Operations Research and Fuzziology 运筹与模糊学, 2024, 14(2), 373-387
Published Online April 2024 in Hans. https://www.hanspub.org/journal/orf
https://doi.org/10.12677/orf.2024.142143

Hans汉斯

结合注意力机制的多智能体深度强化学习的交通信号控制

徐晴晴

上海理工大学光电信息与计算机工程学院，上海

收稿日期：2024年2月9日；录用日期：2024年2月29日；发布日期：2024年4月16日

摘　要

智能交通信号控制方法被越来越多的应用在现实世界中，并且取得了不错的成果。其中，多智能体深度强化学习是一种非常有效的方法，但是，在多交叉口交通信号控制中，大规模的交通网络容易引起严重的维度灾难，而且对于道路环境的特征提取也存在不足。针对以上问题，提出了一种新的多智能体深度强化学习算法，该算法基于双决斗深度Q网络(Double Dueling Deep Q-Network, 3DQN)，消除了传统强化学习算法对Q值的高估问题。引入了平均场(Mean Field, MF)理论大大减少了状态和动作空间的维度，同时融合了注意力机制对道路环境全面观察，使得智能体获得更准确的环境信息。在城市交通模拟器(Simulation Of Urban Mobility, SUMO)中建模了一个交通网络，模拟真实世界中的交通流，对算法进行评估。实验结果表明，提出的算法在奖励方面相较于DQN、DDPG、MA2C分别增加了64.17%、36.40%、32.55%，证明了所提算法的正确性和优越性。

关键词

多智能体深度强化学习，智能交通信号控制，平均场理论，机器学习

Traffic Signal Control Using Multi-Agent Deep Reinforcement Learning Combined with Attention Mechanism

Qingqing Xu

School of Optoelectronic Information and Computer Engineering, Shanghai University of Science and Technology, Shanghai

Received: Feb. 9th, 2024; accepted: Feb. 29th, 2024; published: Apr. 16th, 2024

文章引用: 徐晴晴. 结合注意力机制的多智能体深度强化学习的交通信号控制[J]. 运筹与模糊学, 2024, 14(2): 373-387.
DOI: 10.12677/orf.2024.142143

Abstract

Intelligent traffic signal control methods are increasingly being applied in the real world and have achieved good results. Among them, multi-agent deep reinforcement learning is a very effective method. However, in multi-intersection traffic signal control, large-scale traffic networks are prone to serious dimensional disasters, and there are also shortcomings in feature extraction of road environments. A new multi-agent deep reinforcement learning algorithm is proposed to address the above issues. This algorithm is based on the Double Dueling Deep *Q*-Network (3DQN) and eliminates the problem of overestimation of values in traditional reinforcement learning algorithms. The introduction of Mean Field (MF) theory greatly reduces the dimensions of state and action space, while integrating attention mechanisms to comprehensively observe the road environment, enabling intelligent agents to obtain more accurate environmental information. A traffic network was modeled in the Simulation of Urban Mobility (SUMO) to simulate real-world traffic flow and evaluate the algorithm. The experimental results show that the proposed algorithm has increased rewards by 64.17%, 36.40%, and 32.55% compared to DQN, DDPG, and MA2C, respectively, proving the correctness and superiority of the proposed algorithm.

Keywords

Multi Agent Deep Reinforcement Learning, Intelligent Traffic Signal Control, Mean Field Theory, Machine Learning

1. 引言

近年来，新型城镇化取得重大进展，它带来机遇动力的同时，也使我们面临一些问题和挑战。其中，城镇有限的道路资源和汽车数量的持续增长导致交通效率持续低下、拥堵日益严重。随着大数据和人工智能技术的快速发展，交通信号的智能控制已经在许多大城市推广，并且都获得了令人满意的效果。例如，上海市在浦西世博园区、黄浦泛外滩地区等试点应用的智能交通信号灯管理系统，使道路通行效率平均提升 10% [1]。据统计，近年来上海道路交通事故数、死亡人数、受伤人数持续下降，2018 年同比分别下降了 6.3%、5.2%和 6.8%，创历史新低，日均交通类 110 警情同比下降 12.4% [1]。与传统的交通信号控制方法，如 Webster 固定配时系统[2]、MOVA [3]、SCATS [4]等相比，智能交通信号控制可以根据交通流量和需求自动调整交通信号，实现交通的高效运行。

其中，强化学习(Reinforcement Learning, RL) [5]作为一种最为可行的机器学习技术在智能交通信号控制领域被广泛应用。强化学习可以直接与环境交互，根据学习到的策略选择最佳动作。具体的方法就是将路网的交叉口建模为智能体，智能体观察周围环境的状态(交通状况)，根据状态执行动作(改变交通信号)，此时环境会转移到一个新的状态，同时将刚才动作所获得的奖励(车辆的队列长度、车辆的等待时间、车辆数目等)反馈给智能体，智能体会根据奖励的好坏优化自己的策略。这个过程会被反复执行，直到智能体学习到最优策略。近年来，已经有一些有效的方法将强化学习应用到智能交通信号控制中[6]。一开始大部分研究人员将注意力放在了单个交叉口上，例如，Li 等人[7]将 Q-learning 与深度堆叠自动编码器(Deep Stacked Autoencoders, SAE)神经网络结合，其中，输入状态和奖励是根据进站车道的排队长度来定

义的，仿真结果表明，与传统的强化学习方法相比，深度强化学习方法可以将平均交通延迟减少 14%。Luo 等人[8]提出了一种自适应道路划分策略和基于深度 Q 网络(Deep Q Network, DQN)的改进神经网络模型。改进了传统的交通信号控制模型，考虑到车辆与交通灯之间的距离对智能体决策的影响，提出了一种基于 Fibonacci 序列的自适应道路划分方法，以获得更真实的状态。Wang 等人[9]提出了一种基于 3DQN 的交通信号控制方法，使用了基于高分辨率事件的数据，该方法以两种常用的交通信号控制策略为基准进行对照实验，即固定时间控制策略和驱动控制策略，实验证明所提出的基于事件数据的状态表示优于一般状态表示。虽然上述工作在单交叉口的信号控制上取得了很好的效果，有效地缓解了交通拥堵情况，但是忽略了其他交叉口对当前交叉口的影响，即多交叉口之间的协调问题，也难以在实际中应用。因此，越来越多的研究者开始关注多交叉口信号控制问题。

对于多个交叉口的交通信号控制问题，在强化学习的基础上，一个简单的想法就是将每个交叉口都建模为一个智能体，每个智能体都能以试错的方式与其所在环境和相邻的智能体交互，学习全局最优策略。这种方法称为多智能体强化学习(Multi-Agent Reinforcement Learning, MARL) [10]。在最近的一些成果中，研究者们将深度学习(Deep Learning, DL)与多智能体强化学习结合起来，形成了多智能体深度强化学习(Multi-agent Deep Reinforcement Learning, MADRL)。深度学习能够使用神经网络高效地表示和存储高维状态、动作和奖励，从而使 MARL 快速收敛。Haddad 等人[11]在 DQN 的基础上提出了一种多路口协同控制的方法，每个智能体通过接收相邻智能体的状态、动作和奖励来实现合作。Chu 等人[12]提出了一种新的基于优势演员评论家(Advantage Actor Critic, A2C)的 MARL 算法(MA2C)，通过邻居的指纹信息提高观测性，同时引入了空间折扣因子降低了学习维度，该算法在合成交通网络和现实世界路网中均被证明是有效的。Wu 等人[13]在深度确定性策略梯度(Deep Deterministic Policy Gradient, DDPG)算法的基础上，提出了一种新的多智能体递归深度确定性策略梯度(Multi-agent Recurrent Deep Deterministic Policy Gradient, MARDDPG)算法，它是通过车辆网络收集时空交通信息实现智能体之间的合作，使用集中学习、分散执行的策略，还利用 LSTM 来捕获隐藏的状态信息，实验证明该算法适用于中等规模的流量网络。但是，上述研究中都存在一个问题，即维数诅咒问题——离散状态-动作空间中状态和动作变量的数量呈指数增长。在 MARL 中，由于存在多个智能体，每个智能体都会将自己的变量添加到联合状态－动作空间中，所以 MARL 中的维数诅咒比单智能体强化学习更加严重。除此之外，由于所有的智能体都在同时学习，它们的最优策略都在随着其他智能体的策略变化而变化，所以在 MARL 中还存在着非平稳性问题。最后，探索和利用的权衡问题也应该特别关注，这种权衡是需要在线强化学习算法在利用智能体已学到的知识和为了提升该知识而采取的探索性之间取得平衡。尤其是在 MARL 中，为了适应其他智能体的动作，智能体不仅要探索环境信息，还要探索其他智能体的信息。但是，过多的探索又会破坏其他智能体的稳定，从而使探索智能体的学习任务变得更加困难。

为了应对在 MARL 中面临的上述挑战，Wang 等人[14]采用了基于双 Q 学习和上置信界(Upper Confidence Bound, UCB) [15]算法的探索策略，在保证探索的同时消除了传统独立 Q 学习的高估问题。将多个智能体之间的交互基于平均场理论[16]建模，使智能体能够学习到更好的合作策略，降低了动作空间的维度和计算复杂度。Hu 等人[17]也在双 Q 学习的基础上结合 MFT 降低联合动作空间的维度，其使用 Boltzmann 策略来平衡探索和利用之间的关系。该方法还将递归神经网络与改进的传统交通方法相结合，动态确定相位持续时间，显著提高了交通性能。但是上述研究都没能有效的解决 Q 值的高估问题，只关注到了智能体及其相邻智能体的动作，缺少对交通环境在时间和空间上的观察。

本文在上述研究的基础上提出了一种新的带有注意力机制的平均场双决斗深度 Q 网络算法(MF3DQN-TF)。首先，我们使用 Transformer 模型[18]中的注意力机制对交通环境更全面的观察，为智能体提供更高效的表征结果。然后引入平均场理论，将多个智能体之间的交互近似为单个智能体与其邻居

智能体的平均场虚拟智能体的二元交互，以降低联合动作空间的维度。接着使用 3DQN 算法能够有效的解决先前研究对 Q 值的高估问题。最后使用 SUMO 作为仿真平台，在具有多交叉口的交通路网中模拟真实世界中的交通流，验证本文所提出的算法，根据设计的性能指标与几种最先进的 MARL 算法进行对比实验。综上所述，本文的主要贡献如下：

1) 对 DQN 算法进行一些改进，包括双重网络结构、经验池抽样策略和优势函数，有效地解决了 Q 值的高估问题，从而使得模型快速收敛；

2) 引入了平均场理论，与 3DQN 算法相结合，将先前研究中与所有邻居智能体交换知识(如奖励、动作、状态、Q 值等)改进为与所有邻居智能体的平均知识进行共享，大大降低了智能体的状态动作空间的维度，解决了 MADRL 中收敛困难的问题；

3) 添加了自注意力层分析交叉口处的交通信号灯与各个入站车道之间的关联规则，将分析的结果作为状态输入，有助于智能体从分析后的数据中更快的学习到控制交通信号的策略；

4) 在 SUMO 中建立了包含多个交叉口的交通网络，模拟了更加符合现实世界的交通流，有利于扩展到真实交通网络，实验结果证明，提出的交通信号控制模型优于现有的流行方法。

2. 研究背景

2.1. 多智能体深度强化学习

多智能体强化学习的实质就是一个基于马尔可夫决策过程的随机博弈，多个智能体通过合作或者竞争获取最大利益的过程。假设共有 N 个智能体，多智能体系统可以建模为 $\langle S,A,P,R,\gamma\rangle$，其中，$S=(s_1,s_2,\cdots,s_N)$ 表示系统的联合状态空间，$A=(A_1,A_2,\cdots,A_N)$ 表示系统的联合动作空间，P 表示联合状态转移函数，$R=(R_1,R_2,\cdots,R_N)$ 表示每个智能体的奖励函数，γ 代表折扣因子。与单智能体强化学习类似，它们的目标都是学习到最优策略 π^*，使得累积预期回报最大化。不同的是在多智能体强化学习中，每个智能体都有自己的策略 π，所以每个智能体都必须要考虑其他智能体的动作，在顾全大局的情况下，学习到最优策略，即 $\boldsymbol{\pi}^*=(\pi_1^*,\pi_2^*,\cdots,\pi_N^*)$，这就是纳什均衡[19]。

假设环境的初始状态为 s，那么智能体 k 在联合策略 $\boldsymbol{\pi}=(\pi_1,\pi_2,\cdots,\pi_N)$ 下的价值函数被定义为累积期望折扣奖励：

$$V_k^{\boldsymbol{\pi}}(s)=V_k(s;\boldsymbol{\pi})=\mathrm{E}_{\boldsymbol{\pi},p}\left[\sum_{t=0}^{\infty}\gamma^t r_k(t+1)\middle|s_0=s,\boldsymbol{\pi}\right] \tag{1}$$

根据贝尔曼方程，智能体 k 的动作价值函数，也称为 Q 函数定义如下：

$$Q_k^{\boldsymbol{\pi}}(s,\boldsymbol{a})=r_k(s,\boldsymbol{a})+\gamma\mathrm{E}_{s_{t+1}\sim p}\left[V_k^{\boldsymbol{\pi}}(s')\right] \tag{2}$$

其中，$\boldsymbol{a}=(a_1,a_2,\cdots,a_N)$ 是在当前状态 s 下所有智能体的联合动作，s' 是下一个时间步长时环境的状态，$r_k(s,\boldsymbol{a})$ 为智能体 k 的奖励函数。因此，价值函数 $V_k^{\boldsymbol{\pi}}$ 可通过等式(2)中的 Q 函数表示：

$$V_k^{\boldsymbol{\pi}}(s)=\mathrm{E}_{\boldsymbol{a}\sim\boldsymbol{\pi}}\left[Q_k^{\boldsymbol{\pi}}(s,\boldsymbol{a})\right] \tag{3}$$

由等式(1)和等式(2)可以看出每个智能体的价值函数与所有智能体的联合策略 $\boldsymbol{\pi}$ 有关，根据纳什均衡有：

$$V_k(s,\boldsymbol{\pi}^*)=V_k(s,\pi_k^*,\boldsymbol{\pi}_{-k}^*)\geq V_k(s,\pi_k,\boldsymbol{\pi}_{-k}^*) \tag{4}$$

等式(4)说明在这种情况下，只要其他智能体保持自身策略不变，任何智能体都不能通过改变其策略而获得更好的累积期望奖励。其中，$\boldsymbol{\pi}_{-k}^*$ 表示除了智能体 k 以外，其他智能体的最优联合策略。同时，Q

函数最终也会收敛到纳什 Q 值 $\boldsymbol{Q}^{*}=\left(Q_{1}^{*},Q_{2}^{*},\cdots,Q_{N}^{*}\right)$。

在单智能体强化学习系统中，就因为状态空间和动作空间的维数增加引入了深度学习，利用深度神经网络去近似 Q 函数，形成了深度强化学习。那么在多智能体强化学习系统中，随着智能体数目的增加，系统的状态空间和动作空间的维数都呈指数型增长，环境变得更加复杂，在这种情况下，深度强化学习方法仍然是一种可行的方法。但是，对每个智能体直接近似 Q 函数并不能解决维数灾难问题，反而会使系统难以收敛。因此，本文将引入平均场理论来降低动作空间的维数，以处理 MADRL 中收敛困难的问题。

2.2. 平均场理论

在物理学和概率论学中，平均场理论通过研究一个简单的模型来表示高维随机模型的行为，通过对原始模型的自由度取平均来近似得到。文献[16]最先将 MFT 与多智能体强化学习算法相结合，利用平均场近似方法解决大规模群体问题。其核心思想是将环境中的多个智能体之间的相互作用转化为一个智能体与其相邻智能体的平均值之间的相互作用。该方法大大降低了联合动作空间的维度。具体的，可以将智能体 k 的 Q 函数分解为：

$$Q_{k}\left(s,\boldsymbol{a}\right)=\frac{1}{N_{k}}\sum_{j\in\mathcal{N}(k)}Q_{k}\left(s,a_{k},a_{j}\right)\approx Q_{k}\left(s,a_{k},\overline{a}_{k}\right) \tag{5}$$

其中，$\mathcal{N}\left(k\right)$ 是智能体 k 的邻居智能体的索引集合，$N_{k}=\left|\mathcal{N}\left(k\right)\right|$ 是邻居智能体的个数，$\overline{a}_{k}=\frac{1}{N_{k}}\sum_{j\in\mathcal{N}(k)}a_{j}$ 代表智能体 k 的邻居智能体的平均动作。等式(5)的详细证明过程可以参考文献[16]。

2.3. 平均场双决斗 Q 学习

迄今为止，深度学习和深度强化学习领域都有了最新进展，为了实现等式(5)中的平均场 Q 函数，可以使用深度神经网络去近似。目前存在许多先进的深度强化学习算法，大致可以分为两大类：基于价值的方法和基于策略的方法。具有代表性的如 DQN、DDPG [20]、PPO [21]等。由于本文的动作空间是离散的(关于动作空间的定义将在下文详细说明)，所以本文选择了基于价值的方法。在基于价值的方法中，DQN 被认为是深度强化学习的开山之作，但是原始的 DQN 算法会导致 Q 值的高估，还会影响训练过程的稳定性。因此本文中使用一些技巧去解决上述问题，即双决斗深度 Q 网络。首先利用 Double DQN 去计算智能体 k 目标网络的 Q 值：

$$Y_{k}=r_{k}+\gamma Q_{k}\left(s_{k}^{\prime},\arg\max_{a_{k}}Q_{k}\left(s_{k}^{\prime},a_{k},\overline{a}_{k};\theta\right),\overline{a}_{k}^{\prime};\theta^{-}\right) \tag{6}$$

其中，θ 是智能体 k 主网络的参数，$\overline{a}_{k}^{\prime}$ 为下一个时间步长时智能体 k 的邻居智能体的平均动作，θ^{-} 是智能体 k 目标网络的参数。接着改变网络结构使学习过程更加稳定，即双决斗网络结构：

$$\begin{aligned}Q_{k}\left(s_{k},a_{k},\overline{a}_{k};\theta\right)=&V_{k}\left(s_{k};\theta\right)+A_{k}\left(s_{k},a_{k},\overline{a}_{k};\theta\right)\\&-\frac{1}{\left|A_{k}\right|}\sum_{a_{k}^{\prime}}A\left(s_{k},a_{k}^{\prime},\overline{a}_{k}^{\prime};\theta\right)\end{aligned} \tag{7}$$

最后可以通过最小化损失函数训练平均场双决斗 Q 函数，损失函数定义为：

$$L\left(\theta\right)=\mathrm{E}\left[\left(Y_{k}-Q_{k}\left(s_{k},a_{k},\overline{a}_{k};\theta\right)\right)^{2}\right] \tag{8}$$

3. 结合注意力机制的平均场双决斗深度 Q 网络交通信号控制模型

多智能体系统环境复杂，大多数文献只提取了道路环境的表面特征，缺少对交通特征之间关系的深入分析。因此，利用 Transformer 中的注意力机制捕获每个交叉口的交通特征之间的相关性，并将其作为模型的输入状态，使智能体全面了解环境信息。此外，本节中还定义了模型的动作空间、奖励函数以及所提的 MF3DQN-TF 算法。

3.1. 状态空间

从道路环境中提取的观测值必须能帮助智能体获得足够的信息，以做出正确的决策。对于第 k 个交叉口，它的部分观测值为 $\boldsymbol{O}_k=\left(\boldsymbol{o}_k^1,\boldsymbol{o}_k^2,\boldsymbol{o}_k^3,\boldsymbol{o}_k^4,p_k\right)$，其中 $\boldsymbol{o}_k^i$ 是第 $i\left(i=1,2,3,4\right)$ 条入站道路的观测信息，$\boldsymbol{o}_k^i=\left\{d_k^i[l],v_k^i[l]\right\}$，$d_k^i[l]$ 表示第 i 条入站道路中 l 车道上第一辆车的累积延迟时间(s)，$v_k^i[l]$ 表示第 i 条入站道路中 l 车道上的车辆数目(veh)，p_k 是交叉口 k 当前观测下的交通相位。接下来使用 Transformer 中的自注意力机制(Self-Attention)对交叉口 k 的部分观测值进行处理，结构如图 1 所示。本文将 p_k 作为计算查询向量的输入值，是为了捕获当前相位对于每条入站车道的注意力系数，然后将这些注意力系数分配给每条入站车道。注意力值高的入站车道，就会获得智能体的重点关注。

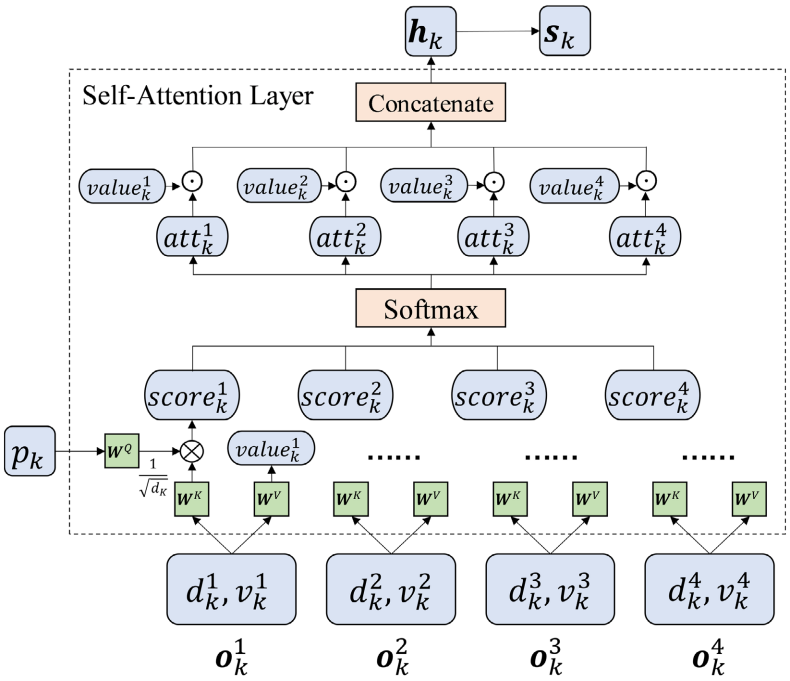

Figure 1. Single intersection observation value processing model based on Self-Attention
图 1. 基于 Self-Attention 的单交叉口观测值处理模型

具体的，假设存在三个线性变换矩阵 $\boldsymbol{W}^Q,\boldsymbol{W}^K,\boldsymbol{W}^V$，交叉口 k 的第 i 条入站道路的观测值 $\boldsymbol{o}_k^i$ 用来计算键值向量和值向量：

$$\boldsymbol{key}_k^i=\boldsymbol{o}_k^i\times\boldsymbol{W}^K \tag{9}$$

$$\boldsymbol{value}_k^i=\boldsymbol{o}_k^i\times\boldsymbol{W}^V \tag{10}$$

查询向量计算方式如下：

$$\boldsymbol{query}_k = \boldsymbol{p}_k \times \boldsymbol{W}^Q \tag{11}$$

交叉口 k 的当前相位对入站道路 i 的注意力得分为：

$$\boldsymbol{score}_k^i = \frac{\boldsymbol{query}_k \times \left(\boldsymbol{key}_k^i\right)^T}{\sqrt{d_K}} \tag{12}$$

其中，d_K 是键值向量 $\boldsymbol{key}_k^i$ 的维度。为了提高可解释性和平滑性，使用 Softmax 函数对所有的注意力得分进行处理，最终得到了所有入站道路的注意力系数：

$$att_k^1, \cdots, att_k^4 = \text{Softmax}\left(score_k^1, \cdots, score_k^4\right) \tag{13}$$

然后将这些注意力系数和值向量点乘，就得到了所有入站道路的注意力值：

$$\boldsymbol{h}_k^i = \boldsymbol{value}_k^i \cdot \boldsymbol{att}_k^i \tag{14}$$

$$\boldsymbol{h}_k = \text{Concatenate}\left[h_k^1, h_k^2, h_k^3, h_k^4\right] \tag{15}$$

这些入站道路的注意力值就作为智能体进行决策的输入状态：

$$\boldsymbol{s}_k = \boldsymbol{h}_k \tag{16}$$

除了将交叉口 k 自身的状态作为输入状态外，本文还将交叉口 k 的所有邻居交叉口的状态的平均值作为附加输入。所以交叉口 k 的联合状态为：

$$\hat{\boldsymbol{s}}_k = \left(\boldsymbol{s}_k, \frac{1}{N_k}\sum_{j \in \mathcal{N}(k)} \boldsymbol{s}_j\right) \tag{17}$$

3.2. 动作空间

根据模拟实验的道路环境，本文设置了如图 2 所示的交通相位。其中，*NSG* 表示南北方向的车辆可以直行和右转，*NSLG* 表示南北方向的车辆可以左转和右转，*EWG* 表示东西方向的车辆可以直行和右转，*EWLG* 表示东西方向的车辆可以左转和右转。为了保证司机在切换相位之前有时间反应，所以在切换相位之前设置了黄灯时间。基于此，定义动作空间为 $A = \{NSG, NSLG, EWG, EWLG\}$，智能体可以直接从 A 中选择合适的相位，同时采用独热编码的形式表示相应的动作。例如，在时间步长 t 时智能体 k 选择了 *NSG* 相位，此时的动作表示为 $a_{k,t} = [1,0,0,0]$。

如果将要选择的动作与当前动作相同，那么继续保持当前动作 τ_g s。如果不相同，则需要先执行 τ_y s 的黄灯时间，再继续执行所选择的动作 τ_g s。

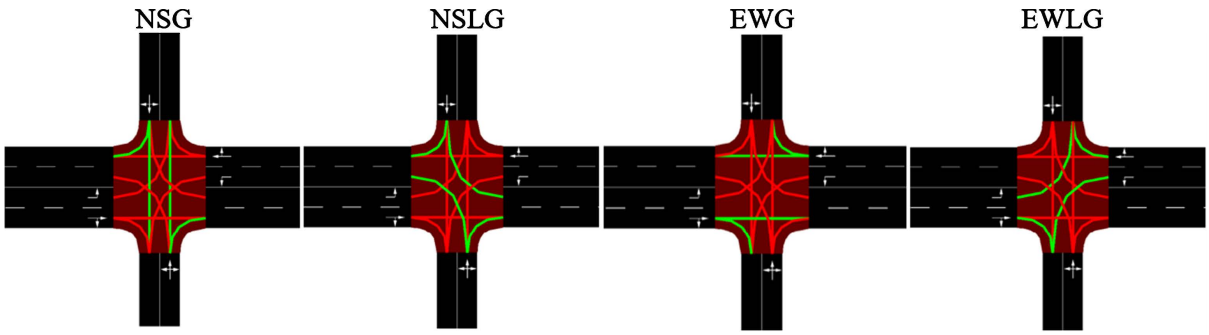

Figure 2. Intersection phase
图 2. 交叉口相位

3.3. 奖励函数

奖励函数的定义非常重要，它可以直接反映所提的算法是否有效。为了直观的展示交通拥堵状况，本文利用入站道路上车辆的排队长度和等待时间定义奖励函数：

$$r_{k,t} = -\sum_{i=1}^{4}\left(queue[i] + \beta \cdot wait[i]\right) \tag{18}$$

其中，$queue[i]$是指入站道路 i 上车辆的队列长度，$wait[i]$是指入站道路 i 上车辆的等待时间。β 是权衡系数，为了使 $queue[i]$和 $wait[i]$在同一范围内。

3.4. 交通信号控制模型

本小节主要介绍基于 MF3DQN-TF 算法的交通信号控制模型，模型的结构如图 3 所示。首先，需要对 Self-Attention Layer 进行预训练，使其具有捕获当前相位对于每条入站道路的注意力系数的能力。然后将从 Self-Attention Layer 获得的交叉口的联合状态输入给智能体，其中每个智能体通过 MF3DQN 学习选择最优动作。MF3DON 将交叉口的联合状态和邻居交叉口的平均动作作为联合状态动作对，其网络结构如图 4 所示，用于估计 Q 值。为了使估计的 Q 值更加精确，本文使用了双估计器，即 Double DQN，此时智能体 k 目标网络的 Q 值为：

$$Y_k = r_k + \gamma Q_k\left(\hat{s}_k', \arg\max_{a_k} Q_k\left(\hat{s}_k', a_k, \bar{a}_k; \theta\right), \bar{a}_k'; \theta^-\right) \tag{19}$$

并且通过最小化损失函数训练智能体 k 的 Q_k 值，使用均方误差定义损失函数：

$$L(\theta) = \frac{1}{M}\sum\left[Y_k - Q_k\left(\hat{s}_k, a_k, \bar{a}_k; \theta\right)\right]^2 \tag{20}$$

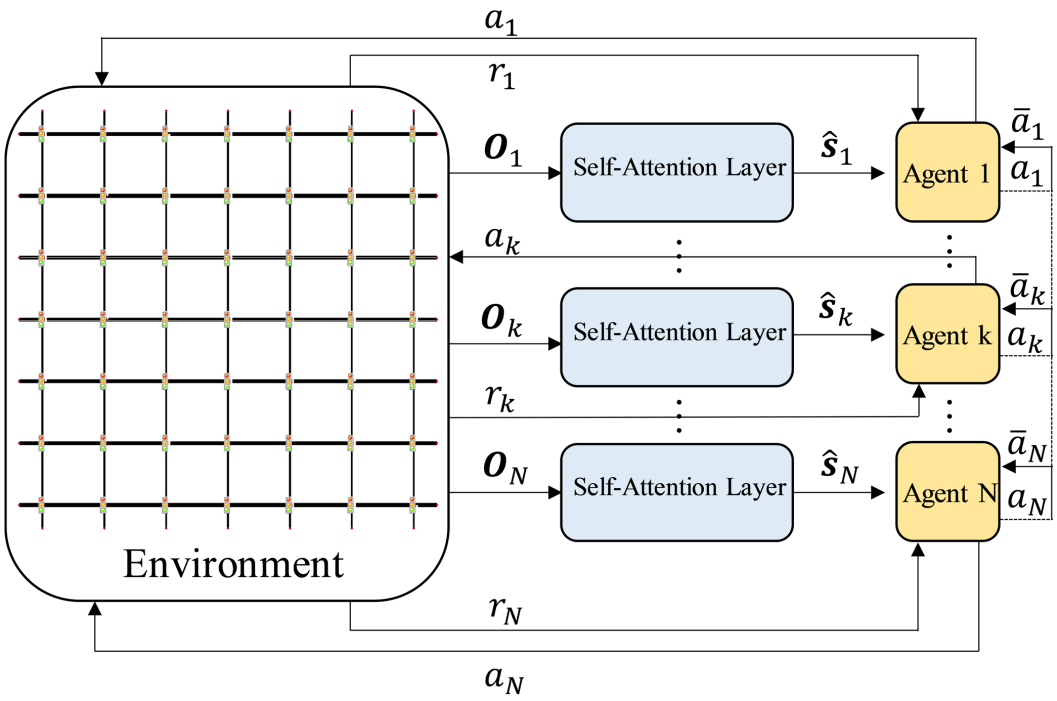

Figure 3. Structure diagram of traffic signal control model
图 3. 交通信号控制模型结构图

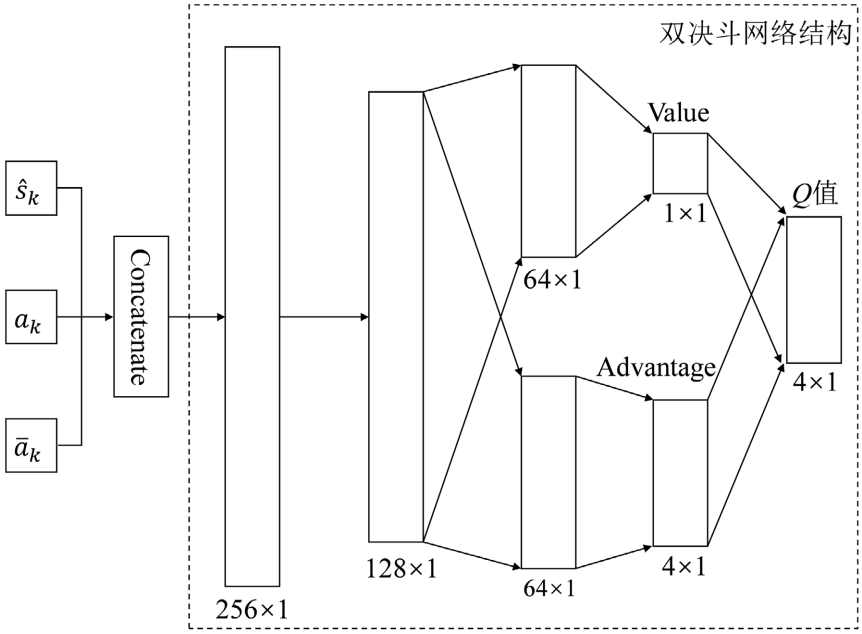

Figure 4. Network Structure of MF3DQN
图 4. MF3DQN 的网络结构

其中，M 是指从经验池中随机抽取的样本数量。为了最小化损失函数，本文使用自适应矩估计(Adaptive Moment Estimation, Adam)优化器更新主网络的模型参数 θ。对于目标网络的参数 θ^-，我们采用软更新的方法，即目标网络的参数缓慢地向主网络的参数靠近，而不是直接复制当前主网络的参数。具体更新方式如下：

$$\theta^- = (1-\tau)\theta^- + \tau\theta \tag{21}$$

其中，$\tau \in (0,1)$，用于控制更新速度，使 θ^- 更加平滑的更新，从而提高模型的稳定性。

另外，在智能体学习的过程中，探索与利用的平衡问题也需要特别关注。一些比较经典的探索与利用算法包括 ϵ 贪婪策略、上界置信算法(Upper Confidence Bounds, UCB)。文献[22]已经证明 UCB 算法在基于多智能体强化学习的交通信号控制上优于 ϵ 贪婪策略，基于此，本文选择使用 UCB 算法去选择智能体 k 执行的动作：

$$a_k = \arg\max_{c \in A_k} \left\{ Q_k(s_k, c) + \sqrt{\frac{\ln R_{s_k}}{R_{s_k,c}}} \right\} \tag{22}$$

其中，R_{s_k} 表示访问状态 s_k 的次数，$R_{s_k,c}$ 表示目前为止在状态 s_k 下动作 c 被选择的次数。

因此，提出了 MF3DQN-TF 算法，伪代码如算法 1 所示。

算法 1：MF3DQN-TF 算法

输入：为所有智能体 $k \in \{1,2,\cdots,N\}$ 初始化主网络参数 θ_k、目标网络参数 θ_k^- 和平均动作 $\bar{a}_k$；初始化经验回放器 $\mathcal{M}$。

a) 预训练 Self-Attention Layer 模块。

b) for $episode = 1$ to E do

c) 初始化所有交叉口的观测值 $\boldsymbol{O}_{k,0}$

d) while $t < T$ do

e) 利用 Self-Attention 处理交叉口的观测值，得到交叉口的联合状态 $\hat{\boldsymbol{s}}=\left(\hat{\boldsymbol{s}}_1,\cdots,\hat{\boldsymbol{s}}_N\right)$

f) 根据公式(22)为每个智能体 k 选择动作 a_k

g) 执行联合动作 $\boldsymbol{a}=\left(a_1,\cdots,a_N\right)$，观察得到的奖励 $\boldsymbol{r}=\left(r_1,\cdots,r_N\right)$ 和交叉口的新观测值 $\boldsymbol{O}'=\left(\boldsymbol{O}'_1,\cdots,\boldsymbol{O}'_N\right)$，并利用 Self-Attention 得到新的联合状态 $\hat{\boldsymbol{s}}'=\left(\hat{\boldsymbol{s}}'_1,\cdots,\hat{\boldsymbol{s}}'_N\right)$

h) 计算平均联合动作 $\overline{\boldsymbol{a}}$

i) 将样本 $\left(\hat{\boldsymbol{s}},\boldsymbol{a},\boldsymbol{r},\hat{\boldsymbol{s}}',\overline{\boldsymbol{a}}\right)$ 存储到经验回放器 $\mathcal{M}$ 中

j) for $k=1$ to N do

k) 从经验回放器 $\mathcal{M}$ 中随机取出 M 个样本

l) 根据公式(19)计算智能体 k 的目标值

m) 通过最小化公式(20)更新主网络的模型参数 θ

n) 每隔 C 次迭代，根据公式(21)更新目标网络的模型参数 θ^-

o) end for

p) end while

q) end for

4. 模拟实验

本节中，在多交叉口的情景下评估本文所提模型，并与其他流行的多智能体深度强化学习算法进行对比实验。实验仿真平台是在交通信号控制领域中广泛使用的 SUMO。

4.1. 实验设置

为了更加准确的验证本文所提模型，本实验参考现实世界中的城市道路网络，在 SUMO 中设计了一个包含 49 个交叉口的交通网络环境，如图 5 所示，其中每条道路的长度为 200 米。每个交叉口详细地车道划分以及交通相位设置如图 2 所示，$\tau_g=5s$，$\tau_y=2s$。网络中车辆的长度为 5 米，加速度为 5 m/s^2，减速度为 10 m/s^2。在仿真环境中，交通流量的设计非常重要，为了模拟现实世界中的车流量，设置了四组不同的交通流，每组交通流的峰值分别为 1100 vec/h、660 vec/h、925 vec/h、555 vec/h。由于实验中每个回合的模拟时间为 60 min，所以每个时间段设置不同的交通流，前两组交通流在前 40 min 每隔 5 min 模拟一次，模拟的交通流分别为峰值的[0.4, 0.7, 0.9, 1.0, 0.75, 0.5, 0.25]倍，后两组交通流在 15~55 min 内每隔 5 min 模拟一次，模拟的交通流分别为峰值的[0.3, 0.8, 0.9, 1.0, 0.8, 0.6, 0.2]倍。

实验中对每一种 MARL 算法都训练了 700 个回合，每个回合包含 3600 个时间步长，详细的算法参数见表 1。除此之外，对于预训练 Self-Attention Layer 模块，采用了 Transformer 的结构去训练。数据集为仿真交通网络中随机生成的一个回合的数据，共 720 × 49 组数据，这些数据按照 8:2 的比例分为训练集和验证集。共训练 200 个回合，在训练的过程中，损失函数使用均方误差函数计算，选择了学习率为 0.0001 的 Adam 优化器。训练好的 Self-Attention Layer 模块用来预测当前相位下每条入站车道的注意力值，为 MF3DQN-TF 的联合状态提供观测值。

4.2. 对比方法

为了证明本文所提模型的正确性和有效性，本实验将 MF3DQN-TF 算法与下列流行的 MARL 算法在相同的流量场景下进行比较。对于所有的算法，状态空间、动作空间和奖励函数的设置和本文定义的完全相同。

DQN：最原始的 DQN 算法，所有超参数的设置和本文完全相同，和 MF3DQN-TF 算法相比，网络架构和目标 Q 值的计算方法都不一样，而且没有注意力模块处理观测值。

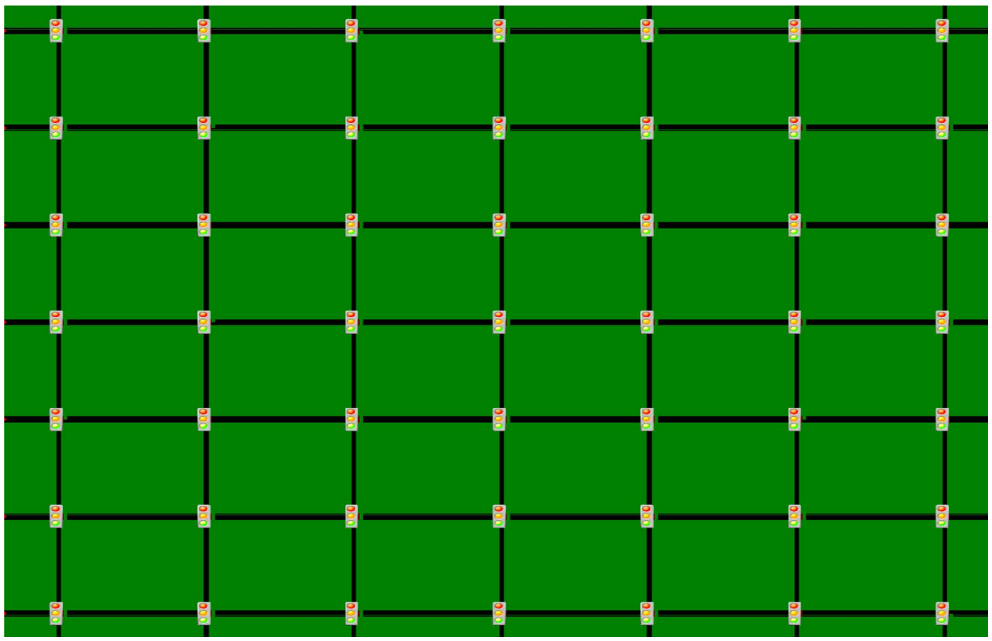

Figure 5. Traffic network simulation environment
图 5. 交通网络仿真环境

Table 1. Parameters of the algorithm
表 1. 算法的参数

参数名称	值
经验回放器 $\mathcal{M}$ 的空间	100,000
训练总回合数 E	700
时间步长 T	3600
智能体总数量 N	49
奖励权衡系数 β	0.2
批样本 M	1024
折扣因子 γ	0.95
目标网络参数更新次数 C	5
软更新参数 τ	0.99
Adam 学习率 lr	1e−4

DDPG：结合了深度强化学习和确定性策略梯度方法，由 Actor 网络和 Critic 网络组成，Actor 用来生成动作，Critic 网络用于估计状态价值函数。除了 Actor 网络的优化器学习率为 0.0001，Critic 网络的优化器学习率为 0.001 外，其余超参数和本文设置的相同。

MA2C [12]：在交通信号控制领域中一种比较先进的基于策略梯度的 MARL 算法，智能体之间通过共享经验以学习全局最优策略。其网络结构与文献[12]中完全一致，其余超参数设置和本文设置的相同。

4.3. 实验结果与分析

在多智能体深度强化学习算法中，除了奖励函数这一评价指标外，本文还设置了车辆等待时间、车

辆行驶速度和交叉口队列长度作为交通性能指标。

图 6 展示了各算法在训练过程中的奖励曲线(阴影部分表示标准差),本文的算法目标是为了最大化奖励函数。从图 6 中可以发现，在复杂的交通环境中，DQN 算法表现最差，整体曲线呈现下降趋势，大部分奖励值一直处于小于−2500 的范围，说明 DQN 算法并没有完全学习到环境的规则，其简单的算法结构无法适应复杂的交通环境。相比之下，DDPG 算法和 MA2C 算法的奖励曲线远远优于 DQN 算法，其中，MA2C 的奖励值范围在−2000 到−1000 之间，DDPG 算法的奖励曲线在整个训练过程中都比较平稳，奖励值一直维持在−1500 左右，也就说明它无法找到更好的策略，遇到更加复杂的环境难以学习最优策略。较为先进的 MA2C 算法相比于 DDPG 并没有明显的优势，这是由于 MA2C 算法对于智能体的数量非常敏感，而且涉及到的超参数也非常多，寻找到所有的最优超参数非常困难。本文提出的 MF3DON-TF 算法在训练开始阶段学习到的奖励就优于其他对比算法，奖励值直接上升到了−1000 左右，此后一直在大于−1000 的范围，虽然在训练后期有些波动，但是这些波动表明了算法仍然在探索，试图寻找更好的策略，从曲线的 650 回合到 700 回合中可以看到，MF3DQN-TF 算法收敛到了−500，证明 MF3DQN-TF 算法确实找到了比之前更好的策略，说明本文提出的算法是有效的，且优于其他对比算法。

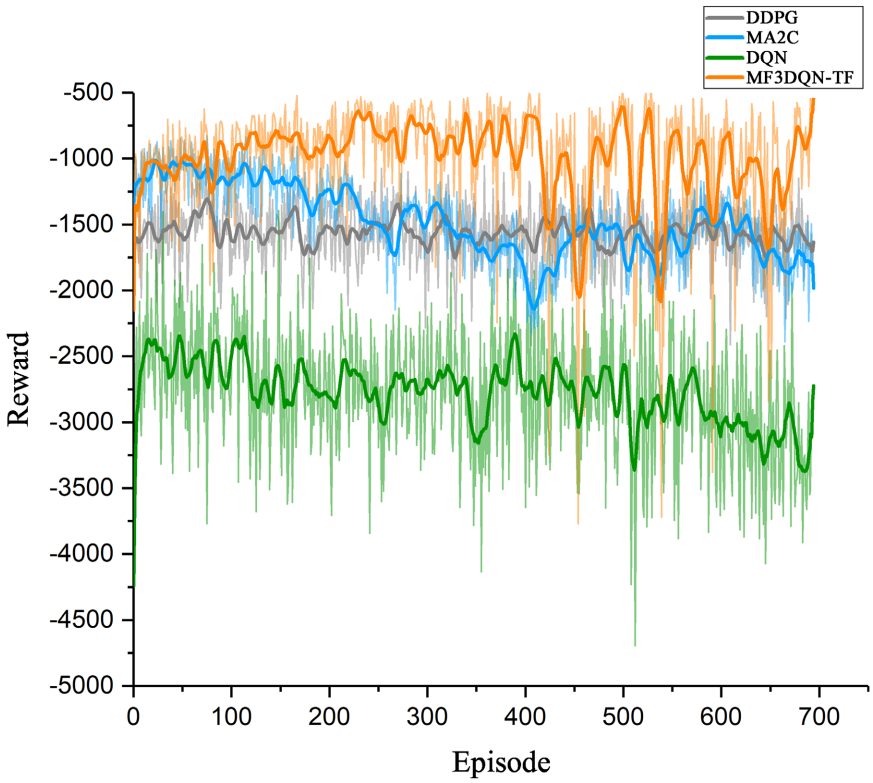

Figure 6. Reward curves of each algorithm during the training process (larger is better)
图 6. 训练过程中各算法的奖励曲线(越大越好)

除了奖励曲线外，一些交通指标也能说明本文所提算法的正确性和有效性。图 7 展示了训练过程中各算法下车辆的等待时间。DQN 算法的表现仍然是最差的，车辆的等待时间超过了 100 s，相较于其他算法来说非常久。DDPG 算法下的车辆等待时间一直维持在 50 s 左右，没有比较明显的改善。MA2C 算法在训练初始阶段时车辆的等待时间比较短，达到了 45 s，但是到了后期，有可能是遇到了比较拥堵的情况，无法执行正确的动作，导致车辆的等待时间增加到了 50 s 以上。虽然 MF3DQN-TF 算法也遇到了

这种情况，但是该算法的学习能力非常强，有了这种经验后，很快就将车辆的等待时间缩短到了正常水平，而 MA2C 算法却很难使车辆再回到较短的等待时间。

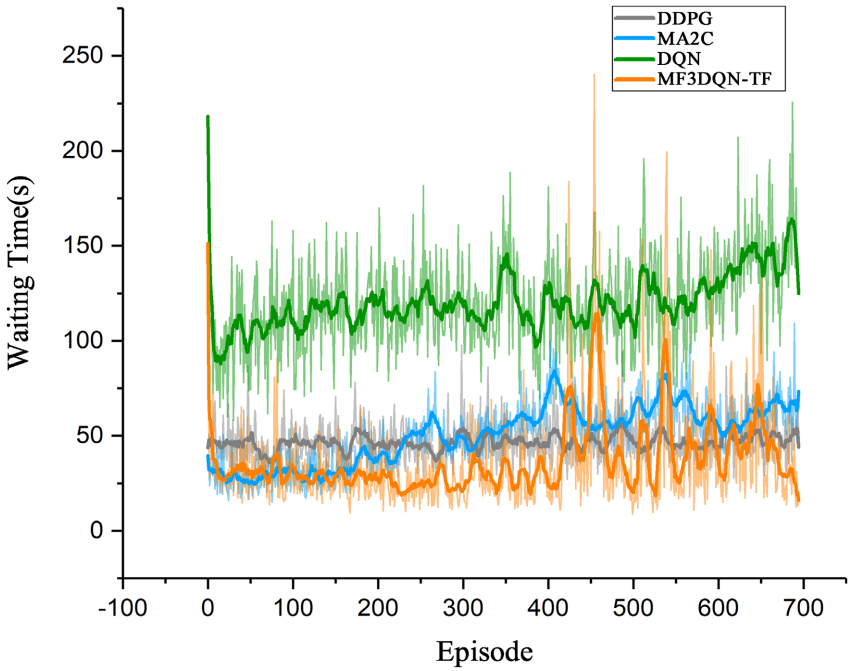

Figure 7. Waiting time curves of each algorithm during the training process (smaller is better)
图 7. 训练过程中各算法的等待时间曲线(越小越好)

此外，本文还总结了各算法的平均交通性能指标，如表 2 所示。本文所提出的 MF3DQN-TF 算法的平均车辆等待时间相较于 DQN 减少了 69.80%，相较于 DDPG 减少了 22.28%，相较于 MA2C 减少了 28.13%。就平均车辆速度来说，MF3DQN-TF 相较于 DQN 增加了 15%，相较于 DDPG 增加了 27.92%，相较于 MA2C 增加了 30%。对于平均队列长度，MF3DQN-TF 相较于 DQN 和 DDPG 减少了 33.33%。

Table 2. Average traffic performance metrics of each algorithm
表 2. 各算法的平均交通性能指标

算法	平均车辆等待时间(s)	平均车辆速度(m/s)	平均队列长度(vehicles)
DQN	120.10	2.04	3
DDPG	46.68	1.73	3
MA2C	50.48	1.68	2
MF3DQN-TF	36.28	2.40	2

5. 结论

本文中，提出了一个基于 MF3DQN-TF 算法的交通信号控制模型。实验结果证明，在 3DQN 算法的基础上引入平均场理论，大大降低了状态动作空间的维数，从而使得模型在训练的过程中同其他算法相比收敛速度加快。同时，在模型中添加的自注意力机制对交通数据之间内在联系的分析，从交通性能指标上看，确实帮助了智能体快速的了解到交通信号控制规则。具体的，本文所提出的算法在奖励值、平均车辆等待时间、平均车辆速度、平均队列长度等指标上均优于 DQN、DDPG、MA2C 算法。对于多交

叉口交通信号控制的研究不仅能在一定程度上有效缓解城市道路的交通拥堵问题，而且会伴随着经济增长和环境改善等优势。未来的工作中，需要继续研究更加先进的、更加实用的多智能体强化学习算法，并应用在现实世界中，为我们不断发展的社会解决城市交通拥堵问题。

参考文献

[1] 政府网站: 国务院. 上海: 道路交通实现智慧治理[J]. 2019-04-03. https://www.gov.cn/xinwen/2019-04/03/content_5379482.htm, 2024-02-15.

[2] Webster, F.V. (1958) Traffic Signal Settings.

[3] Vincent, R.A. and Peirce, J.R. (1988) "MOVA": Traffic Responsive, Self-Optimising Signal Control for Isolated Intersections.

[4] Sims, A.G. (1979) The Sydney Coordinated Adaptive Traffic System. *Engineering Foundation Conference on Research Directions in Computer Control of Urban Traffic Systems*, Pacific Grove, 11-16 February 1979, 12-27.

[5] Kaelbling, L.P., Littman, M.L. and Moore, A.W. (1996) Reinforcement Learning: A Survey. *Journal of Artificial Intelligence Research*, **4**, 237-285. https://doi.org/10.1613/jair.301

[6] Wei, H., Zheng, G., Gayah, V., *et al*. (2021) Recent Advances in Reinforcement Learning for Traffic Signal Control: A Survey of Models and Evaluation. *ACM SIGKDD Explorations Newsletter*, **22**, 12-18. https://doi.org/10.1145/3447556.3447565

[7] Li, L., Lv, Y. and Wang, F.Y. (2016) Traffic Signal Timing via Deep Reinforcement Learning. *IEEE/CAA Journal of Automatica Sinica*, **3**, 247-254. https://doi.org/10.1109/JAS.2016.7508798

[8] Luo, J., Li, X. and Zheng, Y. (2020) Researches on Intelligent Traffic Signal Control Based on Deep Reinforcement Learning. 2020 *IEEE* 16*th International Conference on Mobility*, *Sensing and Networking* (*MSN*), Tokyo, 17-19 December 2020, 729-734. https://doi.org/10.1109/MSN50589.2020.00124

[9] Wang, S., Xie, X., Huang, K., *et al*. (2019) Deep Reinforcement Learning-Based Traffic Signal Control Using High-Resolution Event-Based Data. *Entropy*, **21**, Article No. 744. https://doi.org/10.3390/e21080744

[10] Buşoniu, L., Babuška, R. and De Schutter, B. (2010) Multi-Agent Reinforcement Learning: An Overview. In: Srinivasan, D. and Jain, L.C., Eds., *Innovations in Multi-Agent Systems and Applications*—1, Springer, Berlin, 183-221. https://doi.org/10.1007/978-3-642-14435-6_7

[11] Haddad, T.A., Hedjazi, D. and Aouag, S. (2022) A Deep Reinforcement Learning-Based Cooperative Approach for Multi-Intersection Traffic Signal Control. *Engineering Applications of Artificial Intelligence*, **114**, Article ID: 105019. https://doi.org/10.1016/j.engappai.2022.105019

[12] Chu, T., Wang, J., Codecà, L., *et al*. (2020) Multi-Agent Deep Reinforcement Learning for Large-Scale Traffic Signal Control. *IEEE Transactions on Intelligent Transportation Systems*, **21**, 1086-1095. https://doi.org/10.1109/TITS.2019.2901791

[13] Wu, T., Zhou, P., Liu, K., *et al*. (2020) Multi-Agent Deep Reinforcement Learning for Urban Traffic Light Control in Vehicular Networks. *IEEE Transactions on Vehicular Technology*, **69**, 8243-8256. https://doi.org/10.1109/TVT.2020.2997896

[14] Wang, X., Ke, L., Qiao, Z., *et al*. (2020) Large-Scale Traffic Signal Control Using a Novel Multiagent Reinforcement Learning. *IEEE Transactions on Cybernetics*, **51**, 174-187. https://doi.org/10.1109/TCYB.2020.3015811

[15] Garivier, A. and Moulines, E. (2011) On Upper-Confidence Bound Policies for Switching Bandit Problems. *International Conference on Algorithmic Learning Theory*, Espoo, 5-7 October 2011, 174-188. https://doi.org/10.1007/978-3-642-24412-4_16

[16] Yang, Y., Luo, R., Li, M., *et al*. (2018) Mean Field Multi-Agent Reinforcement Learning. *International Conference on Machine Learning PMLR*, Stockholm, 10-15 July 2018, 5571-5580.

[17] Hu, T., Hu, Z., Lu, Z., *et al*. (2023) Dynamic Traffic Signal Control Using Mean Field Multi-Agent Reinforcement Learning in Large Scale Road-Networks. *IET Intelligent Transport Systems*, **17**, 1715-1728. https://doi.org/10.1049/itr2.12364

[18] Vaswani, A., Shazeer, N., Parmar, N., *et al*. (2017) Attention Is All You Need. *Advances in Neural Information Processing Systems* 30 (*NIPS* 2017), Long Beach, 4-9 December 2017, 232-241.

[19] Pérolat, J., Strub, F., Piot, B., *et al*. (2017) Learning Nash Equilibrium for General-Sum Markov Games from Batch Data. *Artificial Intelligence and Statistics*. *PMLR*, 2017, Fort Lauderdale, 20-22 April 2017, 232-241.

[20] Lillicrap, T.P., Hunt, J.J., Pritzel, A., *et al*. (2015) Continuous Control with Deep Reinforcement Learning.

[21] Schulman, J., Wolski, F., Dhariwal, P., *et al*. (2017) Proximal Policy Optimization Algorithms.

[22] Prabuchandran, K.J., An, H.K. and Bhatnagar, S. (2014) Multi-Agent Reinforcement Learning for Traffic Signal Control. 17*th International IEEE Conference on Intelligent Transportation Systems* (*ITSC*), Qingdao, 8-11 October 2014, 2529-2534. https://doi.org/10.1109/ITSC.2014.6958095