编号_____

南京航空航天大学

毕业论文

题　目　基于深度强化学习智能交通信号调度

学生姓名　陈建

学　　号　SX1916039

学　　院　计算机科学与技术学院

专　　业　计算机科学与技术

班　　级　1318001

指导教师　朱琨

二〇二一年十二月

南京航空航天大学
本科毕业论文诚信承诺书

本人郑重声明:所呈交的毕业论文(题目:基于深度强化学习智能交通信号调度)是本人在导师的指导下独立进行研究所取得的成果。尽本人所知，除了毕业论文中特别加以标注引用的内容外，本毕业论文不包含任何其他个人或集体已经发表或撰写的成果作品。

作者签名:　　年　月　日

(学号):

基于深度强化学习智能交通信号调度

摘 要

本文介绍如何使用NUAA Thesis 文档类撰写南京航空航天大学学位论文。

首先介绍如何获取并编译本文档，然后展示论文部件的实例，最后列举部分常用宏包的使用方法。

关键词：学位论文，模板，NUAA Thesis

NUAA2 Thesis Quick Start and Document Snippets

Abstract

This document introduces NUAA2 Thesis, the LaTeX document class for NUAA Thesis.

First, we show how to get the source code and compile this document. Then we provide snippets for figures, tables, equations, etc. Finally we enforce some usage patterns.

Key Words: NUAA thesis, document class, space is accepted here

目录

第一章 背景和研究意义

1.1 研究背景

交通信号控制是一个重要且具有挑战性的现实问题，其目标是最大化车辆的通行效率通过协调他们在交叉路口的行动。随着汽车制造业的快速发展以及城市化进程的推进，我国汽车保有量在不断的增加，交通拥堵情况也在不断恶化，极大的影响了人们的生活的城市的运作，同时这种拥堵现象也在向中小城市蔓延。为了缓解交通拥堵，很多城市也提出了不同的解决方法，有减少出行车辆数量的“限号”政策，也有通过加快城市道路建设来加大城市交通承载量的方法。其实，交通拥堵通常是由于不同的车流为了争夺同一个“行驶资源”而造成的。这一“行驶资源”通常就是不同道路的交叉口，所以现代城市交通管理中在道路的交叉口安装信号灯并通过简单的策略来调度通过的车流。但是随着车辆数量的不断增加，之前简单的策略已经难以应对现在更加复杂的交通模式。因此，如何制定出更加高效和智能的调度策略显得格外的重要。

智能交通信号控制会根据实时的交通状况做出最优的决策并以此来控制信号灯的变化，已达到最大程度地减轻交通拥堵的目的。传统的交通控制更多的是基于一些既定的规则和一些根据历史数据总结出的经验来控制信号灯，没有考虑实时的交通状况，所以无法很有效的减轻交通拥堵的状况，但是由于其简单以及易于部署的特点，绝大多数城市的信号灯都还在采用这种控制模式。

随着车联网技术的发展，对于实时车辆数据的获取变得越来越容易，利用得到的车辆数据可以获得实时的交通状况，并且如何根据实时的交通状况来制定最优的策略一直是研究的热点。以往多数的研究是采用基于优化的方法，根据车流的状况计算出一个最优的信号灯的相位序列，但是这种方法要求车流的状况是比较简单的，例如服从均匀分布，与现实中的车流情况相比太过理想化，

所以难以部署到实际场景中。伴随着人工智能技术的发展，一些研究者提出利用深度强化学习来控制信号灯，将整个交通信号灯控制建模成一个马尔可夫决策过程（Markov Decision Process）。对于每一次决策，输入当前的交通状况作为状态，输出一个作用在信号灯上的动作，例如变换到下一个相位（phase）。这种方法对于车流的情况没有限制，通过在大量不同的车流状况下进行训练可以得到一个鲁棒的模型，能够应对不同的车流场景并做出最优的决策，并且这种方法在通行效率上也比基于优化的方法和传统的规则控制方法更高。

但是在多数已有的工作中，都是使用车辆的平均行驶时间来量化通行效率，这就会导致一个公平失衡的问题。控制策略在控制信号灯的时候会更加倾向于放行有更多等待车辆的车道（主干道），所以在拥有较少车辆车道（次干道）上的车可能很长时间得不到放行，这对于次干道上的车来说是不公平的。因此如何设计出在最大化通行效率的同时又能保障所有车辆相对公平的策略具有重要的意义。

1.2 国内外研究现状

目前，已有的关于交通信号控制的研究主要可以分为两大类，一类是传统的交通信号控制，另一种是基于学习的交通信号控制。

传统交通信号控制可以细分为两类: 第一类是基于预定义的信号控制（pre-defined signal control）[5]，根据历史交通流量信息预先定义信号的序列和周期。由于没有考虑到实时交通的情况，所以不能够有效地适应动态的交通变化。第二类是基于优化方法的自适应控制，这些方法通常是从优化的角度解决某些交通流模型下的交通信号控制问题，并且根据观测到的数据调整信号的序列和周期。然而，为了让问题易于求解，通常在模型上做出很强的假设 [6]-[8]。例如，为了优化车辆通行时间，Webster 在工作 [9] 中假设车辆的到达率是均匀的。这些假设通常太过理想化，不适用于现实世界。与传统交通信号控制不同的是，基于学习的交通信号控制不需要预先定义的规则和不切实际的假设。具体来说，这类方法直接从与环境的交互中学习，并且根据环境的反馈来完善和改进

控制策略，逐渐达到最优的效果。现有的基于学习的信号控制方法的差异主要表现在三个方面：状态表示，奖励设计和学习算法。交通状态可以使用不同的特征来描述，例如等待车辆的队列长度 [1], [10], [11]、平均延迟 [12], [13] 以及车辆的位置影像 [10], [14]。通常，几个特征被整合起来，以获得对交通状况的全面描述。奖励通常是根据等待时间 [13], [15], [16], 队列长度 [2], [11] 和吞吐量 [10], [17], [18] 来设计的。学习算法一般可分为基于价值的方法 [10], [13], [15]，基于策略的方法 [19], [20] 和基于 Actor-Critic 的方法 [17], [21]。然而，大多数现有的基于学习的方法只侧重于提高交叉口的效率，例如最小化队列长度或最大化吞吐量，而忽略了对公平性的考虑。事实上，这样的目标可能导致学到一个有偏见的策略，控制策略会更加偏向于主干道上的车辆，可能会导致次干道上的车长时间得不到放行（这一情况被称为“饥饿”现象）。另一方面，在最近一段时间，对于多路口交通信号控制的研究，使用分布式强化学习的独立控制方式逐渐取代以往的整体控制方式（对所有路口使用一个智能体来进行学习），不同智能体之间通过显示的通信来交流信息，从而实现协同控制。但是已有的工作对于通信过程中信息选择的研究却有所不足，目前常用的方式就是将智能体自己观测到的局部环境全部传输给相邻路口的智能体。虽然这种做法可以让智能体对全局环境有更加全面的了解，于此同时也增加了通信代价，而且还让智能体的训练难度加大。此外，并不是所有的信息对于其相邻的路口都是有用的，可能有些信息对其上游路口有用，有些影响其下游路口，为特定路口筛选特定的信息来进行交流不仅可以减少通信代价，而且可以让学习的效率提高。

第二章 概述

2.1 交通信号概述

交通信号控制是一个重要而具有挑战性的现实问题，其目的是通过协调车辆在道路交叉口的运动来最小化所有车辆的通行时间。目前广泛使用的交通信号控制系统仍然严重依赖过于简化的信息和基于规则的方法。车联网技术的发展、硬件性能的提升以及人工智能技术的进步使得我们现在有更丰富的数据，更多的计算能力和先进的方法来驱动智能交通的发展。交通信号控制的目的为了方便车辆在交叉路口的安全和高效移动。安全是通过信号灯指定不同车道的车通行来分离相互冲突的运动实现的。为了能够有效地优化通行效率，已有的工作提出了不同的指标来量化通行效率，主要有以下三个：1. 通行时间：在交通信号控制中，车辆的行驶时间被定义为一辆汽车进入系统的时间与离开系统的时间的差值。最常见的优化目标之一是就是减少进过路口的所有车辆的平均通行时间。2. 队列长度：队列长度是指路口等待车辆的数量，越大的队列长度意味着越多的等待车辆，路口的通行效率越低，反之通行效率越高。3. 路口吞吐量：吞吐量是指在一定期间内进过路口完成通行的车辆数量。越大的吞吐量代表着越高的通行效率，所以很多工作将最大化吞吐量作为优化的目标。

2.2 基本术语

- Approach: 指交叉路口的巷道。任何一个交叉路口都有两种 approach, 进入路口的 incoming approach 和离开路口的 outgoing approach。图 2.1(a) 描述了一个典型的有 8 个 approach（四个入口，四个出口）的交叉路口。
- Lane：一个 Approach 是由一组车道组成。与 Approach 的定义类似，车道也分为两种：转入车道（incoming lane）和转出车道 (outgoing lane)。

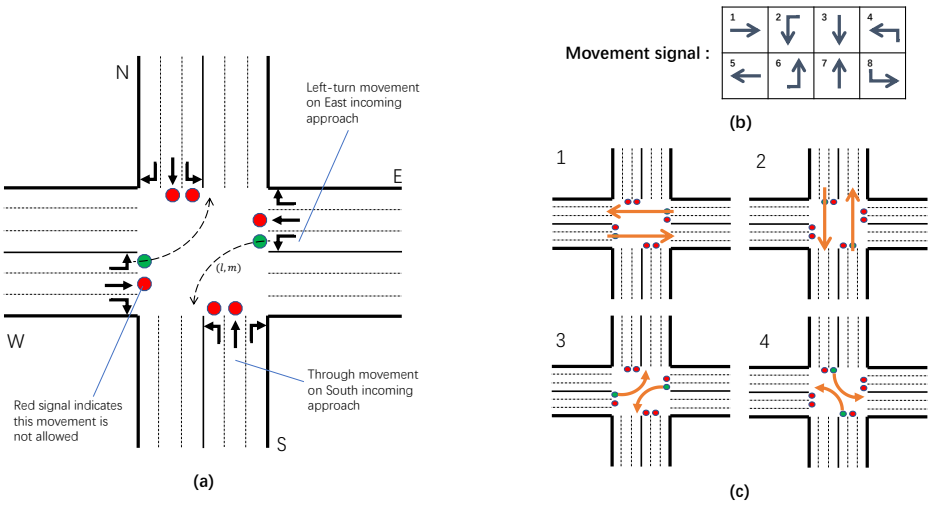

图 2.1 tsc

- Traffic movement：指的是车流从一个 incoming approach 运动到另一个 outgoing approach，表示为 $(r_i \longrightarrow r_o)$，其中 ri 和 ro 分别表示 incoming lane 和 outgoing lane。通常，traffic movement 可以分为左转、直行以及右转三种，在少数特殊的路口也支持 U-turn 的 traffic movement。
- Movement signal: 根据 traffic movement 定义的运动信号，绿色代表可以通行，红色代表禁止通行。根据大多数国家的交通规则，右转的 traffic movement 是可以不受信号约束的。
- Phase: 信号灯的一个 phase(相位) 是指非冲突运动信号的组合，这意味着这些信号可以同时设置为绿色，而不会引起安全冲突。图 2.1(c) 展示了最常用的四相位信号模式。
- Phase sequence: 相序，即一组相位的序列，它定义了一组相位及其变化顺序。
- Signal plan：信号计划，由一组相位序列及其相应的起始时间组成。通常表示为 $(p_1, t_1)\,(p_2, t_2)\ldots(p_i, t_i)\ldots$ 其中 p_i 和 t_i 分别代表相位及其开始时间。

- Cycle-based signal plan：周期性信号计划，与普通的信号计划不同的是其中的相位序列是按循环顺序工作的，可以表示为 $(p_1, t_1^1)(p_2, t_2^1)\ldots(p_N, t_N^1)(p_1, t_1^2)(p_2, t_2^2)\ldots(p_N, t_N^2)\ldots$，其中 $p_1, p_2, \ldots, p_N$ 是重复出现的相位序列，t_i^j 是 j 周期中相位 p_i 的起始时间。具体地，$C^j = t_1^{j+1} - t_1^j$ 是第 j 周期的周期长度，$\left\{\frac{t_2^j - t_1^j}{C^j}, \ldots, \frac{t_N^j - t_{N-1}^j}{C^j}\right\}$ 是第 j 周期中的相位分裂比（phase split ratio），表示每个相位持续时间占总周期长度的比重。现有的交通信号控制方法通常在一天中重复类似的相位序列。

2.3 传统交通控制方法

2.3.1 Webster

对于单个交叉口，交通运输工程领域中的交通信号控制方法通常由三个部分组成：确定信号周期长度，确定信号相位序列以及相位分裂。Webster 是一种广泛使用的计算单个交叉路口的信号周期长度和相位分裂时间的方法。通过假设车流在一段时间内（例如，过去的五分钟或 10 分钟）是均匀到达的，可以计算出确切的最优周期和最佳相位分裂时间，从而最小化车量通行时间。

2.3.2 GreenWave

虽然使用 Webster 可以简单的控制单个交叉路口的交通信号，但是对于相邻的多个交叉路口，不能够简单地直接使用 Webster 来分别优化每一个路口，相邻路口信号灯的信号时间之间的偏移（即相邻路口信号周期起始时间的差值）也需要进行优化，因为对于相距较近的路口来说，一个路口的控制策略可能会影响到其他路口。GreenWave 就是交通运输领域中最经典的协调相邻路口的信号控制方法，它通过优化相邻路口信号时间的偏移来减少车辆在某一方向行驶时的停留次数。这种方法可以形成沿指定交通方向的绿色信号波，在该方向行驶的车辆可以受益于渐进的绿色信号级联，而不会在任何交叉口停留，如下图所示：

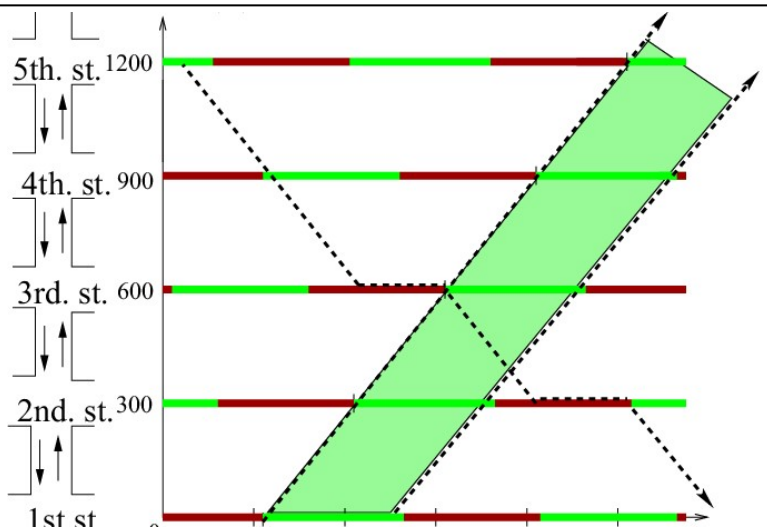

图 2.2 green-wave

2.3.3 Actuated Control

Actuated Control 根据当前相位和其他的竞争相位对绿色信号的请求来决定是否保持或者变化当前的相位。请求触发规则如下：1. 当目前相位的持续时间未达到最小时间周期时，或在当前相位对应入车道上有车辆进入，并且在接近信号的距离内时，就会产生延长绿色信号时间的请求，已让车辆可以直接通过路口。2. 当竞争相位的等待车辆数量大于一个阈值时，就会生成对绿色信号的请求。根据规则的差异，Actuated Control 主要可以分为 Fully-Actuated Control 和 Semi-Actuated Control 两种。

2.3.4 SOTL

Self-Organizing Traffic Light Control(SOTL) 是一种具有附加需求响应规则的 Fully-Actuated Control 方法。它与 Fully-Actuated Control 的主要区别在于当前相位的绿色信号请求定义（虽然它们都需要最小的绿色相位持续时间）：在 Fully-Actuated Control 中，当车辆接近信号灯时，就会产生延长绿色信号的请求，而在 SOTL 中，除非接近信号灯的车辆数量大于不一定是一个阈值，否则就不会产生请求。

2.3.5 Max-Pressure Control

Max-Pressure Control 的目的是通过最小化对应相位的压力（pressure）来平衡相邻路口之间的队列长度，从而降低过饱和的风险，其中压力的概念如下图所示：从形式来看，运动信号的压力可以定义为（交通运动的）传入车道上

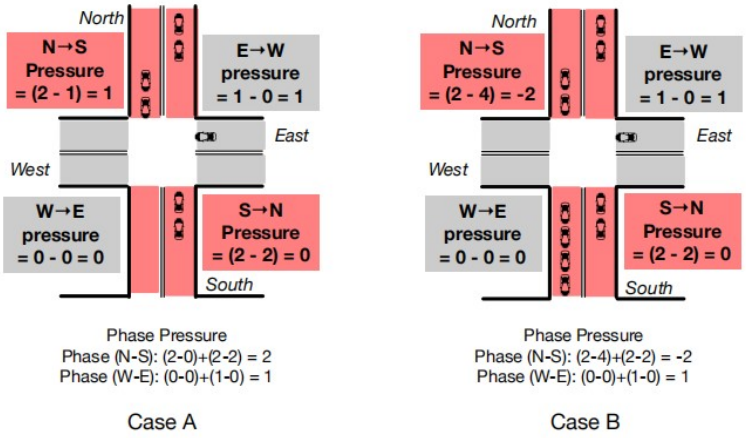

图 2.3 路口敏感性说明

的车辆数减去相应的传出车道上的车辆数；相位的压力定义为传入通道和传出通道上的总队列长度之间的差异。Varaiya 等人证明了当将优化目标设为最小化单个路口的相位压力时，Max-Pressure Control 可以最大限度地提高真个路网的吞吐量。

下表列出了每种方法的限制和要求：

表 2.1 常见的协作策略

方法	先验信息	输入	输出
Webster	相位序列	交通流量	基于周期的单个路口信号计划
GreenWave	信号计划	交通流量、速度限制、车道长度	基于周期的信号计划的偏移量
Actual Control，SOTL	相位序列	交通流量	是否变化到下一个相位
Max-Pressure Control	无	队列长度	所有交叉口的信号计划

2.4 基于强化学习的交通信号控制

最近，人们提出了不同的人工智能技术来控制交通信号，例如遗传算法、群体智能以及强化学习。其中在这些技术中，强化学习在近年来更具趋势。

2.4.1 强化学习概述

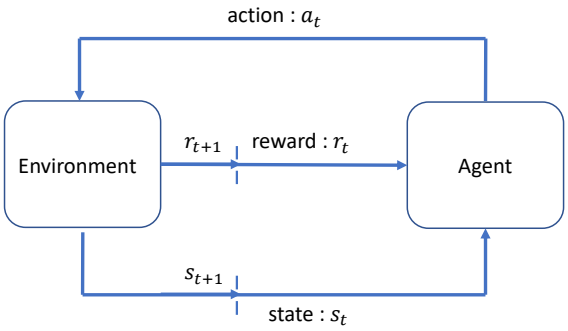

图 2.4 马尔可夫决策过程

通常单智能体强化学习问题被建模成马尔可夫决策过程（Markov Decision Process, MDP, 图 2.4）$< \mathcal{S}, \mathcal{A}, P, R, \gamma >$，其中 $\mathcal{S}, \mathcal{A}, P, R, \gamma$ 分别表示状态集、动作集、概率状态转移函数、奖励函数和折扣因子。具体定义如下：

- $\mathcal{S}$：在时间步骤 t，智能体得到一个观测状态 $s^t \in \mathcal{S}$。
- $\mathcal{A}, P$：在时间步骤 t，智能体采取一个动作 $a^t \in \mathcal{A}$，然后环境根据状态转移函数转移到一个新的状态。

$$P\left(\mathrm{s}^{t+1} \mid \mathrm{s}^t, \mathrm{a}^t\right) : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{S} \tag{2.1}$$

- R：在时间步骤 t，智能体通过奖励函数获得一个奖励 r^t。

$$R\left(s^t, a^t\right) : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R} \tag{2.2}$$

- γ：智能体的目标是找到一种使预期收益最大化的政策，即折扣奖励之

和。折扣因子决定了即时奖励与未来奖励的重要性。

$$G^t := \sum_{i=0}^{\infty} \gamma^i r^{t+i} \tag{2.3}$$

2.4.2 基于强化学习交通信号控制框架

2.4.3 基本要素

使用强化学习来解决交通信号控制问题要先确定以下几个基本要素:

奖励设计：由于强化学习是以最大化累计奖励为目标来学习的，所以奖励的选择决定了学习的方向。在交通信号控制问题中，虽然最终目标是尽量减少所有车辆的通行时间，但由于几个原因，通行时间很难直接作为 RL 的有效奖励。首先，车辆的行驶时间不仅受交通信号灯的影响，还受车辆自由流动速度等其他因素的影响。其次，当交通信号控制器事先不知道车辆行驶目的地（在现实世界中往往是这样），优化道路上所有车辆的通行时间变得特别困难。在这种情况下，车辆的通行时间只能在多个动作完成后车辆完全离开路口后才能测量。已有工作的奖励设计通常是基于一些可以直接在一个动作后测量的指标的加权和。例如，等待车辆的队列长度、车辆等待时间、速度、累计延迟、路口的吞吐量、车辆平均停车次数、信号变化频率（信号在一定时间段内变化的次数，学习到的策略不应该太过频繁的改变信号）以及路口的压力（Max-pressure 中定义的 pressure）等。虽然将奖励定义为几个因素的加权线性组合是现有研究中的一种常见做法，并且取得了不错的效果，但是这种特别的设计存在两个问题。第一，无法保证最大化设计的奖励等价于最优的通行效率，因为它们在交通运输理论中没有直接联系。第二，调整每个奖励函数因子的权重是相当棘手的，在权重设置上的微小差异可能会导致最终的结果有显著的差别。

状态表示：状态表示是以一种数值化的形式来描述路口的交通状况，描述的越全面越有利于学习到最优策略，通常使用多个要素组合来描述交通状况，例如，队列长度、车辆等待时间、车辆数量（包含非等待车辆），车辆速度、车辆位置分布以及当前信号灯的相位等。最近，在基于 RL 的交通信号控制算法中出现了使用更复杂状态的趋势，希望能够更全面地描述交通状况。Mousavi、

Van derPol 以及 Wei Hua 等人在他们的研究工作中提出使用位置图片来当作状态描述。但是，具有如此高维度的状态学习往往需要大量的训练样本，这意味着训练 RL 代理需要很长时间。更重要的是，较长的学习进度不一定会导致显著的性能增益，因为代理可能有更困难的时间从状态表示中提取有用的信息。因此，状态的表示应该简洁且能够充分地描述环境。

动作选择机制：动作选择机制决定了以何种方式来控制信号灯，不同的动作机制有不同的影响。主要可以总结为以下四种方式：

- 确定当前相位时长：在这中动作选择机制下，智能体学习通过从预定义的候选时间段（比如，10 秒、15 秒、20 秒等）中选择来设置当前相位的持续时间。
- 确定基于周期的相位比：这种方式定义的动作为下一个周期的相位分裂比（phase split ratio）通常，给出总周期长度，并预先定义一个包含一些相位比的候选集。
- 保持或改变当前相位：这种方式也是基于周期性的信号计划，通常一个二进制数来定义动作。例如，1 表示保持当前相位，0 表示变换到下一相位。
- 选择下一个相位：这种方式直接从待选相位序列中选择一个相位并变化到该相位，其中相位序列不是预定的。因此，这种信号控制方式更加的灵活，智能体学习在不同的而状态下选择最优的相位，而不假设信号会以循环的方式改变。

学习算法：强化学习发展至今已经提出了很多不同的算法，根据估计潜在奖励和选择动作的不同可以分为以下两种：

- Value-based Methods：基于值的方法近似于状态-值函数或状态-动作值函数（即，如何奖励每个状态或状态-动作对），策略是从学习的值函数隐式获得的。基于于价值的方法（Q-learning 和 DQN 等），直接模拟状态值或状态动作值（例如，在当前交通情况下，如果进行一个动作，平均

速度的增加/减少将生效多少?）。这样，状态和奖励就可以直接输入模型，而不需要额外的处理。然而，这些方法通常与 -greedy 的动作选择方法相结合，因此当 最终衰减到一个很小的数目时，将导致一个几乎确定性的策略（即，在某些状态下的动作是确定性的）。这可能会导致智能体陷入一些看不见的或代表性不足的情况，而没有改进。此外，这些方法只能处理离散的动作，因为它需要对每个动作进行单独的建模过程。

- Policy-based Methods：基于策略的方法直接更新策略（例如，在特定状态下采取行动的概率向量) 参数，以最大限度地实现预定目标 (例如平均预期回报）。基于策略的方法，尝试学习某一状态下不同动作的概率分布。基于政策的方法的优点是，它不要求行动是离散的。此外，它可以学习一个随机策略，并继续探索潜在的更有价值的行动。Actor-Critic 基于策略的方法中广泛使用的框架之一。它包括基于价值的思想来学习行动概率分布的策略，Actor 控制我们的智能体的行为（policy-based)，Critic 衡量所采取的行动有多好 (value-based）。Aslani、Mousavi、Prashanth 以及 Bhatnagar 在他们的工作中使用 Actor-Critic，利用价值函数逼近和策略优化的优势，在交通信号控制问题上表现出优异的性能。

2.5 基于图神经网络的交通流量预测

2.6 图神经网络概述

第三章 单路口场景交通信号调度

3.1 相关工作

最近，强化学习算法在交通信号控制领域表现出了优异的性能，这些算法将当前道路上的交通状况当作状态，并通过与环境交互学习操控信号的策略。现有的基于强化学习的交通信号控制方法之间的差异主要体现在以下这三个方面：状态表示、奖励设计以及学习算法。

状态表示是对当前环境的定量描述。一些常见的状态特征包括：

- 队列长度（Queue length）：队列长度是车道上处于“等待”状态的车辆的数量。对于车辆“等待”状态，有不同的定义。在 [1] 中，速度小于 0.1米/s 的车辆被认为处于等待状态；在 [2,3] 中，等待车辆是指没有移动位置的车辆。
- 等待时间 (Waiting time)：车辆的等待时间定义为车辆处于“等待”状态的时间段。等待期的开始时间的定义可能有所不同，在 [1,4]，他们认为等待时间是从车辆移动的最后一个时间戳开始，而 [5,6] 认为等待时间是从车辆进入路网开始的。
- 交通流量（Traffic volume）：交通流量定义为车道上的车辆数量，等于该车道上处于等待状态的车辆和行驶车辆的总和。
- 相位（Phase）：将相位信息作为状态特征，首先要将其进行量化，[1,4] 用当前相位在预先定义的信号相位组中的索引值来确定。

通常情况下，状态描述会整合多个特征，以获得对交通状况更全面的描述。

在交通信号控制问题中，尽管最终的目标是使所有车辆的行驶时间最小化，但由于一些原因，行驶时间很难直接作为 RL 的有效奖励。首先，车辆的行驶时间不单单受交通信号的影响，还与其他的因素有关，例如车辆的速度。

其次，当交通信号控制器事先不知道车辆的目的地时（现实世界中经常出现这种情况），优化网络中所有车辆的行驶时间变得尤为困难。在这种情况下，只有当网络中的多个交叉路口采取了多项行动时，才能在车辆完成行程后测量车辆的行驶时间。因此，奖励功能通常被定义为下列因素的加权和，这些因素在智能体采取动作后可以被即刻测量出。

- 总队列长度：这里队列长度是所有 incoming lanes 的队列长度之和。[7] 证明了最小化队列长度相当于最小化所有车辆的行驶时间。
- 吞吐量：吞吐量定义为在最后一个动作后的特定时间间隔内通过交叉路口或离开网络的车辆总数。
- 速度：一个典型的奖励是取道路网中所有车辆的平均速度，平均速度越高意味着车辆行驶到目的地的速度越快，时间也就越短。
- 信号变化频率：信号改变的频率被定义为在某一时间段内信号改变的次数。直观地说，学到的政策不应该导致闪烁，即频繁地改变交通信号，因为车辆通过交叉口的有效绿色时间可能会减少。

3.2 已有工作中的不足

虽然目前已经有不少基于强化学习的智能交通信号控制的研究，但是已有的方法更多的只注重提高通行效率，例如最小化队列长度或者最大化吞吐量，而忽略了公平性问题。事实上，这样的目标会导致学习到一个有偏见的策略。例如可能会出现如图 3.1所示的"last-vehicle" 情况：N-S 方向的道路上有源源不断的车辆到达，而 E-W 方向上是该车流的最后一辆车（Last-Vehicle）。显然，为了最大化通行效率，会优先放行 N-S 方向上的车流，而 E-W 车道上的车要等所有的所有的车流通过才能够得到响应，这对 Last-Vehicle 来说是极其不公平的。

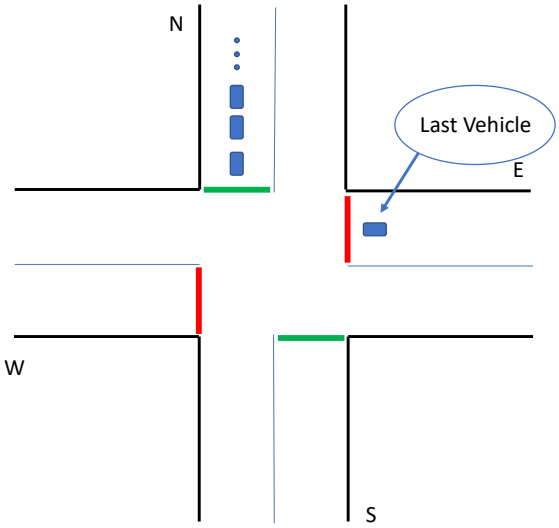

图 3.1 last vehicle

3.3 改进

3.3.1 目标

本工作的目的是在提高通行效率（最小化平均通行时间）的同时，希望每条车道能够有尽可能相同的服务延迟（得到放行所需的时间）。这个目标可以用以下的 Jain Fairness Index(JFI) 指标来量化：

$$\mathcal{J} = \frac{(\sum_{i=1}^{M} \overline{D}_i)^2}{M \sum_{i=1}^{M} \overline{D}_i^2}, \tag{3.1}$$

其中 $\overline{D}_i$ 是第 i 条进近车道的平均延迟。当且仅当每一个 $\overline{D}_i$ 都相等时，这个指标达到最大值，即是 1。所以我们的目标也就是最大化这个指标。

3.3.2 智能体设计

状态表示

在 t 时刻的状态 $S(t)$ 由以下几个部分组成：

1. 交通流量：$\boldsymbol{V}(t) = \{V_1(t), V_2(t), \cdots, V_M(t)\}$。其中 $V_i(t)$ 表示第 i 条进近车道上车的数量。值得注意的是，由于右转不受限于信号灯的特殊性，

这里我们不考虑右车道的交通流量。

2. 平均吞吐量：$\overline{\boldsymbol{L}}(t) = \{\overline{L}_1(t), \overline{L}_2(t), \cdots, \overline{L}_M(t)\}$。其中 $\overline{L}_i(t)$ 表示第 i 条进近车道的平均吞吐量。同上，不考虑右车道的平均吞吐量。
3. 信号相位：$\boldsymbol{P}(t)$ 是当前信号相位的数字化表示，1 表示绿色，可以通行；0 表示红色，禁止通行。

所以 $S(t) = \{\boldsymbol{V}(t)||\overline{\boldsymbol{L}}(t)||\boldsymbol{P}(t)\}$

动作选择

在本文中，动作选择机制是每次选择即将转换的信号相位。之后，交通信号灯将转换到这一新的相位并持续 Δt 的时间。为了安全起见，我们在两个不同的信号相位之间插入了 3 秒的黄色信号和 2 秒的红色信号。如果新选择的相位和当前相位相同，则不插入黄色和红色信号，以确保交通流畅。

奖励函数

受 PFS 分配原则的启发，我们设计了一个可以在效率和公平之间提供良好的平衡的奖励函数，如下所示：

$$r = -\sum_{i=1}^{M} \frac{Q_i(t)}{\overline{L}_i(t) + \delta}, \tag{3.2}$$

其中 $Q_i(t)$ 和 $\overline{L}_i(t)$ 分别是第 i 条进近车道的队列长度和平均吞吐量。在每一次调度后（这里，我们将一次动作选择视作一次调度），$\overline{L}_i(t)$ 按照以下方式进行更新：

$$\overline{L}_i(t) = (1 - \frac{1}{W})\overline{L}_i(t-1) + \frac{1}{W}L_i(t), \tag{3.3}$$

其中 $L_i(t)$ 是此次调度中车道 i 上得到放行的车的数量，W 是一个平衡通行效率和公平性的参数。另外，为了避免公式3.2的分母为 0，我们加上了一个可以忽略不计的正数 δ。

训练过程

如图 3.2所示，这里我们用 DQN 作为学习算法，并且采用经验回放[8]（experience replay）方法定期提取样本来更新模型。

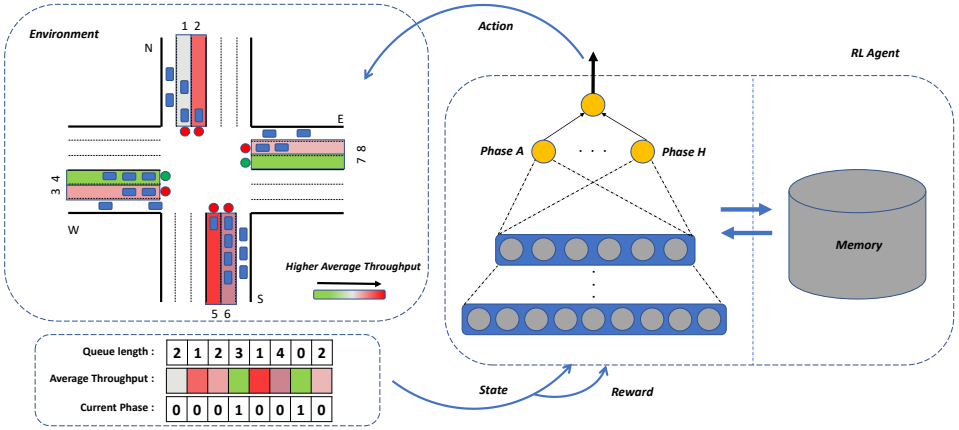

图 3.2　系统模型

3.4　实验

实验在 SUMO（simulation of Urban MObility）[1]仿真平台上进行，利用该模拟器可以方便地实时获取车辆状态，并通过改变交通信号来控制交通运行。我们实现了一个四路交叉口作为我们的实验场景，交叉口与四个 150 米长的路段相连，每条道路有三条引入车道和三条引出车道。

我们将 N-S 方向的道路设置为主干道，车辆到达量更多，将 W-E 方向的道路设置为次干道，车辆到达量较少。车辆到达服从泊松分布，这里我们设置 N-S 方向道路的交通流量比率为 ρ，W-E 方向道路的交通流量比率为 $1-\rho$，ρ 值越高，交通流量不平衡的状况越严重。为了对我们的方法进行综合评价，我们在不同的 ρ 值下进行了实验。注意，为了简化环境，这里我们不考虑行人交通的影响。

3.4.1　评价指标

我们使用以下指标来评估不同方法的效率和公平性表现：

- 行驶时间：车辆行驶时间是指车辆进出路口的时间差。现有的大部分工作都集中在最小化所有车辆通过交叉路口的平均行驶时间。

[1]http://sumo.dlr.de/index.html

- 延误时间：车辆延误时间是车辆通过交叉路口的实际时间与预期时间（以最高限速通过交叉路口所需的时间）之间的差值。
- 驾驶体验得分：此外，我们提出了一种新的评价指标，称为驾驶体验得分（Driving Experience Score，DES），来量化驾驶员的满意度，具体评分标准见下表：事实上，可能有更多的因素需要考虑（如燃油消耗），但是这里的目的是为了缓解车辆的过度延误情况，因此我们这里用延误时间作为评价标准。

表 3.1 驾驶体验得分标准

延误时间（s）	DES
$d \leq 40$	5
$40 < d \leq 80$	4
$80 < d \leq 120$	3
$120 < d \leq 160$	2
$d > 160$	1

3.4.2 比较方法

- FT(Fixed-Time Control[9])：这种方法以预先设定的方式循环改变信号。
- SOTL(Self-Organizing Traffic Light Control[10])：这是一种根据预先设定的阈值来改变信号的自适应方法。如果等待的车辆数量超过了这个阈值，则切换到下一个信号相位。
- LIT[7]：这是一种基于学习的方法，比大多数现有的致力于提高通行效率的方法效果更好。
- FIT(**F**airness-aware **I**ntelligent **T**raffic Light Control)：我们的方法。

3.4.3 性能评价

首先，我们通过实验评估了不同方法的通信效率的表现，为了得到一个综合的结果，我们在不同的 ρ 值下进行了实验，实验结果如图 3.3所示。可以观

察到，我们的方法（FIT）的车辆通过路口的平均行驶时间远低于传统方法（行驶时间越短意味着效率越高），并且仅略低于只注重效率的 LIT 方法。

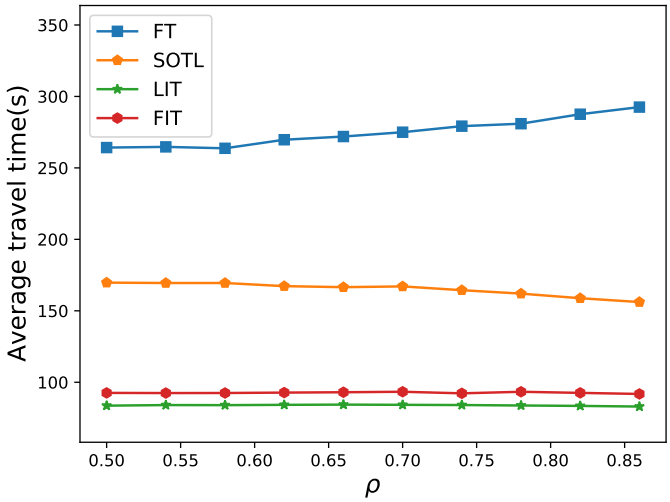

图 3.3　效率

其次，由于我们的主要研究目标是公平性，我们首先分析了在使用不同方法下每条车道的延误情况。这里我们用 Jain Fairness Index(JFI) 来量化公平性指标，JFI 的计算方式如下：

$$\mathcal{J} = \frac{\left(\sum_{i=1}^{M} \bar{D}_i\right)^2}{M \sum_{i=1}^{M} \bar{D}_i^2}, \tag{3.4}$$

其中 $\bar{D}_i$ 是车道 i 的平均延误时间。当每个车道具有相同的平均延误时间时，JFI 的值达到最大值，即 1。图 3.4展示了四种方法在不同 ρ 值下的平均延迟的 JFI 表现。从中我们可以看出 FT 和 LIT 的 JFI 值随着交通不平衡情况的加剧（即 ρ 值越大）而减小，而 FIT 任然能偶保持较高的值，并且高于同样能够保持稳定 JFI 值的 SOTL 方法。

然后，我们更加详细地研究了不同方法的延误情况。下面我们具体分析在主干道和支干道上四种方法在不同的 ρ 值情况下车辆延误时间分布情况。从图 3.5中我们可以看出，在主干道上，基于学习的方法（LIT 和 FIT）比传统

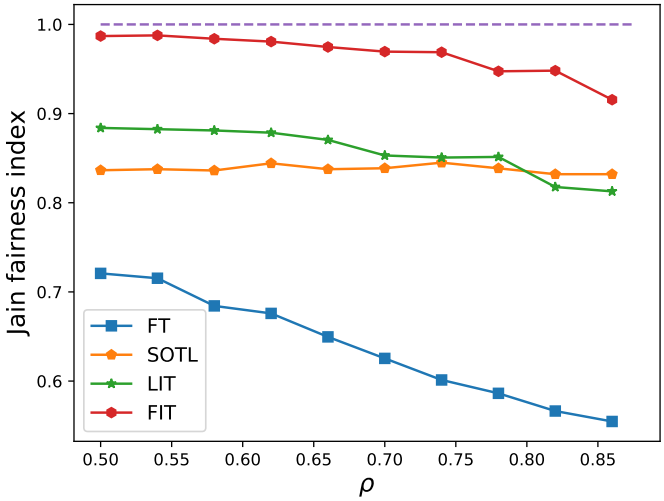

图 3.4　公平

方法（FT 和 SOTL）具有更低的延误时间，虽然 SOTL 方法整体上延迟也比较低，但是会有很多极端值，最高的延误时间甚至超过 800s。

从图 3.6中我们可以看出，在支干道上，随着 ρ 值的增加（即交通不平衡情况的加重），原先在主干道上表现优异的 LIT 方法性能开始恶化 ()，但是 FIT 依然能够保持一个相对低的延迟。

最后我们研究了不同方法的驾驶体验得分情况，图 3.7展示了在 $\rho = 0.75$ 的情况下不同方法的驾驶体验得分分布情况。从中我们可以看出，FT 方法超过半数的驾驶体验的分都是 1 分，由此可以看出该方法的不灵活。对于 SOTL 方法而言，虽然他的 5 分的比例最高，但是其得分分布的方差也是最高的。FIT 的得分分布与 LIT 相似，但 FIT 的方差低于 LIT，在以牺牲少量效率为代价的前提下。

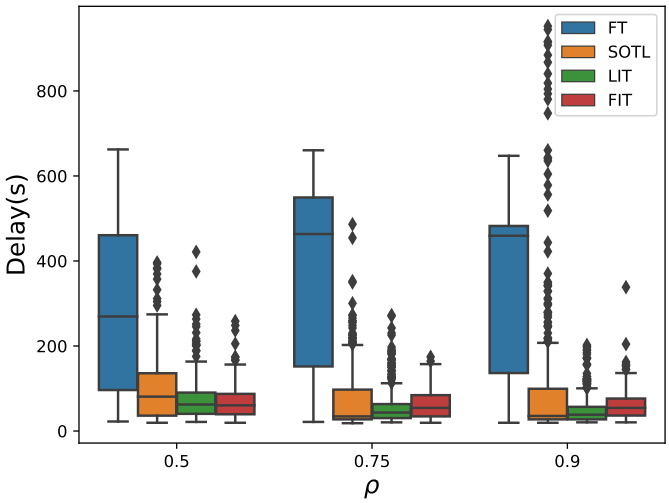

图 3.5　主干道 (N-S 方向) 的车辆延误时间分布

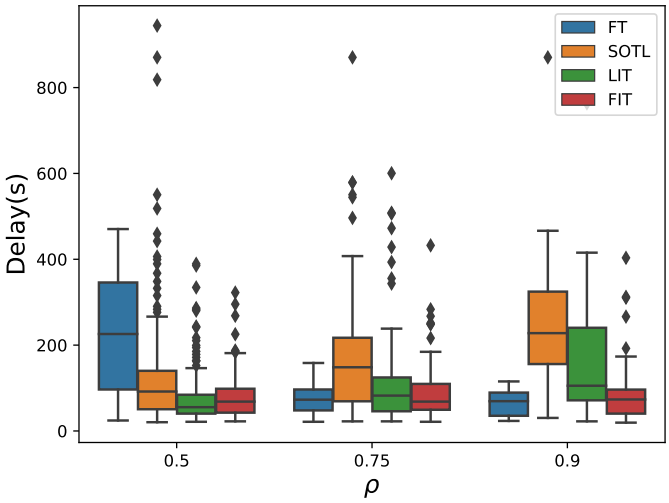

图 3.6　支干道 (N-S 方向) 的车辆延误时间分布

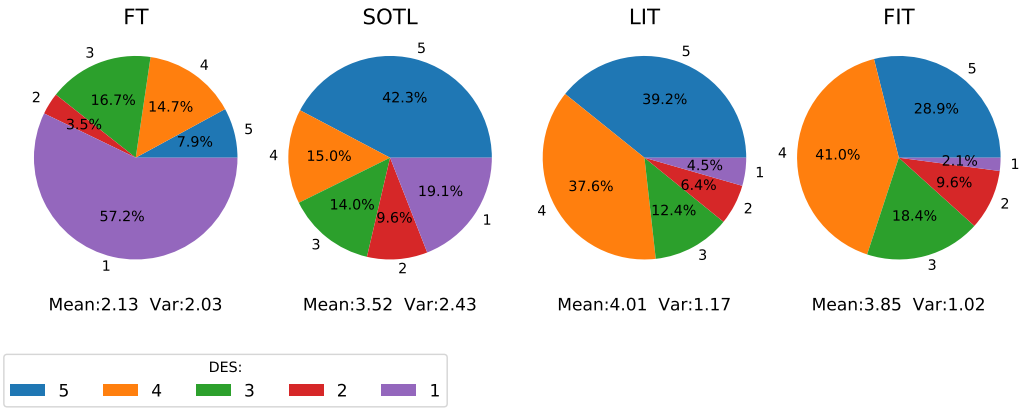

图 3.7　驾驶体验得分统计

第四章　多路口场景下交通信号控制

4.1　相关工作

由于强化学习在单路口交通信号控制上取得了优异的成绩，人们开始致力于使用多智能体强化学习（Multi-Agent Reinforcement Learning, MARL）来解决多路口场景下的交通信号调度。Claus 在 [11] 中将 MARL 分为了两类：联合动作学习（Joint Action Learning）和独立学习（Independent Learning）。

对于多路口信号控制，联合动作学习的思想就是使用一个全局智能体（single global agent）来控制所有的交叉路口，其动作是所有路口动作组合在一起的联合动作，然后通过迭代学习建模多个智能体的联合动作价值函数（Joint Action Value Function）：

$$Q(o_1, o_2, \cdots, o_N, \mathbf{a}) \tag{4.1}$$

其中 o_i 是智能体 i 对路口环境的观测，$\mathbf{a}$ 是所有智能体的联合动作。但是这种方法的缺点是会导致维度灾难（curse of dimensionality），状态动作的联合空间会随着智能体数量的增加呈指数级增长，增加学习的难度。为了缓解这个问题，[4] 使用 max-plus 方法将联合动作价值函数分解为局部子问题的线性组合，如下所示：

$$\hat{Q}\left(o_1, \ldots, o_N, \mathbf{a}\right) = \Sigma_{i,j} Q_{i,j}\left(o_i, o_j, \mathbf{a}_i, \mathbf{a}_j\right) \tag{4.2}$$

其中 i和j 对应于相邻智能体的索引。在 [12–14] 中，将联合 Q 值视为局部 Q 值的加权和：

$$\hat{Q}\left(o_1, \ldots, o_N, \mathbf{a}\right) = \Sigma_{i,j} w_{i,j} Q_{i,j}\left(o_i, o_j, \mathbf{a}_i, \mathbf{a}_j\right) \tag{4.3}$$

其中 $w_{i,j}$ 是预先定义的权重。他们试图通过在单个智能体的学习过程的损失函数中增加一个整形项，并使单个 Q 值的加权和与全局 Q 值的差异最小化，从而确保单个智能体在学习过程中能够考虑到其他智能体的情况。

多路口信号控制的另一条研究路线是使用独立的 RL（IRL）智能体来控制交通信号，其中每个 RL 智能体控制一个路口。与联合动作学习方法不同，每个智能体可以在不知道其他智能体的奖励信号的情况下学习控制策略。根据智能体之间是否进行信息交互进一步分为以下两类：

- IRL without Communication：IRL 单独处理每个交叉口，每个 agent 观察自己的本地环境，不使用显式通信来解决冲突 [6,15–20]。在一些简单的场景中，如动脉网络，这种方法表现良好，可以形成了几个小绿波（Green waves）。然而，当环境变得复杂时，来自相邻 agent 的非平稳影响将被带到环境中，如果 agent 之间没有通信或协调机制，学习过程通常无法收敛到平稳策略。为了应对这一挑战，wei 在 [21] 中提出了一个特定的奖励函数，去描述相邻智能体之间的需求从而实现协调。
- IRL with Communication：这种方法使智能体之间能够就他们的观察进行交流，并作为一个群体而不是个体的集合来完成复杂的任务，在这种情况下，环境是动态的，每个智能体的能力和对世界的可见度是有限的 [22]。典型的方法是直接将邻居的交通状况 [23] 或过去的动作 [24] 加入到自身智能体的观察中，而不是仅仅使用自我观测到的本地交通状况。在这种方法中，不同路口的所有智能体共享一个学习模型，这就需要对相邻的路口进行一致的索引。[25] 试图通过利用图卷积网络的路网结构来消除这一要求，以协作附件的多跳路口的交通，并且通过图卷积网络中定义的固定邻接矩阵来模拟相邻代智能体的影响，这表明他们假设相邻智能体之间的影响是静态的。在其他工作中，[26,27] 提出使用图注意网络来学习相邻智能体和自我智能体的隐藏状态之间的动态相互作用。应该指出的是，利用 max-plus 学习联合行动学习者的方法和利用图卷积

网络学习通信的方法之间有很强的联系，因为它们都可以被看作是学习图上的信息传递，其中前一种方法传递奖励，后一种方法传递状态观测信息。

4.2 已有工作中的不足

目前大多数工作在使用图神经网络 Learn to Communicate 的时候，都是以 intersection 为节点来进行图建模，将每一个路口视作图中的一个节点，每条道路作为连接两个节点的边，很自然地可以将一张交通道路网建模成一个图，如图 4.1所示：在这种建模方式下，每条车道的车辆以及当前的相位将作为该

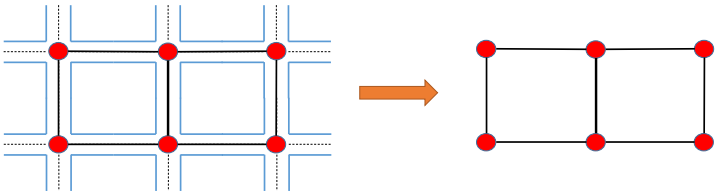

图 4.1 多路口建模成图（路口）

节点的特征。这种建模方式虽然可以很清晰的将多路口场景变成一张图。但是，因为是以一个路口为一个节点，所有车道的状态信息都整合到了一起，有些车道的的信息对目标节点是无用的，如图 4.2所示：路口 B 中只有 2 车道的交通流向与 A 车道有关，1、3 车道的车辆不会行驶到 A 路口。在信息传递的时候，如果将所有的信息都笼统地传递过去，将会增加 A 提取有效信息的难度，从而降低学习的效率。

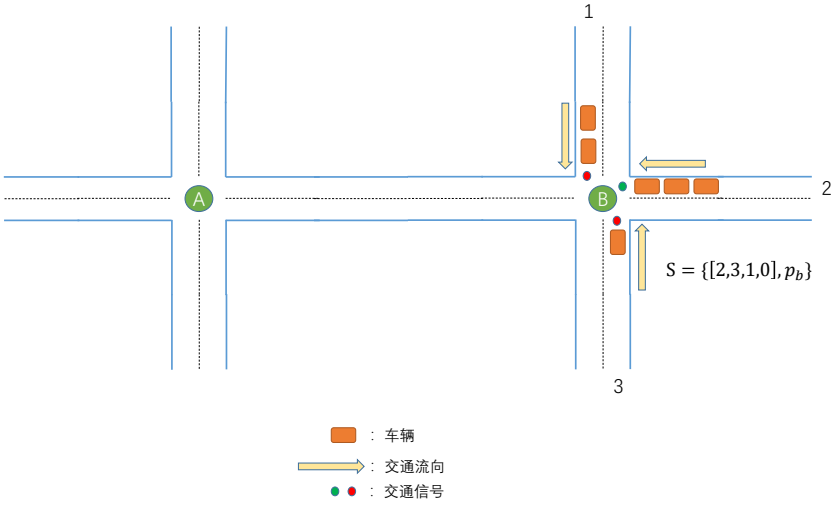

图 4.2　按路口建图模式下信息传递

4.3　改进

本文同样采用 IRL with Communication 的框架，与已有工作不同的是，我们采用不同的建模方式：以道路为节点进行图建模，即一条道路就是一个节点，如图 4.3所示：

此外，我们根据当前的相位对图的边设一个权重。这里我们规定，如果在当前相位下，道路 i 到道路 j 之间的交通是允许通行的，则表示 (i,j) 的状态是’connected’。权重的定义方法如下：

$$w_{i,j} = \begin{cases} 1 & (i,j) \text{ is connected} \\ 0 & \text{otherwise} \end{cases} \tag{4.4}$$

4.4　实验

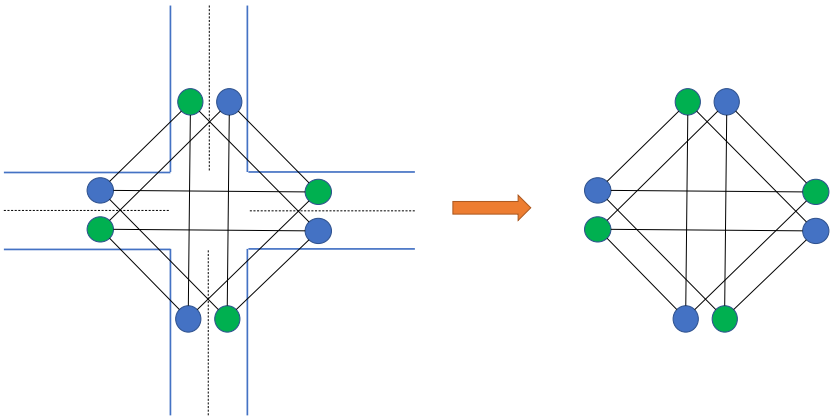

图 4.3　按道路建模成图

参考文献

[1] Wei H, Zheng G, Yao H, et al. Intellilight: A reinforcement learning approach for intelligent traffic light control[C]. Proceedings of Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018. 2496–2505.

[2] Steingrover M, Schouten R, Peelen S, et al. Reinforcement Learning of Traffic Light Controllers Adapting to Traffic Congestion.[C]. Proceedings of BNAIC, 2005. 216–223.

[3] Kuyer L, Whiteson S, Bakker B, et al. Multiagent reinforcement learning for urban traffic control using coordination graphs[C]. Proceedings of Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, 2008. 656–671.

[4] Pol E, Oliehoek F A. Coordinated deep reinforcement learners for traffic light control[J]. Proceedings of Learning, Inference and Control of Multi-Agent Systems (at NIPS 2016), 2016..

[5] Brys T, Pham T T, Taylor M E. Distributed learning and multi-objectivity in traffic light control[J]. Connection Science, 2014, 26(1):65–83.

[6] Pham T T, Brys T, Taylor M E, et al. Learning coordinated traffic light control[C]. Proceedings of Proceedings of the Adaptive and Learning Agents workshop (at AAMAS-13), volume 10. IEEE, 2013. 1196–1201.

[7] Zheng G, Zang X, Xu N, et al. Diagnosing reinforcement learning for traffic signal control[J]. arXiv preprint arXiv:1905.04716, 2019..

[8] Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning[J]. nature, 2015, 518(7540):529–533.

[9] Miller A J. Settings for fixed-cycle traffic signals[J]. Journal of the Operational Research Society, 1963, 14(4):373–386.

[10] Cools S B, Gershenson C, D' Hooghe B. Self-organizing traffic lights: A realistic simulation[M]. . Proceedings of Advances in applied self-organizing systems. Springer, 2013: 45–55.

[11] Claus C, Boutilier C. The dynamics of reinforcement learning in cooperative multiagent systems[J]. AAAI/IAAI, 1998, 1998(746-752):2.

[12] Zhang Z, Yang J, Zha H. Integrating independent and centralized multi-agent reinforcement learning for traffic signal network optimization[J]. arXiv preprint arXiv:1909.10651, 2019..

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

[14] Tan T, Bao F, Deng Y, et al. Cooperative deep reinforcement learning for large-scale traffic grid signal control[J]. IEEE transactions on cybernetics, 2019, 50(6):2687–2700.

[15] Mannion P, Duggan J, Howley E. An experimental review of reinforcement learning algorithms

for adaptive traffic signal control[J]. Autonomic road transport support systems, 2016. 47–66.

[16] Casas N. Deep deterministic policy gradient for urban traffic light control[J]. arXiv preprint arXiv:1703.09035, 2017..

[17] Zheng G, Liu H, Xu K, et al. Learning to simulate vehicle trajectories from demonstrations[C]. Proceedings of 2020 IEEE 36th International Conference on Data Engineering (ICDE). IEEE, 2020. 1822–1825.

[18] Liu X Y, Ding Z, Borst S, et al. Deep reinforcement learning for intelligent transportation systems[J]. arXiv preprint arXiv:1812.00979, 2018..

[19] Calvo J A, Dusparic I. Heterogeneous Multi-Agent Deep Reinforcement Learning for Traffic Lights Control.[C]. Proceedings of AICS, 2018. 2–13.

[20] Gong Y, Abdel-Aty M, Cai Q, et al. Decentralized network level adaptive signal control by multi-agent deep reinforcement learning[J]. Transportation Research Interdisciplinary Perspectives, 2019, 1:100020.

[21] Wei H, Chen C, Zheng G, et al. Presslight: Learning max pressure control to coordinate traffic signals in arterial network[C]. Proceedings of Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019. 1290–1298.

[22] Sukhbaatar S, Fergus R, et al. Learning multiagent communication with backpropagation[J]. Advances in neural information processing systems, 2016, 29:2244–2252.

[23] Xu M, Wu J, Huang L, et al. Network-wide traffic signal control based on the discovery of critical nodes and deep reinforcement learning[J]. Journal of Intelligent Transportation Systems, 2020, 24(1):1–10.

[24] Ge H, Song Y, Wu C, et al. Cooperative deep Q-learning with Q-value transfer for multi-intersection signal control[J]. IEEE Access, 2019, 7:40797–40809.

[25] Nishi T, Otaki K, Hayakawa K, et al. Traffic signal control based on reinforcement learning with graph convolutional neural nets[C]. Proceedings of 2018 21st International conference on intelligent transportation systems (ITSC). IEEE, 2018. 877–883.

[26] Wei H, Xu N, Zhang H, et al. Colight: Learning network-level cooperation for traffic signal control[C]. Proceedings of Proceedings of the 28th ACM International Conference on Information and Knowledge Management, 2019. 1913–1922.

[27] Wang Y, Xu T, Niu X, et al. STMARL: A spatio-temporal multi-agent reinforcement learning approach for cooperative traffic light control[J]. IEEE Transactions on Mobile Computing, 2020..

致　谢

在此感谢对本论文作成有所帮助的人。