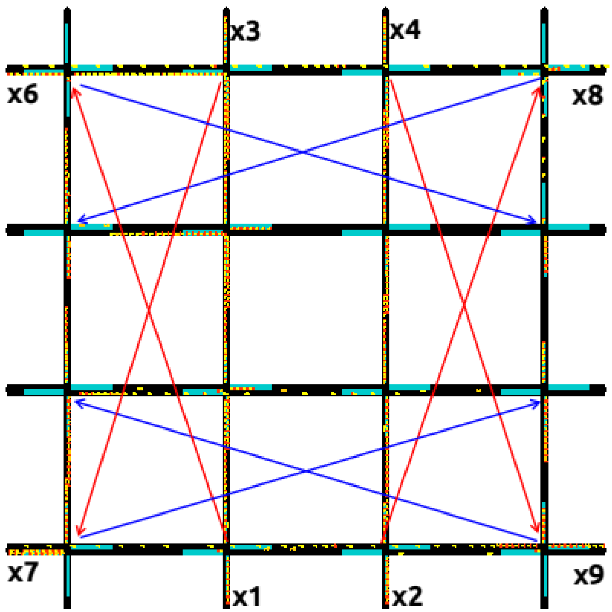
(d) Training curves (4×4 traffic grid)

Table 1: Time-variant traffic flows within the 4×4 traffic grid.

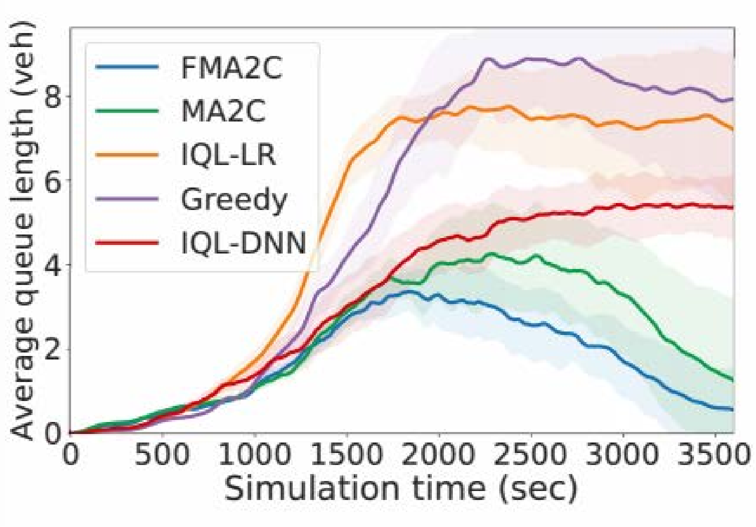
	0	300	600	900	1200	1500	1800	2100	2400	2700	3000	3300	3600	(sec)
f1	264.0	462.0	594.0	660.0	495.0	330.0	165.0	0	0	0	0	0	0	(veh/h)
F1	440.0	770.0	990.0	1100.0	825.0	550.0	275.0	0	0	0	0	0	0	(veh/h)
f2	0	0	0	166.5	444.0	499.5	555.0	444.0	333.0	111.0	0	0	0	(veh/h)
F2	0	0	0	277.5	740.0	832.5	925.0	740.0	555.0	185.0	0	0	0	(veh/h)

合成网络下的实验

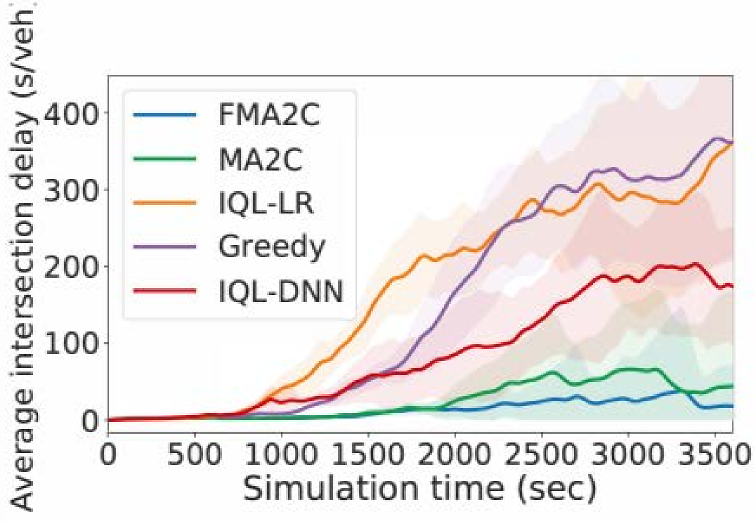
Synthetic Traffic Grid



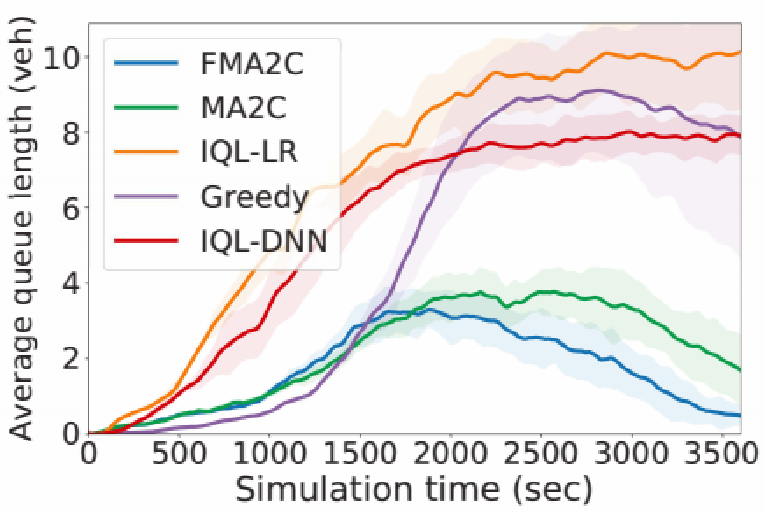
(b) Changing flow direction



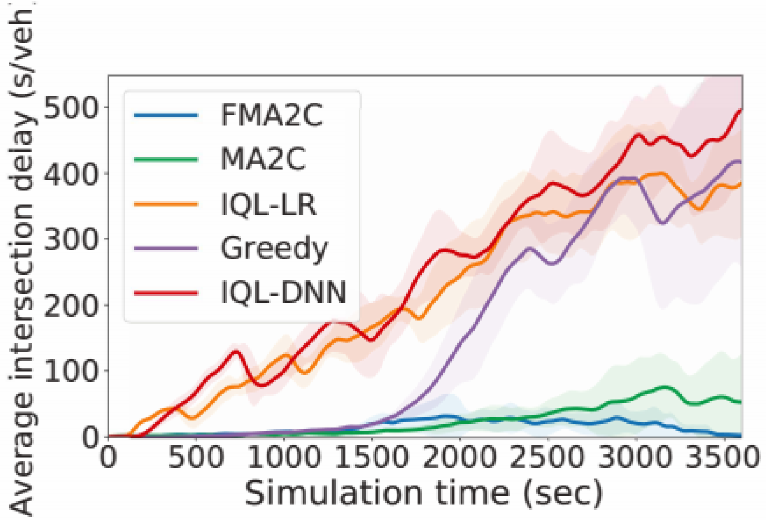
(e) Queue length (increasing flow value)



(f) Delay (increasing flow value)



(g) Queue length (changing flow direction)



(h) Delay (changing flow direction)

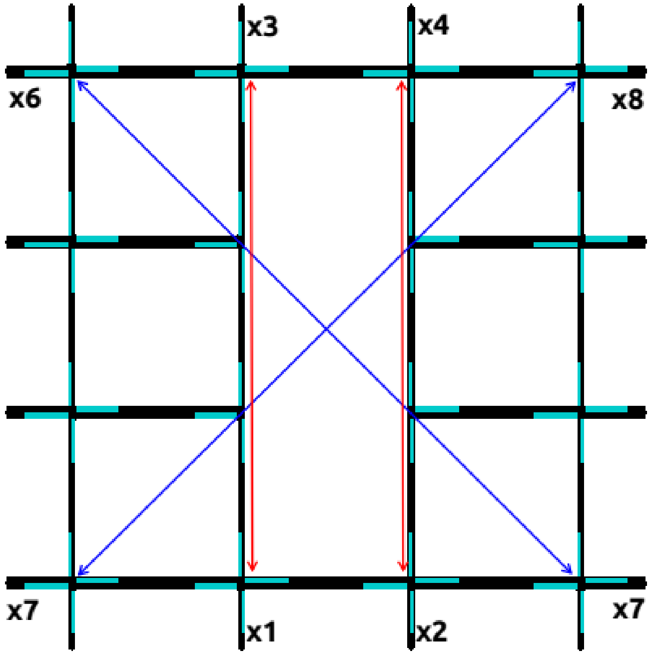
合成网络下的实验

Synthetic Traffic Grid

Metrics	(a) 4×4 traffic grid (increasing flow value)					(b) 4×4 traffic grid (changing flow directions)				
	FMA2C	MA2C	IQL-DNN	IQL-LR	Greedy	FMA2C	MA2C	IQL-DNN	IQL-LR	Greedy
reward	**-310.22**	-467.65	-850.88	-1647.20	-1940.51	**-302.78**	-406.71	-2007.25	-2420.88	-1867.01
avg. queue length [veh]	**1.72**	2.35	3.31	5.02	5.09	**1.69**	2.23	5.51	6.87	4.78
avg. intersection delay [s/veh]	**14.46**	26.18	87.42	168.10	152.15	**15.62**	25.04	247.32	218.15	154.94
avg. vehicle speed [m/s]	**3.80**	3.27	2.77	2.56	2.80	**3.63**	3.09	1.49	1.18	3.43
trip completion flow [veh/s]	**0.81**	0.79	0.42	0.43	0.50	**0.81**	0.76	0.16	0.16	0.56
trip delay [s]	328	398	359	**273**	296	323	374	450	751	**241**

合成网络下的实验

Synthetic Traffic Grid

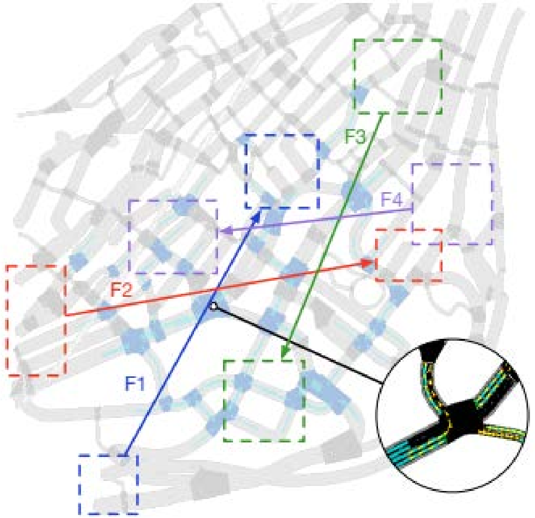


(c) Irregular grid shape

Metrics	(c) 4×4 **traffic grid (irregular grid shape)**				
	FMA2C	MA2C	IQL-DNN	IQL-LR	Greedy
reward	**-105.58**	-138.43	-1527.29	-465.61	-277.27
avg. queue length [veh]	**0.83**	1.21	4.25	2.61	1.08
avg. intersection delay [s/veh]	**3.86**	4.45	179.90	47.31	19.86
avg. vehicle speed [m/s]	**4.76**	4.13	2.15	3.79	4.73
trip completion flow [veh/s]	**0.69**	0.67	0.24	0.57	0.66
trip delay [s]	**216**	296	268	371	225

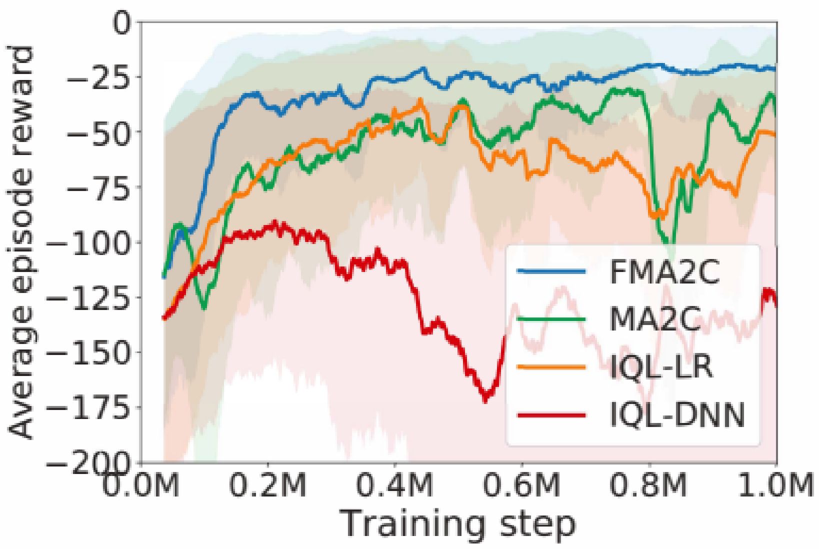
真实网络下的实验

Synthetic Traffic Grid

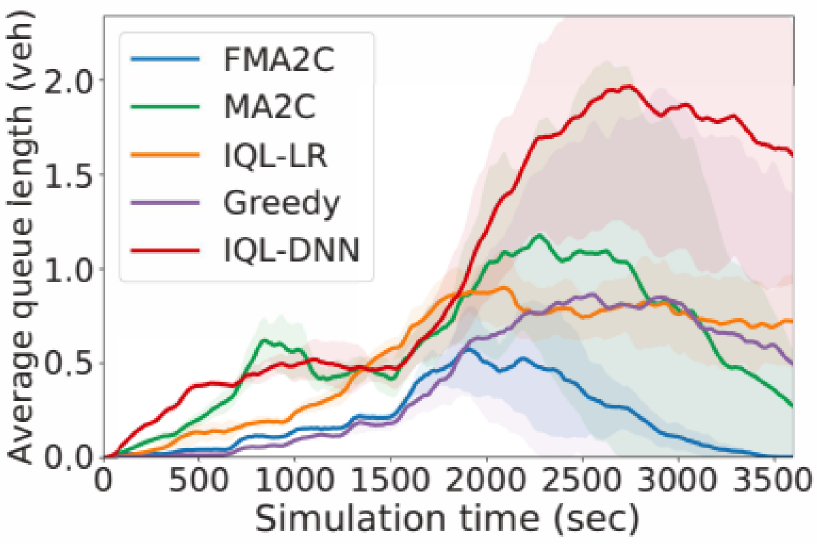


(i) Monaco traffic network

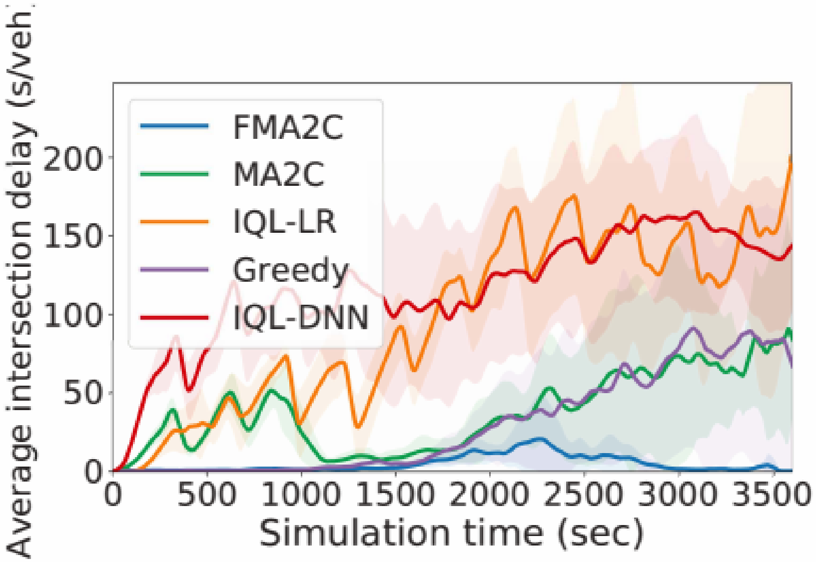
Metrics	(d) Monaco traffic network				
	FMA2C	MA2C	IQL-DNN	IQL-LR	Greedy
reward	**-22.77**	-63.21	-100.29	-53.80	-100.29
avg. queue length [veh]	**0.20**	0.60	1.04	0.54	0.41
avg. intersection delay [s/veh]	**4.58**	38.07	116.61	97.09	29.90
avg. vehicle speed [m/s]	**7.53**	4.88	2.38	4.34	7.38
trip completion flow [veh/s]	**0.68**	0.64	0.54	0.46	0.63
trip delay [s]	**89**	201	267	153	95



(j) Training curves (Monaco network)



(k) Queue length (Monaco network)



(l) Delay (Monaco network)

Part. 4

总结思考

- 总结
- 思考

总结

Conclusion

MA2C
缺乏全局协调

Feudal RL

MA2C

1. 进行等级的划分，分manager和worker；
2. 利用Manager进行全局的协调；
3. Worker负责本地信号灯的策略学习；

思考
Thought

- 多智能体的考虑主要是为了可扩展性的问题，单一智能体无法扩展到大规模的交通路网，多智能体可以对此问题给出解决方案；

- 多智能体存在通信以及无法掌握全局信息的问题，因此建模时要从如何增加全局的协调信息进行考虑；

- 这篇文章划分区域的方法比较简单，还可以更多的从场景上对区域进行划分，譬如商业区、住宅区、学区等等。

东南大学

谢谢大家

汇报人/朱晓璇　　**时间**/2021.04.22