江西财经大学
JIANGXI UNIVERSITY OF FINANCE AND ECONOMICS

学校代码 _____
密　　级 _____
中图分类号 _____
UDC _____

硕士学位论文
MASTER DISSERTATION

论文题目（中文） 基于预测交通流的交通信号控制优化

论文题目（英文） Traffic signal control optimization based on predicted traffic flow

作　　者 占俊卿　　导　　师 邓庆山

申请学位 硕 士　　学院名称 软件与物联网工程学院

学科专业 软件工程　　研究方向 智慧交通

二〇二三年 六月

摘要

近年来，随着我国经济的快速增长与国民生活水平的不断改善，小汽车逐渐地普及到了越来越多的家庭，这种汽车普及的现象也造成了国内私家车持有率大幅增长，并且随着国内城镇化水平不断提高，私家车持有率增速最快的地方基本都集中在城市中，因此也就导致了现有城市交通道路设施建设与升级的速度无法与目前私家车的大量普及而引起的城市路网车流量快速增长相匹配。城市的道路拥堵且通行效率不高，成为大城市道路交通的一个常态。交通拥堵成为我国城市化发展进程中需要亟待解决的问题之一。因此，深入分析研究国内外各大城市交通拥堵现状，采用更高效的交通策略来提高交通路网车辆通行效率解决交通拥堵，也成为国内外交通领域研究的热门话题。

交通信号灯作为在交叉路口控制车辆流通的工具，对整个道路交通的通行能力有着十分重要的作用。近年来，利用深度强化学习（Deep Reinforcement Learning，DRL）方法进行道路交叉口交通信号灯信号控制的研究逐渐受到关注。与传统的固定时间间隔或手动设定的交通信号控制方法相比，利用 DRL 方法进行交通信号控制具有自适应性、实时性和效果优良等优点。

本文提出一种结合预测交通流状态的 DRL 交通信号控制算法，该算法将长短期记忆神经网络（LSTM）与双深度 Q 网络（Double Deep Q-Network）相结合，构建交通信号自适应控制模型。该模型利用 LSTM 预测未来交叉路口的交通状态（交通流状态），通过预测状态并结合 Q-learning 的方式改进 Double DQN 中 Target 目标网络的 Q 值计算方式。该模型还在原有的 Double DQN 网络中加入动态 ε 动作选择策略，该策略改变传统 DQN 类算法使用线性函数来调控动作选择策略的方式，使用线性函数与 Sigmoid 函数构成的分段函数来调控算法中最优动作选择的随机性，该函数相比传统方法更加平滑。本文通过以上方法的改进可以较好地避免传统的 DQN 目标网络过高估计 Q 值的问题以及提高训练模型的可靠性。本文利用 KDD 城市大脑比赛所提供的仿真引擎和模拟实际交通环境的路网数据进行仿真实验，通过构建交叉口与其之间连通的道路所形成的路网为研究对象，在该路网区域内使得交叉路口所能提供的车辆最大承载数量增加，同时最大程度地减少路网车辆的平均延迟。通过与其他一些改进的 DQN 交通信号控制方法进行比较，实验结果表明，该算法能够在不同的路况和交通流量情况下有效地优化交通信号控制策略，提高道路交通效率和安全性。

关键词：深度强化学习；交通信号；动态策略；LSTM；DQN

Abstract

In recent years, with the rapid growth of China's economy and the continuous improvement of national living standards, small cars have gradually become more and more popular among families, and this phenomenon of car popularity has also caused a significant increase in the domestic private car ownership rate, and with the continuous increase in the level of urbanization in China, the fastest growth rate of private car ownership is basically concentrated in the cities, which has led to the construction of existing urban traffic road facilities The speed of construction and upgrading of existing urban transportation road facilities cannot match the rapid growth of traffic on the urban road network caused by the massive popularity of private cars. Urban road congestion and inefficient traffic flow has become a norm for road traffic in large cities. Traffic congestion has become one of the problems that need to be solved in China's urbanization development process. Therefore, in-depth analysis and research on the current situation of traffic congestion in major cities at home and abroad, and the adoption of more efficient traffic strategies to improve the efficiency of the traffic network vehicles to solve traffic congestion, has also become a popular topic of research in the field of transportation at home and abroad.

As a tool to control the flow of vehicles at intersections, traffic signals play a very important role in the overall road traffic capacity. In recent years, research on traffic signal control at road intersections using Deep Reinforcement Learning (DRL) methods has received increasing attention. Compared with traditional traffic signal control methods with fixed time intervals or manual settings, traffic signal control using DRL methods has the advantages of self-adaptability, real-time and excellent results.

In this thesis, we propose a DRL traffic signal control algorithm combining predicted traffic flow states, which combines a Long Short Term Memory Neural Network (LSTM) with a Double Deep Q-Network to build a traffic signal adaptive control model. The model uses the LSTM to predict the future traffic state (traffic flow state) at intersections, and improves the Q-value calculation of the Target network in Double DQN by predicting the state and combining it with Q-learning. The model also adds a dynamic ε-action selection strategy to the original Double DQN network, which changes the traditional DQN-like algorithm using a linear function to regulate the action selection strategy, and uses a segmentation function composed of a linear function and a Sigmoid function to regulate the randomness of the optimal action selection in the algorithm, which is smoother than the traditional method. In this thesis, the

improvement of the above method can better avoid the problem of over-estimation of Q value in the traditional DQN target network as well as improve the reliability of the training model. This thesis conducts simulation experiments based on the simulation engine provided by the KDD City Brain Competition and the road network data simulating the actual traffic environment. By constructing a road network formed by intersections and the roads connected between them, the maximum number of vehicles that can be carried at intersections is increased in the area of the network, and the average delay of vehicles on the network is minimized. By comparing with some other improved DQN traffic signal control methods, the experimental results show that the algorithm can effectively optimize the traffic signal control strategy and improve road traffic efficiency and safety under different road conditions and traffic flow situations.

Key Words: Deep reinforcement learning; Traffic signal; Dynamic Strategy; LSTM; DQN

目 录

1 绪论

1.1 研究背景和意义

1.1.1 研究背景

随着我国不断加快推进城镇化的进程，全国城镇人口的数量也急剧增加。在 2022 年末，我国的城镇常住人口 92071 万人，比 2021 年末增加 646 万人；乡村常住人口 49104 万人，相比上年末减少 731 万人；城镇人口占全国人口比重（城镇化率）为 65.22%，与上年末相比上涨了 0.50%[1]。至 2021 年底，全国城市道路总长度达 53.2 万公里，与 2011 年的 30.9 万公里相比，10 年间我国城市道路总长度翻了接近 1 倍的。城市人均道路面积已达 18.84 平方米，是 2004 年的近 2 倍。城市交通道路的迅猛发展，给传统的交通治理管控模式带来了巨大的挑战。据公安部统计，2022 年全国机动车保有量达 4.17 亿辆，其中汽车 3.19 亿辆。2022 年全国新注册登记机动车 3478 万辆，新领证驾驶人 2923 万人。新注册登记机动车 3478 万辆，新注册登记汽车 2323 万辆[2]。全国机动车保有量扣除报废注销量比 2021 年增加 2129 万辆，增长 5.39%。在汽车保有量不断攀上的前提下，新能源汽车的数量也在急剧上升，截至 2021 年底，全国新能源汽车保有量达 784 万辆，占汽车总量的 2.60%，扣除报废注销量比 2020 年增加 292 万辆，增长 59.25%[3]。虽然我国城市道路建设的步伐不断在加快，但依然无法满足当前过于快速增长的机动车数量的出行需求，因此在我国几乎所有的一、二线城市中都面临着交通拥堵的问题。

在当前 21 世纪，信息和通信技术、计算机技术以及物联网的高速发展促使智能交通系统[4]（intelligent transportation system, ITS）得到快速发展，ITS 的主要目标是为参与者提供安全、有效和可靠的交通系统。为此，最优交通信号控制（TSC）、交通流控制等成为关键研究领域并且在很多实际场景中得到了成功的运用[5]，因此也使得智能交通系统得到了越来越多的关注，越来越多的城市管理者和大众逐渐认识到了智能交通系统的重要性，智能交通系统和人工智能的结合为 21 世纪的交通研究提供了有效的解决方案，其中就包括典型的交通拥堵这样问题。造成交通拥堵的原因一方面除了上述由于私家车辆的急剧增加与城市道路的建设不平衡等原因，另一方面更重要原因在于目前国内很多城市的交通信号控制系统无法发挥与当前日益复杂的交通环境相匹配的交通指挥和调控作用。当前国内许多城市所采用的交叉路口交通信号控制系统大多数是集中式控制系统，比如北京采用的 SCOOT 系统[6]，能够对路网内多个交叉口的交通信号进行协调控制，但该系统对区域路网中单个交叉路口的交通信号控制上却存在许多问题，比如对复杂交通环境适应性较差、调控效果差等。因此从上述举例的交通信

号控制系统在当前交通环境下暴露的问题可以看出，在当前国内交通设施的快速发展和人们对交通需求的不断提高的情况下，城市交通信号控制的重要性将日益突出。

1.1.2 研究意义

随着大数据和人工智能技术的快速发展，城市路网交通信号的智能自适应控制已成为一个热门研究方向。在过去，要获取并处理海量的交通数据是一项艰难的任务，但现在有很多的开放式数据平台为我们提供了丰富且易于获取的详细交通信息数据源。因此可以获得的数据是多样的，例如广泛分布的路控高清摄像头、毫米波雷达等交通信息收集传感器，能够提供的各种车辆信息，包括道路车流量、位置、速度、方向以及各种地图程序和官方交管系统公开的大量真实交通数据。利用这些交通数据我们可以便捷高效地获取交叉路口的车辆信息、道路信息等。高效的数据获取以及其计算处理相关技术的突破，也推动了城市交通控制系统的快速发展。

目前智慧交通在大数据以及人工智能技术突破下快速发展，交通信号控制作为智慧交通中解决交通拥堵这个交通领域重要问题的一个分支系统，在智慧交通系统中具有十分重要的作用，交通的好坏比较直接的表现就在于道路通行效率是否高效，因此更优异的交通信号控制方法对于提高通行效率有着很好的助力。交通拥堵是现代世界的一个棘手问题，道路上的汽车数量不断增加，交通拥堵每年都在恶化，传统的交通信号控制系统比如强化学习、绿波技术、深度学习和遗传算法等方法都常常被应用在交通信号控制方法研究中，但传统的交通信号控制系统在不断增加的交通量方面的不足已经变得显而易见。因此，对于智慧交通系统是迫切需要一种更高效、更智能的交通信号控制系统，而深度强化学习（DRL）方法[7,8]是解决此问题一个很好的途径。DRL 是机器学习的一个子集，它利用人工智能方式让 Agent[9]根据其自身经验学习和做出决策。DRL 算法可以从历史交通数据中学习交通信号控制策略，去优化交通流量、减少拥堵并提高整体交通效率。使用 DRL 方法探索交通信号控制的重要性在于它有可以为交通拥堵提供更有效和高效的解决方案，比如通过利用历史交通数据和实时优化交通信号控制，基于 DRL 算法[10]的系统可以改善交叉路口交通流量并最大限度地减少路口的等待时间。此外，基于 DRL 的交通信号控制系统具有很好的自适应性，可以根据不断变化的交通模式对自身模型的参数进行调整，并相应地调整信号的时序控制策略，使其比传统的固定时间信号系统更具通用性和适应性。高效且智能的交通信号控制系统在提高交通通行效率的同时，还可以减少交通事故以及可以通过减少拥堵和车辆闲置时间来帮助减少空气污染和碳排放等。

总而言之，使用 DRL 方法研究交通信号控制对于开发更高效和智能的交通控制系统至关重要，从而有助于创建更安全和更可持续的交通系统。

1.2 国内外研究现状

交通信号控制作为智慧交通系统的核心功能之一，在世界各国的交通系统中普遍存在并具有非常重要的地位，因此国内外对其运用、控制方法等已经进行了很多的研究。本文将从四个方面来分析交通信号控制方法研究的国内外现状，分别是交通信号控制的发展历程、传统交通信号控制方法、基于强化学习交通信号控制方法以及基于DRL深度强化学习交通信号控制方法。

从19世纪开始，国外就有了对交通信号控制系统的研究，其利用信号灯灯色变换来调控车辆在道路交叉口的分流，并在同时期英国制造了世界第一台交通信号灯然后将其投入到实际环境中使用。到了1949年，世界上第一台数字式电子计算机在美国问世，使得人类在科学技术进入了一个新的时代。依托计算机的发展，利用计算机技术来对交通信号进行控制的系统，成为世界各国在交通领域的研究热点。1969年，英国学者设计的区域信号控制系统TRANSYT[11]，把交通控制技术推向更高的发展阶段。上世纪80年代开始我国也引进了国外的交通信号控制系统，然后并没有经过多久到了90年代我国开始了自主交通信号控制系统的研究，就比如我国自主研发的第一个实时自适应交通信号控制系统NATS[12]。自此，我国的交通信号控制系统的研发进入了蓬勃发展的阶段并开始陆续地涌现了许多各式交通信号控制系统，其中比较具有代表性的比如海信的HiCon交通信号控制系统[13]以及深圳市的SMOOTH智能交通信号控制系统[14]都有着非常不错的控制性能。此外，交通信号控制系统的研究在国内各高校也成为研究热点，在高校中也涌现了一批具有良好性能的信号控制系统，比如吉林大学的NITCS系统[15]、同济大学的TJATCMS系统[16]、天津大学的TICS系统等[17]，都构建了符合我国国情的新一代智能化交通系统。

传统信号配时在交通信号控制系统中运用的较为广泛，比如固定配时的方法、使用绿波技术在控制范围内实现红绿灯控制结合[18]、遗传算法的方法比如PSO[19]以及利用模糊技术[20,21]来进行交通信号配时调控等。但是随着社会现代化进程的加快，更复杂的交通环境使得这些算法缺乏了对多变环境的适应性以及对交通环境的快速响应，这也导致了这些算法变得不再高效快捷，城市的交通拥堵无法得到更进一步的改善。因此利用强化学习（RL）控制交通信号的研究也随着大数据技术的崛起比以前更加备受关注[22,23]。

自20世纪末开始，通过强化学习方法对交通信号控制进行优化的研究成为许多相关技术人员的研究焦点。智能体感知交通环境并使用强化学习相关算法来调控交通信号策略，通过合理的调控交叉口道路区域内信号灯相位配时，以此来尽可能地提高交通通行效率。其实在这方面的研究很早就有学者进行过，国外学者Thorpe和Anderson在1996年首次提出将强化学习的方法运用到交通控制领域当中[24]，从此开启了交通控

制领域新的研究方向。Abdul Hai 等人提出基于 Q-Learning 的单交叉路口控制算法[25]，该算法定义车辆排队长度作为输入的交通状态，以两相位的交叉口为实验对象，通过改变相序的方法来验证控制方法的有效性，Camponogara 等人[26]则基于此方法对交叉口的相序结构进行了扩展改进。Qu 等人提出另一种基于 SARSA(State-Action Reward-State Action)算法的单交叉路口控制算法[27]，该算法考虑了更真实的交通情况，将车辆数量划分为稀疏的离散值，采用随机控制机制。Aberdeen 等[28]运用 Actor-Critic 算法提出了一种更好的信号控制模型，在该模型中为了对交通信号进行优化，其定义交通状态用两种不同的状态组合来表示。沈文等国内学者使用了 Q 学习算法来解决单个交叉口的信号控制问题。通过实时监测交通流量并对其进行调整，该算法最小化了车辆的平均延误时间，并将 Webster 延误模型作为惩罚函数，以惩罚不良策略。王宜举等[29]以交叉口不同相位的最小通行时间，以相位差为状态转换目标建立 Q 学习模型，实现了信号灯的动态智能控制，他们在基于 Q 学习的单路口交通信号智能控制上进行改进，为了克服了传统模糊推理隶属度函数的主观性，通过遗传算法与 Q 学习算法结合，使得其相比于原有的 Q 学习算法在控制效果有了更好的提升。罗杰等人[30]提出了多种方法来改善交通信号控制算法的性能。首先，他们在研究中提出了交通参数融合函数的方法，以降低 Q 学习中状态空间的复杂度，从而提高算法的学习性能。此外，他们还提出了基于交通特征的状态表征和平均函数估计的函数近似强化学习算法，以解决传统 Q 学习中表达交通状态所需的维度过高的问题。然而，随着交通环境日益复杂，基于传统强化学习改进的交通信号控制算法在环境适应性方面面临越来越大的挑战，这种情况有两个主要原因。首先，交通状态空间和动作空间呈指数型增长，导致维数爆炸问题仍然存在。其次，强化学习算法在环境适应能力方面存在局限性，只能在特定场景下取得一定效果。因此，不仅 Q 学习算法，许多传统强化学习算法在面对复杂的交通环境时也会遇到类似的瓶颈问题。

自从 Mnih 等在 2015 年提出深度强化学习的应用概念起[31]，即在强化学习基础上，将 RL 与深度学习相结合，因此也被称为深度 RL[32]。运用深度学习的框架来建立交通控制模型，目前被认为是控制系统中最先进的学习框架。虽然 RL 可以解决复杂的控制问题，但深度学习有助于从复杂的数据集中近似高度非线性的函数，因此该方法能够有效地解决上述传统算法与强化学习所面对的问题。尽管 DRL 方法高效且适应性强，但仍需解决一些问题。例如 DRL 模型在训练过程中可能会出现不稳定性，即同样的代码、数据、参数等情况下，模型的性能可能会有很大的差异。这种不稳定性使得模型的训练和调试非常具有挑战性以及某些深度学习算法采用 Value-Base 方式导致网络估计值偏大等方面。

目前国内外学者在利用深度强化学习对交通信号控制方面已经做了许多研究[33]，提出了很多新的深度强化学习算法以及改进，比如 Majid Raeis 等[34]人提出了基于延迟

的均衡和基于吞吐量的均衡的 DRL 交通信号控制模型。Zheng G 等人提出了基于交通信号控制中相位竞争原理的 DRL 模型，其原理是当两个交通信号发生冲突时，应优先考虑交通流量较大（即需求较高）的信号[35]。其他还有运用比较广泛的比如 DQN[36,37]、DPPO[38]、TRPO[39,40]等算法在交通信号控制研究领域都运用的非常广泛，产生了许多的变种算法，本文也是根据经典 Double DQN 算法[41]进行改进的一种 DRL 变种算法，以便其更好地适应复杂交通环境下的信号控制。

1.3 论文研究内容

面对目前现在各大城市不断建设的城市交通道路，其所构成的庞大交通网络也越来越复杂，大量的道路交叉使得交叉口在城市道路广泛分布，信号灯作为交叉路口的重要调控设备，其控制方法的好坏对交叉口通行效率有着直接的影响。传统信号控制方法在当下复杂的交通环境下已经越来越力不从心。因此，人们正在开发和应用基于深度强化学习的交通信号控制方法。这种方法通过智能体的训练，可以最大限度地缩短车辆通过交叉口所需的时间，并降低车辆的平均延迟时间，从而有效地提高整个交通路网的通行效率。但由于交通环境复杂且随时间在不断地变化，目前的交通信号控制都存在智能体感知的交通环境做出的决策调控存在一定的延迟，无法很好地匹配变化的交通环境，因此提出一种新的模型，该模型中的智能体会感知当前交通状态并利用循环神经网络预测未来交通状态，之后再基于未来环境的交通信号控制策略，提出有一定预见性且符合当下的交通信号调控策略。为了实现提出的模型，本文主要完成以下几个工作内容：

(1) 分析了 DRL 交通信号控制的研究背景、意义以及国内外该领域的研究现状，并分析其他类型交通信号方法的优点和缺点。

(2) 阐述了交通信号控制的相关交通工程学相关理论知识包括交通优化性能评估指标、信号相位等。还介绍了强化学习、深度学习等算法的原理、基本要素以及常用的算法比如 Q-Learning 强化学习算法、梯度策略算法、RNN 网络，并解释了深度强化学习原理以及结构。

(3) 通过分析交通信号控制所遇到的挑战，认识到传统的交通信号控制存在不足之处。因此，提出了一种基于深度强化学习的交通信号控制模型，分析其模型结构并详细介绍了模型结构的参数定义和训练流程。

(4) 针对单个交叉口信号控制问题，在交通环境复杂多变的情况下，传统控制方法已经无法胜任，因此设计了基于 LSTM 网络预测的 Double-DQN 深度强化学习交叉口信号控制算法模型。首先在交叉口设计成智能体结构基础上，分析传统的 Q-Learning 强化学习以及在交通信号控制中，由于交通状态过于复杂，这样会使得算法 Q 值矩阵存在维度爆炸问题以及普通的 DQN 深度强化学习会造成算法估计值过大的问题进而通过

采用基于预测状态以及动态动作选择策略改进 Double DQN 深度强化学习算法作为交叉口智能体的控制方法。

1.4 论文章节安排

第一章为绪论，在该章节中首先介绍了交通信号控制相关的研究背景和意义，然后总结了国内外该领域的研究现状，探讨了多种交通信号控制方法的优缺点。最后，具体介绍了本文研究的主要内容和各章节的安排。

第二章是关于交通信号控制理论、深度强化学习基本理论背景以及循环神经网络相关理论背景的描述。首先，阐述了交通信号灯设置的依据和相应控制策略的主要参数，包括信号相位及其相位周期、相位差、性能评价指标等。其次，阐述了交通信号相位控制中常用的强化学习、深度学习和深度强化学习策略，并探讨了在该领域应用这些策略所需的背景知识。

第三章是关于基于深度强化学习交通信号控制策略模型构建。首先分析了需要解决的交通拥堵问题，然后说明了如何运用深度强化学习算法来构建交通信号控制模型，包括模型的基本结构和模型关键参数的定义，最后再介绍了用于控制交通灯信号相位变换的交通智能体的训练流程。

第四章是关于本文提出的基于预测交通流与动态策略改进的 Double Deep Q Network 交通信号控制策略模型方法的构建。阐述了如何对上一章的基本深度强化学习交通信号控制模型存在一些问题进行优化改进，其优化的方向主要包括 Double DQN 动作选择策略的改进以及利用 LSTM 网络改进 Target 目标网络值计算方式的优化等，以解决传统深度强化学习模型中存在的估值过高以及模型训练参数不准等问题。

第五章是仿真实验和结果分析。本章首先详细介绍了仿真模型中所需要用到的各种数据集，然后解释了相关仿真模拟器和模拟实验参数的设置。然后再进行仿真实验，最后对实验结果进行详细的分析与阐述。

第六章为总结与展望。在本章中，先对算法和实验结果进行了总结，并指出了在算法和实验过程中存在的问题和可以改进的方面。最后，还展望了未来的研究方向。

2 交通信号控制概念及 DRL 相关理论

2.1 引言

目前世界各国的大多数城市都面临着一个共同的问题，这个问题就是交通拥堵。随着经济不断的发展，城市中汽车的数量也在急剧上升，然而大多城市的道路建造、升级速度无法满足当前的需求，因此利用其他方式解决城市交通拥堵问题成为当下研究的热点。交叉口作为道路车辆控制的重要组成，其对交通的运行有着至关重要的作用，然后交叉路口控制车辆运行的核心又在于路口的信号灯相位信号的合理控制，因此优化交叉路口的信号控制策略已成为解决交通拥堵和提高交通运行效率的重要手段。许多国家的研究人员通过对城市交通现状的详细分析，并结合当下发展迅速的人工智能技术，找到了可以通过训练 Agent 模型来进行交通信号的自适应控制的这种方法，该方法不仅可以很好地满足人们日益增长的出行需求，还非常契合当下数字城市的发展。在当下人工智能技术中，DRL 学习算法被广泛地使用在智能体的训练中，该算法模型对复杂外界环境具有很强的适用性，因此十分适合用于交叉口的交通信号控制。

2.2 交通信号控制概念

2.2.1 交通信号控制理论

交通信号控制理论是现代城市交通管理的一个重要理论基础。它涉及设计和实施高效的交通信号系统，以调节交通流量并最大程度地减少拥堵。交通信号控制理论的目标是优化可用道路容量的使用，同时最大限度地减少道路使用者的延误和出行交通消耗时间。当前的交通信号控制理论包含多种交通信号控制系统，比如固定时间、驱动和自适应信号控制系统等，这些交通信控系统使用不同的方法根据交通量和其他因素调整交通信号灯的配时[42]。要实施交通信号控制系统，计算机技术是必不可少的工具。信息化交通信号控制系统的使用允许实时监控交通流量和动态调整信号时间以优化交通流量。在现代交通信号控制理论中采用深度强化学习和长短期记忆网络等先进的机器学习技术来分析交通模式并做出最佳信号配时决策的交通信控系统，在提高交通流量效率和减少城市地区的拥堵方面显示出很好的前景。

总的来说，交通信号控制理论在管理城市交通流量方面起着至关重要的作用。通过使用现代信息化技术比如机器学习、深度学习等，可以设计和实施高效的交通信号系统，优化道路通行能力并最大限度地减少拥堵。

2.2.2 交通参数

(1) 交通流

在交通信号控制中，交通流是交通信号控制中的一个重要参数。交通流可以分为车辆交通流和行人交通流两类，本文主要研究的为车辆交通流。车辆交通流是指在一定时间内通过一条路段或交通设施的车辆数量。车辆交通流通常用单位时间内通过该路段的车辆数（即车流量）来衡量，单位为车辆/小时。交通流对于城市交通规划和交通管理具有重要意义，了解和掌握交通流的特征和规律可以帮助交通规划者和管理者更好地设计和调整交通网络，提高道路运行效率和安全性。因此，对交通流的研究是交通信号控制领域研究的重要课题之一。

(2) 车道

在交通信号控制中，除了路段参数外，车道也是一个非常重要的参数。在控制交通流量时，需要考虑各车道的各种参数指标因素。常用的车道参数包括车道长宽、车道等级、车道类型、车道交通量等。在交通信号控制中，需要考虑多车道的参数，并根据情况进行相应的调整和优化不同的车道。通过交通信号控制参数的合理设置，可以最大程度地提高通行效率，减少交通拥堵和事故的发生。

(3) 信号相位

交通信号控制中的信号相位是指交通信号灯按一定时间间隔依次显示的一组灯光组合，通常包括绿灯、黄灯和红灯。在交通信号控制中，通过调整信号相位的持续时间和灯光显示的顺序来控制车辆和行人的通行。常见的信号相位模式有绿波模式、绿闪模式、黄闪模式等，每种模式都有不同的参数设置。在设计交通信号控制方案时，需要根据路口的交通流量、车道结构等因素，确定各信号相位的持续时间和顺序。同时，还要考虑车辆的通行能力和行人的安全通行。因此，信号相位的设计是交通信号控制中一个非常重要的参数。通过信号相位的优化设置，可以有效提高交通流量和通行效率。常见的信号相位有二相位、四相位、八相位等，如图 2.1 是应用最广泛的四相位。

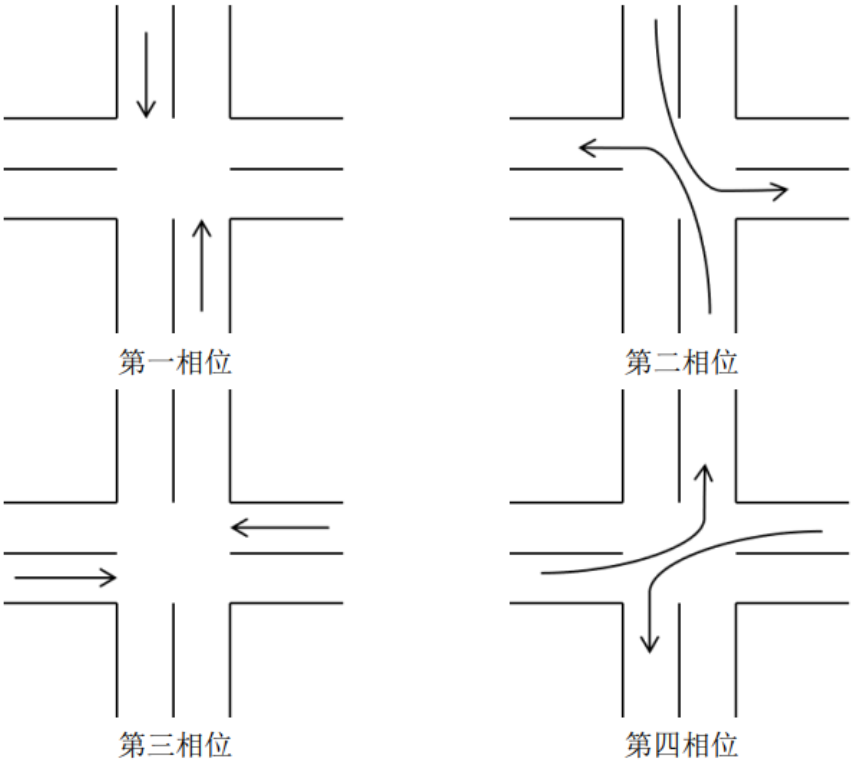

图 2.1 四种相位图

Fig. 2.1 Four-phase diagrams

(4) 信号周期

信号周期是信号灯从一个相位转到下一相位所需的时间长度。它是交通信号控制中一个非常重要的参数，决定了每个相位阶段的持续时间和不同阶段之间的过渡时间。在信号周期内，各阶段的时间分配必须合理，以充分利用道路的通行能力，减少交通拥堵。信号周期的长度通常由交通流量、交叉路口形状和其他交通状况等因素决定。例如，高峰时段的信号周期长度通常比非高峰时段长，因为道路上的车辆较多，需要更多时间让车辆通过。在设计信号周期时，通常需要通过交通流量预测和仿真来确定最佳信号周期长度。同时信号周期也需要根据实际情况动态调整。例如，在交通拥堵的情况下，可以通过减少绿灯时间和增加红灯时间来减少交通拥堵。因此，精确有效的信号周期控制对于优化交通流、减少交通拥堵至关重要。

(5) 相位差

相位差是指不同交通信号灯相互配合关系中两个信号灯之间的时间差。在交通信号控制中，通过控制不同路口信号灯的相位差来调节交通流量。例如，在两个相邻的路口，如果一个路口的绿灯时间与另一个路口的红灯时间相同，则两个路口之间存在零相位差的协调关系。在这种情况下，两个路口的车流可能会相互干扰，造成交通拥堵。因此，通过设置不同的相位差，可以实现不同路口之间交通流的协调，提高整体通行效率。

2.2.3 交通性能评价指标

(1) 车辆平均延误时间

车辆平均延误是用来评价交通信号控制效果的一个重要指标，该指标可用来评价交叉口的交通拥堵程度和通行效率，其定义为通过交叉口的车辆在交通信号控制下与预期通过时间之差的平均值。对于一个交叉路口，其在一个信号相位周期中的平均车辆延误T计算，可以通过一个信号相位周期内的总延误D来计算，具体计算方式如公式 2.1 与公式 2.2 所示。

$$D = \sum_{i=1}^{n} \bar{d}_i p_i \tag{2.1}$$

$$T = \frac{\sum_i \bar{d}_i p_i}{\sum_i p_i} \tag{2.2}$$

其中$\bar{d}_i$表示第i相位每辆车的平均延误时间；p_i表示第i相位的平均交通量。

(2) 车辆通过能力

车辆通行能力是指单位时间内可以通过一个路口的车辆数量，通常用车流量来表示，单位是辆/小时。它是交叉口的一项重要性能指标，反映了交叉口的通行能力和交通拥堵程度。车辆通过能力受交叉口信号控制方式、道路几何条件、交通流构成等诸多因素的影响。在交通规划设计中，需要根据交叉口交通量、车辆通行时间和等待时间等参数，结合实际情况对车辆通行能力进行评价，优化交通信号控制。

(3) 车辆排队长度

车辆排队长度是指车辆在路口或路段等待通过的队列长度，通常用车辆数量或车道长度等单位表示。车辆排队长度的测量通常依赖于交通流量数据采集设备，如交通监控摄像头、车辆传感器等，通过数据分析可以得到排队长度的估计值。排队长度是交通流状态的重要指标之一，可以反映交通流的拥堵情况，对交通信号控制的优化具有重要的参考价值。

(4) 车辆通过时间

车辆通行时间是指车辆从一点行驶到另一点所需的时间，通常用于交通性能评价。对于交叉口，车辆通过时间可以分为两部分，分别是排队延误时间和行驶时间。排队延误时间是指车辆在路口前等待的时间，通常由交通拥堵、信号控制等因素造成。行驶时间是指车辆通过路口的实际行驶时间，通常由以下因素决定如车速、交通量和道路状况。

计算车辆通行时间一般有两种方法。一种是直接观察法，即通过观察车辆通过路口的实际时间来计算车辆通过时间；另一种是排队论方法，即通过排队论计算车辆通过时间。该方法将路口视为一个排队系统，综合考虑车流到达率、服务率、排队长度

等因素，通过排队模型计算车辆通过时间。

2.3 DRL 相关理论

2.3.1 强化学习

(1) 原理概述

强化学习[43]又称作再励学习或者增强式学习，其主要思想来源于条件反射理论和动物学习理论，是一种受生物学启发的仿生算法，是机器学习领域一个非常重要的分支，它主要用于人工智能领域的决策问题，其定义是 Agent 为了适应环境而采取的主动对环境做出试探性的学习。该学习方式是利用智能体对感知环境采取不同动作，以获得环境状态的适应度评价值（即奖励或者惩罚）。如果 Agent 的行为导致环境对其给予正向奖励，那么该智能体在之后采取这种行动策略的倾向就会增强。如果某个动作策略导致环境的负面反馈即惩罚，则 Agent 会将该行为策略趋势减弱，智能体就是通过这种方式不断调整自身的动作选择策略，然后做出最优决策，从而在给定环境中获得最大回报。

(2) 强化学习模型

在强化学习中，学习被视为一种试错的过程，其学习系统的组成主要包括两大部分，即智能体（Agent）和外部环境。在强化学习中，智能体选择一个动作a_t（Action）作用于外部环境，环境接受该动作的影响后发生改变，变成了下一个状态s_{t+1}，同时会产生一个强化信号（即奖励或者惩罚）r反馈给与环境交互的智能体，智能体又会根据强化信号以及现在的环境状态选择下一个动作，其选择动作的策略是使得智能体受到的正奖励概率增大。选择的动作既会影响即时强化值，又会影响环境的下一个时刻的状态和最终强化值。强化学习标准模型如图 2.2 所示，其中r_t和r_{t+1}分别表示两个不同时刻的即时奖励。

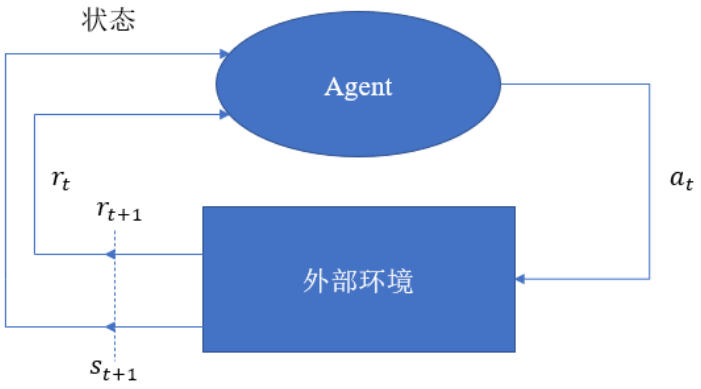

图 2.2 强化学习标准模型

Fig. 2.2 Standard model of reinforcement learning

2.3.2 强化学习基本要素

强化学习系统由 Agent、环境、策略、奖励函数、值函数和环境模型六个基本要素构成。除了智能体和环境这两个主要组成部分外，策略、奖励函数、值函数和环境模型也是不可或缺的，这些基本要素在强化学习中互相作用，智能体会根据当前状态、策略和值函数等信息来做出决策，通过不断地尝试和学习，最终得到最优的策略来最大化长期回报。

强化学习其模型原理主要是基于马尔可夫决策过程[44]（Markov Decision Process, MDP），可以用一个五元组来表示该过程中 Agent 与环境的互动，即(S, A, R, P, γ)。其中，S表示状态集，A表示动作集，R表示回报函数，P表示状态转移概率，γ 表示折扣因子。

• S：所有可能状态空间的集合， s是集合中一种状态($s \in S$)；

• A：所有可能动作空间的集合， a是一个动作($a \in A$)；

• R：在每个状态和动作下，奖励函数定义了智能体会获得多少奖励

• P：它描述了当智能体在某个状态下采取某个动作后，下一个状态是什么的概率分布，其可以表示为如公式 2.3 所示。

$$P_{ss'}^{a} = \Pr\{s_{t+1} = s' \mid s_t = s, a_t = a\} \tag{2.3}$$

• γ：它表示未来奖励的折现率，因此它决定了智能体对于未来奖励的重视程度。

强化学习中的策略是指 Agent 在特定状态下所采取的行为的决策方法。从长远来看，一个好的策略应该使代理人的期望值最大化。在强化学习中，策略可以表示为映射函数$\pi(s,a)$，其s表示当前状态， a表示代理采取的动作。策略可以是确定性的或随机的。确定性策略可以表示为 $\pi(s)=a$，即在给定状态下选择确定性动作。随机策略可以表示为$\pi(a|s)$，即在给定状态下选择具有一定概率分布的行为。强化学习的目标是找到最优策略π^*，让 Agent 在长期的累积奖励中获得最大的期望值。这个过程通常由价值函数辅助，分别是状态价值函数和动作价值函数，两者表示是通过策略π下的状态s和采取行动a的长期累积奖励。两者的具体定义如下：

• 状态价值函数$V^{\pi}(s)$：它表示在当前状态下，智能体能够获得的长期累积奖励的期望值，其计算方式如下式 2.4 所示。

$$V(s_t) = E\left(\sum_{i=0}^{\infty} \gamma_{i+i} r_{t+1}\right) \tag{2.4}$$

上式子展开还可以推导出另一种形式，如公式 2.5 所示。

$$\begin{aligned} V^{\pi}(s) &= E_{\pi}\{r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots \mid s_t = s\} \\ &= E_{\pi}\{r_{t+1} + \gamma V^{\pi}(s_{t+1}) \mid s_t = s\} \end{aligned} \tag{2.5}$$

• 动作价值函数$Q_{\pi}(s)$：它表示在当前状态下采取某个动作后，智能体能够获得的长期累积奖励的期望值，其可以表示为如下式 2.6 所示。

$$Q^{\pi}(s,a) = E[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \cdots \mid s_t = s, a_t = a, \pi] = E\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k} \mid s_t = s, a_t = a, \pi\right] \tag{2.6}$$

其中 γ 折扣因子取值范围在[0,1)，它在累积奖励的计算中起着重要的作用，可以看作是一个“预见度”的调节器。折扣因子的作用是调整未来的奖励，即衰减后续时刻的奖励。在强化学习中，未来的回报往往是不确定的。因此，折扣因子越大，Agent 对未来奖励的权重就越大，反之亦然。同时折扣因子也可以帮助 Agent 克服长期奖励和短期奖励之间的冲突，从而更好地平衡当前动作的即时奖励和未来的累积奖励。

通过递归可得最优动作策略π^*。如果 Agent 已知下一个状态最大 Q 值，则策略π^*会选择下一个状态最大 Q 值所执行的动作。因此，最优$Q(s,a)$可利用下一个状态的最大 Q 值来计算即Q^{π^*}，根据贝尔曼最优方程的公式来计算如下式 2.7 所示。

$$Q^{\pi^*}(s,a) = E_{s'}\left[r_t + \gamma \max_a Q^{\pi^*}(s',a') \mid s,a\right] \tag{2.7}$$

累积奖励是指将近期获得的即时奖励和预估的最优未来奖励相加得到的总回报。因此只要能够求得预估部分的回报，就可以计算出从现在开始到状态结束的总回报。这个方程可以在状态数有限的情况下通过动态规划方法来求解，因为只有在有限状态空间中其计算复杂度才能在可控范围。

强化学习的动作选择策略[45]上目前用的最广泛策略分为两种：贪婪策略(greedy method)和 ε-greedy 策略，本文的后续的优化方向之一也是从调整算法动作选择上进行的。

(1) 贪婪策略

在强化学习中，贪婪策略是指智能体在当前状态下选择具有最大 Q 值的动作。具体而言，Q 值是指执行某个动作后可以获得的累积奖励值，贪婪策略就是在当前状态下选择使 Q 值最大化的动作。贪婪策略的优点是简单易实现，不需要太多的计算资源。但是，它也有一个明显的缺点，即可能会陷入局部最优解而无法找到全局最优解。因此，在某些情况下，为了避免陷入局部最优解，智能体需要使用其他策略进行辅助，在实际应用中，通常将贪婪策略与其他策略结合使用，以实现更好的性能。

(2) ε-greedy 策略

ε-greedy 策略是强化学习中一种常用的策略，其基本思想是在选择动作时，以一定的概率随机选择非最优动作（探索），以另一定的概率选择当前最优动作（利用）。这个概率由参数 ε 控制，ε 通常取一个小的值，比如 0.1 或 0.2，使得在大部分情况下选择最优动作$\max_a Q(s,a)$，但仍有一定概率选择非最优动作，从而保证系统的探索能力。随着时间的推移，探索的概率会逐渐降低，利用的概率会逐渐增加，直到最终只选择

当前最优动作。这种策略的优点在于可以避免陷入局部最优解，同时也能够在训练过程中探索到新的策略。其公式可以用分段函数表示为如下式 2.8 所示。

$$p = \begin{cases} 1-\varepsilon, & \max\limits_{a} Q(s,a) \\ \varepsilon, & \text{other} \end{cases} \tag{2.8}$$

2.3.3 深度学习

深度学习[46]（DL）作为一种机器学习方法，它基于人工神经网络的设计和训练，能够对大规模和复杂的数据进行分析和学习。自 2006 年被正式提出以来，深度学习便与机器学习相结合，更接近于人工智能（AI，Artificial Intelligence）的目标。它通过对样本数据的不断训练学习，掌握其内在规律和表示层次。

深度学习的基础是神经网络，由大量神经元组成。每个神经元接收来自前一层神经元的输入，经过权重计算和激活函数处理后，输出到下一层神经元。通过堆叠多个神经网络层，可以构建深度神经网络，实现更加复杂的任务。

深度学习的核心思想是通过多层神经网络进行特征学习和表示。在深度学习中，通常采用反向传播算法来训练模型。该算法可以计算每个神经元的梯度，从而更新模型参数，使模型更好地拟合训练数据。

深度学习模型的训练需要大量数据和计算资源，通常使用 GPU 进行加速。常见的深度学习模型包括全连接神经网络、卷积神经网络和循环神经网络等，在自然语言处理、推荐系统、图像分类识别等领域都取得了非常好的效果。

除了传统的监督学习，深度学习还可以应用于无监督学习、强化学习等任务。深度无监督学习方法能够学习到数据中的隐含结构，发现数据中的模式和规律。而深度强化学习则可以通过学习如何最大化奖励来实现复杂的决策任务。

2.3.4 深度强化学习

深度强化学习（Deep Reinforcement Learning，DRL）是将深度学习与强化学习结合的一个重要分支。它可以通过自主学习，在没有标签的数据中学习到策略，使智能体（Agent）可以在环境中自主做出好的决策。DRL 是强化学习和深度学习两者的一个结合，既结合了深度学习强大的感知能力，又综合了强化学习优秀的决策能力，因此 DRL 能高效地解决很多深度学习以及强化学习无法解决的问题。

深度强化学习其学习过程可以描述为如下：首先，需要定义状态空间和动作空间。状态空间是指学习模型在模拟环境下所有可能的状态集合，动作空间是指学习模型可以采取的所有动作集合。其次，使用一定的动作选择策略对环境中的训练数据进行采样，这个策略可以是随机抽样，贪婪策略，蒙特卡洛探索等。之后，使用深度神经网络作为强化学习模型的近似函数来进行学习，比如 DQN 算法是利用深度学习来对 Q 值

进行近似。最后，计算损失函数并使用反向传播算法进行优化，以更新深度神经网络中的参数，然后更新目标网络参数以解决深度神经网络中的强相关问题。深度强化学习框架图如图 2.3 所示。

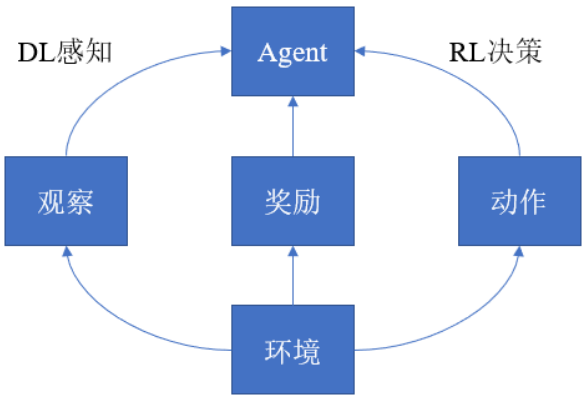

图 2.3 深度强化学习框架图

Fig. 2.3 Framework diagram of deep reinforcement learning

2.3.5 常用学习方法

(1) Policy Base（代表算法有梯度策略）

Policy-Based 方法是直接从状态到动作的映射中学习一个策略。策略本质上是一个函数，输入是当前的状态，输出是应该采取的动作。Policy-Based 方法通过学习最优策略来解决强化学习问题。这个方法的优点是可以处理连续动作空间，但由于直接优化策略本身比较困难，所以常常需要使用策略梯度方法（Policy Gradient）来进行优化。

梯度策略[47]（Gradient Policy）是一种强化学习算法，用于学习策略的参数化表示。梯度策略的基本思想是使用梯度下降法来最大化策略的期望回报。具体来说，策略表示为一个参数化函数，如神经网络，其输入是状态，输出是动作概率分布。梯度策略算法通过计算回报的梯度来更新策略参数，从而最大化期望回报。

(2) Value Base（代表算法有 Q-learning）

Value-Based 方法是通过学习一个价值函数来实现最优策略的计算。价值函数估计的是在当前状态下，采取某种行动的长期累积回报。Value-Based 方法的优点在于可以处理离散和连续的状态和动作空间，同时也能处理部分可观察马尔可夫决策过程（MDP）问题。而且通过使用 Q-Learning 和 Deep Q-Network 等方法，它可以学习到最优的价值函数。缺点是价值函数不一定能直接转化为策略，有时还需要使用 Policy-Based 方法或者 Actor-Critic 方法来进行策略的优化。

Q-learning[48]是一种基于值函数的强化学习算法，用于解决强化学习中的马尔可夫决策过程（MDP）问题。在 Q-learning 中，定义一个 Q 函数，用来评估在给定状态下采取某个动作的价值。Q 函数的值是一个动作价值函数，表示在当前状态下采取某个

动作所能得到的预期回报。Q 函数可以通过迭代的方式进行更新，使得它的值逐渐逼近真实的价值函数。

(3) 深度学习（RNN）

RNN（Recurrent Neural Network）代表递归神经网络[49]，它是一种能够处理序列数据的神经网络模型，与独立处理每个输入的传统神经网络不同，RNN 其关键在于它们使用了循环结构来处理前一个时间步的隐藏状态，可以让它们记住以前的输入并使用该信息来预测未来的输入。RNN 适用于时间序列预测和自然语言处理。

RNN 的基本结构由一个在循环中连接到自身的循环层组成。这个循环允许网络保持一个状态，该状态反映它已经看到的先前输入。当前输入与该状态结合产生输出，然后作为下一个输入反馈回网络，对序列中的每个输入重复该过程， RNN 模型结构如图 2.4 所示。

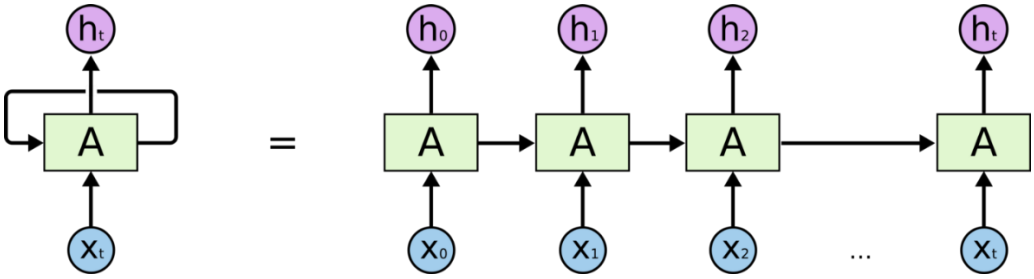

图 2.4 RNN 模型结构图

Fig. 2.4 RNN model structure diagram

其中X_t为输入，A 为隐藏层向量，h_t为网络的输出层向量，在每个时间步，RNN 执行以下操作：

计算当前时间步的隐藏状态，其计算方式如下式 2.9 所示。

$$A_t = f(W_a A_{t-1} + W_x X_t + b_A) \tag{2.9}$$

其中W_a是隐藏状态自连接的权重矩阵，W_x是输入向量X_t到隐藏状态的权重矩阵，b_A是偏置向量。f是激活函数，通常是非线性函数，如 Tanh 或 Relu。

使用隐藏状态进行预测，其方式如下式 2.10 所示。

$$h_t = \mathrm{g}(W_h A_t + b_h) \tag{2.10}$$

其中，W_h是隐藏状态到输出向量h_t的权重矩阵，b_h是偏置向量。g 是输出激活函数，通常是 Softmax 或 Sigmoid 等。

更新隐藏状态：A_t将在下一个时间步变为A_{t-1}。

RNN 的关键在于它对序列数据的处理能力，即会将上一个时间步的隐藏状态传递给当前时间步进行计算，传统 RNN 的一个局限性是梯度消失问题，当损失函数相对于网络参数的梯度变得非常小时，就会出现这种问题。这使得网络很难学习长期依赖关系，这对于许多基于序列的任务来说可能很重要。为了克服这个限制，已经开发了 RNN 的

变体，例如 LSTM（长短期记忆）和 GRU（门控循环单元）

综上所述，Policy-Based 和 Value-Based 是深度强化学习中两种不同的方法，它们各自有优点和缺点，根据问题的不同，可以选择不同的方法来解决强化学习问题。RNN 是 DL 中基于序列处理的神经网络模型，一般用于自然语言以及时间序列相关的数据处理。本论文选择的是基于值函数的 Double DQN 方法并结合 RNN 的改进网络构建的交叉口信号灯调控模型，以学习路口交通信号调控策略。

2.4 本章小结

本章详细地介绍了与交通信号控制相关的原理概念以及用于评价交通信号控制性能对交通效率产生影响相关的评价指标，并说明了它们对交通信号控制研究的重要性，特别是道路网络中所有通过车辆的平均延误指数和可以通过道路网络的最大车辆数量。本章还介绍了智能交通信号控制研究的相关技术主要包括强化学习、深度学习以及结合了两者优点的深度强化学习算法。此外，在本章中还介绍了强化学习中智能体动作选择的方法策略，分析了它们在交通信号控制研究中的优缺点。最后对两种常用于与深度学习结合的强化学习算法以及本文后续需要使用的深度学习算法进行了分析，阐述了算法的主要原理、应用范围以及其各自优势。

3 深度强化学习交通模型

3.1 引言

由前两章内容可以分析了解到传统的机器学习在当前很多复杂环境下，两者都缺乏很好的综合性能。在交通信号控制优化领域中，仅选择使用感知能力强的深度学习算法来进行相位控制则会导致模型整体策略选择能力不足，而结合了强化学习算法的控制方法，虽然在信号控制上拥有较好的策略选择能力，但由于其感知能力的不足以及面对交通复杂场景该算法梯度爆炸的问题，导致对复杂环境不能很好地判断以及适应，会很大影响交通信号控制的有效性。因此，我们本章中先分析所需要解决的交通信号控制问题，再结合深度强化学习中经典的 DQN 算法，提出基于深度强化学习方法的交通信号控制模型，并对模型的参数定义以及训练流程进行相关说明，阐述模型如何获取最优信号控制方案的，来实现对交叉路口的信号相位调控，以提高整个路网的车辆通行效率。

3.2 基本模型结构

结合深度强化学习构建基于交通信号控制的模型，该模型与使用传统信号控制方法的信号调控策略相比，它能够更高效调控交通信号，使得路网中的交叉路口可通行车辆的数量尽可能地增加，同时降低车辆在交叉路口的平均延迟。本文所构建的交通信号控制策略模型主体包括四个部分，分别是评估网络、目标网络、交通环境和经验回放池。在信号控制模型中，几个部分共同协调完成整个深度强化学习交通信号控制模型的训练，并成功地训练出能够高效率调控交通信号的相位变换策略的控制模型，其系统模型的结构如图 3.1 所示。

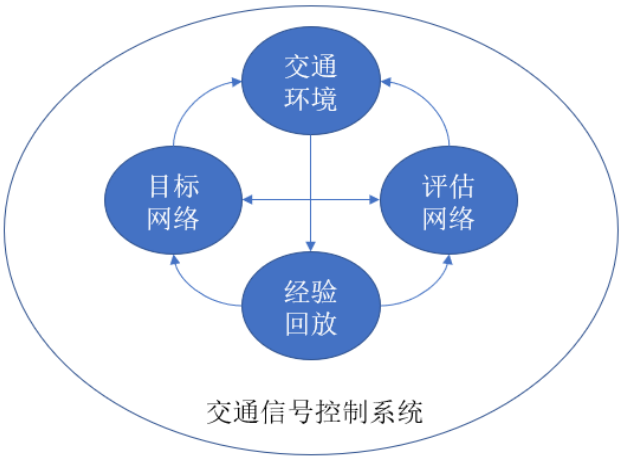

图 3.1 交通信控系统结构图

Fig. 3.1 Structure diagram of traffic information control system

在上图 3.1 中，根据当前交通环境给出最优信号调控策略的是评估网络，它通过从信号灯所处的交通环境中获取到各种反映交通状态的信息，并计算出该状态所能执行的动作空间集合中每个动作的估计 Q 值，并以此为策略，给出最优的信号灯相位切换动作。其获取到的交通状态信息包括交叉路口位置坐标、车辆数量、道路位置坐标、车道平均车流量等。在算法模型中其表现形式就是把这些交通状态信息作为评估网络的输入并进行相应的计算而得到相位切换动作的策略，之后将策略信息发送给信号灯，信号灯会按照相应调控策略切换到指示的相位。在这个过程中，评估网络的主要作用就是获取交通环境信息并根据该信息为路口交通信号灯调控提供所需切换的相位策略。在模型训练中，目标网络主要用于协助评估网络完成其参数的训练，它的作用类似于神经网络训练中的标签值。在估计当前状态的 Q 值时，目标网络所输入的交通状态信息与评估网络是同样的样本数据。后续的系统训练中用于更新评估网络参数的损失函数采用的就是以目标网络所计算的 Q 值为真实值与评估网络的值所构成的误差函数。交通环境指的是交叉路口车道上获取到的各种交通感知信息，本文所建立的模型是基于最广泛的四路交叉路口，因此交通信息指的就是每条车道上的各种参数。经验回放是深度强化学习算法常用的优化训练的方法，用于改善深度强化学习算法训练的稳定性，在传统的强化学习算法中，智能体与环境的交互是连续的、实时的，每次决策都会对当前的状态和动作产生反馈。这种交互方式有助于提高学习速度和精度，但同时也会导致样本的高度相关性，容易造成训练不稳定和过拟合等问题。经验回放[50]通过将智能体与环境交互得到的经验存储在经验池（Experience Replay Memory）中，因此经验池储存的是大量可以提供给模型进行训练的有效样本数据，可以随机抽取部分经验进行训练。这种方式可以打破样本之间的相关性，减小样本的方差，使得深度神经网络的训练更加稳定和高效。

该深度强化学习模型训练的执行过程如图 3.2 所示。

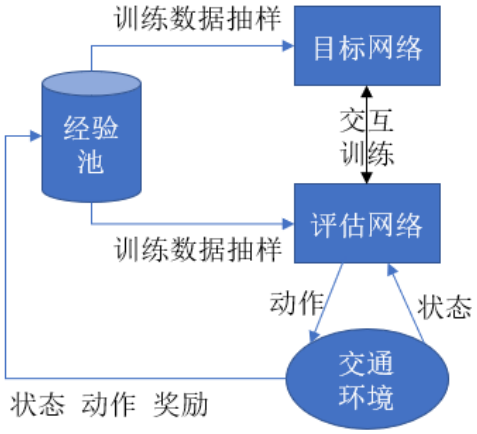

图 3.2 模型训练图

Fig. 3.2 Model training diagram

图中，评估网络模型把交通环境作为交互对象，从经验池获取抽样训练数据，根据数据的交通状态并为其给出相应信号调控策略动作，而环境也会将执行动作后的状态反馈给模型。与此同时，交通环境由于执行动作策略，环境的改变会不断产生新的训练样本并存入经验池，而这些样本同时被评估网络模型和目标网络模型抽样用来进行训练。

3.3 模型中参数定义

在本文中，是一个以城市道路网络中分布最为广泛的四向交叉路口来构建系统模型的。本文的研究目的是为了研究能够协调控制交通信号，提高十字路口中最大可顺畅通行的车辆数量，并使得交通延迟在可以接受的范围内。基于上文所提到的深度强化学习交通策略模型，本文建立了该研究所需的交通智能体模型，其模型图如图 3.3 所示。

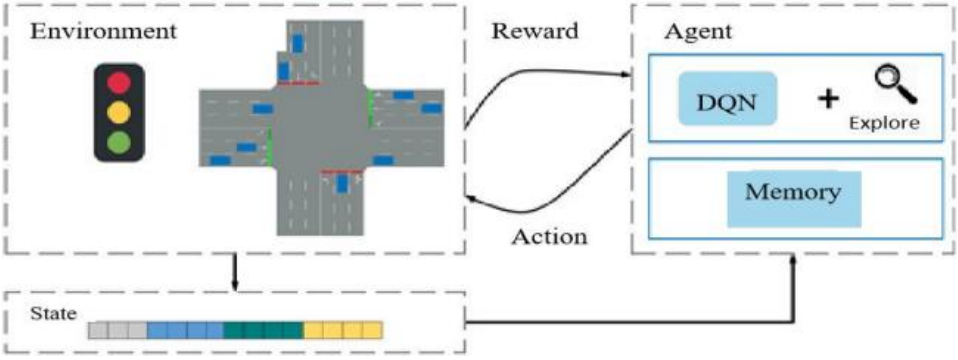

图 3.3 交通智能体模型图

Fig. 3.3 Traffic agent model diagram

从图中可以看到，整个智能交通信号控制模型中所需要的参数主要是：状态（State）、奖励（Reward）、行为动作（Action），这构成交通模型三个非常重要的要素。

(1) 状态

本文模型每间隔一个信号相位周期会对路网中的交叉口交通状态进行一次感知来获取状态信息，其获取的交通状态信息包括车辆级别信息（例如位置、速度、方向等）和车道级别信息（例如每条车道的平均速度、车道平均车流量等）。本文构建的四向车道分为二十四个车道，每个方向十二条，总共八个相位，因此构建的交通状态是一个多维的向量，这些向量记录着每条车道观察到的交通信息。状态的定义对网络模型进行动作的选取有很大的影响，而选择不同的动作也将直接影响交通信号灯的控制效果。

(2) 奖励

在强化学习中，奖励是一种机制，用于评估智能体在执行某个动作后所获得的反馈。智能体在某一时刻会接收到当前的环境状态信息，并选择执行一个动作。执行该动作后，智能体将接收到一个奖励信号，它表示该动作的优劣程度。奖励是强化学习

算法中非常重要的一部分，它为智能体提供了一种有效的反馈机制，使其能够逐步调整策略，从而更好地适应环境。因此，合理有效地定义奖励是十分重要的，这能够帮助交通智能体获取最佳行动策略并执行。

在本文所设计的交通智能体信号控制模型，其最主要的目的就是通过提升交叉路口的车辆通行效率来缓解路网中道路的交通拥堵状况，在本文第二章也介绍了一些交通性能评价指标，其中对效率进行衡量的重要指标主要包括交叉路口的车辆通行能力以及平均车辆延迟，而车辆通行能力常以最大通行车辆数来表示。因此，这里采取一种交通学的最大压力法来定义模型的奖励，其定义的奖励就是在切换相位后，观察到的交叉路口所有出口方向的车辆数减去进口方向的车辆数。

(3) 动作空间（行为动作）

在本文中，交通信号灯需要对自身所处交叉口的交通状态进行感知，选择合适的相位变换策略来更有效地调控交叉路口的车辆通行。交通智能体在感知到当前的交通环境状态后，根据训练好的模型给出动作空间中各动作的评价值（Q 值）来选择最佳的相位调整动作，并通过改变信号相位以实现对路口交通流的调度。本文的四向十字路口总共有八个不同的非冲突相位，因此动作空间是由这八种相位组成的八维的多维向量。

3.4 训练流程

本文结合深度强化学习的方式，搭建交通信号控制模型，因此模型的训练方式也是按照标准的 DRL 方式来训练的。在本文交通信号控制模型中，我们采用了 Double DQN 的深度强化学习算法，该算法的训练流程如下：首先初始化神经网络及其他参数，即 Q 评估网络和目标网络，并设置其他超参数，例如经验池大小、学习率、折扣率等。然后选择相位动作，通过输入状态，从 Q 评估网络中选择当前状态下的最优动作。之后执行动作并观察结果，观察执行后的交通环境状态及获得的奖励，并将当前的状态、动作、奖励、下一状态等信息构成四元组形式存储到经验池 Memory 中，这个过程是生成训练样本数据。然后从经验池中随机采样一批经验，用于训练 Q 评估神经网络并更新其参数。采样完成再计算目标 Q 值，使用目标网络计算下一状态的最优动作，之后根据 Q-learning 算法计算目标 Q 值并使用 Q 评估网络计算该动作的 Q 值。最后构建估计 Q 值与目标 Q 值的损失函数，利用反向传播算法更新 Q 评估网络的参数，以及根据设定每间隔若干次训练更新目标网络，目的是为了减少目标值的震荡，交通智能体的训练流程图如图 3.4 所示。

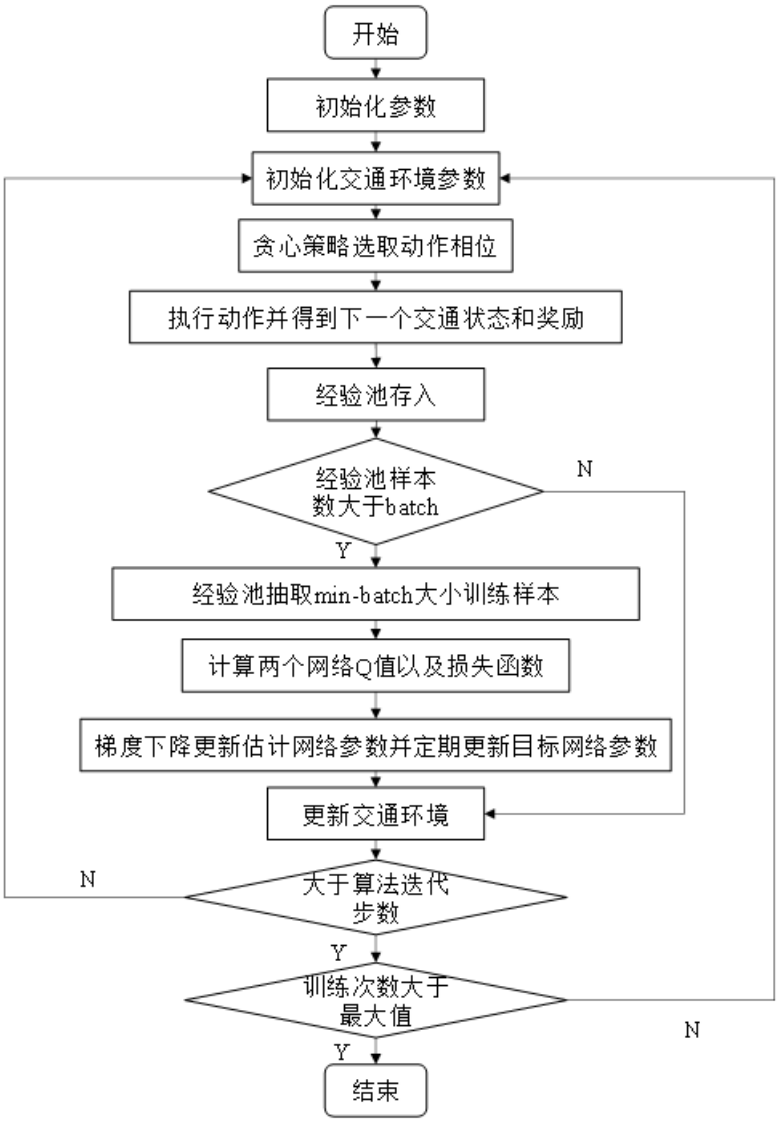

图 3.4 交通智能体模型训练流程图

Fig. 3.4 Traffic agent model training flowchart

3.5 本章小结

本章节对整个信号控制的研究过程基于深度强化学习算法模型的基本结构、模型参数的定义、模型构建以及训练过程进行了相关详细的描述。首先，本章节整体交通信号控制模型的基本结构进行了说明。其次，基于深度强化学习的参数定义方式结合交通路口环境描述了本模型相关参数的定义。最后，构建出基于深度强化学习算法的交通智能体模型，并介绍构建的交通信控系统按照 DRL 模型训练流程的训练过程，为后文基于 DQN 模型的优化提供基础条件。

4 基于预测交通流的 DRL 交通信号控制

4.1 引言

通过第三章所提出的基于 DRL 交通信号控制策略模型能够看出，目前的交通智能体模型采用的是基于 Q-Learning 算法改进的 DQN 算法交通信号控制策略。这种策略目前来说虽然较为成熟，但仍然存在着一些缺点。DQN 算法的提出是为了解决原来 Q 学习算法在复杂环境下 Q 表维度爆炸的问题，因此将该算法与深度学习结合，利用神经网络模型去计算 Q 值代替原有的 Q 表结构，但是 DQN 由于其单一网络结构的原因，会导致其计算的 Q 值会出现偏大的问题，因此提出了 Double DQN 对其进行改进。本文的算法模型也是基于双 Q 深度学习网络，通过对该的算法进行一定改进，使得整个模型相比原来的算法模型在交通信号控制中有一定的提升。

本章的主要内容包括三个部分，在第一小节中首先阐述了 LSTM 网络的结构原理，然后提出基于 LSTM 网络结构如何构建适配传统四车道交叉路口的交通流预测模型以及其训练的损失函数的选择。在第二小节中提出了一种以线性函数与非线性函数结合构成的分段函数来动态调整原 Double DQN 的动作选择策略，使得模型训练相比原有的调整方式达到更好的效果。在第三小节中提出了基于 LSTM 预测的交通流状态去优化原有的双 Q 网络中目标网络的计算方式，用预测的交通流作为交通状态通过 Q-learning 的更新方式计算 Q 值代替原目标网络计算中估计的最大 Q 值部分，达到用真实值代替预测部分，以提高模型目标网络作为评估网络预测结果标签值的精准度。

4.2 交通流预测

4.2.1 LSTM 长短期记忆神经网络

长短期记忆神经网络 LSTM[51]（Long Short-Term Memory）是一种用于处理序列数据的深度学习模型，其主要解决的问题是长期依赖性（Long-term dependency）问题。在传统的循环神经网络（RNN）中，每个时间步的输出都是基于上一个时间步的隐藏状态和当前输入计算得到的。由于每个时间步的隐藏状态都会受到前面时间步的影响，因此在序列数据很长或者存在时间间隔比较大的情况下，传统的 RNN 模型很容易出现梯度消失或者梯度爆炸的问题，导致无法学习到长期依赖关系。

LSTM 是一种基于门控机制的改进 RNN 模型，它是由记忆单元构成，每个结构单元内有三个门（输入门、遗忘门、输出门）来控制每个时间步的输入、输出和隐藏状态的信息流动，从而解决了长期依赖性问题。具体来说，LSTM 在每个时间步都会维护一个隐藏状态h_t和一个记忆内部状态C_t，其中h_t类似于传统 RNN 的隐藏状态，用于保

存当前时刻的重要信息，C_t则用于保存历史信息。

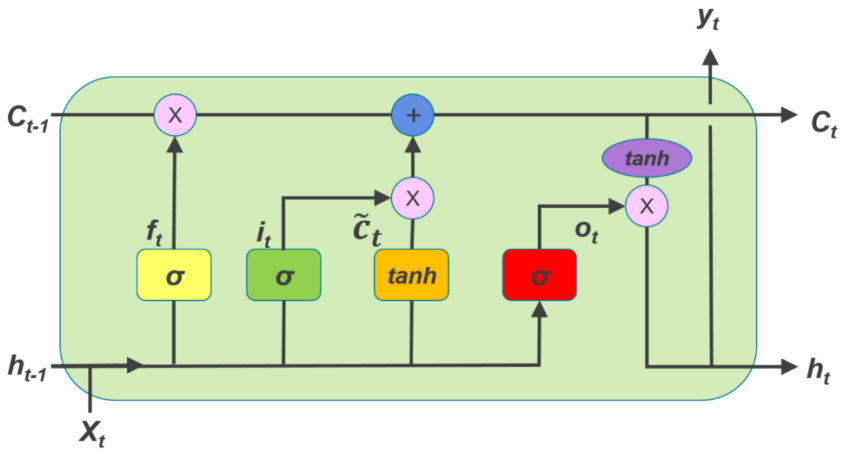

图 4.1 LSTM 记忆单元结构模型图

Fig. 4.1 Structural model diagram of LSTM memory unit

其中X_t为当前状态的输入值，σ 为 Sigmoid 函数，y_t为网络的输出值，$\tilde{C}_t$为输入门需要记忆信息的中间状态，该中间状态是两部分的结合，为当前输入与上个单元隐藏信息两者结合要保留的历史信息，其结构如上图 4.1 所示。LSTM 网络主要通过上述三个门来完成，在三个门中完成输入信息的遗忘，该单元重要信息的记录以及要长久保留的历史信息，这三个门的具体功能以及更新方式如下：

遗忘门（Forget Gate）控制内部状态的遗忘。遗忘门通过一个 Sigmoid 函数将当前时刻的输入和前一时刻的隐藏状态进行计算，得到一个介于 0 和 1 之间的数值f_t，该值反映了当前时刻需要保留多少前一时刻的内部状态信息。然后，遗忘门将前一时刻的内部状态向量和该权重值相乘，得到一个经过遗忘的向量，其更新方式如公式 4.1 所示。

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, X_t] + b_f\right) \tag{4.1}$$

输入门（Input Gate）控制新的输入数据对内部状态的更新。输入门通过一个 Sigmoid 函数将当前时刻的输入X_t和前一时刻的隐藏状态h_{t-1}进行计算，得到一个介于 0 和 1 之间的数值i_t，该值反映了当前时刻输入的重要性。然后，输入门根据这个权重值和当前时刻的输入与h_{t-1}经过 Tanh 函数得出中间值$\tilde{C}_t$计算出一个向量，这个向量就是当前单元要保留的历史信息，再将其与前一时刻的内部状态向量相加，得到新的内部状态，其更新方式如公式 4.2 与公式 4.3 所示。

$$\begin{cases} i_t = \sigma(W_i \cdot [h_{t-1}, X_t] + b_i) \\ \tilde{C}_t = \tanh\left(W_C \cdot [h_{t-1}, X_t] + b_C\right) \end{cases} \tag{4.2}$$

$$C_t = f_t \times C_{t-1} + i_t \times \tilde{C}_t \tag{4.3}$$

输出门o_t（Output Gate）其主要功能就是将$t-1$时刻传递过来并经过了前面遗忘门与记忆门选择后的细胞状态C_t，与$t-1$时刻的输出信号h_{t-1}和t时刻的输入信号X_t整合到一起作为当前时刻的输出信号。整合的过程如公式 4.4 所示。

$$\begin{cases} o_t = \sigma(W_o[h_{t-1}, X_t] + b_o) \\ h_t = o_t \times \tanh(C_t) \end{cases} \tag{4.4}$$

4.2.2 交通流预测 LSTM 模型

交通流量是解决交通拥堵问题的一个重要特征。随着道路上车辆数量的不断增加，传统的交通信号控制系统已被证明无法应对不断增加的交通量，导致交通拥堵、十字路口等待时间过长以及整体交通效率下降。本文根据 LSTM 网络构建交叉路口交通流预测模型，该模型是使用从十字路口收集的历史交通流量数据集进行训练，LSTM网络利用这些数据训练的模型可以预测交叉口的未来若干个时间步长交通流量，然后利用预测的交通流作为后续算法的交通状态去优化后文DRL模型，使得改进后的DRL模型能够更高效地实时调整交通信号灯时间，从而改善交通流量并减少十字路口的等待时间。

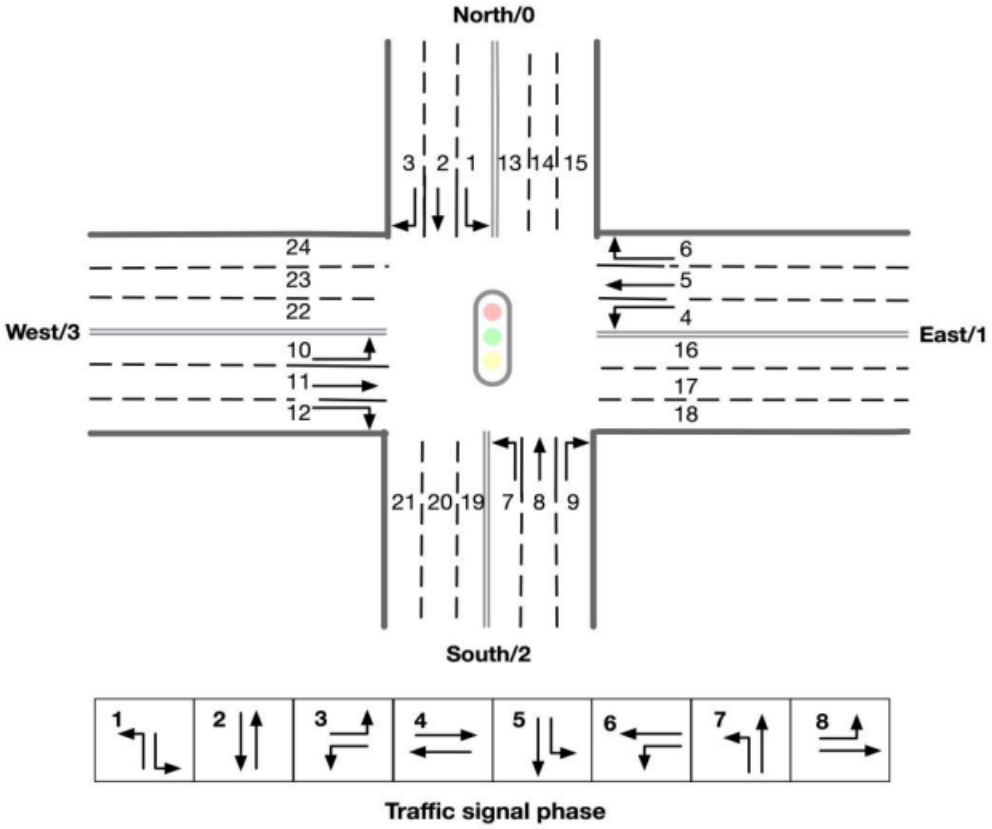

图 4.2 四道路交叉口和八种非冲突相位图

Fig. 4.2 Four-way intersection and eight non-conflicting phase diagrams

在我国很多城市十字交叉路口其通行车道如图 4.2 所示，交叉口由南北和东西两个大方向车道构成，两个大方向车道又分别由 12 条通行车道构成，总计 24 条并分别用 1 到 24 的数字给车道进行标号，道路中南北、东西方向的左右两边车道都包括右转、直行、左转三条车道。用于控制交通信号配时方案所包含的八种非冲突相位分别是南北

左转相位、南北直行相位、东西左转相位、东西直行相位、北直行及左转相位、东直行及左转相位、南直行及左转相位以及西直行及左转相位。由于目前世界各国中交通规则中存在着两种驾驶方式分别是左驾驶和右驾驶，本文是以国内道路为实验对象，我国采取的是左驾驶方式，因此右转相位是持续可通行的，所以各个相位中没有右转相关的相位。在十字路口，当信号相位变换到这八种相位之一，与其相应车道的可以获得车辆通行权利，LSTM 模型图如图 4.3 所示。

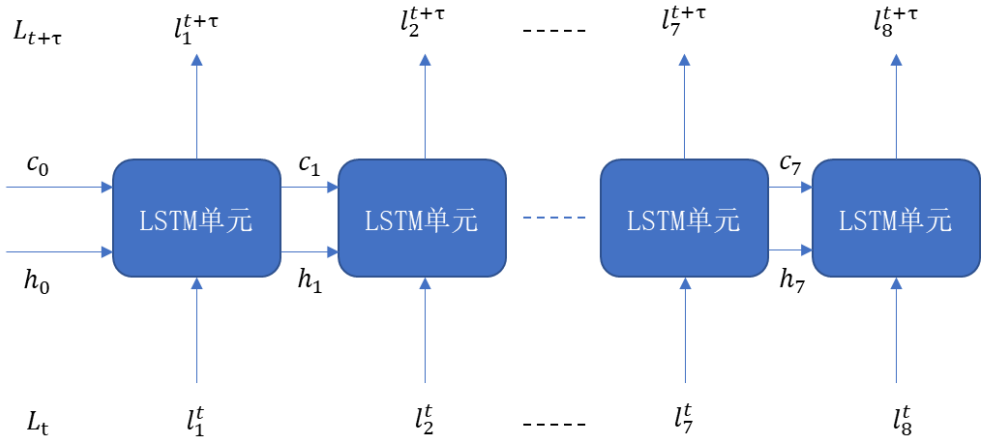

图 4.3 LSTM 交通流预测模型结构图

Fig. 4.3 Structural diagram of LSTM traffic flow prediction model

短期交通预测[52]具有时空复杂性，下一时刻的预测结果基于当前状态和先前知识，其中包括目标道路网络之间的相互作用，这在交叉路口也是一样，不同相位之间各车道车流量也会对其他车道通行相位的时间造成一定的影响，因此本文构建的 LSTM 网络输入向量可以定义为L_t（l_1^t，l_2^t，l_3^t，l_4^t，l_5^t，l_6^t，l_7^t，l_8^t）。如上图所示其中L_t为t时刻在一个信号周期内交叉口各个相位下的在相位持续时间内车道平均车流量构成的向量，l_i^t为各个相位在相应相位的平均车流量，τ为交叉路口相位周期即一个路口所有信号灯完整的信号控制变换时间，相位调控策略主要是调控各个相位持续时间，以达到整个路口通行效率的最优，各个相位的变换都是有序的时间序列，因此可以利用路口历史相位车流数据和 LSTM 网络对下一个信号周期内各相位车流进行预测，使用反向传播方式更新网络内部参数向量，LSTM 采用的激活函数主要是 Sigmoid 函数和 Tanh 函数，损失函数采用平均相对误差（MRE），其定义如公式 4.5 所示。

$$\mathrm{MRE}=\frac{1}{n}\sum_{i=1}^{n}\left|\frac{l_i^{t+\tau}-\varphi_i}{\varphi_i}\right| \tag{4.5}$$

其中n是相位总数，i是相位号，$l_i^{t+\tau}$是预测结果，φ_i是$t+\tau$时刻开始在信号相位周期内各相位车流量平均值。

4.3 动态 ε 策略改进 DRL 算法

4.3.1 Double DQN

Double DQN(Double Deep Q-Network)是深度强化学习中使用的 DQN 算法的扩展。DQN 算法使用神经网络来近似给定状态下每个动作的最佳 Q 值。然而，在 DQN 算法中，使用一个神经网络来估计每个动作的 Q 值，但是这个估计值往往会被过高估计。这是因为在更新 Q 值时，选择最大 Q 值的动作时使用了同一个神经网络估计，而这个估计本身存在一定的不准确性，因此会导致估计出来的 Q 值过高。这种过高估计的问题会导致训练不稳定，甚至出现过拟合的情况，从而导致次优策略。Double DQN 算法则是通过使用两个神经网络来解决这个问题。一个神经网络（称为“行动评估网络”上文中被称为估计网络）用于选择最优的动作，而另一个神经网络（称为“目标网络”即 Target 网络）用于计算目标 Q 值[53]。在更新 Q 值时，目标 Q 值（即真实值）的计算使用目标网络估计，而选择最优动作时使用行动评估网络估计。这种方法能够减少过高估计的问题，从而提高算法的稳定性和性能。Double DQN 背后的思想是通过使用 Q 值估计网络来评估动作选择网络选择的动作的价值，从而将最佳动作的选择与其值的估计解耦。这有助于减少 Q 值的高估并训练出更准确的策略。双 DQN 已应用于各种领域，包括游戏、机器人和自动驾驶。它已被证明在几个基准任务上优于原始 DQN 算法。

Double DQN 算法主要是利用强化学习 Q-Learning 算法的更新方式进行目标网络的 Q 值计算，公式 4.6 为目标网络利用 Q-Learning 计算t时刻目标 Q 值。

$$Q(s_t,a_t)=Q(s_t,a_t)+\alpha[r_t+\gamma maxQ(s_{t+1},a_{t+1})-Q(s_t,a_t)] \tag{4.6}$$

但是在实际训练时其更新方式是对上式进行简化，因此将上式展开可得如公式 4.7 所示。

$$Q(s_t,a_t)=(1-\alpha)Q(s_t,a_t)+\alpha[r_t+\gamma maxQ(s_{t+1},a_{t+1})] \tag{4.7}$$

并将上式经过多次迭代可得如公式 4.8 所示。

$$\begin{aligned}Q(s_t,a_t)&=(1-\alpha)^nQ(s_t,a_t)+\\&[1-(1-\alpha)^n][r_t+\gamma maxQ(s_{t+1},a_{t+1})]\end{aligned} \tag{4.8}$$

上述式中，s_t是t时刻的状态，a_t是t时刻采取的行动，s_{t+1}是$t+1$时刻的状态，a_{t+1}是$t+1$时刻采取的动作，$Q(s_t,a_t)$是s_t状态下采用行动a_t的值函数，α是学习率，r_t是t时刻获得的即时奖励，γ是衰减因子。由于$\alpha\in(0,1)$，因此$1-\alpha\in(0,1)$，$\lim_{n\to\infty}(1-\alpha)^n$趋近于 0，从而简化得到 Q 值的计算方式如公式 4.9 所示。

$$Q(s_t,a_t)=r_t+\gamma maxQ(s_{t+1},a_{t+1}) \tag{4.9}$$

上述式即是常用于目标网络 Q 值计算的方式，在 DQN 算法中目标网络就是根据其进行计算的，目标网络根据当前状态的下一个状态的最大动作的 Q 值计算得到当前状态目标网络的目标 Q 值，该值也是另一个需要训练网络的标签值。在 Double DQN 中由于目

标网络与 Q 评估网络采用两个不同的网络，因此其 Target 网络 Q 值计算需要进行相应的调整，其计算方式如公式 4.10 所示。

$$Q(s_t, a_t) = r_t + \gamma Q'\left(s_{t+1}, \arg\max_a Q(s_{t+1}, a_{t+1})\right) \tag{4.10}$$

在该式子中Q'表述的含义是另一个参数与评估网络不同的网络，也是作为算法的目标网络，$\arg\max_a Q(s_{t+1}, a_{t+1})$表示的含义是 Q 网络在$s_{t+1}$状态最大 Q 值选择的动作，因此用目标网络也执行相同的动作计算 Q 值，该值会作为真实 Q 值， Q 评价网络其计算值的标签构造损失函数并利用梯度下降方式优化更新评估网络参数，其损失函数如公式 4.11 与公式 4.12 所示。

$$Q_{Target} = r_t + \gamma Q'\left(s_{t+1}, \arg\max_a Q(s_{t+1}, a_{t+1})\right) \tag{4.11}$$

$$\text{LOSS} = \left(Q_{Target} - Q(s_t, a_t, \theta)\right)^2 \tag{4.12}$$

4.3.2 动态 ε 动作选择策略

DQN 算法在使用经验回放从先前的经验中学习，并使用目标网络来降低训练中的方差。在这个过程中，DQN 使用贪婪策略来选择当前状态下最优的动作。但是，如果仅使用贪婪策略，那么算法很容易陷入局部最优解，而无法探索更多的状态和动作。因此，DQN 引入了一定程度的探索，即在一定概率下，随机选择动作而不是选择当前最优动作。这个概率称为探索率（epsilon 在本章中对应着 1-ε）。随着训练的进行，探索率会逐渐减小，从而逐渐依赖于贪婪策略。探索率的逐渐减小使得模型可以在早期阶段更多地探索状态空间，以发现潜在的高回报区域。然后在训练的后期，模型将更加倾向于选择已知的高回报动作，以优化性能。因此，在 DQN 中引入探索率的原因是为了在权衡探索和利用之间找到一个平衡点，从而更好地学习和优化策略。

对于探索率的取值来说，在实际操作中一般是会给定一个变化概率值。随着模型的训练次数的不断增加，神经网络模型的参数趋近于稳定，DQN 需要使用更加贪婪的策略来选取最优 Q 值的动作。此时探索率的值也会趋近于 0，DQN 训练时会以更大的概率选取最优 Q 值动作，更小概率选取随机动作。因此，在其他的一些 DQN 改进方法中，常常会将探索率按照一个线性函数的形式进行变化调整，但是这种固定变化的调整，在训练后期时可能会因为前期线性函数斜率 k 的选择上过大或者过小，导致动作选择策略在后期变化过快或者过慢，使得训练结果不够好，还有些会利用探索率自身叠乘的方式去迭代其变化，但是这两者都需要设定一个探索率的下限。因此，在到达下限时算法的探索率虽然已经很小，但是会保持恒定不变。这会导致只要训练次数会很

大时，还是会有较大概率选择到随机动作，但此时训练参数已经趋近于稳定。

在本文中利用 Sigmoid 函数[54]其在函数大于 0 的区间内，其值会无限趋近于 1 的特性，构建一个分段函数形式对 Double DQN 算法中对 Agent 按照 ε 动作策略选取动作的方式进行调整，分段函数图像如图 4.4 所示。

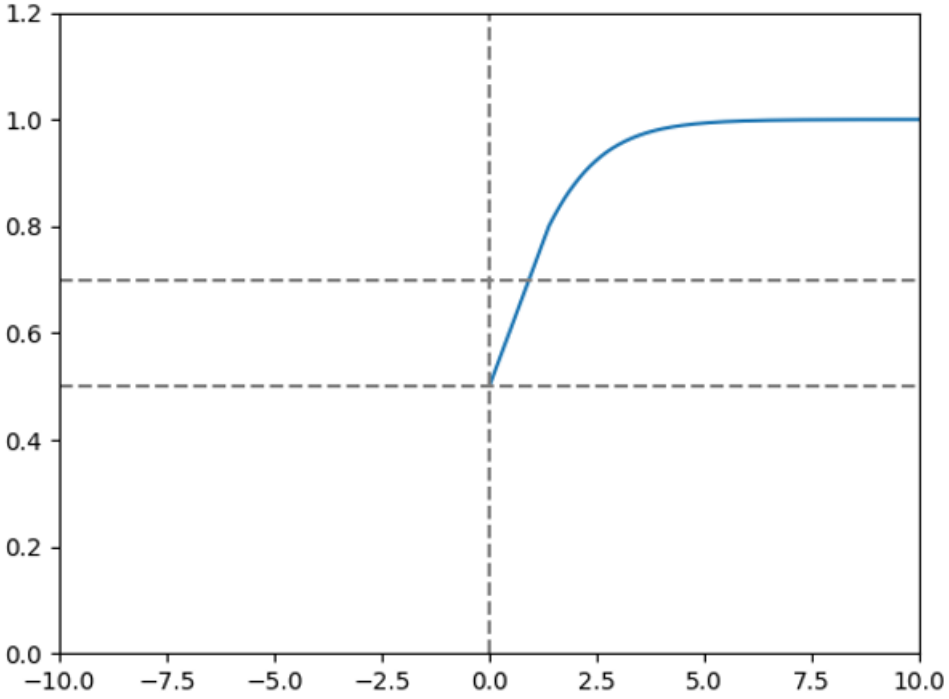

图 4.4 改进动作选择策略分段函数图

Fig. 4.4 Segmented function diagram of improved action selection strategy

该分段函数是通过线性函数与 Sigmoid 函数结合形成的，函数的前半段是一个线性函数，后半段为 Sigmoid 函数，在训练前期使用常用的线性函数调整 ε 策略，使得前期有较大概率选择随机动作；在训练的后期 Q 网络参数稳定时，Sigmoid 函数会无限地接近于 1 的概率，Q 网络会尽可能选择最优动作，其分段函数如公式 4.13 所示。

$$\varepsilon(x)=\begin{cases}\dfrac{0.3x}{ln\,4}+0.5, & 0<x<\ln 4\\ Sigmoid(x), & x\geq \ln 4\end{cases} \tag{4.13}$$

然后利用该分段函数对 DQN 算法中以 ε 策略选取智能体所执行的动作，具体操作方式可以表示为：在算法流程中会以 Random 函数随机数生成的方式，生成 0 到 1 之间的随机数x，当随机数大于 ε，会选择一个随机动作；当随机数x小于 ε，则选择 Q 值最大的，之后通过梯度下降反向传播的方式来更新训练网络的参数。

4.4 结合 LSTM 预测状态的改进 Double DQN

4.4.1 方法概述

单个路口智能体训练不仅需要考虑当前状态对交通信号决策的影响，还需要考虑未来交通状态决策的影响。智能体会考虑未来可能决策以及当前状态来做出相应的信号控制策略，这样做出的决策不仅符合当前交通状况的调控，也会考虑到未来的决策而做提前的准备（即未来对现在的影响）。

在 Double DQN 算法中目标网络的目标 Q 值计算是根据 Q-Learning 算法更新方法来求得目标 Q 值，作为训练网络的优化目标，其计算方式是通过在t时刻状态s_t，在当前模型根据动作策略执行动作后所得到奖励值加上执行完动作之后进入s_{t+1}状态下其在目标网络下对应的最大 Q 值乘以折扣因子的总和，这里采用的方法与 TD 算法的思想是一致的，将当前即时奖励值加上未来剩余过程的价值来表示当前状态该动作所具有的价值。TD 算法更新与 Double DQN 中按照 Q-Learning 更新对比如公式 4.14 与公式 4.15 所示。

$$V(s_t) \leftarrow V(s_t) + \alpha[R_{t+1} + \gamma V(s_{t+1}) - V(s_t)] \tag{4.14}$$

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha\left[R_t + \gamma Q'\left(s_{t+1}, \arg\max_a Q(s_{t+1}, a_{t+1})\right) - Q(s, a)\right] \tag{4.15}$$

根据本文第二章描述的状态价值函数以及动作价值函数的定义，我们可知两者表示的是即时期望值与未来期望值之和，并在未来期望值上乘上一个折扣因子，表示未来对当前影响的一个折扣，更远的未来乘上的折扣因子会更小，其影响也会越小。在 DQN 的目标网络的部分的计算，是一个即时可知的奖励加上未来根据目标网络预测的奖励乘上折扣因子，因此该部分是加上了一个预测值与折扣因子的乘积，而且在实际操作中，目标网络与评估网络两者参数是一致的，在经过多次定期迭代后评估网络的参数才会复制给目标网络，因此在前期的计算中目标网络的计算结果不能足够反映实际 Q 值，所以将目标网络中预测值延后，更多真实值添加到目标 Q 值的计算中（在 DQN 算法中利用 Q-Learning 公式计算的结果为真实值）有利于提高训练出模型的精准性和有效性。

4.4.2 结合 LSTM 的改进目标网络计算

为了实现上述目标，本文引入 LSTM 网络对交通流进行预测，以得到更未来的交通流特征，将更未来的特征作为目标网络计算依据，来达到将模型训练中目标网络计算使用预测值延后的效果，其具体流程如下：当网络模型获取当前交通环境的状态s_t（本文是以车流量信息为交通状态输入并且每增加一个时刻为增加一个相位周期），并

根据 LSTM 交通流预测模型 $\rho(S)$预测下一个信号周期的交通状态s_p。因此目标网络原本预测值$Q'\left(s_{t+1}, \arg\max_a Q(s_{t+1}, a_{t+1})\right)$部分改用可以用 Q-Learning 更新方式计算，其式可以表示为如公式 4.16 所示。

$$Q'\left(s_{t+1}, \arg\max_a Q(s_{t+1}, a_{t+1})\right) = R_{t+1} + \gamma' max Q'(s_p, a_p) \tag{4.16}$$

改进原有的 Target 网络 Q 估计值的计算方式，构造新的目标网络优化的 Q 值策略，其计算方式可以表示如公式 4.17 所示。

$$Q_{Target} = R_t + \gamma\left[\omega R_{t+1} + \gamma' max Q'(s_p, a_p)\right] \tag{4.17}$$

其中ωR_{t+1}计算的是目标网络执行$\arg\max_a Q(s_{t+1}, a_{t+1})$的奖励值，$\gamma'$代表预测折扣因子，$\gamma$为实际折扣因子。在训练模型中，$s_t$执行动作之后进入$s_{t+1}$状态，$s_{t+1}$状态执行的动作可能会改变，这是由于预测的状态计算部分是在上一次评估网络的参数下进行的。由于在交通环境中动作空间是有限集合，训练模型时可以在样本中寻找R_{t+1}与R'_{t+1}两者相关性ω，R_{t+1}为s_{t+1}下真实动作的即时奖励，R'_{t+1}代表s_{t+1}执行动作进入s_p的即时奖励，在本文由于通过使用交通流作为主要状态特征，所以考虑到实际操作性将ω参数用s_{t+1}与s_p两者比值作为它的取值。完成目标 Q 值计算后与评估网络 Q 值构建损失函数并利用梯度下降法更新优化网络参数，可将式表示为如下公式 4.18 所示。

$$\theta' = \theta + E\left[Q_{Target} - Q(s_t, a_t, w_t)\right]\nabla Q(s_t, a_t, \theta) \tag{4.18}$$

其中$Q(s_t, a_t, w_t)$表示评估网络所预测的 Q 值，Q_{Target}为目标网络 Q 值，θ原参数，θ'更新参数。

4.4.3 系统模型以及算法流程

本文通过 LSTM 网络预测交通流状态并结合改进的 DRL 算法对交通信号进行控制，对交通信号变换策略进行优化，整体的算法流程较为复杂。表 4.1 中给出了本文算法伪代码。本文所提出的算法其优化目标是根据交叉路口道路环境做出最优的交通信号控制策略。通过对算法模型不断进行训练，使得训练好的智能体模型能依据其所处交通环境状态控制交通信号灯采取有效的相位切换，从而使得交叉路口通行效率提高，可通行车辆数以及车辆延迟得到优化。本文所提出的交通信号控制策略模型在训练初期，

其在交通环境状态下是随机选择信号相位切换动作并迭代一定的次数，这样做的目的是为了让经验回放池可以获得足够多的样本来对神经网络进行训练。每个经验样本的优先级在没有进行神经网络训练之前都是保持相同的，模型会按照指定批量的方式从经验回放池中随机抽取经验样本进行训练，模型的整体结构图如图 4.5 所示。

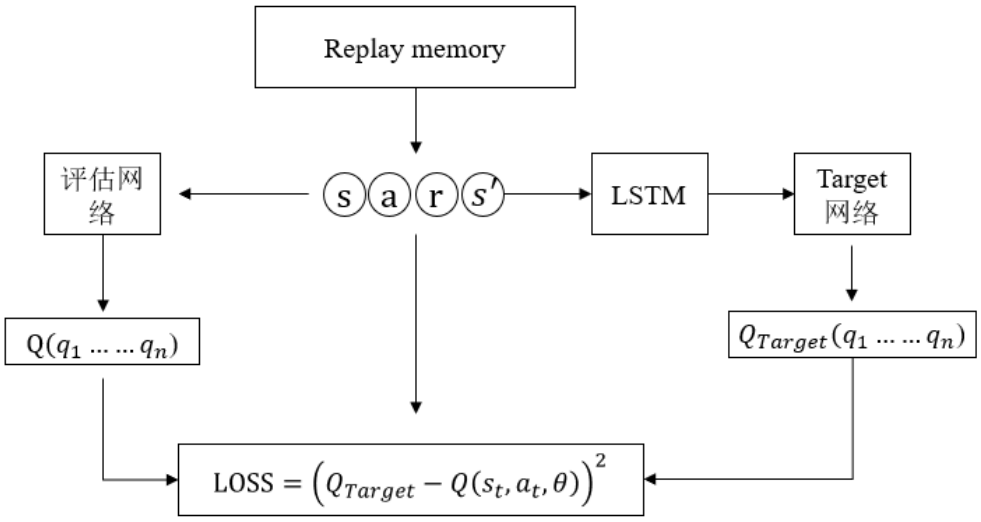

图 4.5 基于预测改进 Double DQN 模型整体结构图

Fig. 4.5 The overall structure of the improved Double DQN model based on prediction

表 4.1 基于 LSTM 预测的 Double DQN 交通信号控制伪代码

Tab. 4.1 Double DQN traffic signal control pseudocode based on LSTM prediction

算法：基于预测交通流状态的交通信号控制算法
初始化训练总轮数 T，学习率 α，折扣系数 γ，γ′动作ε参数，关系参数 ω，选择批量大小 B，预先训练次数 P
初始化训练轮次 episode =1
初始化经验回放池 D 动作空间 A
初始化神经网络 Double-DQN 评估网络 Q1 并随机赋予参数
初始化 LSTM 网络、目标网络 Q2 其参数与 Q1 相同
for episode=1 to T do
初始化的路口交通环境中得到状态s
预先按照优化ε策略选取动作训练 P 次并存放样本进入经验池
随机从经验池抽取批量大小为 B 的样本，样本数据格式为$< s，a，R，s' >$
s输入 Q1 网络计算预测 Q 值, s'输入 LSTM 预测s_p
s'匹配样本与经验池的样本中数据的第一项s，获取其奖励值R'、参数 ω，γ′
计算目标网络真实值$Q_{Target} = R + \gamma[\omega R' + \gamma' maxQ'(s_p, a_p)]$
构建损失函数，梯度下降反向传播更新网络参数
end for

4.5 本章小结

本章提出了一种基于预测改进的 Double DQN 交通信号控制策略模型。首先我们基于 LSTM 网络构建了交通流的预测模型。然后，构建一个更平滑的分段函数来控制算法动作选择。最后，结合预测的交通流状态改进了原有 Double DQN 目标网络的计算方式并给出了本文改进的模型结构以及算法伪代码，为后文的仿真实验提供了理论基础。

5 仿真实验

5.1 引言

本章将对 DQN、Double DQN 以及基于预测交通流改进 Double DQN 这三种交通信号控制策略分别进行仿真实验，并与传统信号配时以及这三种优化方式的结果来比较几种交通信号控制策略的实际调控效果。同时也对上章所提出基于预测 LSTM 交通流的交通信号控制策略模型的优化效果进行验证。在对上述几种方案进行仿真后，得到在几种策略调控下其基于交通性能评价指标的结果，然后再对所得到的实验结果进行相关分析解释。

5.2 仿真环境与实验数据

5.2.1 仿真环境

本文的仿真实验是基于 KDD 比赛提供 CBEngine 仿真器上进行的，该仿真模拟器能够支持城市尺度路网的微观交通仿真，并且它可以支持对包含大量交叉路口以及车辆的路网交通环境进行快速仿真。该引擎是由云栖工程院团队所开发，同时该团队也提供了相应的模拟环境设置指南。其模拟仿真的交通路网环境如图 5.1 所示。

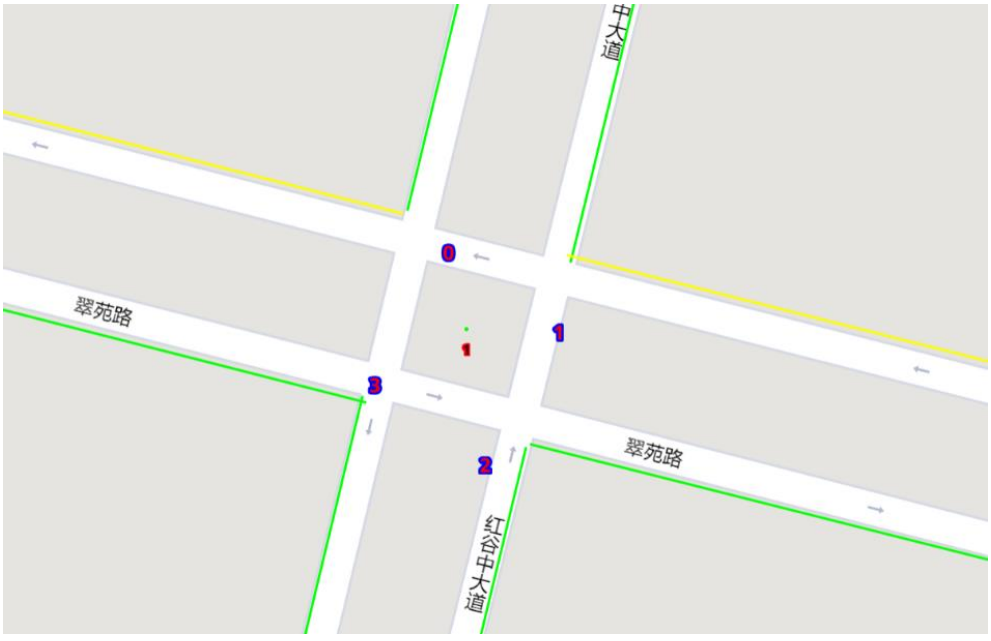

图 5.1 交通路网仿真模拟图

Fig. 5.1 Simulation diagram of traffic road network

5.2.2 路网文件

本文数据格式包含两大类，分别是路网文件格式和车流文件格式。其中路网文件格式中包含四类数据集，分别是交叉口数据集、道路数据集、交通信号数据集以及路网数据集。

(1) 交叉口数据集

交叉路口数据包括每个路口的标识、坐标和信号灯安装状态。图 5.2 展示了一个交叉口数据集的片段。

```
28.674579650000002 115.847174925 42167350403 1
28.675813249999997 115.841737175 42167350405 1
28.67896945 115.84839492500001 42167350420 1
28.6825176 115.84938435000001 42167350438 0
28.682649750000003 115.83705922499999 42167350445 1
```

图 5.2 交叉口数据集片段

Fig. 5.2 Intersection dataset fragment

表 5.1 交叉口数据集属性

Tab. 5.1 Intersection dataset properties

属性名	示例	描述
纬度	28.6825176	路口坐标纬度
经度	115.84938435000001	路口坐标经度
交叉口 ID	42167350438	交叉口标号
信号灯	0	如果安装了交通信号灯则为 1，否则为 0

(2) 道路数据集

道路数据集包含有关路网中道路段的信息。其中，每个路段有两个方向，其主要包括交叉口 ID、限速、车道数、车道编号、车道长度等数据。道路数据集的片段示例如图 5.3 所示。

```
3012
22296635640 41704581960 1016.0 20 3 3 1 2
1 0 0 0 1 0 0 0 1
1 0 0 0 1 0 0 0 1
42266617929 41704581960 771.0 20 3 3 3 4
1 0 0 0 1 0 0 0 1
1 0 0 0 1 0 0 0 1
```

图 5.3 道路数据示例

Fig. 5.3 Road data example

表 5.2 道路数据集属性

Tab. 5.2 Road dataset properties

属性名	示例	描述
上游交叉口 ID	22296635640	上游交叉路口的编号
下游交叉口 ID	41704581960	下游交叉路口的编号
长度（米）	1016.0	道路长度
限速（米/秒）	20	道路限速
方向 1 车道数	3	方向 1 的车道数量
方向 2 车道数	3	方向 2 的车道数量
方向 1 路段 ID	1	方向 1 路段（边）编号
方向 2 路段 ID	2	方向 1 路段（边）编号
dir1_mov	100010001	每 3 位数字组成一个方向 1 车道的允许移动指示器，100 表示仅左转内车道，010 表示仅通过中间车道，011 表示共享直通和右转外车道
dir2_mov	100010001	同 dir1_mov

(3) 交通信号数据集

该数据集描述了交叉路口和路段之间的连通性。在本文的仿真实验中，以每个交叉路口的入口不超过四个为标准，从北边出口开始以顺时针方向，具有方向 1 到 4 的四个大方向。数据片段中的-1 表示缺少相应的进场，一般表示三段式交叉路口。交通信号数据集展示以及交叉口示例图，如图 5.4、5.5 所示。

```
859
42266617929 327 3 3984 3985
21858133810 10 5531 4279 8
21858133803 3162 4306 9 3160
42495943806 75 1892 -1 1896
```

图 5.4 交通信号数据集片段

Fig. 5.4 Traffic signal dataset fragments

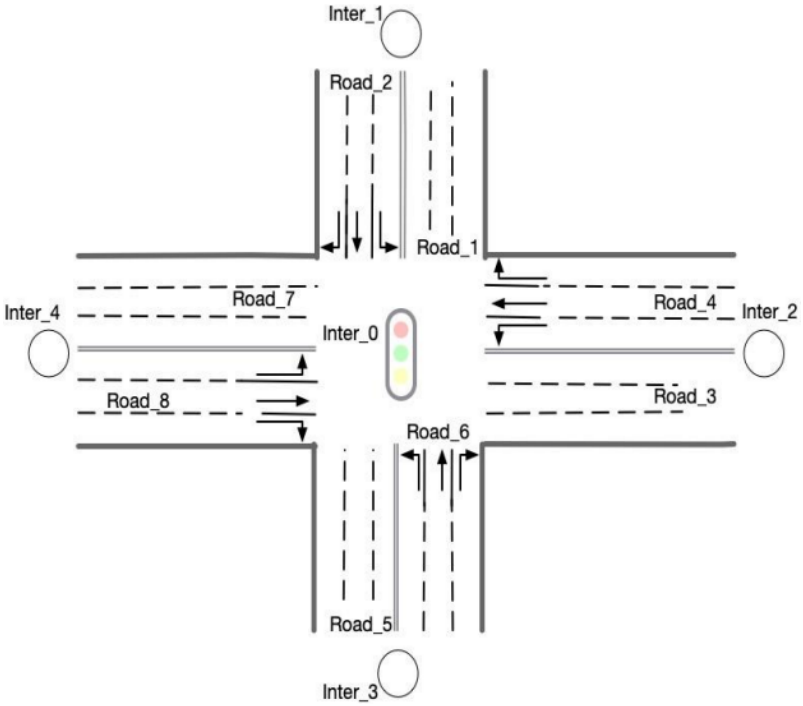

图 5.5 1×1 路网示例图

Fig. 5.5 1×1 Road network diagram

表 5.3 交通信号数据集属性

Tab. 5.3 Traffic signal dataset properties

属性名	示例	描述
交叉口 ID	21858133810	交叉路口编号
方向 1_ID	10	北出口的路段（边缘）ID（示例中的 Road_1）
方向 2_ID	5531	东出口的路段（边缘）ID（示例中的 Road_3）
方向 3_ID	4279	南出口的路段（边缘）ID（示例中的 Road_5）
方向 4_ID	8	西出口的路段（边缘）ID（示例中的 Road_7）

（4）路网数据集图 5.6 是一个示例 1×1 的路网数据集片段。

```
5
30 120 0 1
31 120 1 0
30 121 2 0
29 120 3 0
30 119 4 0
4
0 1 30 20 3 3 1 2
1 0 0 0 1 0 0 1 1
1 0 0 0 1 0 0 1 1
0 2 30 20 3 3 3 4
1 0 0 0 1 0 0 1 1
1 0 0 0 1 0 0 1 1
0 3 30 20 3 3 5 6
1 0 0 0 1 0 0 1 1
1 0 0 0 1 0 0 1 1
0 4 30 20 3 3 7 8
1 0 0 0 1 0 0 1 1
1 0 0 0 1 0 0 1 1
1
0 1 3 5 7
```

图 5.6 1×1 路网数据集片段

Fig. 5.6 1×1 Road network data set fragments

5.2.3 车流文件

本文的车流文件由若干流组成。每个流通过元组<起始时间，结束时间，vehicle_interval，路线>的形式表示，它表示在每隔 vehicle_interval（车辆间隔时间）秒的时间段内，从起始时间到终止时间，道路会有一辆车按照这条路线在路网中行驶。车流文件包含以下两个部分。图 5.7 是一个车流文件的展示。

在车流文件示例中，第一行是表示车流的数量，从后面每三行指示每个流的配置，每个流有 3 个配置行。第一行是由起始时间、终止时间、vehicle_interval 组成，第二行是该车流路线的路段数 k，第三行是该车流路线的经过的路段，比如下图中 2、3 代表就是路段 ID，并且车流的路线是根据道路进行定义的。

```
12
0 1800 40
2
2 3
0 1800 60
2
2 5
0 1800 70
2
2 7
0 1800 60
2
4 5
0 1800 50
2
4 7
0 1800 40
2
4 1
0 1800 50
2
6 7
0 1800 40
2
6 1
```

图 5.7 车流数据集片段

Fig. 5.7 Traffic flow data set fragments

5.3 仿真实验及结果

5.3.1 相关参数设置说明

本文进行的仿真实验想要达到的主要目标是让交通路网中交叉路口可通行的车辆数最大化即使得交叉路口在每个相位周期内的车流量最大化。通过利用 CBEngine 仿真引擎（城市尺度路网交通仿真的微观引擎）实现对路网模型的快速仿真，该模型可以支持对较为复杂交通环境的路网交通进行快速模拟。本文使用的 CBEngine 仿真引擎提供了相应的接口来获取交通信号灯所处的交叉路口道路信息并且还可以通过相应接口给仿真环境中的信号灯发送相应交通信号相位切换的指令来控制模拟环境下交通灯相位的变换。在实验中的仿真数据文件包括了上千个交叉路口、道路、交通信号灯以及十多万辆车辆的轨迹信息。算法模型利用仿真环境不断地迭代训练，并在合适的时刻给出切换交通信号灯相位的策略以最大化每小时内交通路网中可通行的车辆数并使得延迟降低且在可接受范围。代码运行环境为 Tensorflow2.1 及以上，本文提出的改进模

型初始化网络参数如表 5.4 所示，所给出的参数是 KDD 提供的参数设置，但本文由于提出了动态探索策略，因此在官方原模型的 DQN 方式进行信号控制与 Double DQN 模型的实验中使用原有参数，在本文提出的基于 LSTM 改进的动态策略 Double DQN 模型中使用本文提出的函数去控制探索率的自动变化，其他参数则保持相同。LSTM 网络在训练时，根据引擎仿真收集的历史车流数据进行训练，其训练用数据集的 80%作为训练数据，20%的比例作为训练的测试集，采用 Adam 的方法来优化神经网络。图 5.8 展示的是基于交通路口交通流 LSTM 网络预测模型情况。

表 5.4 模型初始化参数

Tab. 5.4 Model initialization parameters

参数名	Value
Replay memory size M	3600
Batch_size	1800
Discount factor γ	0.9
Learning rate	0.005

5.3.2 仿真结果

（1）本文的实验主要分为两部分，分别是 LSTM 预测模型的训练和 DRL 交通信号控制模型的训练，图 5.8 展示了预测模型使用测试集的预测情况。

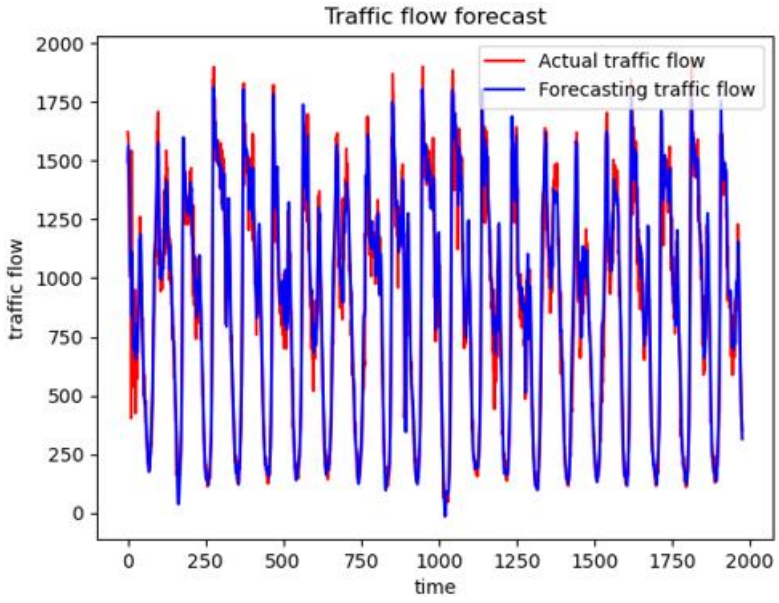

图 5.8 LSTM 网络预测模型训练情况

Fig. 5.8 LSTM network prediction model training situation

（2）本文实验的另一部分 DRL 模型的训练部分，在实验中我们分别训练的 DQN、Double DQN 以及 LSTM-Double DQN 三种模型，并将其分别放入仿真引擎建立的仿真

交通环境中，比较各自对交通信号控制的效果，下文的图 5.9、图 5.10 以及图 5.11 分别是三者的训练过程中在迭代过程里其损失函数的变化情况。图 5.12 则是本文提出的模型在 100 轮次的测试训练中其 Q 值的变化趋势图。

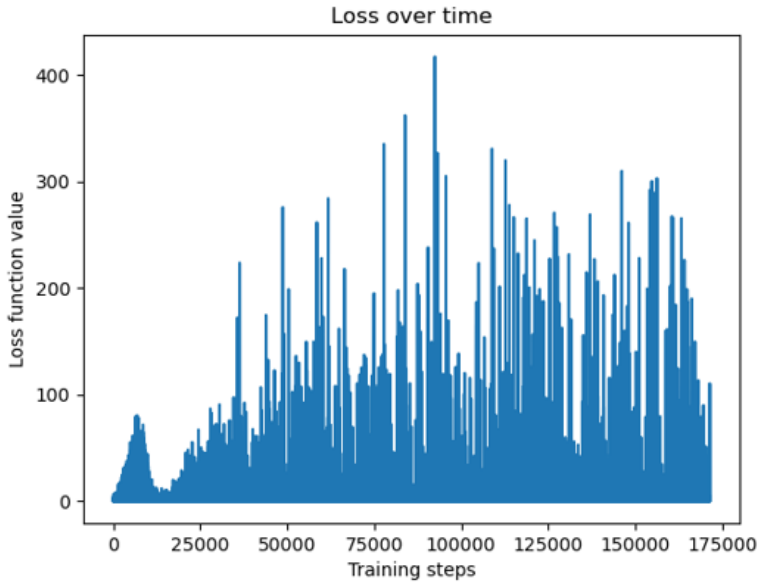

图 5.9 DQN 训练损失函数值变化图

Fig. 5.9 DQN training loss function value change diagram

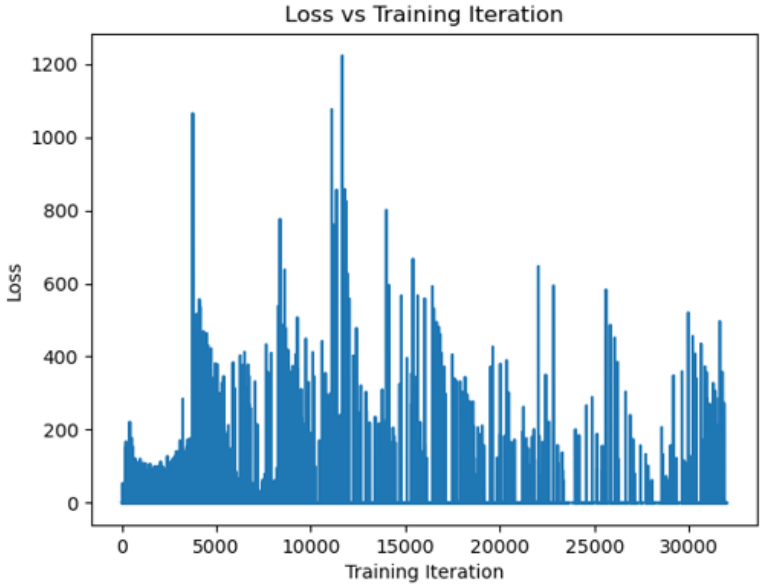

图 5.10 Double DQN 训练损失函数值变化图

Fig. 5.10 Double DQN training loss function value change diagram

图 5.11 LSTM-ε Double DQN 训练损失函数值变化图

Fig. 5.11 LSTM-ε Double DQN training loss function value change diagram

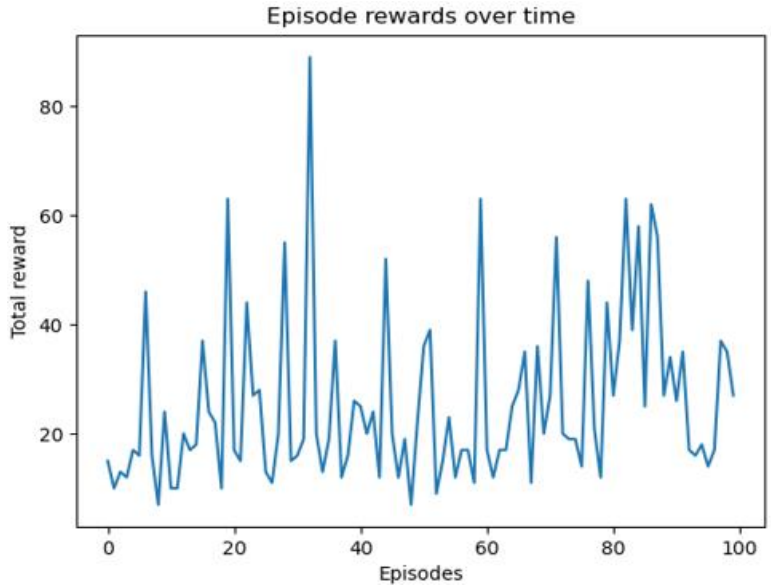

图 5.12 交通信号控制模型预测 Q 值图

Fig. 5.12 Prediction Q value diagram of traffic signal control model

这里 Q 值预测使用了提高经验回放池的容量，选择较小 Batch_size 大小，以较小的 Episodes 提高每轮的训练次数的方式进行。最后将上述三个模型分别运用于 CBEngine 仿真引擎模拟的交通环境中，该模型的结果会通过车辆数来反映模型对交叉口调控的效果，其服务车流数据结果如表 5.5 所示，其延迟计算方式是利用公式 2.1 与公式 2.2 进行的。从表中我们可以看出结合预测状态特征与利用分段函数动态调控探索率的改进 Double DQN 模型相比于其他的模型，其服务车辆数都要更多，并且延迟也会相对更小，

因此在调控交通信号控制以提高交通路口的效率，缓解道路拥堵等方面，相比传统的 DRL 模型是有所提高的，能够更好地服务交通信号调控，该实验证明了本文提出的模型的有效性。

表 5.5 交通模型仿真结果

Tab. 5.5 Traffic model simulation result

Model	服务车辆数（辆）	延迟指数/s
Fixed time	28698	1.72363
DQN	70875	1.64055
Double DQN	77267	1.62435
LSTM-ε Double DQN	84896	1.61031

5.4 本章小结

本章主要工作是结合上一章所提出的算法模型进行仿真实验并分析实验所得到的结果。首先介绍了实验所使用的仿真环境以及仿真模拟器，然后对所使用的实验数据以及模型相关参数设置进行了说明，并展示了本文模型以及对比模型的实验结果。通过最终的实验结果可以看到在交通信号控制优化中，结合 LSTM 预测与改进的动态探索率的 Double DQN 算法控制模型能够在提升交叉路口的通行车辆数以及降低车辆延迟方面优于其他对比的信号控制模型。因此，从该实验中可以总结得出本文提出的有关算法的改进，对交通信号的控制优化具有一定的改进效果，可以更好地利用仿真环境中交通状态信息去优化交叉路口的交通信号控制。

6 总结与展望

6.1 总结

本文首先对国内城市的交通现状以及问题进行阐述，找到了当前城市交通普遍拥堵的原因。然后结合当下人工智能技术快速发展的状况，对利用人工智能技术解决交通领域中交通拥堵问题的必要性以及紧迫性进行了相关分析。通过查询相关资料我们发现，交叉路口在城市交通控制中有很重要的作用，车辆的拥堵不仅仅是在其他路段，交叉口的交通拥堵以及车辆通行效率低也会间接地影响到交叉口之间道路车辆的通行，因此解决城市交通拥堵的关键之一就是通过更加合理的方式调控交叉路口的车辆通行效率。然而传统交通信号控制的方法已经对目前复杂多变的交通环境状况所带来的问题力不从心，因此本文在研究信号控制方法前，学习了解了许多研究学者对城市交叉路口交通信号控制问题的深层次研究，发现越来越多的研究指出通过与现代迅速发展的人工智能技术相结合是目前能够有效解决传统交通信号控制不能自适应调控信号配时的问题并提高交叉路口车辆通行效率。目前在人工智能技术这方面，由于交通路口环境的复杂性，传统的人工智能方法不太能适应这种状况，所以通过将当前人工智能领域中具有更强自适应的深度强化学习方法去结合到交通信号控制中，依靠该方法强大的学习以及自适应能力能够有效地解决城市交通因传统信号控制方法效率低而造成交通拥堵问题。传统的人工智能方法主要以强化学习方法为主比如经典的 Q-Learning 算法，将其与交通信号控制结合在提高交叉路口通行效率以及解决道路拥堵问题的表现来看，基于该方法的交通信号控制策略相对于利用传统交通学原理的控制策略来说是一种更为高效的方法。然而，经典的 Q-Learning 算法本身具有一定的缺陷，造成这种缺陷的原因是因为 Q-Learning 算法其学习的本质是通过建立 Q 表形式来存储自身学习经验，然后这种建立 Q 表的方式在面对复杂场景时就会出现许多问题比如存在复杂环境由于状态维度过高导致的维度爆炸、目标网络的 Q 值估计过高以及 Q 值估计时学习不稳定等，进而也导致模型学习到的策略不佳。这种算法缺陷在如今日益复杂的交通环境下更加突出，也使得基于该算法的交通信号控制策略在实际应用时效果达不到预期。因此，通过引入对强化学习算法缺点进行改进的深度强化学习的方法来使得交通信号控制模型能更好地适应当前环境。之后本文又在对深度强化学习技术进行了多方面了解以及调研的基础上，提出了一种基于 LSTM 交通流预测以及分段函数动态控制探索策略 ε 的改进 Double DQN 网络的交通信号控制优化模型，目的在于通过更好的信号控制策略模型实现对交叉路口交通流进行高效分流，从而提高道路整体通行效率，解决交通拥堵问题。从第五章的实验结果来看，本文所提出的基于预测交通流的交通信号控制优化模型能够使最大可通行车辆数量得到一定的提升并降低车辆的平均延迟，

达到了提升路网整体运行效率以及提高路网通行能力的目的，能一定程度的缓解交通拥堵的现状，并且相比于传统方法以及其他 DRL 方法也具有明显的优势。因此，在本文中提出的结合 LSTM 交通流预测状态的改进深度强化学习交通信号控制方法，在交通信号控制优化方面的研究上，具有一定的实用价值和现实意义。

6.2 展望

本文所提出的基于 LSTM 预测交通流和分段函数控制动态探索策略的改进 Double DQN 交通信号控制优化模型从仿真效果来看具有相比于一些其他 DRL 算法有更好的效果，但本文对比的算法数量还是有所欠缺，还有更多优于 Double DQN 的算法可以进行对比、结合以及改进。因此在提升交通性能方面上，本文模型还有更进一步的优化的空间。因此本文可以做出以下几点展望：

（1）在研究基于 LSTM 预测交通流的交通信号控制算法时，可以考虑改进模型训练样本的取样方法、优化模型参数设置以及调整神经网络的结构等来提升算法的性能以及模型的鲁棒性，还可以使用更好的方法找到预测状态奖励与实际奖励的关联关系。

（2）对于交通状态定义与预测方面，可以采用更优秀的神经网络模型以及更多维的状态定义，使得模型具有更多的可预测交通特征而不仅仅是交通流，以提高训练出来的交通信号控制模型具有更强学习能力、自适应性以及时效性。

（3）目前本文提出模型都是在仿真环境下进行的实验，在实际场景的运用效果上还存在着不确定性。在后续的工作中可以尝试加入更多的实际环境影响因素来提高算法的实际运用能力。

参考文献

[1] 数说 2022[J]. 中国总会计师, 2023, No.234(01): 21.

[2] 商讯. 公安部交管局：截至 2022 年 9 月底，全国汽车保有量达 3.15 亿辆，其中新能源汽车占比 3.65%[J]. 商用汽车, 2022, No.380(10): 7.

[3] 2021 年全国车驾业务统计数据[J]. 公安研究, 2022, No.328(02): 96.

[4] Qureshi K N, Abdullah A H. A survey on intelligent transportation systems[J]. Middle-East Journal of Scientific Research, 2013, 15(5): 629-642.

[5] Haydari A, Yılmaz Y. Deep reinforcement learning for intelligent transportation systems: A survey[J]. IEEE Transactions on Intelligent Transportation Systems, 2020, 23(1): 11-32.

[6] 赵冬斌, 刘德荣, 易建强. 基于自适应动态规划的城市交通信号优化控制方法综述[J]. 自动化学报, 2009, 35(06): 676-681.

[7] Arulkumaran K, Deisenroth M P, Brundage M, et al. Deep reinforcement learning: A brief survey[J]. IEEE Signal Processing Magazine, 2017, 34(6): 26-38.

[8] Li Y. Deep reinforcement learning: An overview[J]. arXiv preprint arXiv:1701.07274, 2017.

[9] Nwana H S, Ndumu D T. An introduction to agent technology[J]. Software Agents and Soft Computing Towards Enhancing Machine Intelligence: Concepts and Applications, 2005: 1-26.

[10] Ivanov S, D'yakonov A. Modern deep reinforcement learning algorithms[J]. arXiv preprint arXiv:1906.10025, 2019.

[11] Robertson D I. TRANSYT: a traffic network study tool[J]. 1969.

[12] 张继锋, 陈云. NATS 城市交通信号控制系统[J]. 中国公共安全: 智能交通, 2007(008): 50-54.

[13] 朱中, 管德永. 海信 HiCon 交通信号控制系统[J]. 中国交通信息产业, 2004(10): 52-55.

[14] 李群祖, 夏清国, 巴明春, et al. 城市交通信号控制系统现状与发展[J]. 科学技术与工程, 2009(24): 7436-7442.

[15] 杨兆升, 蹇峰, 胡坚明. 城市交通流诱导系统理论模型和实施技术研究——智能运输系统重要研究内容[J]. 道路交通与安全, 2003(6): 9-14.

[16] 杨晓光, 林瑜, 同济大学智能交通系统研究中心 上. TJATCMS-同济先进的交通控制与管理系统[C]. 科技部, 2008.

[17] 王桂珠, 贺国光, 马寿峰. 一种新型的自学习智能式城市交通实时控制系统[J]. 自动化学报, 1995(04): 424-430.

[18] Wu X, Deng S, Du X, et al. Green-wave traffic theory optimization and analysis[J]. World Journal of Engineering and Technology, 2014, 2(3): 14-19.

[19] Celtek S A, Durdu A, Alı M E M. Real-time traffic signal control with swarm optimization

methods[J]. Measurement, 2020, 166: 108206.

[20] Garud K S, Jayaraj S, Lee M Y. A review on modeling of solar photovoltaic systems using artificial neural networks, fuzzy logic, genetic algorithm and hybrid models[J]. International Journal of Energy Research, 2021, 45(1): 6-35.

[21] Bernal E, Lagunes M L, Castillo O, et al. Optimization of type-2 fuzzy logic controller design using the GSO and FA algorithms[J]. International Journal of Fuzzy Systems, 2021, 23: 42-57.

[22] Wei H, Zheng G, Gayah V, et al. A survey on traffic signal control methods[J]. arXiv preprint arXiv:1904.08117, 2019.

[23] Yau K-L A, Qadir J, Khoo H L, et al. A survey on reinforcement learning models and algorithms for traffic signal control[J]. ACM Computing Surveys (CSUR), 2017, 50(3): 1-38.

[24] 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.

[25] El-Tantawy S, Abdulhai B. An agent-based learning towards decentralized and coordinated traffic signal control[C]. 13th International IEEE Conference on Intelligent Transportation Systems, 2010: 665-670.

[26] Camponogara E, Scherer H F. Distributed optimization for model predictive control of linear dynamic networks with control-input and output constraints[J]. IEEE Transactions on Automation Science and Engineering, 2010, 8(1): 233-242.

[27] Wen K, Qu S, Zhang Y. A stochastic adaptive control model for isolated intersections[C]. 2007 IEEE International Conference on Robotics and Biomimetics (ROBIO), 2007: 2256-2260.

[28] Richter S, Aberdeen D, Yu J. Natural actor-critic for road traffic optimisation[J]. Advances in neural information processing systems, 2006, 19.

[29] 马跃峰, 王宜举. 一种基于 Q 学习的单路口交通信号控制方法[J]. 数学的实践与认识, 2011, 41(24): 102-106.

[30] 刘成健, 罗杰. 基于参数融合的 Q 学习交通信号控制方法[J]. 计算机技术与发展, 2018, 28(11): 48-51.

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

[32] François-Lavet V, Henderson P, Islam R, et al. An introduction to deep reinforcement learning[J]. Foundations and Trends® in Machine Learning, 2018, 11(3-4): 219-354.

[33] Lin Y, Wang P, Ma M. Intelligent transportation system (ITS): Concept, challenge and opportunity[C]. 2017 ieee 3rd international conference on big data security on cloud (bigdatasecurity), ieee international conference on high performance and smart computing (hpsc), and ieee international conference on intelligent data and security (ids), 2017: 167-172.

[34] Raeis M, Leon-Garcia A. A deep reinforcement learning approach for fair traffic signal control[C]. 2021 IEEE International Intelligent Transportation Systems Conference (ITSC), 2021: 2512-2518.

[35] Zheng G, Xiong Y, Zang X, et al. Learning phase competition for traffic signal control[C]. Proceedings of the 28th ACM international conference on information and knowledge management, 2019: 1963-1972.

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

[37] Mnih V, Kavukcuoglu K, Silver D, et al. Playing atari with deep reinforcement learning[J]. arXiv preprint arXiv:1312.5602, 2013.

[38] Schulman J, Wolski F, Dhariwal P, et al. Proximal policy optimization algorithms[J]. arXiv preprint arXiv:1707.06347, 2017.

[39] Schulman J, Levine S, Abbeel P, et al. Trust region policy optimization[C]. International conference on machine learning, 2015: 1889-1897.

[40] Yang J, Zhang J, Wang H. Urban traffic control in software defined internet of things via a multi-agent deep reinforcement learning approach[J]. IEEE Transactions on Intelligent Transportation Systems, 2020, 22(6): 3742-3754.

[41] Van Hasselt H, Guez A, Silver D. Deep reinforcement learning with double q-learning[C]. Proceedings of the AAAI conference on artificial intelligence, 2016.

[42] 晏松. 智能网联环境下复杂交叉口信号控制研究[D]. 中国人民公安大学, 2019.

[43] Wiering M A, Van Otterlo M. Reinforcement learning[J]. Adaptation, learning, and optimization, 2012, 12(3): 729.

[44] Puterman M L. Markov decision processes[J]. Handbooks in operations research and management science, 1990, 2: 331-434.

[45] Cruz F, Wüppen P, Fazrie A, et al. Action selection methods in a robotic reinforcement learning scenario[C]. 2018 IEEE Latin American Conference on Computational Intelligence (LA-CCI), 2018: 1-6.

[46] Lecun Y, Bengio Y, Hinton G. Deep learning[J]. nature, 2015, 521(7553): 436-444.

[47] Silver D, Lever G, Heess N, et al. Deterministic policy gradient algorithms[C]. International conference on machine learning, 2014: 387-395.

[48] Clifton J, Laber E. Q-learning: Theory and applications[J]. Annual Review of Statistics and Its Application, 2020, 7: 279-301.

[49] Grossberg S. Recurrent neural networks[J]. Scholarpedia, 2013, 8(2): 1888.

[50] Schaul T, Quan J, Antonoglou I, et al. Prioritized experience replay[J]. arXiv preprint arXiv:1511.05952, 2015.

[51] Graves A, Graves A. Long short-term memory[J]. Supervised sequence labelling with recurrent

neural networks, 2012: 37-45.

[52] Zhao Z, Chen W, Wu X, et al. LSTM network: a deep learning approach for short-term traffic forecast[J]. IET Intelligent Transport Systems, 2017, 11(2): 68-75.

[53] 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.

[54] Ramachandran P, Zoph B, Le Q V. Searching for activation functions[J]. arXiv preprint arXiv:1710.05941, 2017.