中图分类号：TP371　　论文编号：*****16 22-S038
学科分类号：080605

硕士学位论文

基于深度强化学习的智能交通信号调度研究

研究生姓名　　×××
学科、专业　　计算机科学与技术
研 究 方 向　　网络与分布计算
指 导 教 师　　×××

×××

×××

二〇二一年十二月

×××
×××
×××

Traffic Signal Control Based on Deep Reinforcement Learning

A Thesis in

Compute Science and Technology

by

×××

Advised by

×××

Submitted in Partial Fulfillment

of the Requirements

for the Degree of

Master of Engineering

December, 2021

承诺书

本人声明所呈交的硕士学位论文是本人在导师指导下进行的研究工作及取得的研究成果。除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得×××或其他教育机构的学位或证书而使用过的材料。

本人授权×××可以将学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。

（保密的学位论文在解密后适用本承诺书）

作者签名：_____

日　　期：_____

摘　要

随着近些年来车辆数量的不断增加，交通拥塞情况已经变得越来越严重，并且极大地影响了人们的日常生活和城市的运作。传统的交通信号控制方法由于受限于严苛的假设条件以及没有考虑实时的交通状况，难以在现在更为复杂的交通模式下起到很好的作用，如何设计出能够进一步提高通行效率的智能交通信号调度方法是一个亟需解决的问题。

随着人工智能技术的不断发展以及对实时交通数据的获取变得更加容易，使得根据实时交通状况动态调整信号这一想法成为可能。一些研究工作提出使用强化学习实现交通信号控制，与传统的方法相比，这类方法的性能更加出色。然而，现有的方法仍然有一些需要改进的地方。例如，在单路口场景下，大多数基于深度强化学习的交通控制方法只注重于提高路口的通行效率，而忽略了对公平性的考虑，这会导致学习到一个有偏见的策略，即优先服务交通流量大的车道，而忽略交通流量较小的车道上的车辆；在多路口场景下，现有的方法在通过信息交互实现协调控制的过程中，笼统地将自己路口的所有信息传递给目标路口，这使得目标路口难以挖掘出对自身有用的信息，增加了学习的难度。在本文中，我们对已有工作在这两种场景下的不足进行了改进。

首先，对于单路口场景下的智能交通信号调度，通过引用无线网络中的比例公平调度方法（Proportional Fair Scheduling，PFS），我们提出了一个具有公平感知能力的基于深度强化学习的智能交通信号调度模型，这个调度模型可以在效率和公平性之间提供一个良好的权衡，并且可以有效地解决小交通流量的“饥饿等待”问题。为了验证了模型的效果，我们进行了大量的实验，并与已有方法进行对比，进一步阐述了我们模型在公平性方面的性能提升。

其次，在多路口场景下，路口之间进行信息交互可以有效地实现协调控制。为了解决已有工作在信息交互过程中出现的信息冗余情况，我们提出了一种新的路网建图方式并在此基础上设计了一种新的基于图神经网络的信息交互模块，可以有效地剔除数据交互过程中邻居节点的无效信息。为了展示我们的模型与已有方法相比在通行效率和学习速度上的提升，我们在仿真环境中分别就合成交通数据和真实交通数据进行了实验。

关键词：交通信号控制，强化学习，公平性，协调控制，图神经网络

ABSTRACT

With the increasing number of vehicles in recent years, traffic congestion has become more and more serious, and has greatly affected people's daily life and urban operation. The traditional traffic signal control methods have been difficult to play a good role in the more complex traffic mode because they are limited by strict assumptions and do not consider the real-time traffic conditions. How to design an intelligent traffic signal scheduling method that can further improve the traffic efficiency is an urgent problem to be solved.

With the continuous development of artificial intelligence technology and the easier acquisition of real-time traffic data, it is possible to dynamically adjust the signal according to the real-time traffic conditions. Several studies have proposed to use reinforcement learning (RL) for traffic signal control and achieved superior performance compared with the traditional methods. However, existing methods still have something to be improved. For example, in the single intersection scenario, most traffic control methods based on deep reinforcement learning only focus on improving the traffic efficiency of the intersection and ignore the consideration of fairness, which will lead to learning a biased strategy, that is, giving higher priority to serving the lanes with large traffic flow and ignoring the vehicles in the lanes with small traffic flow. In the multi intersection scenario, in the process of realizing coordinated control through information interaction, the existing methods generally transfer all the information of their own intersection to the target intersection, which makes it difficult for the target intersection to mine useful information for themselves and increases the difficulty of learning. In this thesis, we improve the existing work in these two scenarios.

For intelligent traffic signal scheduling in single intersection scenario, by referring to the Proportional Fair Scheduling (PFS) method in wireless network, we propose an intelligent traffic signal scheduling model based on deep reinforcement learning with fairness perception, which can provide a good trade-off between efficiency and fairness, also, this model can effectively solve the "hungry waiting" problem of small traffic flow. To verify the effect of our model, we conduct comprehensive experiments, and compare with the existing methods to further illustrate the performance improvement of our model in terms of fairness.

In the multi intersection scenario, information interaction between intersections can be effective for coordinated control. In order to solve the information redundancy in the information interaction process of existing work, we propose a new method of transforming road structure into graph, and design a

new information interaction module based on graph neural network, which can effectively eliminate the invalid information of neighbor nodes in the process of information interaction. In order to show the improvement of traffic efficiency and learning speed of our model compared with existing methods, we conduct comprehensive experiments on synthetic traffic data and real traffic data in the simulation environment.

Keywords: Traffic Signal Control, Reinforcement Learning, Fairness-Aware, Coordination Control, Graph Neural Network

目　录

图表清单

注释表

$\mathcal{S}$	状态集	$\mathcal{A}$	动作集
P	状态转移函数	R	奖励函数
γ	折扣因子	G^t	累积（折扣）奖励
ϵ	随机动作选取概率	$\mathcal{L}$	损失函数

缩略词

缩略词	英文全称
PFS	Proportional Fair Scheduling
DQN	Deep Q-Learning
SOTL	Self-Organizing Traffic Light Control
MDP	Markov Decision Process
RNN	Recurrent Neural Network
CNN	Convolutional Neural Networks
GNN	Graph Neural Networks
PPMI	Positive Pointwise Mutual Information
GCN	Graph Convolutional networks
GAT	Graph Attention Network
RL	Reinforcement Learning
DRL	Deep Reinforcement Learning
MARL	Multi-Agent Reinforcement Learning
IRL	Independent Reinforcement Learning

第一章　绪论

1.1　研究背景及意义

1.1.1　研究背景

近些年来，经济的快速增长推动了城市化发展的进程，人民生活水平日益提高，但与此同时一些“城市病”也逐渐显现出来，例如城市交通拥堵。在过去五年内，中国机动车辆的保有量在逐年增加，如表1.1所示，在2016年，中国汽车保有量超过一百万的城市数量有49个，而截至2020年汽车保有量超过一百万的城市数量已经突破了70，其中汽车保有量突破三百万的城市更是达到了13个。不断恶化的交通拥堵情况，极大的影响了人们的日常生活，人们花费在交通通勤的时间变长也越来越常态化并且这种现象已经开始向中小城市蔓延。为此，很多城市开始研究不同的解决方案以缓解不断恶化的交通拥堵情况。对于基础建设已经趋于饱和的大城市而言，例如北京，通过出台“限号”政策来减少出行路面上的车辆数量，而对于正在发展中的中小城市而言，更多的是选择通过加快城市道路建设来加大城市交通的承载量。虽然这些方法在一定程度上使交通拥堵得到了缓解，但是并没有从根本上解决问题。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，在城市道路中，这一“行驶资源”通常就是车辆所处的交通路口，所以现代城市交通管理的主要方法是在道路汇合的交叉路口安装信号灯并通过简单的策略来调度通过的车流，已到达减少交通拥堵的目的。但是随着车辆数量的不断增加，之前传统的交通信号控制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交通信号调度策略显得格外的重要。

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，因此如何根据实时的交通状况来制定最优的调度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分布是相对稳定的，例如服从均匀分布，这一假设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（Deep Reinforcement Learning，DRL）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。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，例如变换到下一个信号相位。这种方法对于车流的分布情况没有限制，通过在大量不同的仿真车流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。

表 1.1　2016-2020 年汽车保有量城市数据统计 (数据来源：公安部)

城市汽车保有量	单位	2016 年	2017 年	2018 年	2019 年	2020 年
汽车保有量超百万城市数量	个	49	53	61	66	70
汽车保有量超两百万城市数量	个	18	24	27	30	31
汽车保有量超三百万城市数量	个	6	7	8	11	13
汽车保有量超五百万城市数量	个	1	1	1	2	3

1.1.2　研究意义

(1) 传统交通信号控制方法的不足

传统的交通信号控制方法在当下已经难以有效地减轻交通拥堵的情况。对于使用预先定义信号方式的控制方法来说，由于其信号规则是根据历史交通数据计算出来的，没有考虑实时的交通状况，所以难以有效地提高系统的通行效率。而对于使用优化方法的信号控制策略来说，由于其假设条件太过严苛，导致难以应用到实际场景中。

(2) 基于深度深度强化学习的智能交通信号调度的优点

随着最近强化学习技术的发展，我们看到学术界对使用强化学习来改善交通信号控制的热情越来越高涨，并且也提出了很多基于深度强化学习的智能交通信号控制方法。与传统的交通信号控制方法相比，基于深度强化学习的智能交通信号调度方法有诸多优点。首先，它能够根据实时的交通状况更改交通信号，并且可以在与环境的迭代中不断地修改和提升自身调度策略的性能，从而最大程度上的提高路口的通行效率。其次，强化学习是一个无模型的方法，智能体能够在训练中进行自我学习，而不需要人为地在外部监督和操作，也不需要环境的先验知识。再者，它不需要预先定义一个能够涵盖环境中影响系统性能的所有因素的模型。

(3) 目前基于深度强化学习的智能交通信号调度工作的不足

虽然目前已经有不少关于深度强化学习控制交通信号的研究工作，并且与传统方法相比也取得了不错的性能提升，但是任然有一些问题值得深入研究，例如公平性问题和协调通信问题。

公平性问题是指，不同车辆通过同一个路口所需的通行时间可能有很大的差别，因为信号灯可能为了提高整体通行效率而牺牲一些车辆，让这些车辆多等待一些时间，即便这些车辆可能是先进入路口的，这种做法对这些车来说是不公平的。一个好的控制策略应该在提高通行效

率的同时能够保证每辆车所需的通行时间大致相同，也就是说，车辆通行时间的方差应该越小越好。但是已有的工作都是使用车辆的平均通行时间来衡量通行效率，很自然的忽略了公平性问题。

协调通信问题是指在多路口交通信号控制问题中，路口之间通过信息交互来实现多个路口的协调控制控制，从而提高整个路网的通行效率。但是已有的工作在进行信息交互时，笼统地将自身路口所有的信息传递给目标路口，会导致目标路口难以提取出有效的信息，从而增加学习的难度，甚至学习出错误的策略。

1.2 国内外研究现状

交通信号控制是城市交通管理中最有效的方法之一，为每个交叉路口提供更顺畅、更安全的交通流。从简单的自动信号控制器问世以来，交通信号控制系统一直在不断地改进，以解决造成交通信号控制障碍的因素。图 1.1展示了近 20 年来国内外关于交通信号调度的研究趋势，

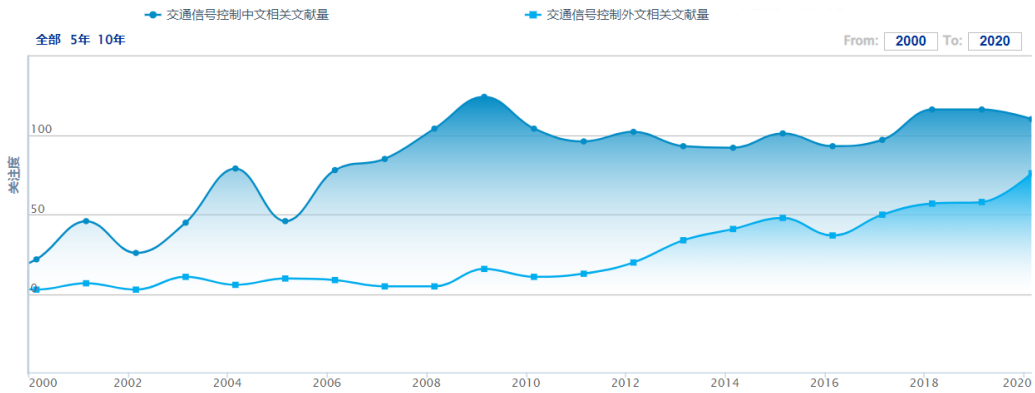


图 1.1 交通信号调度国内外研究趋势（来源：中国知网研究指数统计）

从中可以看出，相较于国外而言，早期国内对于交通信号调度研究的关注相对较少，但是随着近些年来国内交通拥堵情况的加重，国内学者对这方面的研究越来越感兴趣。

目前，已有的关于交通信号调度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智能交通信号控制。

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)[1]，这类方法根据历史交通流量信息预先计算好交通信号的序列和周期长度，当部署到交通信号灯上就循环的在之前定义的信号序列上变换。由于没有考虑实时的交通情况，所以这类方法不能够有效地适应动态的交通变化，因此通行效率相对较低。第二类是基于优化方法的自适应控制，这类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[2–4]。例如，为了优化车辆通行时间，Webster 在工作 [5] 中假设车辆的到达率是均匀

的，由于这些假设条件通常太过理想化，所以难以适用于现实世界的交通流量。此外，这类方法仅考虑当前（或）瞬时的交通状况，例如，它根据车道上时候有车辆来决定是否激活绿色信号，而不是长期的交通状况（如，队列长度）。因此，传统的交通信号控制方法缺乏对当前和长期的交通状况的同时考虑。目前，人们已经开始研究能够优化交通信号调度和计时的交通信号控制，如在适当的时候延长绿灯时间，以改善中载或重载交通的单个或多个交叉路口的交通瓶颈。

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过多的假设条件，此外，强化学习作为一种自适应的方法，能够对当前和长期的交通状况的动态性做出反应。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自身的控制策略，逐渐达到最优的效果。2003 年，Abdulhai[6] 首次提出了使用强化学习来自适应地改变交通信号的控制思路，为后续的研究工作指明了方向。在研究初期，智能体的状态表示都是定义成离散值的形式[6–10]，例如车辆的位置或者等待的车辆数量，这种做法的副作用就是离散的状态-动作值函数矩阵需要巨大的存储空间，这使得这些方法无法用于大型状态空间问题。为了解决这个问题，Li[11] 和 Van der Pol[12] 使用连续状态表示的深度强化学习算法 DQN（Deep Q-learning）来降低模型存贮的难度。这一类方法通过学习一个价值函数来将状态和动作对映射到奖励空间上，他们的主要区别体现在状态表示（例如，队列长度[13]、平均延误时间[14?] 和图像特征[12,15,16] 等）以及奖励设计（例如，平均延误时间[12,17]、平均通行时间[12,16] 和队列长度[13] 等）。然而，这些方法都使用了相对静态的交通环境，因此与实际交通场景想去甚远。为此，Wei 在 [18] 工作中提出了一个新的基于强化学习的信号控制模型并首次在真实的交通数据集上进行了实验。

随着深度强化学习在单路口场景下的交通信号控制问题中取得了出色的表现，人们开始将问题拓展到多路口场景下。相对于单路口而言，多路口场景交通信号控制问题中需要考虑可拓展性问题和协调控制问题。针对可拓展性问题，Mannion 在 [13] 工作中使用隔离处理的方法，即对于路网下的每一个路口使用单独的交通信号控制方法。这种方法可以轻松地应对路网规模的变化，但是由于忽略了相邻路口的交通状况，所以无法实现协调控制。关于协调控制，一种方法是通过对一个区域内的所有控制智能体进行集中优化[12] 以确保最优性。然而，随着路网规模的扩大，联合动作空间也会呈指数增长，阻碍了这种方法在大规模交通信号控制中的应用。还有一类方法试图通过去中心化的方式，即每个智能体单独控制一个路口，同时通过适当的奖励和状态设计来实现协调控制[19,20]。Chen 在 [21] 中通过将智能体之间的参数进行共享实现了在仿真环境中城市级别的交通信号控制。

目前，虽然已经有不少基于强化学习的智能交通信号调度研究，但是任然有一些问题没有得到解决，有待于进一步地研究。

1.3 论文主要工作

近年来，基于深度强化学习的智能交通信号调度已经逐渐称为研究热点。然而，目前的工作对于单路口场景下的公平性调度以及多路口场景下的协调控制的研究还不够充分。本文分析了在单路口和多路口这两个不同信号调度场景下已有工作的不足，并在已有工作的基础上，分别对这两个场景下存在的问题进行了改进。本文的主要工作如下：

1. 在单路口交通信号调度场景下，为了提高路口的整体通行效率，强化学习模型会偏向于优先放行交通流量大的车道，而忽略交通流量小的车道，从而导致部分车辆出现“饥饿等待”的情况。本文引入无线通信网络中 Proportional Fair Scheduling(PFS) 调度策略，并以此设计了一个能够在效率和公平性之间取得良好平衡的强化学习模型。通过在不同的交通负载情况下进行的实验结果表明，与传统的交通信号控制方法相比，本文提出的模型能够更大程度地提高通行效率，缓解交通路口的压力；与已有的基于深度强化学习的交通信号控制方法相比，我们的模型在保证系统通行效率的同时，还能够确保相对的公平性，有效地缓解的车辆“饥饿等待”的情况。
2. 在多路口交通信号调度场景下，为了实现多路口的协调控制，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，笼统地传递所有观测信息会导致目标路口难以从这些冗余的信息中挖掘出对自身路口有效的信息，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。本文在已有工作的基础上提出了一种新的针对道路网的建图方式，在这种方式下，目标路口在聚合邻近路口交通信息时可以自动剔除与自身无关的交通流量信息，从而快速的实现信息聚合。通过在合成交通数据和真实交通数据上进行实验，结果表明这种新的建图方式可以有效提高系统的通行效率以及学习模型的收敛速度。

1.4 论文组织结构

论文总共分为五章，其组织结构如图 1.2所示，各章节的安排如下：

第一章：绪论。结合交通拥堵情况恶化的严峻趋势，提出了高效的交通信号调度方法的重要性，并进一步给出了使用强化学习解决交通信号调度问题的重要意义。再次基础上，总结了基于强化学习的智能交通信号调度的研究现状，介绍了本文的主要研究工作以及具体的研究内容。

第二章：相关概念和技术。介绍了传统的交通信号控制方法以及基于深度强化学习的智能

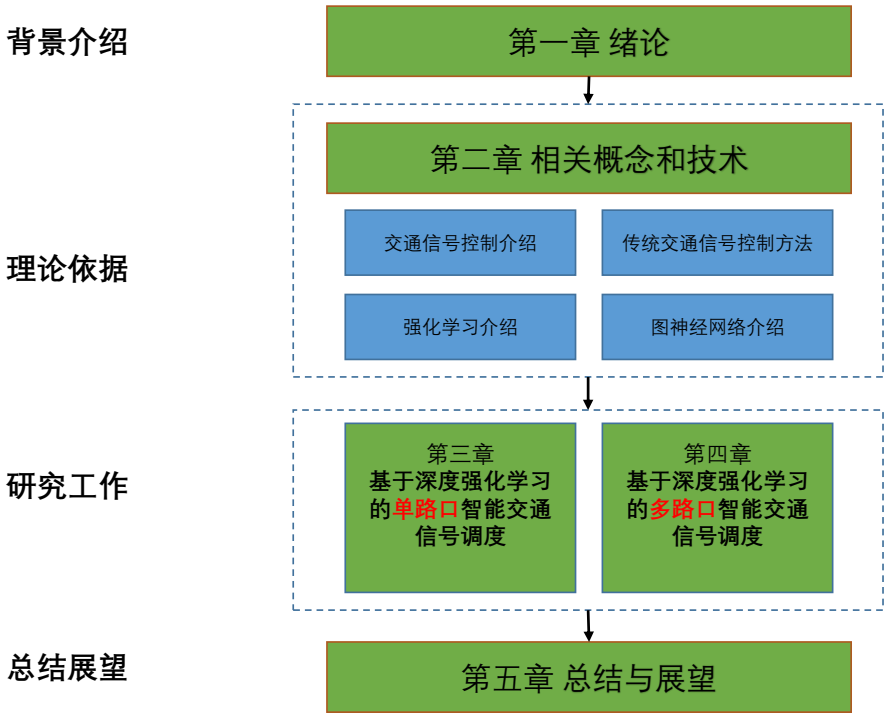


图 1.2　论文组织结构

交通信号调度的技术背景，包括深度强化学习以及图神经网络。

第三章：基于深度强化学习的单路口智能交通信号调度。针对为了提高系统通行效率而忽视小交通流量导致的车辆“饥饿等待”问题，引用无线网络中的 Proportional Fair 调度方法，提出了一种新的在提高通行效率的同时保证整体相对公平的基于强化学习的智能交通信号调度方法，最后给出了仿真实验结果并与已有的方法进行了对比分析。

第四章：基于深度强化学习的多路口智能交通信号调度。针对多路口协调控制中出现的信息冗余问题，提出了一种基于新的路网建图方式的多路口交通信号调度模型来自动剔除对目标路口无关的交通流量，从而降低提取有效信息的难度，最后在仿真环境中分别就合成数据和真实数据进行了实验，并对比分析了提出的模型的性能。

第五章：总结与展望。总结了本文在基于深度强化学习的智能交通信号调度方面所作的工作，并提出了下一步智能交通信号调度的研究方向。

1.5　本章小结

本章交通信号控制的背景入手，分析的智能交通信号调度的重要性，给出了研究基于深度强化学习的智能交通信号调度的研究意义。并在此基础上，总结了基于深度强化学习的智能交通信号调度的研究现状，给出了本文的主要研究工作。最后，介绍了论文的结构安排。

第二章　相关概念和技术

2.1　交通信号控制概述

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：

- 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。
- 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。
- 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。

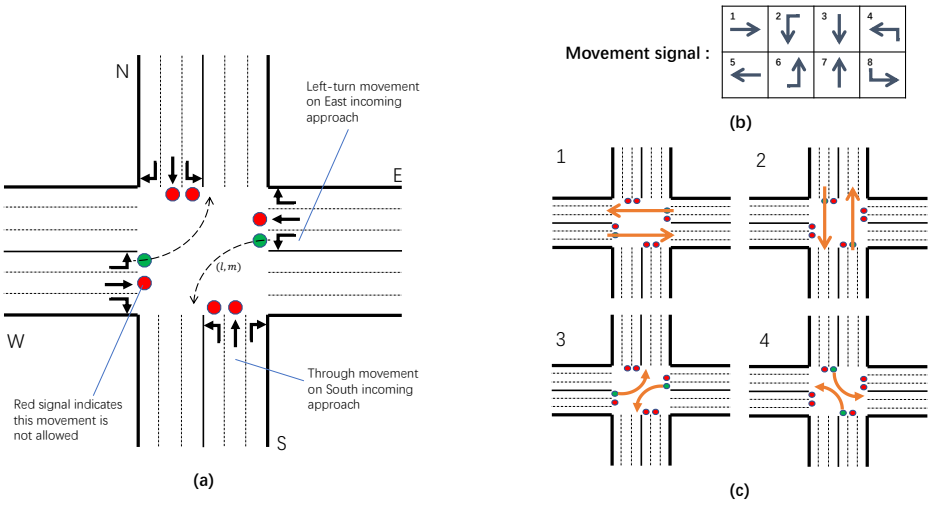


图 2.1　交通信号术语图示

2.1.1 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, 即转入巷道（incoming approach）和转出巷道（outgoing approach）。图 2.1(a) 描述了一个典型的有 8 个 approach（四个转入巷道，四个转出巷道）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)。
- 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, 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，即一组信号相位的序列，它定义了一组信号相位及其变化顺序。
- 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 $(p_1, t_1) \ldots (p_i, t_i) \ldots$，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。

2.2 传统交通信号控制方法

2.2.1 Fixed-Time

Fixed-Time[1] 是使用最为广泛的一种交通信号控制方法，它按照事先定义的信号序列不断的循环，不依赖于任何类型的检测器，例如行人按钮或车辆检测装置，按照特定的顺位为所有道路和运动提供服务。即使没有车辆或行人，信号也会改变，这种方法在交通较为稳定的情况下能够起到不错的效果，但是当交通变化很大时效率会变低。由于不需要安装任何检测器，所以这种方法的成本效益高。

2.2.2 Webster

Webster[22] 方法是一种广泛使用的针对单路口场景的交通信号控制方法。对于单路口场景，传统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 的做法就是根据交通流量计算信号周期长度和相位分裂时间。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长度和最佳相位分裂时间，从而最小化车辆通行时间。周期长度计算依赖于以下方程：

$$C_{des}\left(V_c\right)=\frac{N\times t_L}{1-\frac{V_c}{\frac{3600}{h}\times PHF\times(v/c)}} \tag{2.1}$$

其中 N 是信号相位的数量（既有多少种不同的信号相位），t_L 是每个相位下的总共延误时间，可以看作是一个与全红信号相位（全是红灯，所有的交通都不允许通行）时间以及车辆的加减速相关的参数。参数 h 则是饱和间隔时间（单位：秒/车），指前后车辆通过道路上某一点的最小时间间隔。PHF 是一个交通高峰时段因子，是衡量高峰时段交通需求波动的参数。v/c 是期望的流量与容量比，用来描述路口的繁忙程度。这些参数在不同的交通条件下会有不同的值，通常是根据经验观察和机构标准来选择。在确定周期长度后就可以算出绿色信号比（即在一个周期时间内，绿色信号所占的时间比重），具体计算方式如下：

$$\frac{t_i}{t_j}=\frac{V_c^i}{V_c^j} \tag{2.2}$$

t_i和t_j 分别表示相位 i 和相位 j 的时长。V_c^i 表示相位 i 的临界车道流量，其中临界车道是指在一个相位中交通流量与饱和流量之比最高的车道，通常与队列长度有关。

2.2.3 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[23] 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。其中路口之间的偏移量通过以下公式计算：

$$\Delta t_{i,j}=\frac{L_{i,j}}{v} \tag{2.3}$$

其中 $L_{i,j}$ 是路口 i 和路口 j 之间的道路长度，v 是道路上车辆的预期行驶速度。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，如图图 2.2所示：

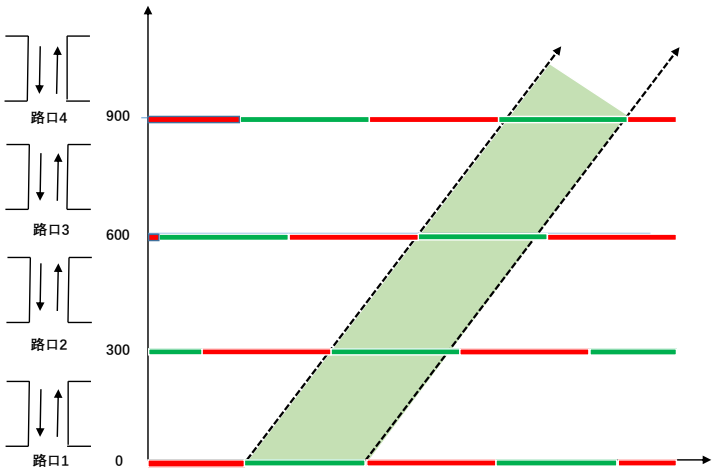


图 2.2　GreenWave 方法的绿色信号波示意图

2.2.4　MaxBand

MaxBand[24] 是一种以优化邻居路口的通行效率为目标的传统交通信号调度算法，与 GreenWave 不同的是，它旨在根据整个路网系统内各个路口的信号计划找到一个最大的带宽（maximal bandwidth ）来减少沿两个相反方向行驶的车辆的停靠次数。其中带宽是指同步绿波在一个周期长度内所持续的部分时间，通常较大的带宽意味着沿一个方向的更多交通可以不间断地通过。与 GreenWave 相同的是，MaxBand 同样要求所有路口具有相同的信号周期长度（通常设置为所有信号周期长度最大值）。在遵循以下约束条件的情况下，MaxBand 会通过使用一个混合整数线性规模模型来计算出一个对称且宽度统一的带宽。

- 单个路口的带宽限制：首先对于每个方向（包括入站方向和出站方向，入站方向指的是进入交叉路口的方向，出战方向则是离开交叉路口的方向）来说，绿色信号的时间应该大于带宽的长度：

$$w_i + b \leq 1 - r_i, w_i > 0 \tag{2.4}$$

$$\hat{w_i} + \hat{b} \leq 1 - \hat{r_i}, w_i > 0 \tag{2.5}$$

这里 w_i和$\hat{w-i}$ 分别表示入站和出站方向上红色信号结束可带宽开始之间的时间间隔，b

值得是带宽，r_i和$\hat{r_i}$ 分别表示入站和出站方向的红色信号时间。其次对于两个不同的方向，入站和出战的带宽要相同：

$$b = \hat{b} \tag{2.6}$$

- 路口 i 和路口 j 之间时间和空间的偏移量约束：首先时间上要满足下列约束条件:

$$\theta(i,j) + \hat{\theta}(i,j) + \delta_i - \delta_j = m(i,j) \tag{2.7}$$

这里 $\theta(i,j)$**和**$\hat{\theta}(i,j)$ 分别表示路口入站和出站的偏移量，δ 为入站和出站的红色信号时间偏移量，$m(i,j)$ 是一个整数变量，表示路口 i 和路口 j 的信号周期长度的一个倍数。其次空间上，车辆从一个路口出发到另一个路口的行驶时间要满足以下约束条件：

$$\phi(i,j) + 0.5 * r_j + w_j + \tau_j = 0.5 * r_i + w_i + t(i,j) \tag{2.8}$$

$$\hat{\phi}(i,j) + 0.5 * \hat{r}_j + \hat{w}_j = 0.5 * \hat{r}_i + \hat{w}_i - \hat{\tau}_i + \hat{t}(i,j) \tag{2.9}$$

其中 t 和 $\hat{t}$ 分别表示交叉路口之间的进站和出战的通行时间，与道路的长度和车辆行驶速度有关；τ和$\hat{\tau}$ 表示交叉路口的队列清理时间，与从其他路口转入以及自身的交通流量有关。

最后 MaxBnad 的优化目标定义如下所示：

$$\begin{aligned} &\textbf{maximize } b \\ &\textbf{subject to } (2.4\text{-}2.9) \end{aligned} \tag{2.10}$$

2.2.5 Actuated Control

Actuated Control 根据当前信号相位和其他的竞争信号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，根据规则的差异，Actuated Control 主要可以分为 Fully-Actuated Control 和 Semi-Actuated Control 两种。具体的请求触发规则定义如表 2.1所示：

2.2.6 SOTL

SOTL[25]（Self-Organizing Traffic Light Control）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信号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预先定义的一个阈值，否则就不会产生绿色信号请求。

表 2.1　Actuated Control 信号相位请求触发规则

来自当前信号相位的请求	来自其他信号相位的请求	动作
是	是	**Fully-actuated control**：如果当前信号相位的持续时间大于阈值，则切换到下一信号相位；否则，保持当前信号相位。 **Semi-actuated control**：保持当前信号相位。
	否	保持当前信号相位。
否	是	切换到下一信号相位。
	否	**Fully-actuated**：保持当前信号相位。 **Semi-actuated control**：切换到默认信号相位（通常设置为主干道的绿色信号）。

2.2.7　Max-Pressure Control

Max-Pressure Control[26] 的目的是通过最小化对应信号相位的压力（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，其中压力的概念如图图 2.3所示：其中运动信号的压力是指转入车道上的车辆数减去相应的转出车道上的车辆数，而信号相位的压力定义为转入巷道和转出巷道上的总队列长度之间的差异。Varaiya[26] 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。Max-Pressure Control 算法流程如算法 1所示：

Algorithm 1 Max-Pressure Control 算法流程

输入： t：当前信号相位时间。

t_{min}：最小信号相位持续时间。
for all timestamp **do**
　$t = t + 1$。
　if $t \geq t_{min}$ **then**
　　计算每个信号相位 i 的压力 P_i。
　　选择压力最大的信号相位 $i = \arg\max_i P_i$。
　　重置信号时间 $t = 0$。
　end if
end for

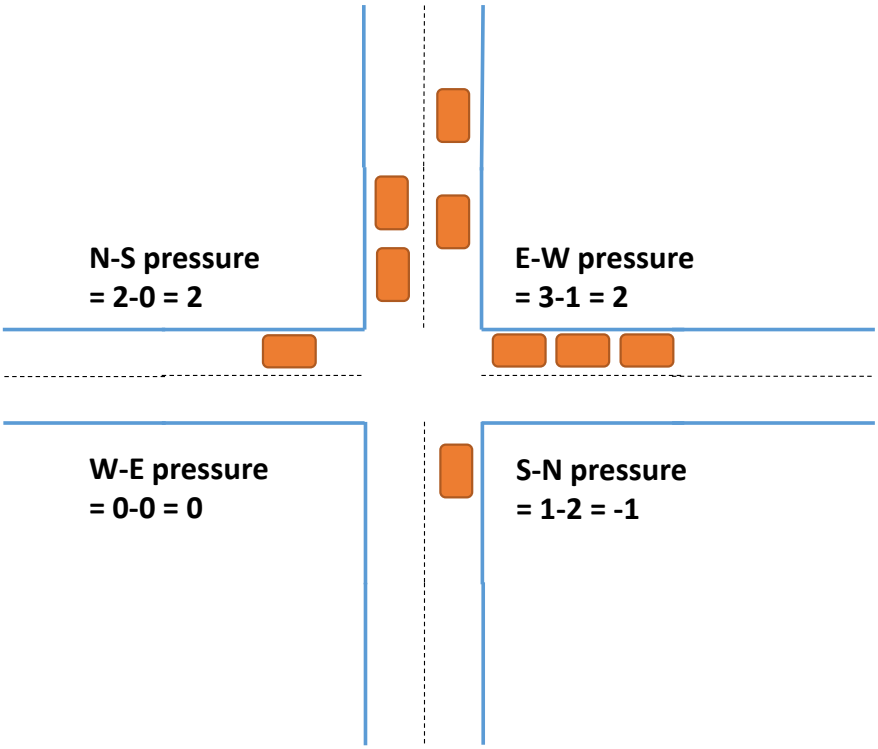


图 2.3　Max-Pressure 的压力图示

2.2.8　SCATS

SCATS[27] 通过将预先定义的信号计划（包括信号周期长度、信号相位分裂比以及偏移量）作为输入，并根据事先定义的性能指标来从这些反复选择，其性能指标饱和度的定义如下：

$$DS = \frac{g_E}{g} = \frac{g - (h' - h \times n)}{g} \tag{2.11}$$

其中 g 是可用的绿色信号时间，g_E 是有车辆通过路口的有效绿色信号时间，等于可用绿色信号时间减去浪费的绿色信号时间，而浪费的绿色信号时间是通过检测来计算的，h' 是检测到的总时间差，n 是检测到的车辆数量，h 是车辆之间的单位饱和间隔，它表示连续车辆通过一个点的最小时间间隔。信号相位分裂比、信号周期长度以及偏移量是通过近似机制从实现定义的信号计划中选择的。以信号相位分裂比的选择为例，SCATS 首先计算当前信号计划下的饱和度，然后使用以下公式估计出在当前信号周期内为被应用的其他信号计划的饱和度。

$$\hat{DS}_p^j = \frac{DS_p \times g_p}{g_p^j} \tag{2.12}$$

其中 DS_p和g_p 分别是信号相位 p 在当前信号计划下的饱和度以及绿色信号时间，g_p^j 是信号相位 p 在信号计划 j 下的饱和度，具体的算法流程如算法 2所示。

Algorithm 2 SCATS 中信号相位分裂比的选择算法流程

输入： $A = a_j | j = 1, 2, \cdots, N$：候选的信号计划集合

1: a_c ： 当前的信号计划。
2: **for** textbfall $pin 1, \cdots, P$ **do**
3: 计算每个信号相位 $p \in 1, \cdots, P$ 在当前信号计划 a_c 下的饱和度 DS_p。
4: **end for**
5: **for** textbfall a_j in A **do**
6: 计算每个信号相位 p 在信号计划 a_j 下的饱和度 $\hat{DS}_p^j$。
7: **end for**

表 2.2列出了上述几种传统交通信号控制方法的要求和输出结果：

表 2.2 传统交通信号控制方法总结

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 MaxBand	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
SCATS	所有路口信号计划	交通流量	调整后的信号计划

2.3 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（Reinforcement Learning，RL）。其中在这些技术中，强化学习在近年来更具趋势。

2.3.1 强化学习概述

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, 如图 2.4所示）$< \mathcal{S}, \mathcal{A}, P, R, \gamma >$，其中 $\mathcal{S}, \mathcal{A}, P, R, \gamma$ 分别表示状态集、动作集、状态转移函数、奖励函数和折扣因子。具体定义如下：

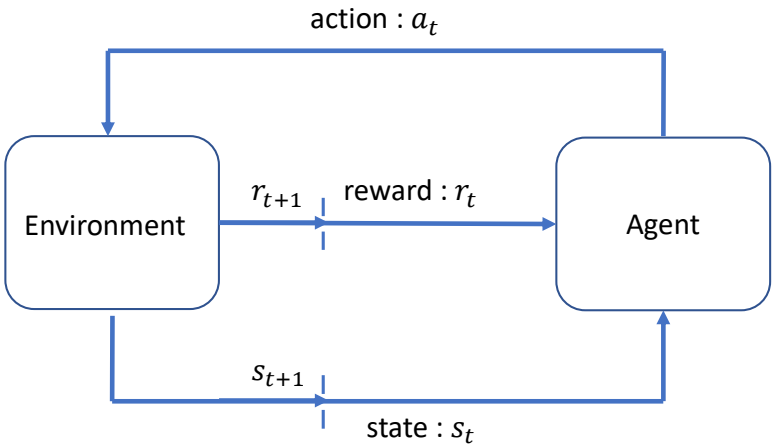


图 2.4　马尔可夫决策过程示意图

- $\mathcal{S}$：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。

$$P\left(\mathrm{s}^{t+1} \mid \mathrm{s}^t, \mathrm{a}^t\right) : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{S} \tag{2.13}$$

- R：在时间步骤 t，智能体通过奖励函数获得一个奖励 r^t。

$$R\left(s^t, a^t\right) : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R} \tag{2.14}$$

- γ：智能体的目标是找到一种使预期收益最大化的策略，即累积（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。

$$G^t := \sum_{i=0}^{\infty} \gamma^i r^{t+i} \tag{2.15}$$

解决一个强化学习任务意味着要找到一个能够使预期收益最大化的最优策略 π^*，一般来说，我们难以直接找到这个最优策略，更多的是比较若干个不同的策略然后从中选出较好的那个作为局部最优解。而策略的筛选是通过比较其对应的价值函数来实现的，即通过寻找较优的价值函数来筛选出较优的策略。价值函数是对未来奖励的期望，根据输入的不同可以分为状态价值函数和动作价值函数。状态价值函数的定义如下：

$$V_\pi(s) = \mathbb{E}_\pi\left[G_t \mid S_t = s\right] \tag{2.16}$$

其描述的是当在 t 时刻处于状态 s 的预期收益。动作价值函数的定义如下：

$$Q_\pi(s, a) = \mathbb{E}_\pi\left[G_t \mid S_t = s, A_t = a\right] \tag{2.17}$$

其描述的是当在状态 s 下采取动作 a 的预期收益。

最优策略 π^* 对应的是最优状态价值函数和最优动作价值函数。最优状态价值函数定义为 $V_\pi^*(s) = \Updownarrow\neg\S_\pi V^\pi(s)$，它满足以下贝尔曼最优方程：

$$V^*\left(s^t\right) = \max_{a^t \in \mathcal{A}} \sum_{s^{t+1} \in \mathcal{S}} P\left(s^{t+1} \mid s^t, a^t\right)\left[r + \gamma v_*\left(s^{t+1}\right)\right], \forall s^t \in \mathcal{S} \tag{2.18}$$

最优动作价值函数定义为 $Q^*(s,a) = \Updownarrow\neg\S_\pi Q^\pi(s,a)$，其满足以下贝尔曼最优方程：

$$Q^*\left(s^t, a^t\right) = \sum_{s^{t+1} \in \mathcal{S}} P\left(s^{t+1} \mid s^t, a^t\right)\left[r^t + \gamma \max_{a^{t+1}} Q^*\left(s^{t+1}, a^{t+1}\right)\right], \forall s^t \in \mathcal{S}, a^t \in \mathcal{A} \tag{2.19}$$

2.3.1.1 分类

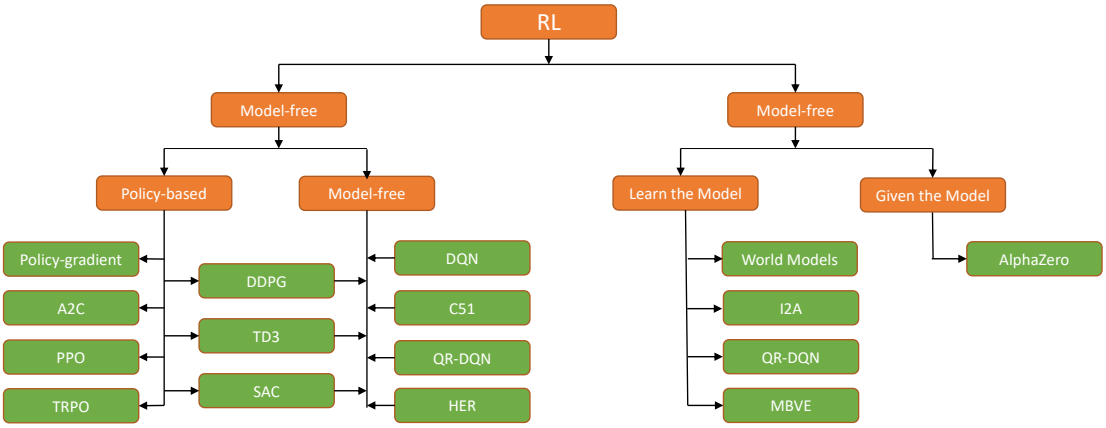


图 2.5　强化学习分类及常见算法

强化学习主要可以分为两大类：Model-based（有模型）和 Model-free（无模型）。其中 Model-based 又可以分为基于策略函数（Policy-based）和值函数（Value-based）两大类，而 Model-free 可以分为模型学习（Learn the Model）和给定模型（Given the Model）两大类，具体如图 2.5所示。

Model-free 方法是指在不知道状态转移函数函数的情况下，通过采样大量的经验来学习策略函数或者价值函数。其中 Policy-based 的方法直接将拟学习的策略参数化，在以最大化奖励的目标下，直接对策略函数进行优化。而 Value-based 的方法则是学习一个最佳策略对应的动作价值函数的近似。当然，Policy-based 和 Value-based 并非无法兼容的，有一类方法（Actor-Critic）就是融合了两中方法的思想。这类方法同时使用策略和价值评估来做出决策，其中，智能体会根据策略做出动作，而价值函数会对做出的动作给出评分，这样可以在原有的基于策略的方法的基础上加速学习过程，取得更好的效果。代表算法如图 2.5中的 DDPG[28]、TD3[29] 以及 SAC[30] 等。

Model-based 方法是指在已知状态转移函数的情况下，学习用一个模型去模拟环境，然后用这个模拟的环境去预测接下来可能会发生的所有所有情况并从中选择最有利的情况。

2.3.2 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强化学习智能体的有效奖励。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快速学习到最优策略，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 智能体需要很长时间。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，因为智能体可能需要花费更多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。

2.4 图神经网络

近些年来，图神经网络 (Graph Neural Networks, GNN) 在交通领略得到了广泛的应用，包括交通流量预测[31,32]、出行需求预测[33,34] 以及轨迹预测[35–37] 等。在交通信号控制方面，Wei[20] 和 Van der Pol[12] 也提出使用图神经网络来学习路口之间的信息交互，从而实现协调控制。

2.4.1 图神经网络概述

神经网络的成功推动了模式识别和数据挖掘领域的研究，许多机器学习任务（例如如对象检测、数据翻译和语言识别等这些曾经严重依赖于手工特征工程的任务），已经被各种端到端的深度学习范式（例如，卷积神经网络、循环神经网路和对抗生成网络等）所取缔。深度学习在过去几年得到巨大发展的原因一方面归功于诸如 GPU 等支持并行计算的硬件资源的发展，另一方面得益于其方法本身能够有效地从欧式数据（如图片、文本和视频数据等）中提取出潜在特征。但是在某些领域，数据更多的是以非欧式结构（如图结构）的形式出现的。例如，在电子商务中，为了提高推荐系统的准确率通常将用户和产品之间的交互建模成图；在化学领域，研究药物分子的生物活性时会将分子建模成图以便于新的药物发现；这些任务由于特殊的数据结构导致难以直接使用原有的机器学习方法进行解决。与欧几里得数据相比，图结构数据的任意性更强，图中的节点之间并没有固定的先后顺序，并且不同的节点的结构信息也不尽相同（例如不同的邻居节点数目），这些因素直接导致一些原本在欧几里得域下容易计算的操作（例如图像上的二维卷积）很难直接应用到图领域。此外，基于欧式数据结构的机器学习方法的核心假设是实例之间是相互独立的，这一点在图数据上是不成立的，因为每个实力（节点）都通过不同类型的链接与其他实例相关，如引用关系、互动关系或者友谊关系等，下列给出关于图的定义：

定义 2.1（图）：

图表示为 $G = (V, E)$，其中 V 是顶点或者节点的集合，E 是边的集合。$v_i\ inV$ 表示点集合中的一个节点 i，$e_{ij} = (vi, vj) \in E$ 表示边集中的一条从 v_j 指向 v_i 的边。节点 v 的领域定义为 $N(v) = u\ inV|((v, u) \in E)$。图可能有节点属性 X，其中 $\mathbf{X} \in \mathbf{R}^{n\times d}$ 是图的节点特征矩阵，$x_v\ \mathbf{R}^d$ 表示节点 v 的特征向量。同理，图可能有边属性 X^e，其中 $\mathbf{X}^e \in \mathbf{R}^{m\times c}$ 是图的边特征矩阵，$\mathbf{x}^e_{v,u} \in \mathbf{R}^c$

表示边 (v,u) 的特征向量，通常如果图的节点或者边具有属性，那么这个图被称为属性图。

定义 2.2（有向图）：有向图是一个所有边都从一个节点指向另一个节点的图。无向图可以看作是有向图的一个特殊情况，如果两个节点相连，就会有一对方向相反的边。当且仅当邻接矩阵是对称的时候，一个图是无定向的。其中邻接矩阵 A 的定义如下：

$$A_{i,j} = \begin{cases} 1 & e_{ij} \in E \\ 0 & e_{ij} \notin E \end{cases} \tag{2.20}$$

定义 2.3（时空图）：时空图是一种属性图，其中节点属性随时间动态变化。时空图的数学表示为：$G^{(t)} = \left(\mathbf{V}, \mathbf{E}, \mathbf{X}^{(t)}\right)$ 其中$\mathbf{X}^{(t)} \in \mathbf{R}^{n \times d}$。

Sperduti[38] 等人在 1997 年首次将神经网络应用于有向无环图，这促进了人们对图神经网络的早期研究。在 2005 年，Gori[39] 首次提出了一种针对图领域的神经网络模型来捕捉图数据中的拓扑信息，并将其命名为图神经网络。这份工作在 2008 年被 Scarselli[40] 等人延续，并对其中提出的神经网络模型进行了更加细致的设计。早期在研究图神经网路的时候，为了使网络能够收敛，通常要借助于循环神经网络来根据图的拓扑结构来聚合节点的邻居信息，并在迭代过程中对网络进行更新直到网络趋于稳定。

图神经网络可以分为：图卷积网络（Graph Convolutional networks），图注意力网络（Graph Attention Network），图自编码器（Graph Auto-encoder），图生成网络（Graph Generative Network）和图时空网络（Graph Spatial-Temporal Network）。

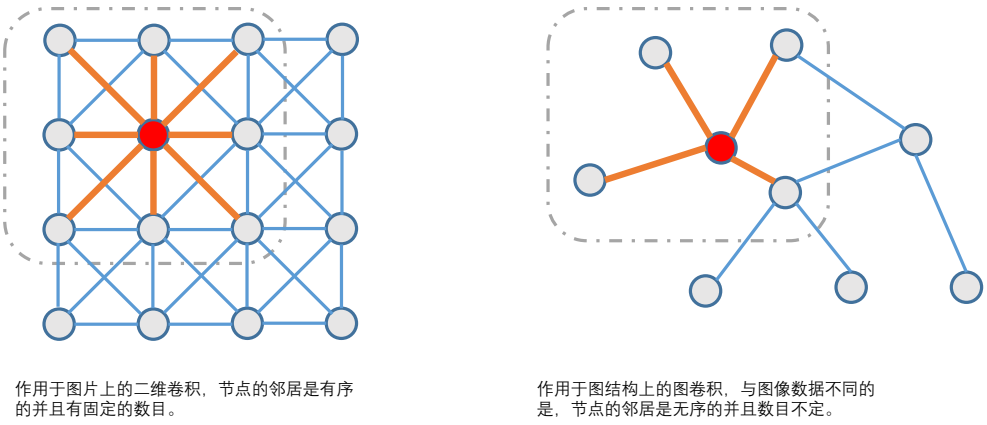
2.4.2 图卷积网络

图卷积网络 (Graph Convolutional networks, GCN) 启发于深度学习中的卷积神经网络（Convolutional Neural Networks，CNN），如图 2.6所示，它将原本应用于诸如图片和视频等传统数据上的卷积操作拓展到了更加复杂的图数据上。这种方法秉承的思想是根据图的拓扑结构来聚合一个节点及其邻居节点的特征信息并生成一个新的特征表示。由于其在图数据潜在特征提取上的有效性，图卷积网络常常被用作其他图神经网络的基础构建模块。根据卷积操作定义的不同，图卷积网络主要可以分类两类：基于谱（spectral-based）和基于空间（spatial-based）的方法。基于谱的图卷积网络从信号处理的角度引入滤波器来定义图卷积操作，在这种方式下，图卷积操作可以看作是一种降噪操作。

基于空间的图卷积操作则是受启发于深度学习中的卷积神经网络对图像的卷积操作，与之不同的是基于空间的图卷积网络是使用图中节点之间的位置关系来进行图卷积操作的。

2.4.3 图注意力网络

图注意力网络是将注意力机制引入到基于空间域的图神经网络。图神经网络不需要使用拉普拉斯等矩阵进行复杂的计算，仅通过邻居节点的表征来更新目标节点的特征。由于能够放大



作用于图片上的二维卷积，节点的邻居是有序的并且有固定的数目。

作用于图结构上的图卷积，与图像数据不同的是，节点的邻居是无序的并且数目不定。

图 2.6　图卷积示意图

数据中最重要部分的影响，注意力机制已经广泛应用到很多基于序列的任务中，图神经网络也受益于此，在聚合过程中使用注意力整合多个模型的输出。主要方法包括：

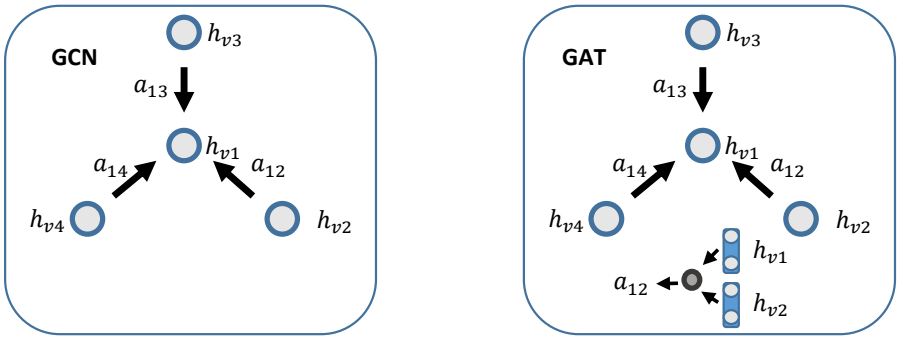


图 2.7　GCN 与 GAT 聚合信息的区别示意图

- Graph Attention Network(GAT)[41]：本质上 GAT 是一种基于空间的图卷积网络，它与 GCN 的主要区别在于对邻居节点信息的聚合方式不同（图 2.7）。GCN 在在聚合过程中显式地为节点 v_i 的邻居 v_j 赋予一个非参数静态权重 $a_{ij} = \frac{1}{\sqrt{\deg(v_i)\deg(v_j)}}$。而 GAT 则是通过使用一个端到端的神经网络架构隐式地捕捉权重 a_{ij}，以便更重要的节点获得更大的权重，具体操作如下：

$$\mathbf{h}_i^t = \sigma\left(\sum_{j\in\mathcal{N}_i} \alpha\left(\mathbf{h}_i^{t-1}, \mathbf{h}_j^{t-1}\right)\mathbf{W}^{t-1}\mathbf{h}_j^{t-1}\right) \tag{2.21}$$

其中 $\alpha(\cdot)$ 是一个注意力函数，它可以动态地调整邻居节点 j 对目标节点 i 的贡献。通常

为了学习不同子空间中的注意力权重，GAT 会使用多个注意力函数（即多头注意立机制，Multi-head Attention）：

$$\mathbf{h}_i^t = \|_{k=1}^K \sigma\left(\sum_{j\in\mathcal{N}_t} \alpha_k\left(\mathbf{h}_i^{t-1}, \mathbf{h}_j^{t-1}\right) W_k^{t-1}\mathbf{h}_j^{t-1}\right) \tag{2.22}$$

- Gated Attention Network(GAAN)[42][43]：GAAN 除了采用多头注意力机制（self-attention mechanism）外，还引入了自注意机制来更新节点的隐藏状态。自注意机制可以为每个注意力头计算出一个额外的注意分数：

$$\mathbf{h}_i^t = \phi_o\left(\mathbf{x}_i \oplus \|_{k=1}^K g_i^k \sum_{j\in\mathcal{N}_t} \alpha_k\left(\mathbf{h}_i^{t-1}, \mathbf{h}_j^{t-1}\right)\phi_v\left(\mathbf{h}_j^{t-1}\right)\right) \tag{2.23}$$

其中 ϕ_o和ϕ_v 是反馈神经网络，而 g_i^k 是第 k 个注意力头的权重。GAM 是一种基于循环神经网络模型的图神经网络，它主要应用于图数据的分类任务，即输入一个图结构数据，输出这个图所属的类别。这种方法会以自适应的方式逐个处理图中重要节点，并根据以下方式计算这些节点潜在特征：

$$\mathbf{h}_t = \mathbf{f}_h\left(\mathbf{f}_s\left(\mathbf{r}_{t-1}, \mathbf{v}_{t-1}, g; \theta_s\right), \mathbf{h}_{t-1}; \theta_h\right) \tag{2.24}$$

其中 $\mathbf{f}_h(\cdot)$ 是循环神经网络模型，通常使用 LSTM，f_s 是一个步进网络（Step Network），他会优先聚合当前节点 v_{t-1} 的邻居节点中优先级高的节点信息。

2.4.4 图自编码器

图自编码器是一类图嵌入方法，其目的是利用神经网络将图的顶点表示为低维向量。典型的解决方案是利用多层感知机作为编码器来获取节点嵌入，其中解码器重建节点的邻域统计信息，如 PPMI（Positive Pointwise Mutual Information）或一阶和二阶近似值。主要包括基于 GCN 的自编码器，如 Graph Autoencoder（GAE）[44] 和 Adversarially Regularized Graph Autoencoder (ARGA)[45], 以及 Network Representations with Adversarially Regularized Autoencoders (NetRA)[46]、Deep Neural Networks for Graph Representations (DNGR)[47]、Structural Deep Network Embedding (SDNE)[48] 和 Deep Recursive Network Embedding (DRNE)[49]。DNGR 和 SDNE 学习仅给出拓扑结构的节点嵌入，而 GAE、ARGA、NetRA、DRNE 用于学习当拓扑信息和节点内容特征都存在时的节点嵌入。图自动编码器的一个挑战是邻接矩阵 A 的稀疏性，这使得解码器的正条目数远远小于负条目数。为了解决这个问题，DNGR 重构了一个更密集的矩阵，即 PPMI 矩阵，SDNE 对邻接矩阵的零项进行惩罚，GAE 对邻接矩阵中的项进行重加权，NetRA 将图线性化为序列。

2.4.5 图生成网络

图生成网络的目标是在给定一组观察到的图的情况下生成新的图。图生成网络的许多方法都是特定于领域的。例如，在分子图生成中，一些工作模拟了称为 SMILES 的分子图的字符串

表示。在自然语言处理中，生成语义图或知识图通常以给定的句子为条件。这种方法通常使用GCN 或者其他框架作为基础构建模块，其中使用 GCN 构建的方法有：

- Molecular Generative Adversarial Networks (MolGAN)[50]：MolGAN 整合了图卷积网络、图注意力网络和强化学习以生成具有预期属性的图。MolGAN 由生成器和判别器组成，相互竞争以提高生成器的真实性。在 MolGAN 中，生成器尝试提出一个假图及其特征矩阵，而鉴别器旨在将假样本与经验数据区分开来。此外，还引入了一个奖励网络来促使生成器能够按照外部的评估生成具有特定属性的图。
- Deep Generative Models of Graphs (DGMG)[51]：利用基于空间的图卷积网络来获得现有图的隐藏表示。生成节点和边的决策过程是以整个图的表示为基础的。简而言之，DGMG 递归地在一个图中产生一个节点，直到达到某个停止条件。在添加新节点后的每一步，DGMG 都会反复决定是否向添加的节点添加边，直到决策的判定结果变为假。如果决策为真，则评估将新添加节点连接到所有现有节点的概率分布，并从概率分布中抽取一个节点。将新节点及其边添加到现有图形后，DGMG 将更新图的表示。

使用其他架构作为基础模块的图生成网络有：

- GraphRNN[52]：通过两个层次的循环神经网络 (Recurrent Neural Network, RNN) 的深度图生成模型。图层次的 RNN 每次向节点序列添加一个新节点，而边层次 RNN 生成一个二进制序列，指示新添加的节点与序列中以前生成的节点之间的连接。为了将一个图线性化为一系列节点来训练图层次的 RNN，GraphRNN 采用了广度优先搜索（BFS）策略。为了建立训练边层次的 RNN 的二元序列模型，GraphRNN 假定序列服从多元伯努利分布或条件伯努利分布。
- NetGAN[53]：Netgan 将 LSTM 与 Wasserstein-GAN 结合在一起，使用基于随机行走的方法生成图形。GAN 框架由两个模块组成，一个生成器和一个鉴别器。生成器尽最大努力在 LSTM 网络中生成合理的随机行走序列，而鉴别器则试图区分伪造的随机行走序列和真实的随机行走序列。训练完成后，对一组随机行走中节点的共现矩阵进行正则化，我们可以得到一个新的图。

2.4.6 图时空网络

在现实生活中，很多应用场景下的图数据都是动态的，包括拓扑结构以及特征属性。而图时空网络就是一种能够有效的处理动态图数据的模型。图时空网络的方法分为两类，一种是基于 RNN 的方法，另一种是基于 CNN 的方法。

大多数基于 RNN 的方法通过过滤输入和使用图卷积传递给循环单元的隐藏状态来捕获时空依赖性。但是基于 RNN 的方法存在耗时的迭代传播和梯度爆炸或者消失的问题。作为替代解决方案，基于 CNN 的方法以非递归的方式处理空间-时间图，具有并行计算、稳定梯度和低

内存需求的优势。基于 CNN 的方法将一维卷积层和图卷积层交织在一起，分别学习时间和空间的依赖关系。

2.4.7 任务分类

以图结构和节点特征信息作为输入，根据输出的类别，可以将 GNN 的分析任务分为以下几类：

- 节点级别：节点级的输出与节点回归（Node Regression）和节点分类（Node Classification）任务相关。如图卷积网络可以通过信息传播和图卷积操作提取出节点的潜在表示。使用多感知器或 softmax 层作为输出层，GNN 能够以端到端的方式执行节点级任务。
- 边级别：边级别的输出与边分类（edge classification）和链接预测（link prediction）任务相关。以 GNN 的两个节点的潜在表示作为输入，可以利用相似性函数或神经网络来预测一个边的标签或者连接强度。
- 图级别：图级别的输出与图分类任务相关。通常 GNN 会与池化（pooling）和读出（read-out）操作相结合，以获得在图级别上的紧凑表示。

2.5 本章小结

本章总结了交通信号控制中的一些一本定义以及一些传统的交通信号控制方法，然后介绍了基于深度强化学习的智能交通信号调度的技术背景，包括强化学习、不同场景下的控制框架以及图神经网络。

第三章　基于深度强化学习的单路口智能交通信号调度

3.1　引言

随着城市道路上汽车数量的日益增长，交通拥堵情况变得越来越严重，极大地影响了人们的日常生活以及城市交通的运作。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，而这些“行驶资源”就是不同道路汇集的交叉路口。交通信号调度问题研究的就是如何高效以及安全的调度交叉路口的不同交通流量。早期，由于技术的不成熟以及成本的约束，城市交通管理通常使用基于预先定义规则的调度方法来控制交叉路口的信号灯，这种做法的好处是可以保证调度的安全性以及不需要更多的人为干预成本，但是随着城市汽车数量的增加，这种方法已经难以高效地调度交通流量，因为它在调度时没有考虑路口的实时交通状况，只是按照预先定义的规则以固定的顺序改变信号灯的状态。

传统方法的不足促使了人们对智能交通信号调度方法的研究。与传统方法不同的是，智能交通信号调度方法能够根据实时的交通状况制定出最优的策略，从而最大程度地缓解交通拥堵。在这些方法中，基于深度强化学习的智能交通信号调度方法在近些年逐渐成为了研究热点。首先神经网络的发展以及强化学习理论的成熟为这类方法的研究提供了技术支持。其次，强化学习的无模型特性使得它能够在训练中进行自我学习，而不需要外部的人为干涉，与那些需要人为定期修改控制规则的调度方法相比，降低了模型部署之后的管理成本。

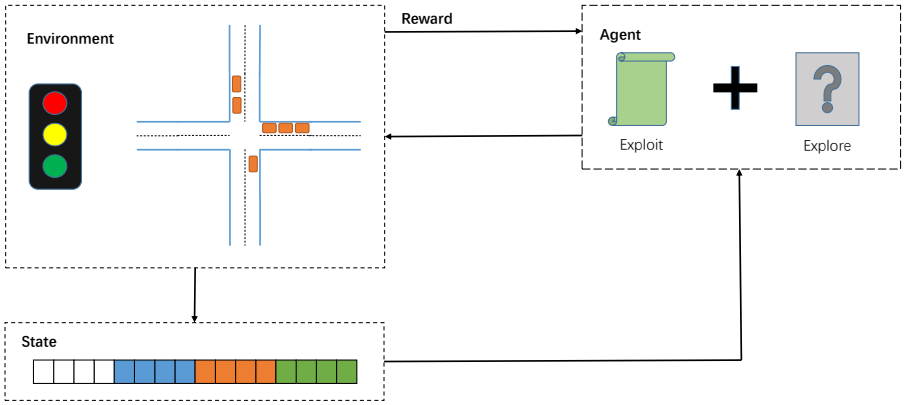


图 3.1　基于深度强化学习的单路口交通信号控制框架

3.2 相关工作

目前有很多基于强化学习的智能交通信号调度研究工作,这些工作大多数都是使用如图3.1所示的控制框架，在这个框架下，环境是道路上的交通状况，智能体要做的是控制交通信号灯，在每个调度时刻 t，获取环境的状态描述 s^t（例如，当前的信号相位，车辆的等待时间，排队长度，车辆的位置），然和根据这个状态描述对下一步采取的行动作出预测，以使预期收益最大化，然后该动作将在环境中执行，并且产生一个奖励 r^t。通常，在决策过程中，智能体采取的策略结合了对所学策略的利用和对新策略的探索。已有共组的差异主要体现在智能体的设计方面，包括以下三点：

（1）状态表示

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 [18] 中，速度小于 0.1米/s 的车辆被认为处于等待状态；在 [54,55] 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 [12,18]，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 [56,57] 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，[12,18] 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。

（2）奖励设计

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。[58] 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[18,59]。

- 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[12,60]。
- 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[12,18]。
- 停车次数：使用网络中所有车辆的平均停车次数作为奖励。直观地说，停车次数越少，交通就越顺畅[12]。
- 路口压力：Wei 在 [61] 工作中，将路口压力定义为每个交通运动的绝对压力之和。直观地说，较高的压力表明进站车道和出站车道数量之间的不平衡程度较高。

(3) 学习算法

根据估计潜在奖励和选择动作的不同，现有基于强化学习的智能交通调度方法可以分为两类，分别是 Value-based 以及 Policy-based，具体如表 3.1所示。

表 3.1　基于深度强化学习的智能交通信号调度方法分类

方法	相关工作	优点
Value-based	[12,17,18,20,58,61–64]	使用策略评估和策略控制来逼近最优策略，更易于理解。
Policy-based	[59,65–69]	直接学习控制策略，收敛速度更快。

Value-based 的方法通过近似价值值函数或动作价值函数的方式来隐式地学习控制策略。通常，为了避免模型陷入局部最优，这类方法在选择动作时会使用 $\epsilon - greedy$ 的选择机制，从而提高探索效率，随着 ϵ 的减小，模型也将趋于稳定。此外，这类方法只能处理离散工作。

Policy-based 的方法直接对控制策略进行建模并在迭代过程中对其进行更新。这类方法的优势是它不要求动作必须是离散的。此外，它自身就可以学习随机策略并不断探索可能更有价值的动作，因此不需要使用 $\epsilon - greedy$ 的动作选择机制。

3.3　研究目标

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，即为了最大化系统的通行效率，控制模型会倾向于优先调度交通流量大的车道（即将这些交通流量对应的信号变为绿色），而忽略那些交通流量小的车道，因为这样做对整个路口的通行效率增益更高，但是这中做法会导致那些在交通流量小

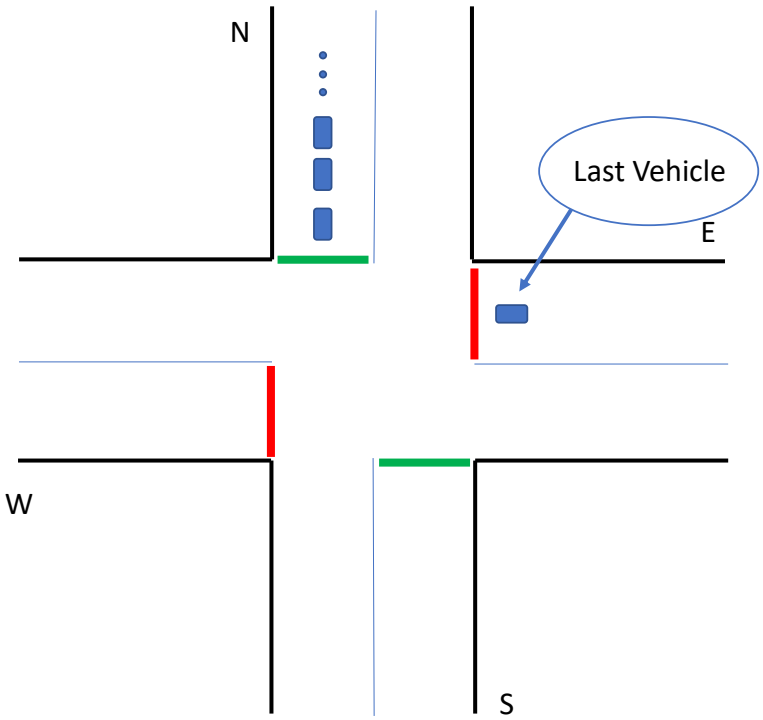


图 3.2　"last-vehicle" 情况示意图

的车道上的车辆出现“饥饿等待”的情况，甚至在一些极端情况下会导致这些车辆要等待很长的时间才能够得到通过路口的机会。更极端一点会出现如图 3.2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。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 E-W 方向上的车来说是极其不公平的。

本工作的研究目标是在提高系统通行效率的同时（最小化平均通行时间），保证调度的相对公平性，即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）。

3.4　系统设计

本工作同样使用基于深度强化学习的智能交通信号调度框架，针对我们的研究目标，我们在已有方法的智能体设计上进行了改进，并给出了我们的系统框架。

3.4.1　智能体设计

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

（2）动作选择机制

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。

（3）奖励函数设计

为了能够在效率和公平性之间得到一个良好的平衡，我们借鉴了 PFS（Proportional Fair Schedule）的调度思想。PFS 是一种应用于无线通信领域的调度算法，它能够在最大化吞吐量的同时保证每个用户的比例公平，其中比例的定义如下：

定义 3.1（比例公平）：一个调度方案 P 是比例公平的当且仅当任何其他调度方案 S 满足下列条件：

$$\sum_{i \in U} \frac{R_i^{(S)} - R_i^{(P)}}{R_i^{(P)}} \leq 0 \tag{3.1}$$

其中 U 是服务的用户集合，$R_i^{(S)}$ 表示在调度方案 S 下，用户 i 的平均速率。

上述的约束表明，一个用户在分配上的任何正向变化必须导致系统的平均负向变化。在单载波无线网络情况下 [20]，每次只允许一个用户传输，比例公平性的实现是通过每次将唯一的载波（或信道）按照以下规则分配给用户来实现的。

$$i^* = \arg\max_{i \in U} \frac{I_i}{\overline{T}_i} \tag{3.2}$$

其中 I_i 是根据实时信道条件估计的用户 i 瞬时吞吐量，T_i 是用户 i 的平均吞吐量。这一原则意味着具有更高瞬时吞吐量的用户应具有更高的优先级来访问信道资源，以提高系统的整体吞吐量，同时，其平均吞吐量越小越有可能被调度，以保证一定程度的公平性。我们将这一思想拓展到交通信号控制领域，然后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：

$$r = -\sum_{i=1}^{M} \frac{Q_i(t)}{\overline{L}_i(t) + \delta} \tag{3.3}$$

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：

$$\overline{L}_i(t) = (1 - \frac{1}{W})\overline{L}_i(t-1) + \frac{1}{W}L_i(t) \tag{3.4}$$

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3.3的分母为 0，我们加上了一个可以忽略不计的正数 δ。

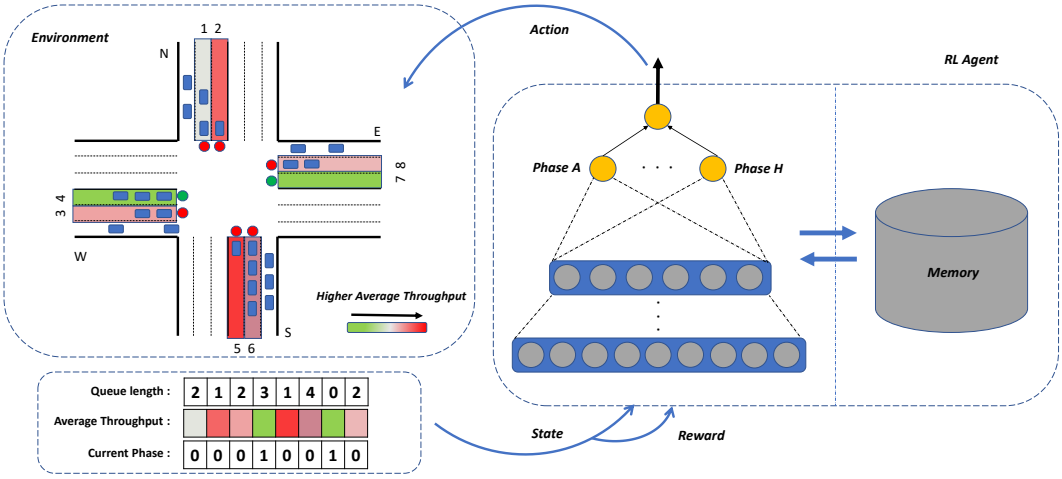


图 3.3　单路口场景下信号调度智能体与环境交互示意图

3.4.2　系统框架

如图 3.3所示，这里我们使用 DQN 作为学习算法，并且通过经验回放[70]（experience replay）机制定期提取样本来更新学习模型，具体算法过程如算法 3所示。之所以使用 DQN 作为学习算法是因为，一方面 DQN 使用了神经网络来拟合动作价值函数，可以应对连续状态空间的控制问题，并且降低了模型存储的难度。此外，DQN 作为一种离线学习方法，可以更加高效地利用历史数据；另一方面，强化学习算法更新迭代速度很快，而我们重点是研究针对 PFS 的调度思路而设计的框架对调度公平性的提升。当然，在我们的框架下，将 DQN 替换成其他更加先进的算法，例如，DDPG、TD3 等，可能会带来进一步的性能提升。

Algorithm 3 基于深度强化学习的智能交通信号控制模型训练流程

输入： E：学习片段数

T：每个学习片段的步数

b：学习经验数

ϵ：随机选择动作概率

γ：折扣因子

Δt：信号维持时间

for $episode = 1, E$ **do**
  初始化环境。
  生成车辆。
  for $t = 1, T$ **do**
    从环境中获取状态观测 s_t。
    生成一个 0 到 1 之间的随机数 $rand$。
    if $rand < \epsilon$ **then**
      从动作空间中随机采样一个动作 a_t。
    else
      使用 DQN 模型选择动作：$a_t = \arg\max_a Q(s_t, a; \theta)$。
    end if
    将当前信号更改为 a_t 并维持 δt 秒时间。
    更新每条车道的平均吞吐量。
    环境转移到新的状态 s_{t+1} 并返回一个奖励 r_{t+1}。
    将经验 $(s_t, a_t, r_{t+1}, s_{t+1})$ 存储到经验回放池 M 中。
    if $|M| > b$ **then**
      从经验回放池 M 中随机采样 b 条经验数据。
      计算损失函数 $\mathcal{L}_j$: $\mathcal{L}_Q = [r_{t+1} + \gamma \arg\max_{a'} Q(s_{t+1}, a'; \theta) - Q(s_t, a_t; \theta)]^2$。
      更新 DQN 参数。
  end if
end for
end for

3.5 实验

3.5.1 环境介绍

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

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

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。

3.5.2 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。
- 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 3.2 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

3.5.3 比较方法

- FT(Fixed-Time Control[1])：这种方法以预先设定的方式循环改变信号。
- SOTL(Self-Organizing Traffic Light Control[25])：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[58]：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。

3.5.4 性能评估

（1）通行效率评估

首先，我们通过实验评估了不同方法的通行效率的表现，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 3.4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。

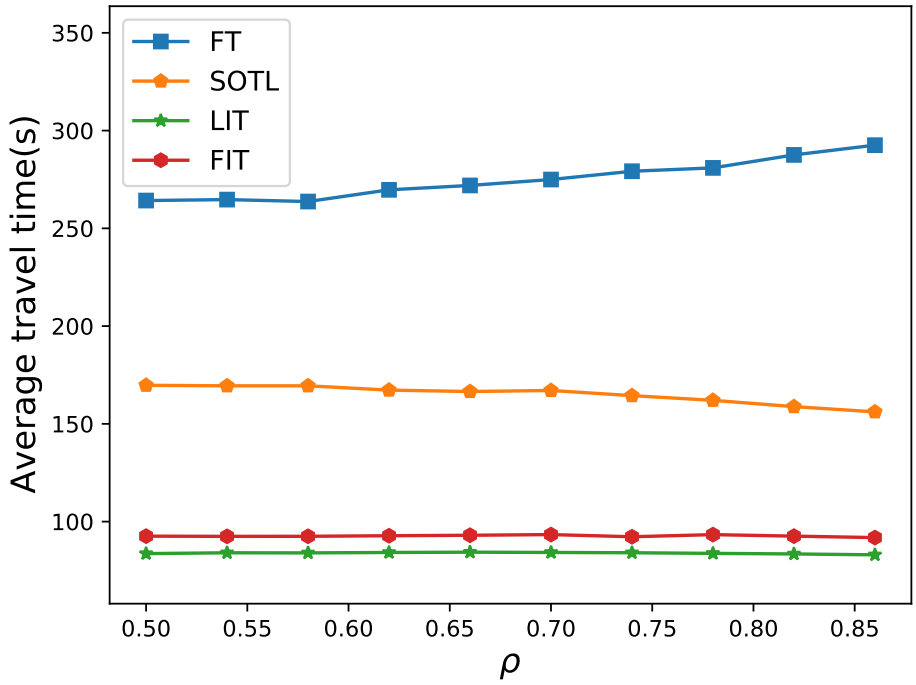


图 3.4　不同交通负载情况下四种方法的路口通行效率表现

（2）公平性评估

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 JFI（Jain Fairness Index）来量化公平性指标，JFI 的计算方式如下：

$$\mathcal{J} = \frac{\left(\sum_{i=1}^{M} \bar{D}_i\right)^2}{M \sum_{i=1}^{M} \bar{D}_i^2} \tag{3.5}$$

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 3.5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 3.6中我们可以看出，在主干道上，基于学习的方法（LIT 和 FIT）比传统方法（FT 和 SOTL）具有更低的延误时间，虽然 SOTL

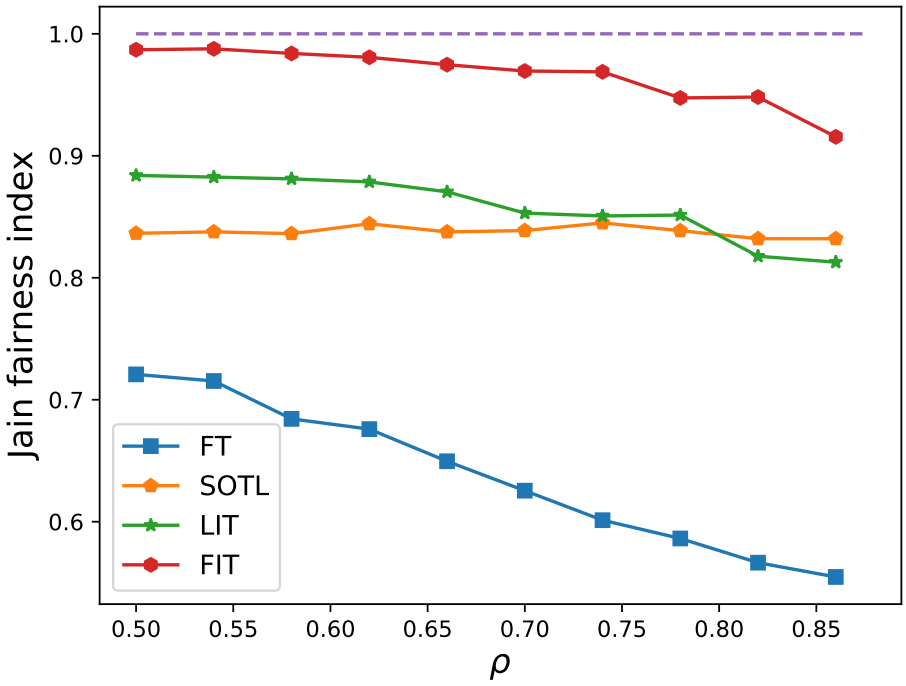


图 3.5　不同交通负载情况下四种方法的调度公平性表现

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 800 秒。

从图 3.7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。

（3）驾驶体验评估

我们研究了不同方法的驾驶体验得分情况，图 3.8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。

（4）W 参数影响分析

最后，我们研究了式 3.4中 W 参数的取值对模型性能的影响，因为 W 是用来平衡效率和公平性的，不同的值会导致学习到不同的策略，从图 3.9中我们可以看出 W 值越高，越有利于调度的公平性。相反，W 值越小，路口的通行效率就越高。具体来说，当 $W = 1$ 时，我们方法 FIT 的整体性能（包括通行效率和公平性）接近于 LIT。

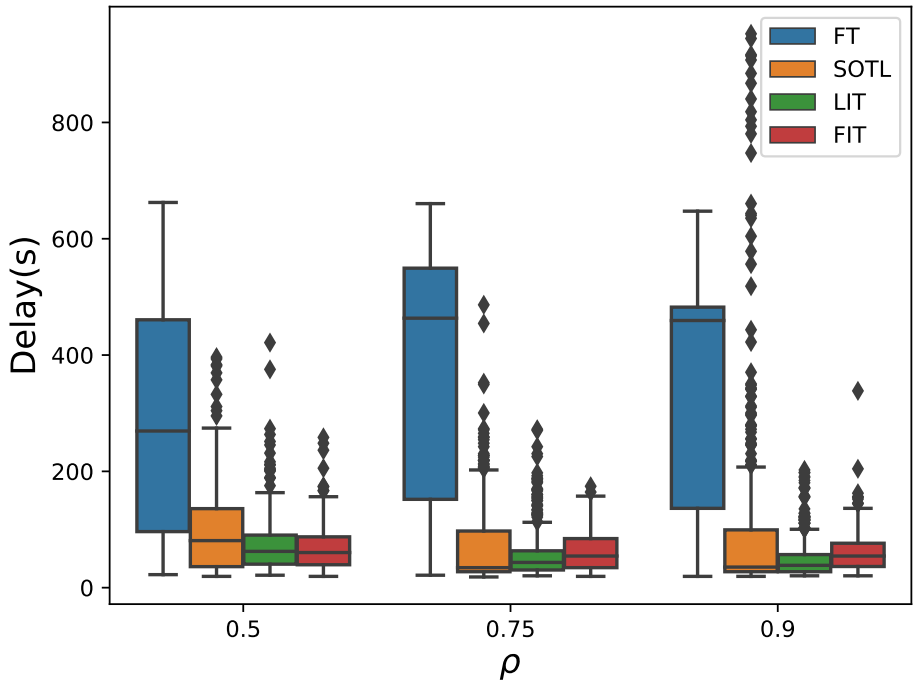


图 3.6　不同交通负载情况下四种方法在主干道 (N-S 方向) 的车辆延误时间分布

3.6　本章小结

本章首先对比了传统交通信号调度方法和基于深度强化学习的智能交通信号调度方法的优缺点，并总结了近些年来基于深度强化学习的智能交通调度研究工作，然后分析了这些工作的不足并进一步提出了本章工作的研究目标。通过引入无线通信网络中的 PFS 调度方法并配合深度强化学习，我给出了一个具有公平感知能力的模型，这个模型在提高通信效率的同时能够确保调度的相对公平性。最后，我们在仿真环境中进行了大量的实验来验证我们模型的效果，并与已有方法进行对比，进一步阐述了我们模型在公平性方面的性能提升。

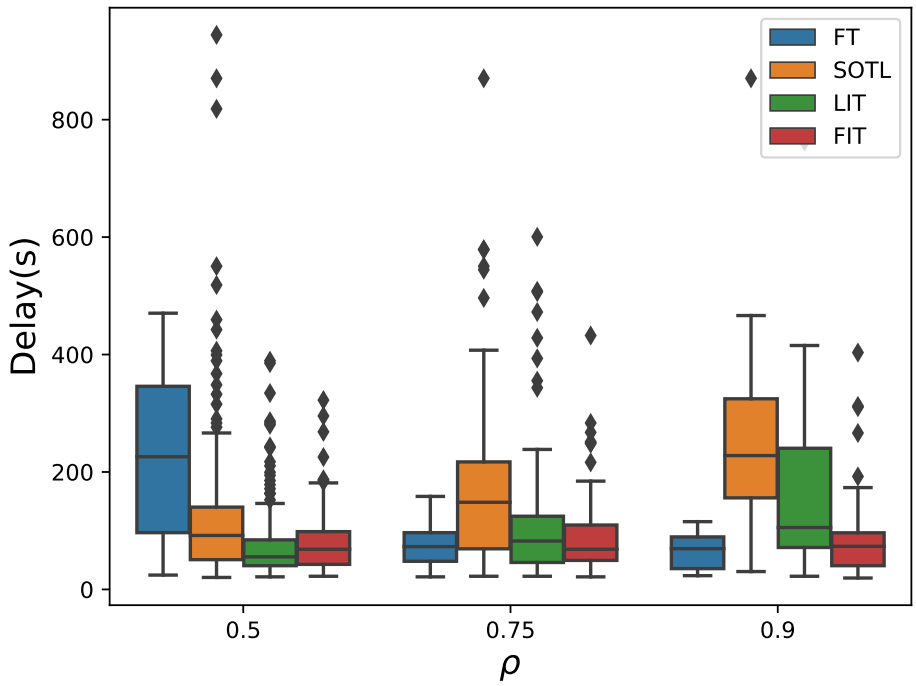


图 3.7　不同交通负载情况下四种方法在支干道 (N-S 方向) 的车辆延误时间分布（单位：秒）

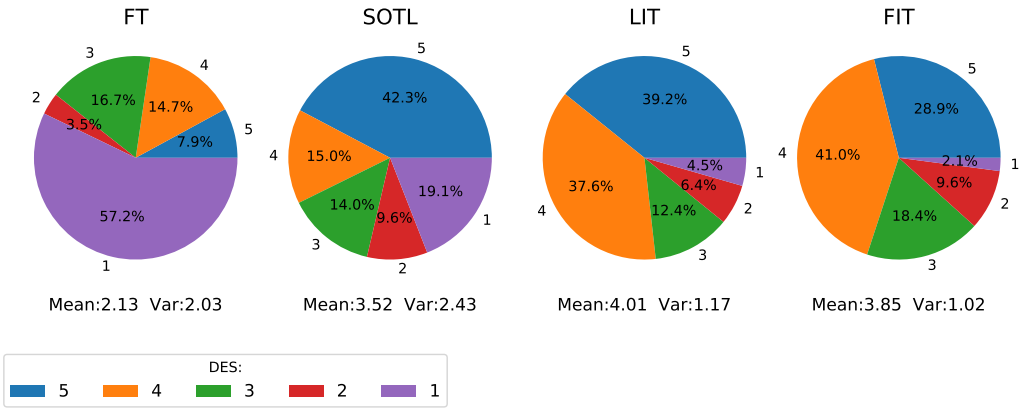
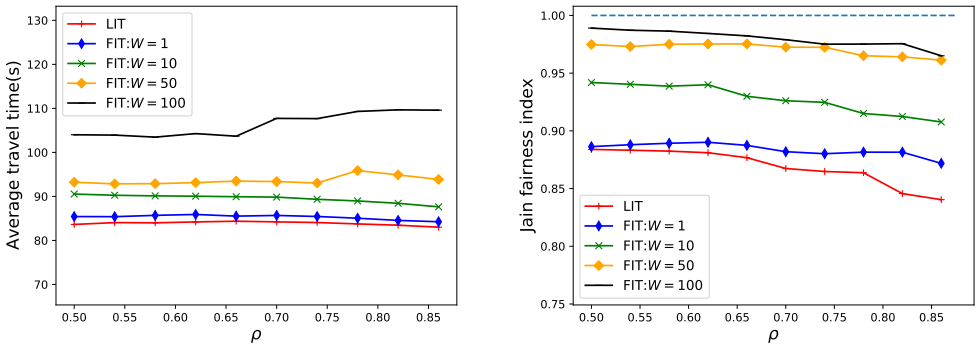


图 3.8　在 $\rho = 0.75$ 的交通负载情况下四种方法的驾驶体验得分分布



(a) 式 3.4中的参数 W 的取值对通行效率的影响　(b) 式 3.4中的参数 W 的取值对调度公平性的影响

图 3.9　式 3.4中的参数 W 的取值对通行效率和调度公平性的影响

第四章　基于深度强化学习的多路口智能交通信号调度

4.1　引言

相较于单路口的交通信号调度，多路口场景下的交通信号调度更注重对信号协调的研究。信号协调是指交通流量能够不间断地通过连续的交叉路口，即在行驶过程中没有因为遇到红色信号而停靠，那么这些路口的交通信号就是实现了协调。在现在交通协调系统中，这种信号协调只有在交通流量相对稳定的单行道上才能够实现，但为了缓解高峰时刻的交通压力，更多的是使用双向道。此外，交叉路口之间的相互干扰造成的交通拥堵也会致使信号协调失败，因此实现交叉路口之间的协调控制显得格外重要，因为信号灯的行动可能会相互影响，特别是当交叉路口紧密相连时，交通信号灯之间的良好合作能够使车辆更快地通过。

传统的交通协调方法[22]和系统[27,71,72]通常通过修改连续交叉路口之间的偏移量（即绿灯开始的时间间隔）来协调交通信号，并要求各交叉路口具有相同的信号周期长度，但这种方法只能优化某些预先定义方向的交通流[73]，而实际上，协调道路网络中交通信号的偏移量并不是一件容易的事。对于网络级的协调控制，Max-pressure[26]方法是一种最先进的信号控制方法，它在假设下游车道具有无限容量的情况下，贪婪地采取使网络吞吐量最大化的动作。其他交通控制方法如TUC[74]也使用优化技术在某些假设下（例如，在特定时间内的交通流量是均匀的）最小化车辆行驶时间或在多个交叉路口的停车次数。然而，这些假设在网络级的道路环境中往往不成立，因此阻碍了这些方法的广泛应用。

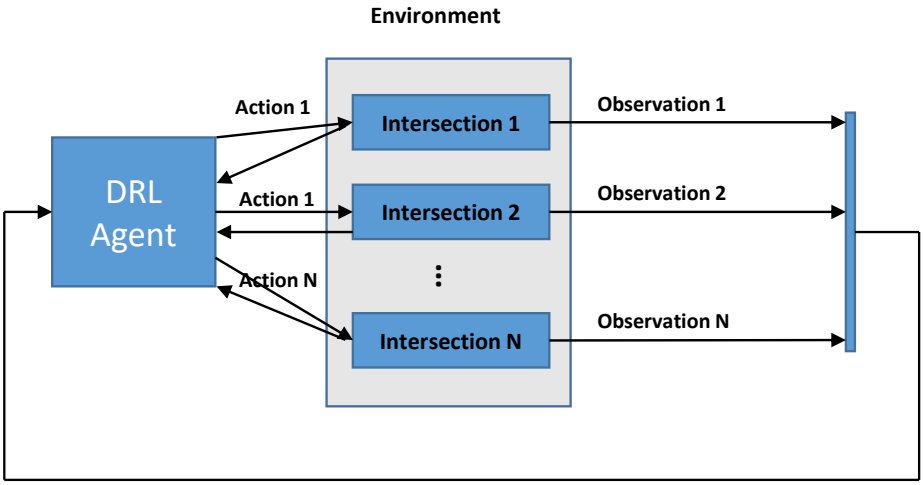


图 4.1　基于深度强化学习的多路口交通信号控制框架

4.2 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。图 4.1描述了使用多智能体强化学习解决多路口交通信号调度问题上的基本思路。智能体被定义为环境中 N 个路口的信号控制者，目标是学习每个智能体的最优策略，以优化整个环境中所有路口的通行效率。在每个调度时刻 t，每个智能体 i 观察环境的一部分作为观察点 o_i^t，智能体将对接下来要采取的行动 α^t 做出预测，这些动作将在环境中执行，并产生奖励 r_i^t，其中奖励可以在环境中的单个路口或一组路口的层面上定义。Claus 在 [75] 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，[12] 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 [67,76,77] 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。

多路口智能交通信号调度的另一条研究路线是使用独立的强化学习（Independent Reinforcement Learning, 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 [13,57,60,78–81]。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 [61] 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 [82]。典型的方法是直接将邻居的交通状况 [83] 或过去的动作 [84] 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。[19] 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，[20,85] 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。

4.3 研究目标

目前大多数工作在使用图神经网络进行信息交互的过程中，都是以交叉路口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 4.2所示：在这种建模方式下，每条车道的车辆以及当前

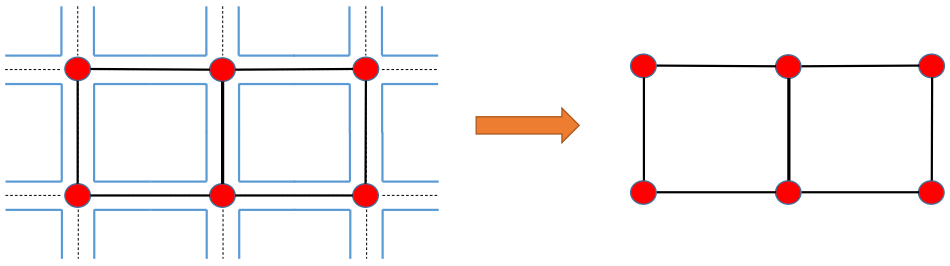


图 4.2　以路口为节点的建图方式示意图

的相位将作为该节点的特征。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，在进行信息交互的时候会出现信息冗余的情况，即当前路口下有些车道的的信息对目标路口是没有价值的，如图 4.3所

示：路口 B 中只有车道 2 的交通流向与路口 A 有关，车道 1 和车道 3 的车辆不会行驶到路口

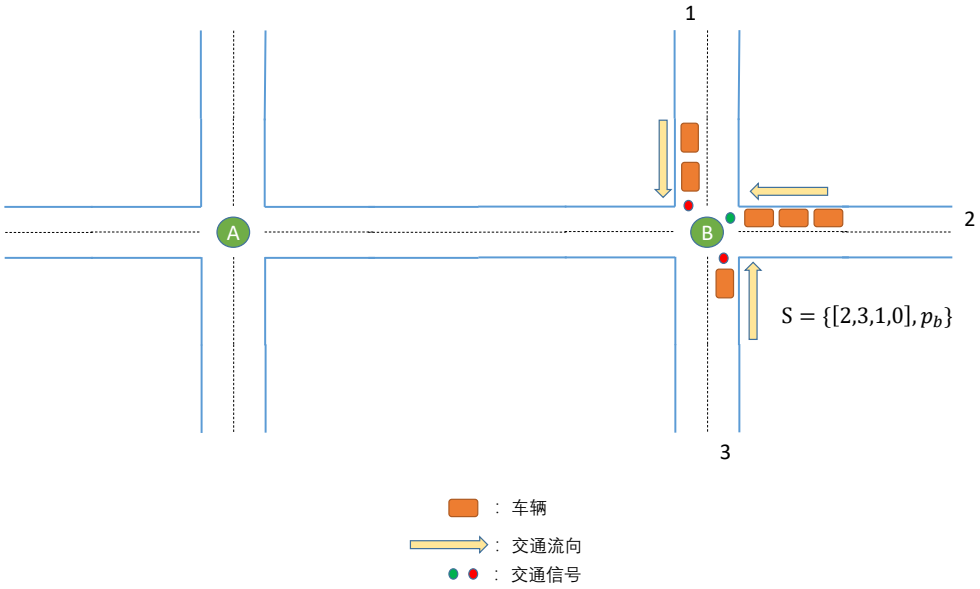


图 4.3　在以路口为节点的建图方式下的信息传递示意图

A。在信息交互的时候，如果将路口 B 所有的信息都笼统地传递过去，将会增加路口 A 的智能体提取有效信息的难度，从而降低学习的效率。

因此本工作的研究目标是在进行信息交互的时候能够剔除与目标节点无关的信息，从而降低目标节点聚合邻居节点信息的难度，提高学习效率。

4.4　系统设计

本文沿用已有的工作[20] 采用 IRL with Communication 的协调控制策略。首先提出了一种新的道路网建图方式，在此基础上，我们设计了一个新的信息交互模块，并与基于深度强化学习的智能交通调度模块想相结合从而实现协调控制。

4.4.1　基于道路的图建模方式

为了缓解智能体之间信息交互过程中的信息冗余问题我们提出了一种新的图建模方式：以道路为节点进行图建模，即一条道路就是一个节点，如图 4.4所示：

此外，我们根据当前的信号相位对图中的每边赋予了一个权重。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

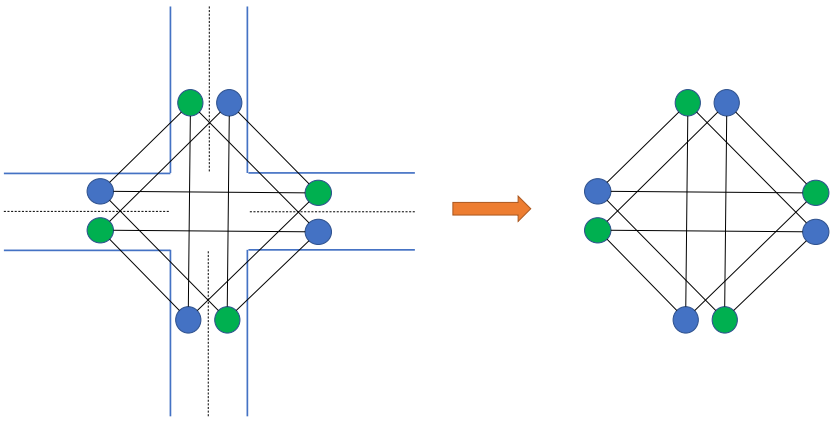


图 4.4　以道路为节点的建图方式示意图

的具体定义方法如下：

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

这个权重 $w_{i,j}$ 信息将被用于在信息交互过程中无效信息的剔除。

4.4.2　信息交互模块设计

这里我们沿用 [20,85] 工作使用图注意网络来学习信息交互，但是与之不同的是，我们不是用来学习目标节点和邻近节点的隐藏状态之间的动态相互作用，而是用来做节点回归（Node Regression），即估计目标节点在下一个时间点的特征。具体包括以下几个过程：

- 重要性计算：为了了解节点 j（源节点）的信息在确定节点 i（目标节点）下一时刻的特征的重要性，我们首先嵌入两个节点的特征，然后计算他们之间的相关系数 e_{ij}（节点 j 在确定节点 i 的特征时的重要性），具体操作如下：

$$e_{ij} = a\left(\left[Wh_i \| Wh_j\right]\right) \tag{4.5}$$

  其中 W 是一个共享参数，用来进行特征增强，然后用 $[\cdot\|\cdot]$ 对节点 i 和节点 j 增强后的特征进行拼接，最后使用 $a(\cdot)$ 将拼接后的高维特征映射到一个实数上。
- 特征筛选：由于当前交通信号的影响，道路 j 的车辆不会进入到道路 i，即节点 j 的信息在确定节点 i 的下一时刻的特征时是无用的，因此我们要筛选掉对目标节点无用的信息。

这里我们通过之前介绍的边的权重来实现：

$$e_{ij} = e_{ij} * w_{i,j} \tag{4.6}$$

如果 $WE_{ij} = 1$（即道路 j 到道路 i 的交通在当前相位下是可以通行的），e_{ij} 将维持之前的计算结果。反之，如果 $w_{i,j} = 0$，将清除节点 j 对节点 i 的影响。

- 注意力分布计算：为了重新确定源节点和目标节点之间的注意力值，我们进一步将目标节点 i 和其邻近节点之间的交互等分进行归一化：

$$\alpha_{ij} = \text{softmax}\left(e_{ij}\right) = \frac{\exp\left(e_{ij}/\tau\right)}{\sum_{j\in\mathcal{N}_i}\exp\left(e_{ij}/\tau\right)} \tag{4.7}$$

其中 τ 是一个温度系数，$\mathcal{N}_i$ 是目标节点 i 邻近范围的节点集合。

- 特征回归：为了确定目标节点在下一个时刻的特征，这里我们将其所有邻近节点的信息按照各自的重要性进行组合：

$$h_i' = \sigma\left(\sum_{j\in\mathcal{N}_i}\alpha_{ij}Wh_j\right) \tag{4.8}$$

其中 $\sigma(\cdot)$ 是激活函数。h_i 是 i 节点融合了邻域信息后的新特征。

- Multi-Head Attention：进一步，我们使用多头注意力机制（Multi-Head Attention）来关注不同相关性下的信息，如下所示：

$$h_i'(K) = \sigma\left(\frac{1}{K}\sum_{k=1}^{k=K}\sum_{j\in\mathcal{N}_i}\alpha_{ij}^k W^k h_j\right) \tag{4.9}$$

其中 K 是注意力头的数量，例如当 $k = 3$ 时，结构如图图 4.5所示：

对于每个路口，我们要估计每条引进道路在下一调度时刻的状态，即我们要对四个节点进行计算，如图 4.6中 A 路口的四个红色节点。通过在目标节点（红色节点）的及其邻近节点（绿色节点）构成的子图（如图图 4.6中右上角的图）上对目标节点进行上述的计算。由于我们是以道路为节点进行建图的，所以即便是单个路口也可以表示成一个图，所以当有多个路口的时候，会产生一张很大的图。这里每个节点维护自己路口的子图，包括更新子图中的节点特征以及边的权重（根据当前路口的信号相位确定）。

4.4.3 系统框架

如图 4.7所示，对于单个路口来说，有两个模型，一个是用来进行交通信号控制的 DQN 模型 Q，另一个是用来预测下一调度时刻状态的 GAT 模型 G。对于 DQN 模型来说，每次根据输入的状态来选择动作，其中状态由以下几部分组成：

- Queue length：当前路口每条车道的队列长度。
- Traffic volume：当前路口每条车道的车辆数。

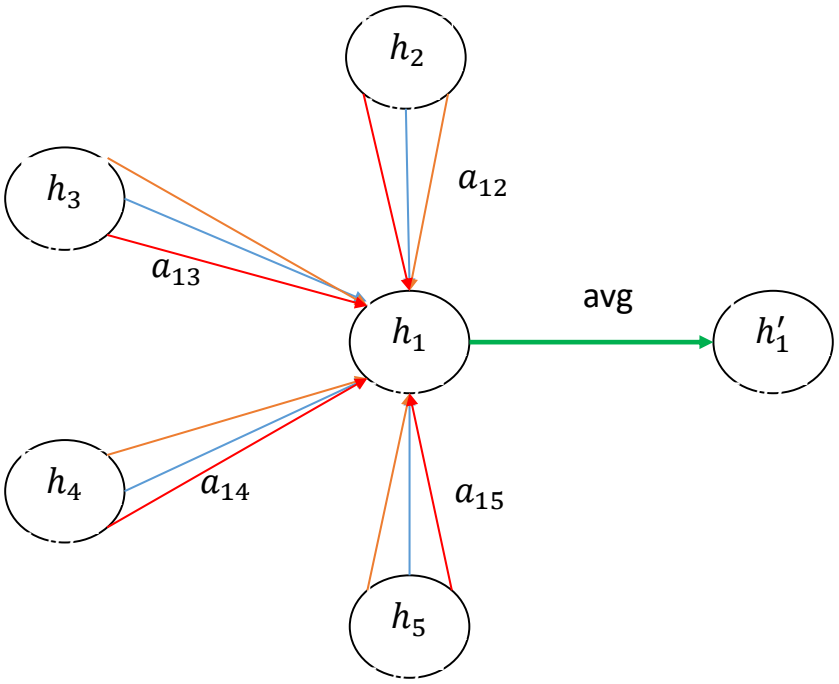


图 4.5　多头注意力计算示意图（$k = 3$）

- Current Phase：路口当前的相位。
- Next Queue length：通过节点回归估计的下一调度时刻的 Queue length。
- Next Traffic Volume：通过节点回归估计的下一调度时刻的 Volume。

其损失函数如式 4.10所示：

$$\mathcal{L}_t = [r_{t+1} + \gamma \arg\max_{a'} Q(s_{t+1}, a'; \theta) - Q(s_t, a_j; \theta)]^2 \tag{4.10}$$

对于 GAT 模型来说，首先我们规定节点的特征 $f_t = \{q_t, v_t\}$，其中 q_t 和 v_t 分别时队列长度（Queue length）和交通流量（Traffic volume）。当环境转移到新的状态 s_{t+1} 时，更新节点特征（$f_{t+1} = \{q_{t+1}, v_{t+1}\}$）以及边的权重，其损失函数如下所示：

$$\mathcal{L}_G = [f_{t+1} - f'_{t+1}]^2 \tag{4.11}$$

其中 f'_{t+1} 是在 t 调度时刻，利用邻近节点信息预估的下一时刻的节点特征，是预测值，而 f_{t+1} 是 $t+1$ 时刻的真实节点特征。

具体的算法流程如算法 4所示。

Algorithm 4 多路口协调控制中单一路口智能调度策略训练流程

输入： E：学习片段数

T：每个学习片段的步数

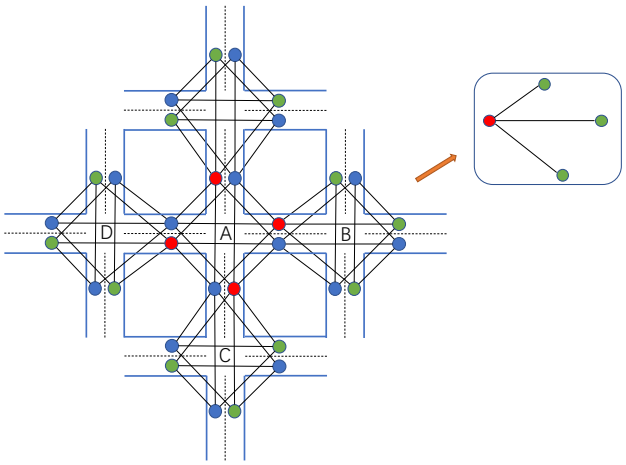


图 4.6　节点回归计算示意图

b：学习经验数

ϵ：随机选择动作概率

γ：折扣因子

Δt：信号维持时间

for $episode = 1, E$ **do**

　初始化环境。

　for $t = 1, T$ **do**

　　从环境中获取当前状态观测 $s_t = q_t, v_t, p_c$。

　　使用 GAT 估计下一调度时刻的节点特征 $f'_{t+1} = \{q'_{t+1}, v'_{t+1}\}$。

　　生成一个 0 到 1 之间的随机数 $rand$。

　　if $rand < \epsilon$ **then**

　　从动作空间中随机采样一个动作 a_t。

　　else

　　使用 DQN 模型选择动作：$a_t = \arg\max_a Q((s_t || f'_{t+1}), a; \theta)$。

　　end if

　　将当前信号更改为 a_t 并维持 δt 秒时间。

　　环境转移到新的状态 s_{t+1} 并返回一个奖励 r_{t+1}。

　　更新节点特征 $f_{t+1} = q_{t+1}, v_{t+1}$。

　　计算 GAT 损失函数 $\mathcal{L}_G$: $\mathcal{L}_G = [f_{t+1} - f'_{t+1}]^2$

更新 GAT 模型参数。
将经验 $(s_t, a_t, r_{t+1}, s_{t+1})$ 存储到经验回放池 M 中。
if $|M| > b$ **then**
从经验回放池 M 中随机采样 b 条经验数据。
计算 DQN 损失函数 $\mathcal{L}_Q$: $\mathcal{L}_Q = [r_{t+1} + \gamma \arg\max_{a'} Q(s_{t+1}, a'; \theta) - Q(s_t, a_t; \theta)]^2$
更新 DQN 模型参数。
end if
end for
end for

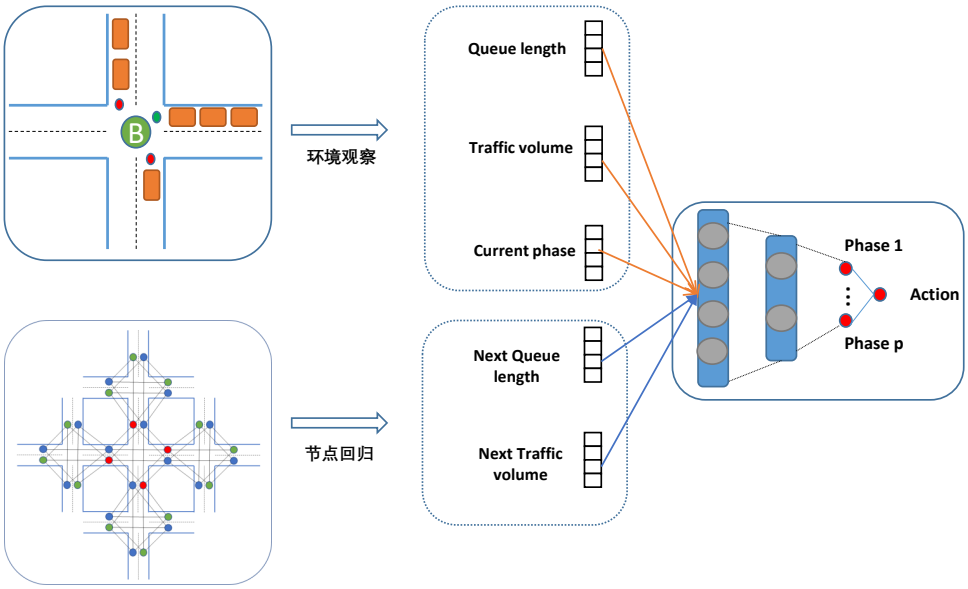


图 4.7　多路口场景下信号调度智能体与环境交互示意图

4.5　实验

4.5.1　环境介绍

我们分别在两个合成的数据集和两个真实数据集上对我们的方法进行了实验，以评估我们方法的性能。由于实验是对针对多路口场景，我们这里的仿真工具使用的是 CityFlow[1]，其对大规模交通信号控制的支持更加出色。

合成数据：延用 [20] 的做法，我们使用了以下以下两种人为合成的交通数据：

[1]http://cityflow-project.github.io

- $Grid_{6x6}-b$：6x6 的路网结构，其中每个路口有四个方向（西 → 东、东 → 西、南 → 北以及北 → 南），每个方向有三条车道（长 300 米，宽 3 米）。车辆在西 ↔ 东方向的生成速率是 300 辆/车道/小时，在南 ↔ 北方向的生成速率是 90 辆/车道/小时。
- $Grid_{6x6}$：与 $Grid_{6x6}$ 的路网结构相同，但是只有单向的车辆流动，即只有西 → 东和北 → 南的单向交通流动。

真实数据：我们还使用了杭州和济南两个城市某个路段下采集到的真实数据[1]进行了实验，表 4.1统计了这两个数据集的关键信息。

表 4.1　真实数据集（杭州和济南）的相关信息

数据集	路口数量	车辆到达率（300 辆/s）			
		均值	方差	最大值	最小值
$D_{Hangzhou}$	16	526.63	86.70	676	256
D_{Jinan}	12	250.70	38.21	335	208

- $D_{Hangzhou}$：这个数据集中有 16 个路口，其中交通数据是由路侧监控摄像头拍摄产生，每条数据包含时间、摄像头 ID 和车辆信息。通过使用摄像头位置分析这些记录，记录车辆通过道路交叉口时的轨迹。我们以通过这些路口的车辆数作为实验的交通量。
- D_{Jinan}：与 $D_{Hangzhou}$ 类似，数据集中包含了 12 个路口。

:

4.5.2 比较方法

我们将我们的方法与传统交通控制方法和基于强化学习的几种方法进行了对比：

- FT[22]: 这种方法以预先设定的方式循环改变信号。
- MaxPressure[26]：传统交通领域最先进的网络级的交通信号控制方法，每次调度时，选择压力最大的相位。
- Individual RL[18]：一种基于深度强化学习的交通信号控制方法，不考虑邻居信息。每个路口由一个智能体控制，智能体之间不共享参数，而是独立更新自己的网络。
- GCN[12]：一种基于深度强化学习的交通信号控制方法，使用 GCN 提取相邻路口的交通特征，不过其图建模方式是以路口为节点。
- Colight[20]：与 GCN 方法类似，都是以路口为节点的图建模方法，不过其使用 GAT 来学习不同路口之间的动态交互。
- GAT-Road：我们的方法，基于一种新的图建模方式（以道路为节点）。

[1]https://traffic-signal-control.github.io/#open-datasets

4.5.3 性能评估

表 4.2 不同方法在合成数据集和真实数据集上关于平均通行时间的表现

方法	所有路口的车辆平均通行时间（单位：秒）			
	$Grid_{6x6}-Uni$	$Grid_{6x6}$-b	$D_Hangzhou$	D_Jinan
Fixedtime	209.68	209.68	728.79	869.85
MaxPressure	186.07	194.96	422.15	361.33
Individual RL	314.82	261.60	345.00	325.56
GCN	205.40	272.14	768.43	625.66
CoLight	173.79	170.11	297.26	291.14
GAT-Road	**156.32**	**161.56**	**290.44**	**279.67**

我们在合成数据和真实数据上进行了大量实验，并和已有方法进行了对比，然后在通行效率和模型收敛性上进行了分析。

（1）效率比较：

表 4.2列出了所有方法在合成数据和真实数据上路口的平均通行时间，可以看出，无论是在合成数据还是真实数据集上，我们的方法都取得了最好的表现。进一步的观察数据我们可以看出：与传统交通控制中最先进的方法 (MaxPressure) 相比，我们的方法在合成数据和真实数据上都取得了一致的性能改进，其中在合成数据上的平均改进率是 16.6%，在真实数据上的平均改进率是 26.9%。

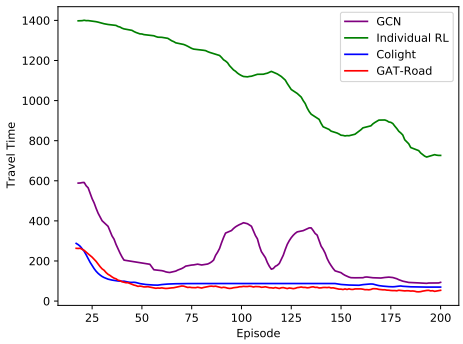
在合成数据上，Fixedtime 的表现相对比较出色，但是在真实数据上，和基于学习的方法的性能差异变得明显。这是因为人为合成的数据是相对比较规则的数据，车辆流动的随机性没有真实数据中强，交通相对更加稳定。

GCN 在真实数据上的表现并不出色是因为其在聚合邻居节点的信息时，是根据预先定义的静态权重处理邻居节点的信息，而不是根据实时的交通情况进行动态调整，当面对到变化性较强的真实数据时，这种静态处理方式会导致无法正确地聚合信息。

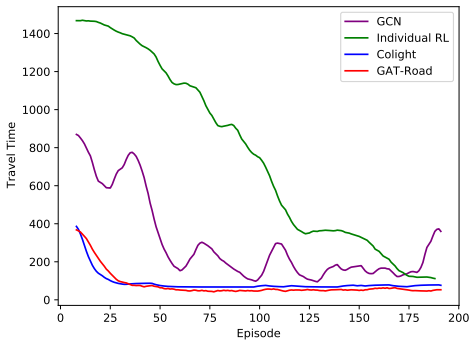
虽然 CoLight 也是使用 GAT 学习邻居节点的动态交互，但是由于其是以路口为节点来进行建图的，在数据交互的时候传递的是路口的所有交通信息，导致目标节点在聚合信息时难以挖掘有效的信息。而我们的方法 GAT-Road 以道路为节点进行建模，并且根据相位信息剔除掉了对目标节点无效的信息，因此表现更加出色。

（2）收敛性比较：

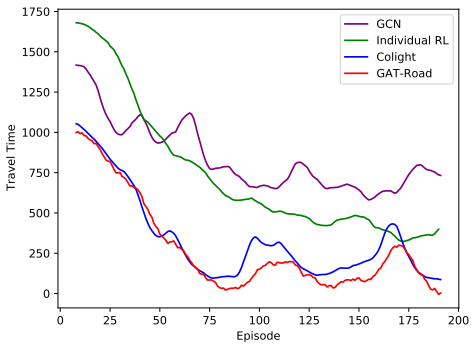
我们将我们的方法 GAT-Road 与其他三种基于深度强化学习的方法（Individual、GCN 和



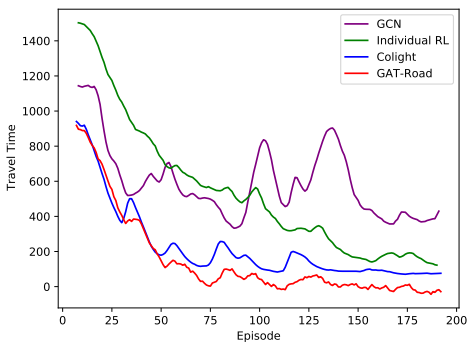
(a) 四种基于深度强化学习的方法在数据集 $Grid_{6x6}$ 上的训练收敛性表现



(b) 四种基于深度强化学习的方法在数据集 $Grid_{6x6}$-b 上的训练收敛性表现



(c) 四种基于深度强化学习的方法在数据集 $D_{Hangzhou}$ 上的训练收敛性表现



(d) 四种基于深度强化学习的方法在数据集 D_{Jinan} 上的训练收敛性表现

图 4.8　四种基于深度强化学习的方法在合成数据集以及真实数据集上的训练收敛性表现

Colight）分别合成数据和真实数据上的学习收敛速度进行了比较，结果如图 4.8所示通过对比我们可以看出：与同样是使用 IRL with Communication 框架的 GCN 和 Colight 相比，我们的方法的收敛速度更快，这得益于我们提出的新的建图方式，在这种建图方式下，节点在进行信息聚合时可以剔除与目标节点无关的信息，从而降低了学习的难度，模型更容易收敛。

GCN 方法在合成数据上收敛效果较好，但是在真实数据上收敛效果差的原因和上一节效率结果分析中的一样，静态的处理邻居路口的交通导致其难以聚合准确的信息。

虽然 Individual 最终也能收敛到最佳性能，但是由于其是独立优化单个路口的策略，没有考虑到周围路口环境的交通信息，因此其刚开始时的性能表现相较于其他三种方法更差，并且收敛速度更慢。

（3）信息交互模块边权重研究

在本工作中，我们提出了一种新的路网建图方式，并且我们根据当前的信号相位对图中的边赋予了一个权重，用来剔除在当前相位下对目标节点无效的交通流量。这里，我们研究了这个权重对信息交互模块的影响。

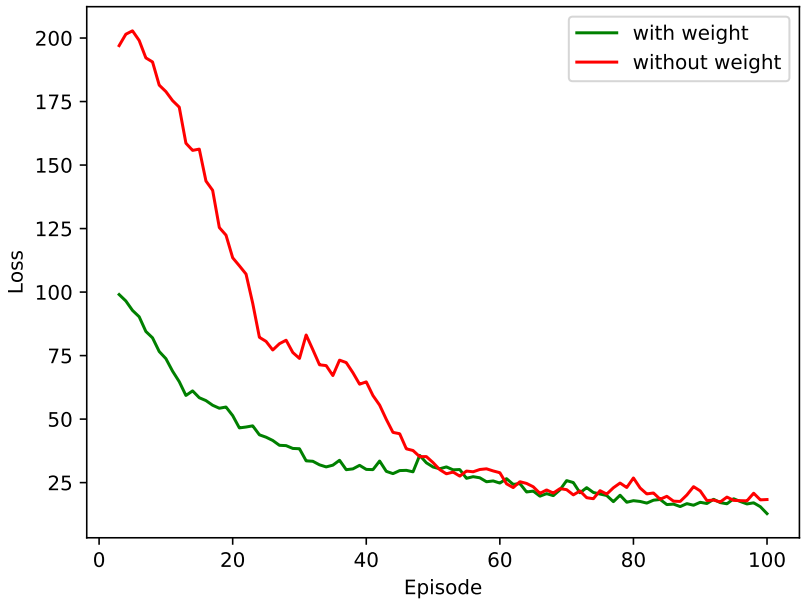


图 4.9　信息交互模块在边有权重和无权重情况下的收敛结果

图 4.9展示了我们的信息交互模块在边有权重和权重情况下的收敛结果，其中红色曲线代表边有权重情，绿色曲线代表边无权重，横坐标代表训练集数，纵坐标代表模型的损失，这里我们使用信息交互模块来对图的节点做节点回归，即估计目标道路在下一调度时刻的交通状况。从图中我们可以看出，在为边加了权重之后，信息交互模块的收敛速度明显提高，这是因为我们的权重是根据相位信息确定的，在加了权重后，可以过滤掉在当前信号相位下对目标节点（道路）无效的交通流量。

4.6　本章小结

本章首先给出了多路口场景下交通信号调度的主要研究目标，然后总结了近些年来基于深度强化学习的多路口交通信号调度的研究工作，通过分析这些工作的不足，我们给出了本章工作的研究目标。我们提出了一个新的路网建图方式并在此基础上设计了一种新的信息交互模块，可以有效地剔除数据交互过程中邻居节点的无效信息。最后我们在仿真环境中分别就合成交通数据和真实交通数据进行了实验，并与已有方法进行对比，进一步阐述了我们的模型在通行效率和学习速度上的提升。

第五章　总结与展望

5.1　工作总结

本文主要研究了基于深度强化学习的智能交通信号调度问题，包括单路口场景以及多路口场景下的信号调度问题。首先我们总结了一些传统的交通信号控制方法，这些方法要么受制于一些严苛的假设条件导致难以应用于实际交通控制中，要么缺乏对实时交通状况的考虑导致无法高效地缓解交通拥堵的情况。然后我们总结了一些已有的使用深度强化学习解决智能交通信号调度问题的工作，在不同的调度场景下（单路口场景和多路口场景），我们对这些已有工作的进行了系统的分析，并指出他们存在的不足，最后我们针对这些不足提出新的解决方案。

对于单路口场景下的交通信号控制，已有的基于学习的方法更多的注重于提高通行效率，而忽略了公平性问题。在本文中，我们提出了一个新的模型可以在保证通行效率的同时，兼顾到对公平性的考虑。最后，我们在仿真环境中进行了大量的实验来验证我们模型的效果，并与已有方法进行对比，进一步阐述了我们模型在公平性方面的性能提升。

对于多路口场景下的交通信号控制，我们使用了 IRL with communication 的协调控制框架，并提出了一种新的将道路网建模成图的建模方式，在此基础上，我们设计了一种新的信息交互模块，通过这个信息交互模块，智能体在提取邻近节点的信息时可以剔除那些对自己无用的信息。最后我们在仿真环境中分别就合成交通数据和真实交通数据进行了实验，并与已有方法进行对比，进一步阐述了我们的模型在通行效率和学习速度上的提升。

5.2　未来展望

虽然目前有很多使用强化学习来解决智能交通信号调度的工作，并且也取得了不错的效果，但是任然有一些问题值得被深入研究：

（1）信用分配问题

信用分配问题是强化学习领域中被广泛研究的问题之一，它考虑的是对一个动作的成功（或失败的惩罚）的信用分配，即一个动作对于最终任务的实现有多大的贡献度，或者对于最终任务的失败应该承担多大的责任。在交通信号控制问题中，交通状况是交通信号控制器所采取的若干行动的结果，这带来了两个问题。(1) 一个行动可能在几步行动后仍有效果；(2) 每个时间戳的奖励是之前几个动作的组合结果。在常规的强化学习应用环境中（例如，Atari 游戏或围棋），有时会把一个训练集的最终得分分配给与这一集相关的所有动作，而交通信号控制问题中的动作从执行到产生影响可能有一个时间间隔，这个时间间隔可能是动态变化的，需要进一步研究。

（2）Bus-Priority 问题

Bus-Priority 是指在交通信号调度中，公共交通车辆相较于普通车辆应该具有更高的调度优先级，以鼓励人们在日常出行中乘坐公共交通用具。但是目前很多工作为了简化控制难度，都是无区别的对待道路上的交通流量。

（3）实地测试问题

目前有很多使用强化学习来解决智能交通信号调度的工作都只停留在仿真阶段，即实验场景的搭建以及效果的验证都是通过仿真器完成的，要想将这些模型部署到现实生活中的信号灯上还需要更多的研究和实地测试。但是随着车联网技术的发展使得我们可以实时地获取道路上车辆的信息，加上最近人工智能技术的崛起，一些基于学习的交通信号调度算法可以在与环境的交互中不断的提升自身的性能，这些因素为进一步实现真实道路场景下的智能交通信号控制提供了技术支持。

（4）安全性问题

如何使强化学习智能体在物理环境中可以被接受是未来研究的一个重要方向。虽然强化学习方法从试错中学习，但在现实世界中，学习成本可能是关键的，甚至是致命的，因为交通信号的故障可能导致很严重的事故。因此，如何在强化学习中采用风险管理从而防止在智能体的学习过程中和学习结束后出现不必要的行为是十分关键的问题。

参考文献

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

[2] Bång K, Nilsson L. Optimal control of isolated traffic signals[J]. IFAC Proceedings Volumes, 1976, 9(4):173–184.

[3] Silcock J. Designing signal-controlled junctions for group-based operation[J]. Transportation Research Part A: Policy and Practice, 1997, 31(2):157–173.

[4] Haddad J, De Schutter B, Mahalel D, et al. Optimal steady-state control for isolated traffic intersections[J]. IEEE Transactions on automatic control, 2010, 55(11):2612–2617.

[5] Webster F. Traffic signals[J]. Road research technical paper, 1966, 56.

[6] Abdulhai B, Pringle R, Karakoulas G J. Reinforcement learning for true adaptive traffic signal control[J]. Journal of Transportation Engineering, 2003, 129(3):278–285.

[7] Abdoos M, Mozayani N, Bazzan A L. Holonic multi-agent system for traffic signals control[J]. Engineering Applications of Artificial Intelligence, 2013, 26(5-6):1575–1587.

[8] Bakker B, Whiteson S, Kester L, et al. Traffic light control by multiagent reinforcement learning systems[M]. . Proceedings of Interactive Collaborative Information Systems. Springer, 2010: 475–510.

[9] El-Tantawy S, Abdulhai B, Abdelgawad H. Multiagent reinforcement learning for integrated network of adaptive traffic signal controllers (MARLIN-ATSC): methodology and large-scale application on downtown Toronto[J]. IEEE Transactions on Intelligent Transportation Systems, 2013, 14(3):1140–1150.

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

[11] Li L, Lv Y, Wang F Y. Traffic signal timing via deep reinforcement learning[J]. IEEE/CAA Journal of Automatica Sinica, 2016, 3(3):247–254.

[12] 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..

[13] 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.

[14] Genders W, Razavi S. Using a deep reinforcement learning agent for traffic signal control[J]. arXiv preprint arXiv:1611.01142, 2016..

[15] Gao J, Shen Y, Liu J, et al. Adaptive traffic signal control: Deep reinforcement learning algorithm with experience replay and target network[J]. arXiv preprint arXiv:1705.02755, 2017..

[16] Liu M, Deng J, Xu M, et al. Cooperative deep reinforcement learning for tra ic signal control[C]. Proceedings of Proc. 23rd ACM SIGKDD Conf. Knowl. Discovery Data Mining (KDD), 2017.

[17] Arel I, Liu C, Urbanik T, et al. Reinforcement learning-based multi-agent system for network traffic signal control[J]. IET Intelligent Transport Systems, 2010, 4(2):128–135.

[18] 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.

[19] 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.

[20] 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.

[21] Chen C, Wei H, Xu N, et al. Toward a thousand lights: Decentralized deep reinforcement learning for large-scale traffic signal control[C]. Proceedings of Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, 2020. 3414–3421.

[22] Koonce P, Rodegerdts L. Traffic signal timing manual.[R]. Technical report, United States. Federal Highway Administration, 2008.

[23] Roess R P, Prassas E S, McShane W R. Traffic engineering[M]. Pearson/Prentice Hall, 2004.

[24] Little J D, Kelson M D, Gartner N H. MAXBAND: A versatile program for setting signals on arteries and triangular networks[J]. 1981..

[25] 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.

[26] Varaiya P. The max-pressure controller for arbitrary networks of signalized intersections[M]. . Proceedings of Advances in Dynamic Network Modeling in Complex Transportation Systems. Springer, 2013: 27–66.

[27] Lowrie P. Scats, sydney co-ordinated adaptive traffic system: A traffic responsive method of controlling urban traffic[J]. 1990..

[28] Lillicrap T P, Hunt J J, Pritzel A, et al. Continuous control with deep reinforcement learning[J]. arXiv preprint arXiv:1509.02971, 2015..

[29] Fujimoto S, Hoof H, Meger D. Addressing function approximation error in actor-critic methods[C]. Proceedings of International Conference on Machine Learning. PMLR, 2018. 1587–1596.

[30] Haarnoja T, Zhou A, Abbeel P, et al. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor[C]. Proceedings of International conference on machine learning. PMLR, 2018. 1861–1870.

[31] Zhang Q, Jin Q, Chang J, et al. Kernel-weighted graph convolutional network: A deep learning approach for traffic forecasting[C]. Proceedings of 2018 24th International Conference on Pattern Recognition (ICPR). IEEE, 2018. 1018–1023.

[32] Guo S, Lin Y, Feng N, et al. Attention based spatial-temporal graph convolutional networks for traffic flow forecasting[C]. Proceedings of Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, 2019. 922–929.

[33] Shi H, Yao Q, Guo Q, et al. Predicting origin-destination flow via multi-perspective graph convolutional network[C]. Proceedings of 2020 IEEE 36th International Conference on Data Engineering (ICDE). IEEE, 2020. 1818–1821.

[34] Hu J, Yang B, Guo C, et al. Stochastic origin-destination matrix forecasting using dual-stage graph convolutional, recurrent neural networks[C]. Proceedings of 2020 IEEE 36th International

Conference on Data Engineering (ICDE). IEEE, 2020. 1417–1428.

[35] Li J, Ma H, Zhang Z, et al. Social-wagdat: Interaction-aware trajectory prediction via wasserstein graph double-attention network[J]. arXiv preprint arXiv:2002.06241, 2020..

[36] Sun Y, He T, Hu J, et al. Socially-aware graph convolutional network for human trajectory prediction[C]. Proceedings of 2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC). IEEE, 2019. 325–333.

[37] Monti A, Bertugli A, Calderara S, et al. Dag-net: Double attentive graph neural network for trajectory forecasting[C]. Proceedings of 2020 25th International Conference on Pattern Recognition (ICPR). IEEE, 2021. 2551–2558.

[38] Sperduti A, Starita A. Supervised neural networks for the classification of structures[J]. IEEE Transactions on Neural Networks, 1997, 8(3):714–735.

[39] Gori M, Monfardini G, Scarselli F. A new model for learning in graph domains[C]. Proceedings of Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005., volume 2. IEEE, 2005. 729–734.

[40] Scarselli F, Gori M, Tsoi A C, et al. The graph neural network model[J]. IEEE transactions on neural networks, 2008, 20(1):61–80.

[41] Veličković P, Cucurull G, Casanova A, et al. Graph attention networks[J]. arXiv preprint arXiv:1710.10903, 2017..

[42] Lee J B, Rossi R, Kong X. Deep graph attention model[J]. arXiv preprint arXiv:1709.06075, 2017..

[43] Zhang J, Shi X, Xie J, et al. Gaan: Gated attention networks for learning on large and spatiotemporal graphs[J]. arXiv preprint arXiv:1803.07294, 2018..

[44] Kipf T N, Welling M. Variational graph auto-encoders[J]. arXiv preprint arXiv:1611.07308, 2016..

[45] Pan S, Hu R, Long G, et al. Adversarially regularized graph autoencoder for graph embedding[J]. arXiv preprint arXiv:1802.04407, 2018..

[46] Yu W, Zheng C, Cheng W, et al. Learning deep network representations with adversarially regularized autoencoders[C]. Proceedings of Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018. 2663–2671.

[47] Cao S, Lu W, Xu Q. Deep neural networks for learning graph representations[C]. Proceedings of Proceedings of the AAAI Conference on Artificial Intelligence, volume 30, 2016.

[48] Wang D, Cui P, Zhu W. Structural deep network embedding[C]. Proceedings of Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining, 2016. 1225–1234.

[49] Tu K, Cui P, Wang X, et al. Deep recursive network embedding with regular equivalence[C]. Proceedings of Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining, 2018. 2357–2366.

[50] De Cao N, Kipf T. MolGAN: An implicit generative model for small molecular graphs[J]. arXiv preprint arXiv:1805.11973, 2018..

[51] Li Y, Vinyals O, Dyer C, et al. Learning deep generative models of graphs[J]. arXiv preprint arXiv:1803.03324, 2018..

[52] You J, Ying R, Ren X, et al. GraphRNN: Generating Realistic Graphs with Deep Auto-regressive

Models[J]. 2018..

[53] Bojchevski A, Shchur O, Zügner D, et al. NetGAN: Generating Graphs via Random Walks[J]. 2018..

[54] 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.

[55] 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.

[56] 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.

[57] 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.

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

[59] Aslani M, Mesgari M S, Wiering M. Adaptive traffic signal control with actor-critic methods in a real-world traffic network with different traffic disruption events[J]. Transportation Research Part C: Emerging Technologies, 2017, 85:732–752.

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

[61] 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.

[62] Abdoos M, Mozayani N, Bazzan A L. Traffic light control in non-stationary environments based on multi agent Q-learning[C]. Proceedings of 2011 14th International IEEE conference on intelligent transportation systems (ITSC). IEEE, 2011. 1580–1585.

[63] Abdoos M, Mozayani N, Bazzan A L. Hierarchical control of traffic signals using Q-learning with tile coding[J]. Applied intelligence, 2014, 40(2):201–213.

[64] Zang X, Yao H, Zheng G, et al. Metalight: Value-based meta-reinforcement learning for traffic signal control[C]. Proceedings of Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, 2020. 1153–1160.

[65] Aslani M, Mesgari M S, Seipel S, et al. Developing adaptive traffic signal control by actor–critic and direct exploration methods[C]. Proceedings of Proceedings of the Institution of Civil Engineers-Transport, volume 172. Thomas Telford Ltd, 2019. 289–298.

[66] Aslani M, Seipel S, Mesgari M S, et al. Traffic signal optimization through discrete and continuous reinforcement learning with robustness analysis in downtown Tehran[J]. Advanced Engineering Informatics, 2018, 38:639–655.

[67] 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.

[68] Rizzo S G, Vantini G, Chawla S. Time critic policy gradient methods for traffic signal control in complex and congested scenarios[C]. Proceedings of Proceedings of the 25th ACM SIGKDD

International Conference on Knowledge Discovery & Data Mining, 2019. 1654–1664.

[69] Zheng G, Xiong Y, Zang X, et al. Learning phase competition for traffic signal control[C]. Proceedings of Proceedings of the 28th ACM International Conference on Information and Knowledge Management, 2019. 1963–1972.

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

[71] Hunt P, Robertson D, Bretherton R, et al. SCOOT-a traffic responsive method of coordinating signals[R]. Technical report, 1981.

[72] Hunt P, Robertson D, Bretherton R, et al. The SCOOT on-line traffic signal optimisation technique[J]. Traffic Engineering & Control, 1982, 23(4).

[73] Gartner N H, Assman S F, Lasaga F, et al. A multi-band approach to arterial traffic signal optimization[J]. Transportation Research Part B: Methodological, 1991, 25(1):55–74.

[74] Diakaki C, Papageorgiou M, Aboudolas K. A multivariable regulator approach to traffic-responsive network-wide signal control[J]. Control Engineering Practice, 2002, 10(2):183–195.

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

[76] 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..

[77] 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.

[78] 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.

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

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

[81] 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.

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

[83] 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.

[84] 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.

[85] 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。

在学期间的研究成果及学术论文情况

攻读硕士学位期间发表（录用）论文情况

1. 以后可能会在这里也用上 biber
2. 不过目前还需要手写论文全称

研究生期间参与的科研项目

1. 国家自然科学基金 (No.12345678)