Computer Science and Application 计算机科学与应用, 2023, 13(1), 113-125
Published Online January 2023 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2023.131012

Hans汉斯

基于LSTM Dueling Double DQN的车联网分布式资源管理算法

陈志鹏

天津工业大学，天津

收稿日期：2022年12月16日；录用日期：2023年1月12日；发布日期：2023年1月29日

摘　要

世界卫生组织(World Health Organization, WHO)指出，为了防止每年造成数百万人死亡的交通事故，车辆与其他车辆、基础设施或行人之间的信息交换(Vehicle-to-Everything, V2X)通信至关重要。V2X使车辆能够广播安全信息，包括其位置、速度、碰撞警告信息，以减少道路事故。且随着无线通信技术的进步和智能驾驶技术的发展，实时交付、低网络成本和通信信息的稳定交付(可靠性)成为各类智能交通系统高速发展的基础和核心。然而，在V2X通信中，数据传输通常是通过一个基站来分配频率和时间资源给车辆进行，传输的稳定性并不能总是得到保证，且车联网中无线网络的通信特征和车辆的移动性，给车联网通信服务带来了巨大挑战，在大规模时态数据分发场景中，传统的分布式服务框架无法仅凭基站服务，满足高发状态下的异构网络资源分配和通信资源共享，因为基站的覆盖范围有限且车辆处于高速移动中，所以服务请求通过路测单元和车辆协作来完成服务。针对以上的问题，本文从服务框架、算法设计的角度，对车联网信息服务技术进行了系统的研究。针对V2X通信环境下的大规模数据分发进行了研究，首先提出了一种分布式资源调度的多智能体时态信息服务框架，并建立了分布式资源分配模型，在此基础上，提出了多智能体强化学习的分布式资源分配算法。通过车辆检网络实时交互，智能决策动态环境下无线资源的预留和重用，使得资源选择能够适应车辆周边环境的动态变化，显著降低了分组碰撞概率。

关键词

车联网，分布式资源，LSTM，Dueling-Double-DQN

Distributed Resource Management Algorithm Based on LSTM Dueling Double DQN in V2V Communication

Zhipeng Chen

Tiangong University, Tianjin

文章引用: 陈志鹏. 基于 LSTM Dueling Double DQN 的车联网分布式资源管理算法[J]. 计算机科学与应用, 2023, 13(1): 113-125. DOI: 10.12677/csa.2023.131012

Received: Dec. 16th, 2022; accepted: Jan. 12th, 2023; published: Jan. 29th, 2023

Abstract

The World Health Organization (WHO) points out that in order to prevent traffic accidents that cause millions of deaths every year, the Vehicle-to-Everything (V2X) communication between vehicles and other vehicles, infrastructure or pedestrians is essential. V2 enables vehicles to broadcast safety information, including its position, speed and collision warning information, so as to reduce road accidents. With the progress of wireless communication technology and the development of intelligent driving technology, real-time delivery, low network cost and stable delivery of communication information have become the foundation and core of the rapid development of various intelligent transportation systems. However, in V2X communication, data transmission is usually carried out through a base station to allocate frequency and time resources to vehicles, and the stability of transmission cannot always be guaranteed. Moreover, the communication characteristics of wireless network and the mobility of vehicles in vehicle networking bring great challenges to vehicle networking communication services. In large-scale temporal data distribution scenarios, the traditional distributed service framework can't only rely on base station services to meet the high-incidence heterogeneous network resource allocation and communication resource sharing, because the coverage of base stations is limited and vehicles are moving at high speed. In view of the above problems, this paper systematically studies the information service technology of vehicle networking from the perspective of service framework and algorithm design. The main achievements include the following aspects, the large-scale data distribution in V2X communication environment is studied. Firstly, a multi-agent temporal information service framework for distributed resource scheduling is proposed, and a distributed resource allocation model is established. On this basis, a distributed resource allocation algorithm for multi-agent reinforcement learning is proposed. Through the real-time interaction of vehicle inspection network, the reservation and reuse of wireless resources in dynamic environment are intelligently decided, which makes the resource selection adapt to the dynamic changes of the surrounding environment of vehicles, and significantly reduces the probability of group collision.

Keywords

Vehicle Networking, Distributed Resources, LSTM, Dueling-Double-DQN

1. 引言

随着无线技术的发展，华为将 5G 技术突破后应用到我们生活的各个方面，由于 5G 的低时延的优秀特性成为车联网通信的主要网络之一，通过实时的车与车之间位置、速度、方向和路况信息的共享，提高了交通效率。

通常情况下，V2V 的资源分配通过基站的协调，即集中式的资源分配，由车向基站申请通信资源，在收到基站的资源分配消息之后才可以进行 V2V 通信，这种方式的优点在于，由于资源分配由基站决定，一定程度上避免了资源的碰撞，不过集中式资源分配只能在基站覆盖的区域内实行，随着智能设备的发

展，用户通信量增加，基站覆盖范围有限，在通讯过程中车辆的高速移动性会频繁加入或者离开基站范围，导致通信会出现不稳定，而分布式 V2V 通信由单个车辆自主执行，使得在 V2V 通信中实现多对多广播成为一种可能。

在分布式 V2V 通信中，协作感知消息(CAM) [1]时车辆之间周期性交换信息的而一种基本信息，为了支持 CAM 通信车辆采用半静态调度 SPS 算法分配无线资源，即允许车辆自主选择无线资源并通过 PC5 接口进行直连通信，车辆首先通过感知无线资源的质量，然后从可用资源中随机选择一个 CAM 资源进行预留，但是在车辆的高速移动和高业务负载的网络拥挤的动态环境下容易发生传输碰撞，如相互靠近的车辆很可能会选择相同的资源，并在随后的一系列传输中相互干扰，导致这些车辆发送的 CAM 重复解码[2]。特别是在拥塞场景下，对 SPS 算法提出了更高的挑战，影响可靠通信，因此需要配合使用一定的碰撞检测机制。

所以我们的目标是生成一种分布式资源分配策略来克服 SPS 感知算法存在的碰撞问题，近年来，在深度学习中，通过训练出有效的函数对图像、文本的上下文特征进行有效的信息融合和特征提取，使得强化学习可以通过深度学习生成逼近策略或价值函数，从而应用于高维状态和动作空间。深度强化学习(DRL)的研究主要集中在单智能体环境下，即单个智能体解决静态任务[3]。然而，许多问题，例如自动驾驶车辆、交通灯控制、任务和资源分配等，涉及多个代理。在多智能体环境中，考虑到环境中其他智能体的存在，要求智能体仅根据其局部观察执行动作，以最大化其个体报酬或总体目标[4] [5] [6]。因此，V2X 通信中的分布式资源分配也可以在多代理设置中建模。选择频率和时间资源进行传输的车辆应考虑其他车辆的存在，以最大化其数据包接收率或设计奖励函数。多智能体强化学习已被应用于各种问题，以提供分布式解决方案，即基于局部信息的决策。随着车辆数量的增加，V2V 之间连接的组合数量也会增加，因此该方法在计算上无法扩展[7]。我们将 V2V 连接问题描述为一个多智能体问题，将每辆车的 V2V 决策分组为单个决策，以克服这一问题。此外，我们只训练车辆共享的一个模型，这样车辆也可以学习其他车辆的经验。

2. 研究现状

通常情况下，V2V 的资源分配通过基站的协调，即集中式的资源分配，由车向基站申请通信资源，在收到基站的资源分配消息之后才可以进行 V2V 通信，这种方式的优点在于，由于资源分配由基站决定，一定程度上避免了资源的碰撞，不过集中式资源分配只能在基站覆盖的区域内实行，随着智能设备的发展，用户通信量增加，用户等待资源池分配的时间较长，影响用户体验并带来较高时延，且在通讯过程中车辆的高速移动性会频繁加入或者离开基站范围，导致通信会出现不稳定，随着车辆用户的提高，车载智能设备的增多，导致服务请求越来越多，导致基于基站分配网络资源的算法复杂度越来越高，所以以车为移动基站，将原始基站的任务分布在一个个车辆上，既减缓了基站压力，也提升了数据服务效率，所以分布式资源分配方法应运而生。

在分布式 V2V 通信中，基站不需要获取全局信息进行集中控制，协作感知消息(CAM)时车辆之间周期性交换信息的而一种基本信息，为了支持 CAM 通信车辆采用半静态调度 SPS 算法分配无线资源，即允许车辆自主选择无线资源并通过网络接口进行直接通信，当多辆车处于同一环境下，感知其他车辆使用的资源，并确定可用的传输资源，由于空间–局部感知结果相似，附近的车辆很可能选择相同的资源进行传输，从而导致传输碰撞，使得数据传输的可靠性降低，所以需要对于分布式数据传输需要配合响应的数据碰撞检测机制。

针对上述数据可靠性问题，不少学者通过改进 SPS 算法的资源选择方式来减小碰撞和提高包接收率，文献[8]提出基于 SPS 的资源交替选择(Resource Alternative Selection, RAS)算法，采用两个预留资源

进行交替分配，能够缓解连续碰撞，文献[9]提出使用 Lookahead 的半持续调度(Lookahead Based Semi Persistent Scheduling, Lookahead-SPS)算法，在广播消息的控制字段增加车辆的 SPS 参数信息，减少由于缺少邻居预留信息而产生的碰撞，但需要额外的信令开销。文献[10]通过强化学习来解决车联网中分布式资源分配和功率控制问题，极大提高了数据传输的可靠性和提升车联网数据服务的传输效率。而在上述方法中，资源的预留和重用是固定的，不能有效反映信道质量动态变化的影响。

而现有的分布式资源管理方法是感知型 SPS 算法，针对 SPS 算法中存在的碰撞问题，文献[11]通过多种性能网络混合方法，来控制 SPS 算法的发包周期，即针对不同场景下的不同数据采用不同的网络进行发送，如数据爆发情况下 5G 网络为数据传输提供快速、可靠的传输通道，极大保证了数据传输质量，减少了数据碰撞。文献[12]使用计数的方法对资源池中的数据进行标号，如果两辆车选择广播的车辆具有系统的标号数字，则强制改变一方的标号，用以接解决标号相同数据可能发生的碰撞问题。

还有研究者在调度中考虑网络环境对通信干扰的影响，文献[13]通过对同方向和逆方向车辆进行区分，仅仅同向车辆数据进行数据传输，配合使用 SPS 算法，极大减少了数据碰撞的概率，增加了数据服务的可靠性。文献[14]利用深度强化学习算法学习周围环境实时状态，构造深度神经网络拟合信道状态，提出车联网网络资源优化算法来提升数据的接收率。

近年来，多智能体深度强化学习被用于各种资源分配问题，比如频谱资源分配问题等，所以可以考虑将多智能体强化学习应用到车联网的分布式网络资源分配问题上，首先，分布式任务中存在的每辆车的终端设备可以作为一个个智能体，通过与环境交互来获得调整数据分发策略，从而使得车辆做出下一步收发任务，对于 SPS 算法中的碰撞问题，可以通过多智能体深度强化学习的训练过程不断生成合理的策略函数，从而减小数据的碰撞问题，而且可以对不同的环境生成不同的数据传输策略，适合应用于环境多变的 V2V 数据传输任务。

所以本文利用强化学习框架，基于竞争网络的深度学习算法，研究多智能体在无基站情况下生成合理的资源分配方案，要求每个智能体仅根据其局部观察执行动作，以最大化其个体报酬或总体目标。因此，V2X 通信中的分布式资源分配也可以在多代理设置中建模。选择频率和时间资源进行传输的车辆应考虑其他车辆的存在，以最大化其数据包接收率或设计奖励函数。

3. 系统模型和问题描述

3.1. 系统模型

在本章，对所提问题进行定义，在车联网中，车辆需要和周围环境中多个车辆进行状态信息的周期性交换，，车辆定期生成传输合作感知信息(CAMs)，用于对周围车辆进行广播。其中每辆车的 CAMs 信息中包含自身的邻居表和其他信息，如位置信息、身份信息、速度大小、方向以及加速度等，使得接收到该 CAMs 信息接收器根据解码信息，做出合理的计算判断出合适的反应，从而提高通信效率。

首先我们定义车辆集合用 $N=\{1,2,3,\cdots,n\}$，车辆的处于一直运动的状态，且每辆车都有一个半双工的收发器，半双工是指设备在请求信息的同时无法发送信息，反之亦然。本章为验证分布式算法在解决信息碰撞方面的优越性，故而本次建模不设置基站，即所有车辆都不与基站通信。本章核心任务为分布式资源的调度，定义所有车辆可从可用资源池 $K=\{1,2,3,\cdots,k\}$ 选择任一资源进行广播或者请求。

对于资源的调配和通信，此次模型设计一个基于时间分配资源的系统，即所有车辆在 t 时刻同时对可用资源 K 进行调度，同时需要每辆车都试图最大化解码其包的邻居数量，所以设计了一种可靠性评价标准，通过包接受比例(PRR) [15]来表示，包接受比例可定义为 t 时刻车辆 i 发送的 CAMs 包被成功接收的次数占总邻居数 N 的比例。

在模型中无服务的车辆周期性发送 CAM 数据包，周围有服务请求的车辆对广播数据进行接收解码，

所以可以对资源的选择对 PRR 进行建模，每辆车 i 在 t 时刻对资源 k 的选择表示为动作 $a_t^i = k$，所有车在 t 时刻选择的动作可表示为 $a_t = \left(a_t^1, a_t^2, \cdots, a_t^N\right)$。

对于问题的表述和奖励的设计，我们假设了一个只反映路径损失的简化信道模型，但忽略了快慢衰落，干扰建模仅限于来自其他车辆的传输[16]。而且如果多辆车辆选择发送同一种资源，则接收器只选择信噪比最高的车辆进行解码。

在以上基础上，通过计算 $PRR_t^1\left(a_t\right)$ 表示每辆车 i 在 t 时刻所发送的数据包被解码数占其邻居的比例，公式如下：

$$PRR_t^i\left(a_t\right) = \frac{1}{N_t^i}\sum_{j=1}^{N_t^i} 1\left\{P_{eer}\left(\gamma_t^{i,j}\right) < X \sim U\left([0,1]\right)\right\} \tag{1}$$

其中 N 表示在 t 时刻车辆 i 的相邻车辆数目，γ 是车辆 i 的数据包在车辆 $j \in N$ 处的信干噪比(SINR)，公式如下：

$$\gamma_t^{i,j} = \frac{P\left|H_t^{i,j}\right|^2}{\sigma^2 + \sum P\left|H_t^{k,j}\right|^2}, k \in \frac{H_t^c\left(a_t\right)}{\{i\}} \tag{2}$$

P 函数是基于固定的调制编码方案(MCS)计算给定 SINR 下的误块率。如果 P 等于或小于一个介于 0 和 1 之间的随机数，则数据包将被成功解码。P 是所有车辆固定的发射功率，σ^2 是加性高斯白噪声的功率，N 是由 t 时刻的动作决定的干扰数据包集合，$\left|H_t^{i,j}\right|$ 和 $\left|H_t^{k,j}\right|$ 分别是发送者 i 和接收者 j 之间以及发送者 k 和接收者 j 之间的信道增益。

3.2. 问题定义

车辆的目标是通过 PRR 来最大化收到 CAM 信息的邻居数量。根据 PRR 的定义，我们制定了一个优化问题来选择策略 π，该策略将车辆的动作设定为 $\pi(t)$ 使得 PRR 最大化。

$$\begin{aligned} &\max_{\pi} \sum_{i=1}^{N} PRR_t^i\left(a_t^{\pi}\right) \\ &\text{subject to } a_t^i \in \left(1,2,\cdots,K\right) \end{aligned} \tag{3}$$

对于非拥堵情况，如果所有车辆都选择单独的资源，则式(3)很容易满足。然而，在拥堵的情况下，也就是说，必须使用一个策略来动态调整行动，因为车辆的通道收益和邻居的数量因车辆的流动性而不同。最大化所有车辆的平均 PRR 的策略称为最优策略 π。

对于分布式场景，获取最优策略 π^* 具有挑战性，因为没有中央调度器可以考虑到每个车辆的通道增益和彼此相对的位置。在集中式解决方案中，基站协调车辆传输，采用资源空间复用，保持通信可靠，即高 PRR，具体来说，当车辆之间的距离大于最小重用距离时，基站将相同的资源分配给车辆[17]。但是，我们考虑车辆位于基站覆盖之外的情况。因此，每辆车都只根据本地观测自主选择行动。

由于问题的动态多维性质，我们将其描述为一个多智能体深度强化学习问题，这样远的车辆会被激励选择相同的资源，而近的车辆使用不同的资源。简单地说，我们研究了以分布式方式分布式调度的概念去最大化总体包接收比。

4. 算法框架

强化学习(Reinforcement Learning, RL) [18]，是人工智能领域中一类特定的机器学习问题。RL 从统计学、控制理论和心理学等多学科交叉发展而来。深度学习(Deep Learning, DL)作为机器学习研究中的重要领域，近年来随着硬件平台计算能力的长足发展，在图像、自然语言、语音等诸多应用领域取得了瞩目

的成绩[18]。自然而然，人们同样会期望 RL 能够借助 DL 来解决以往难以处理的问题，例如直接读取像素来玩视频游戏等。谷歌 Deepmind 于 2015 年发表于 Nature 的文章使用深度强化学习首次实现了在 Atari 游戏中达到与人类同等甚至更高的水平，向世人展现了 DRL 的巨大潜力。

本小节主要侧重于强化学习中的多智能体领域，介绍多智能体强化学习的基本建模框架及算法。

4.1. D2DQN

在本小节中，我们提出对偶式分布式多智能体深度强化学习资源分配算法(Dueling Double DQN, D2DQN)。本文的创新之处在于用独特的状态表示来处理多智能体学习系统中的非平稳性(Non-Stationarity)，从而实现分布式资源分配。即让每辆车从自己的视角来观察道路上其他车辆的位置，并创建基于视图的位置分布向量。视场常识通过基于视图的位置分布的观测而产生，车辆可以从自己的观测推断其他车辆的观测。更具体地说，车辆 A 和 C 是距离较近的两辆车，车辆 A 可以从自己的观测中推断出 C 的观测，从而预测到车辆 C 将会选择哪些资源选择，从而，车辆 A 可以为传输 CAM 合理分配资源。目标的合作性指的是车辆需要合作进行资源分配，使第 i 辆车目标函数值最大化，集中训练使车辆能够根据常识制定完全分散的策略[19]。在 MADRL 系统[20]多采用集中训练，分散执行的框架。在集中训练部分，通过访问车辆的行动来确定每个代理的奖励。训练部分完成后，每个车辆作为代理利用训练过的策略只需根据本地观察做出决策。集中式训练使 DQN 的参数与所有智能体共享，有利于实现参数共享。因为每个代理具有同质性[21]，我们设置代理间允许共享参数，包括共享相同的奖励效用、状态和动作空间。

为了更好地实现资源配置，我们对于参数共享的算法做了进一步的改进，来实现稳定学习和改进策略。我们采用 Dueling Double DRQN，即将动作值函数分为状态值函数和优势函数并采用长短期记忆方法的 Double DQN 的增强版本，旨在解决解决文献中所提 Double DQN 针对本模型收敛慢、适配性(效果)差等问题。主要思想大致可以描述为，当前车辆移动过程中，智能体左右移动对车辆并没有影响，说明动作对 Q 值没有影响，但是状态对 Q 值很有影响，Dueling Double DQN 中的对偶网络思想符合本次建模的场景的设定，状态和动作分别用来训练，增强网络参数的有效性和加速模型的收敛[22]。在 double DQN 中利用目标网络计算下一个动作的 Q 值，即 a' 计算损失，并与评估网络的参数定期更新。此外，我们存储每个代理的经验，即 $e_t^i=\left(s_t^i,a_t^i,r_t^i,s_{t+1}^i\right)$ 存储在经验重放记忆用于训练。通过在训练过程中对经验回放 $(s,a,r,s')\sim U(D)$ 进行随机采样，这样也使得数据分布变化平滑，缩短了经验回放的大小，并保持一个 FIFO 缓冲区，其大小与代理的数量成比例，以避免非平稳性，而 Dueling Double DRQN 与 Double DQN 输入一样，均为状态信息，但是输出却有所不同。Dueling DQN 算法的输出包括两个分支，分别是该状态的状态价值 V (标量)和每个动作的优势值 A (与动作空间同维度的向量)。利用卷积网络提取特征获取特征向量，输出时会经过两个全连接层分支，分别对应状态价值和优势值，最后将状态价值和优势值相加即可得到每个动作的动作价值[23]。

4.2. 状态和奖励设计

作为 Dueling-double DRQN 深度神经网络结构的一部分，我们使用长期短记忆(LSTM)网络作为第一层，根据位置分布预测车辆的移动模式。LSTM 层维持一个内部状态，并随着时间的推移将观测结果结合起来。我们近似 $Q\left(s_t,a_t,h_{t-1};\theta\right)$ 值与递归神经网络，其中 h_{t-1} 是前一步骤中代理的隐藏状态。隐藏状态 $h_t=\mathrm{LSTM}\left(s_t,h_{t-1}\right)=\mathrm{LSTM}\left(s_{t-(L-1)},\cdots,o_t\right)$，其中 L 为观察次数。LSTM 网络之后是全连接的前馈网络层，用于计算每个动作的值。虽然我们只训练一个 DQN，但代理表现仍然不同，因为每个代理都根据不同的观察而演化出自己的隐藏状态，这使得代理尽管共享相同的 DQN 也能表现不同。之后对状态和动作空间进行设计[24]。

1) 状态和动作空间

车辆 i 的状态向量 s_t^i 由代理 i 在时隙 t 采取的先前动作组成，并且该向量表示从车辆 v_t^i 的角度来看其他车辆的位置分布，即函数 $f\left(p_t^i,B,R\right)$ 的输出。

基于视图的位置分布(View-Based Positional Distribution, VPD)函数 $f\left(p_t^i,B,R\right)$ 利用时隙 t 时代理 i 的邻居表 p_t^i 、基于视图的观察向量的粒度的整数 $B\in\mathbb{Z}^+$ 和指示代理 i 的观察半径的整数 $R\in\mathbb{Z}^+$ 来获取到其他车辆的位置。当前邻居表被附带到 CAM 消息中，因为每辆车都共享它们的位置信息。

在这项工作中，代理车辆使用所有可用的传输频率快，因此 1 个时隙和 1 个子信道代表一个资源块。动作空间变成了 $\mathcal{A}=\{a\mid a=k,k\in\mathcal{K}\}$。

2) 奖励设计

为进一步分析所提出的方法在拥塞和非拥塞情况下的性能。每辆车被鼓励选择不同的资源，但是当拥挤的情况，远处的车辆被鼓励使用相同的资源。每个代理的奖励计算如下：

$$r_t^i\left(a_t^i\mid s_t^i\right)=\begin{cases}1 & N_t^c=1\\ 0 & \text{if } dist(c)>r_{reuse}N_t^c=2\\ -N_t^c & \text{else},N_t^c=2\\ -N_t^c & N_t^c>2\end{cases}\tag{4}$$

其中向量 c 是相同资源上的干扰的代理，即 $c=[i,k,\cdots,l]$， $c\subset N$ 、 $N_t^c=|c|$ 是相撞车辆的总数。在时隙 t 的总回报的平均值被加到单个车辆的回报上，以加强合作行为 $r_t^i\left(s_t^i\mid a_t^i\right)=r_t^i\left(s_t^i\mid a_t^i\right)+\dfrac{\sum_{j=1}^{N}r_t^i\left(s_t^i\mid a_t^i\right)}{N}$。

我们不需要额外的反馈信道来通知发送的分组是否被成功解码以计算回报。我们仅基于车辆的位置分布和资源分配来对模型进行设计：

算法 3-1：基于视图的位置分布

```
Function F(p_t^i, B, R, thre, txID)
    初始化一个列表向量 D
    for e in p_t^i do
        if e[date] < thre then     # 始终准备更新车辆
                        dist = calculate_dist(e["pos"], p_t^i, ["txID"]["pso"])
            If dist < R  then # 如果接收机在发射机的范围内
                D.push(dist)
            end if
        end if
    end for
    D =D.sort()  // Sort the distance in ascending order
    H = Histogram(D, B, [-R, R])
    v_t^itxID = H/len(D)      //Calculate the distribution from histogram
    return v_t^itxID
end function
```

之后我们对 double DQN 算法结构进行设计，具体步骤和算法如下：

Step 1：

输入层：DRQN 的输入 S 是一个大小为$|A|+B$ 的向量，其中$|A|$是动作空间的大小，表示可用资源(子

帧)的数量。每辆车都添加在前一个时间戳 A 中采取的动作作为状态的一部分。B 是每一个代理划分其观测视图的间隔数，它描述了观测的粒度，并对所有车辆的这个参数固定。

Step 2：

隐藏层：第一层为 LSTM 层。多智能体学习环境打破了 MDP 假设，因为智能体的下一个观测状态不仅由智能体本身决定，还由其他车辆采取的行动决定。在我们的工作中，我们不观察其他车辆的通道选择，因为我们使用车辆的位置分布。然而，部分观测是基于其他车辆未知的移动模式，由其他车辆的位置分布引起的。文献[25] [26] [27]的几项研究表明，LSTM 网络可以成功地用于移动预测。LSTM 层保持内部状态，并随着时间的推移组合观测值。LSTM 层与全连接前馈网络层相连。

Step 3：

输出层：DRQN 的输出是一个向量，其大小$|A|$表示对于给定的状态输入，每个动作的值。如果策略是贪婪的，车辆选择价值最高的动作。

Step 4：

Double DQN 如果车辆选择和评估与同一网络的行动，由于过分估计使得算法的性能下降。因此，采用双 Q 学习实现动作选择与评价的解耦。

算法 3-2：Double DQN 算法伪代码

初始化：两个 dueling DDRQN 网络 Q 和 Q'

初始化：价值-奖励函数 Q，随机化生成权重 θ

初始化：目标价值-奖励函数 Q'，初始化权重 $\theta' = \theta$

//在经验池中随机选择一个 episode

for episode $e = 1, \cdots, M$ do

// *episode 中随机选择一个时间点，从这个点一直运行到 episode 结束*

for $t = 1, \cdots, T$ do　　*//对所有车辆进行迭代*

for $v \in [1, \cdots, N]$ do

//获取当前代理车辆的状态向量 s_t^i in $\langle s, a_R, a_B, \gamma, s' \rangle$

if $random(0,1) > \varepsilon$ do

在动作池中随机选择一个动作 a_t^i

else $a_t^i = \arg\max_{a_t^i \in A} Q(s_t^i, a_t^i; \theta)$

end for

#获得有车辆所获得的奖励

$$r_t = (r_t^1, r_t^2, \cdots, r_t^N)$$

$D = (s_t, a_t, r_t)$ *#将获得的* s_t, a_t, r_t *对 D 进行更新*

end for

在 D 中随机采样一组状态信息 s_t, a_t, r_t, s_{t+1}

计算 $y_t = \begin{cases} r_t \text{ if episodes endsat } t+1 \\ r_t + \gamma Q\left(s_{t+1}, \arg\max_{a'} \hat{Q}'(s_t, a', \hat{\theta})\right) \text{ otherwise} \end{cases}$

根据公式 $(y_t - Q(s_t, a', \theta))^2$ 计算关于权重值 θ 的损失值。

对隐藏层的权重值 θ 进行更新。

步长设置为 K 对 episodes 的 $\theta' = \theta$

更新 ε

end for

在 Double DQN 网络中，输入为该时刻智能体的状态(用向量表示)，使用 LSTM 网作为基本单元时，利用环境状态转移模型进行采样，使用连续四个时刻的状态序列作为网络的输入(s_1,s_2,s_3,s_4)。

每个车辆作为一个智能体，不断重复交互、做出动作、调整策略，将会设置一个在记忆库用于存储每一组数据的执行情况，可以表示为元组$\langle s,a_R,a_B,\gamma,s'\rangle$，其中 s、s' 分别表示智能体车辆当前所处环境状态和执行动作之后所处状态，a_R、a_B 表示车辆周围任意车辆采取的动作，γ 为当前的奖励函数值，训练过程中网络参数作为不断变化的策略函数，根据当前策略对智能体的做出动作，再根据损失函数 loss 计算出的损失值进行反馈，即梯度下降法对网络参数进行更新，不断重复以上步骤，从而达到不断完善策略函数(网络参数)的目的，从而使得达到最终奖赏值最大化。

由于多智能体的状态特征量是具有连续、多个维度等特性，所以利用深度学习网络中的非线性激活函数比如 ReLU 来不断生成目标函数$Q(s,a)$，可以解决连续空间中的维度灾难问题。

与 Double DQN 和 DQN 网络不同，Dueling Double DQN 网络将网络分成了两部分，价值函数$V(s,\omega,\alpha)$和优势函数$A(s,\omega,\alpha,\beta)$，网络的最终结果包含两部分：

$$Q(s,a,\omega,\alpha,\beta)=V(s,\omega,\alpha)+A(s,a,\omega,\beta) \tag{5}$$

其中 ω 为是深度学习网络中的权重参数，属于隐藏层参数，α、β 分别是价值函数和优势函数的独立参数。

在模型中，每个车辆作为一个智能体，智能体的状态由 n 个向量组成，因此神经网络的输入是以那个向量作为输入，再输出每个向量所对应的奖励即价值，中间层称为隐含层，将长短期网络单元作为隐含层的基本神经元，并对网络的输入层进行改进，如此使得最终获得策略更优且使得能加快网络收敛。

5. 仿真结果

我们首先在一个轻量级测试模拟器中评估所提出的方法的性能，以更快地分析各种实验。测试模拟器中设计一个简单的通道模型，这样可以确保如果在接收器的范围内有多个车辆使用同一资源块，接收器解码较近的车辆的数据包。但是，在现实的实时测试环境中的需要综合考虑快速和缓慢衰落、预编码和接收滤波方法等问题。表 1 总结了在轻量级测试模拟器下算法对于每一个代理车辆的训练和网络参数设置。在表 2 从车辆数量和可用资源的角度评估 Dueling Double DQN 算法在不同配置、不同场景中在的性能。

5.1. 训练参数设置

将每一个车辆作为一个代理放入网络中训练时候，训练环境中的参数如模拟总时长、激活函数、步长、衰退ε等设置如表 1 所示：

Table 1. Parameter settings
表 1. 参数设置

参数	值	参数	值
总时长	250,000	载波频率	5.9 GHz
LSTM 步长	6	系统带宽	10 MHz
学习率 rate	0.0001	衰退系数 ε	0.999
折扣系数 γ	0.7	CAM 数据包大小	300 bytes
激活函数	ReLU	周期长度	100 ms

Table 2. Evaluation scenarios
表 2. 评估场景

序号	车辆数	资源数	公路长度	速度
1	4	3	100 m	<35
2	6	5	250 m	<35
3	8	10	500 m	<35
4	10	10	500 m	<35

5.2. 训练效果

我们从图 1 所示的一个简单的例子开始评估，并与 DIRAL 算法进行对比，DIRAL 采用传统 Double DQN 算法，算法步骤如算法 1 所示，在本次场景中，每个时间步有 $N = 4$ 辆车和 $K = 3$ 个资源。每辆车都以随机的速度向同一方向行驶。我们使用了式 4 中的奖励函数，并对场景 1 和 2 进行了轻微修改。我们给出了中立的奖励，最远的车辆使用相同的资源，以更好的观察所需的行为。所以，在每个时间步中，对于场景 1，系统的最大奖励是 2。每一 episode 的长度是 25 个单位时间，我们在每一 episode 对模型进行重复迭代训练，为了便于观察 Dueling Double DQN 在本次建模情况下的效果，本文仅仅每 100 个 episode 的奖励值进行展示，防止两种算法波动值彼此覆盖影响对比效果。图 1 说明了这种方法的收敛性，相较于 DIRAL 算法，本文所提算法可以更早收敛。说明所提出的方法在本次仿真场景下达到了最优策略。

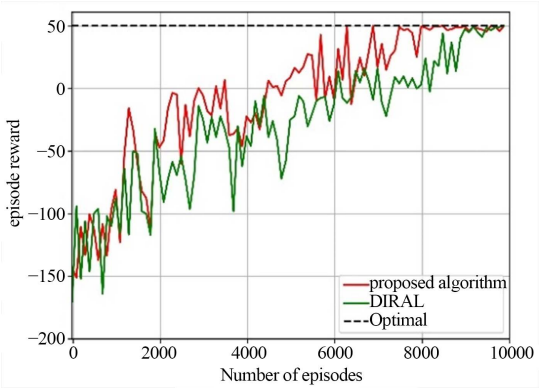

Figure 1. 4 vehicles and 3 resource scenarios
图 1. 4 车辆、3 资源场景

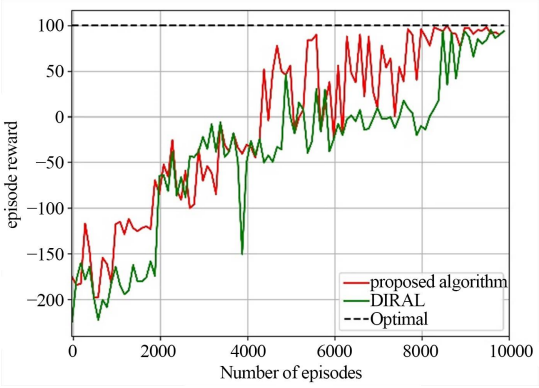

Figure 2. 6 vehicles and 8 resource scenarios
图 2. 6 车辆、8 资源场景

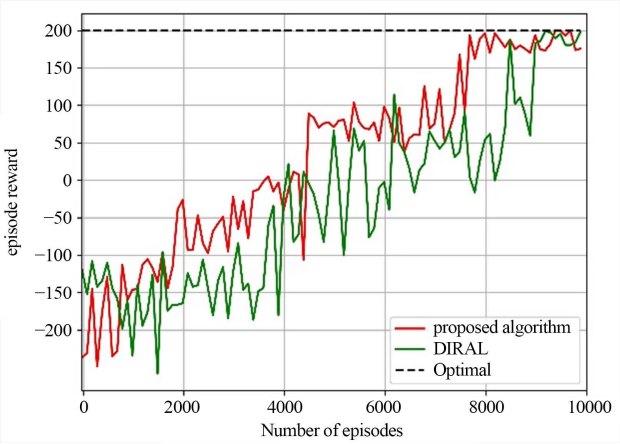

Figure 3. 8 vehicles and 10 resource scenarios
图 3. 8 车辆、10 资源场景

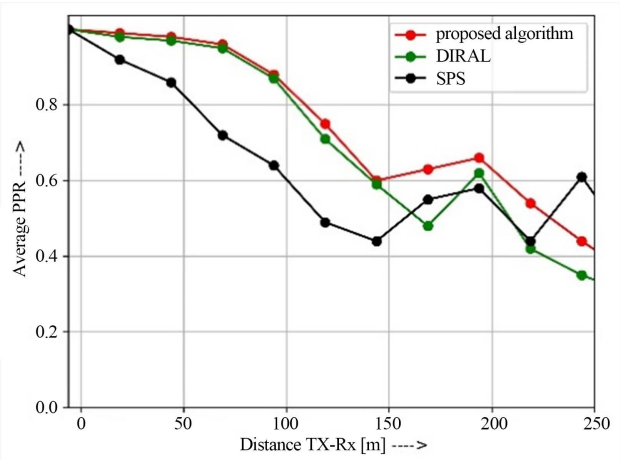

Figure 4. RealNeS analysis
图 4. RealNeS 分析

对于其他场景，我们也做了相同的实验测试，将 N 和 K 分别设置为 6、8，如图 2 所示，与 DIRAL 对比来说，本文采用 dueling 对偶式网络结构对于网络收敛的效果更好，结合 LSTM 体系结构的 Dueling Double DQN 能够找到较好的 q 值估计值。为了证明了本文所提算法适用于大规模场景下，进行了如图 3 场景下的实验，设置车辆数和资源数分别设置为 10、10，从图中可以看出，在更大规模场景下，本文所提算法也可更快收敛至最优策略。我们又将所有训练策略应用于网络模拟器，并将所提算法与 DIRAL、SPS 进行比较，SPS 算法由感知、选择和重选组成。对于选择，期望 RSSI 低于阈值的资源形成资源池，并且从共享池中随机选择资源。如果资源池的大小小于所有共享资源的 20%，则阈值增加 3 dB，并重复选择过程。在选择过程之后，车辆利用由重选计数器设置的相同资源进行后续的~[5,15]传输。在每次传输之后，重选计数器减 1，并且当它达到零时，车辆以 0.8 的概率继续使用相同的资源或者选择新的资源。对于场景一如图 4 所示，描绘了作为发射机和接收机之间距离的函数的 PRR，与 SPS 和 DIRAL 算法相比，该系统为附近车辆实现了非常高的 PRR 值。从而证明本文所提在收敛性和包接收率上都有较好表现。

6. 结论

在这项工作中，我们提出了一种新的算法，基于 Dueling Double DQN 的分布式资源分配。该算法针

对由于车辆的移动性而自然上升的拥堵问题提出了一种新颖的解决方案。我们展示了我们的建议将拥挤场景中的 PRR 提高了多达 20%。该算法使得 V2V 能够更可靠地用于基站非覆盖场景。通过在一个简单的基于 Python 的仿真器中进行训练，证明了我们所提算法的优越性。

参考文献

[1] Zhang, Q. and Li, H. (2007) MOEA/D: A Multiobjective Evolutionary Algorithm Based on Decomposition. *IEEE Transactions on Evolutionary Computation*, **11**, 712-731. https://doi.org/10.1109/TEVC.2007.892759

[2] Almeida, J., Alam, M., Ferreira, J. and Oliveira, A.S. (2016) Mitigating Adjacent Channel Interference in Vehicular Communication Systems. *Digital Communications and Networks*, **2**, 57-64. https://doi.org/10.1016/j.dcan.2016.03.001

[3] Bernstein, A.V., Burnaev, E.V. and Kachan, O.N. (2018) Reinforcement Learning for Computer Vision and Robot Navigation. In: Perner, P., Ed., *Machine Learning and Data Mining in Pattern Recognition*. *MLDM* 2018. *Lecture Notes in Computer Science*, Vol. 10935, Springer, Cham, 258-272. https://doi.org/10.1007/978-3-319-96133-0_20

[4] Molina-Masegosa, R., Gozalvez, J. and Sepulcre, M. (2020) Comparison of IEEE 802.11p and LTE-V2X: An Evaluation with Periodic and Aperiodic Messages of Constant and Variable Size. *IEEE Access*, **8**, 121526-121548. https://doi.org/10.1109/ACCESS.2020.3007115

[5] Zhao, Q., Tong, L., Swami, A. and Chen, Y. (2007) Decentralized Cognitive MAC for Opportunistic Spectrum Access in Ad Hoc Networks: A POMDP Framework. *IEEE Journal on Selected Areas in Communications*, **25**, 589-600. https://doi.org/10.1109/JSAC.2007.070409

[6] Nasir, Y.S. and Guo, D. (2019) Multi-Agent Deep Reinforcement Learning for Dynamic Power Allocation in Wireless Networks. *IEEE Journal on Selected Areas in Communications*, **37**, 2239-2250. https://doi.org/10.1109/JSAC.2019.2933973

[7] Cui, J., Liu, Y. and Nallanathan, A. (2019) Multi-Agent Reinforcement Learning-Based Resource Allocation for UAV Networks. *IEEE Transactions on Wireless Communications*, **19**, 729-743. https://doi.org/10.1109/TWC.2019.2935201

[8] Wijesiri N.B.A., G.P., Haapola, J. and Samarasinghe, T. (2019) A Markov Perspective on C-V2X Mode 4. 2019 *IEEE* 90*th Vehicular Technology Conference* (*VTC*2019-*Fall*), Honolulu, 22-25 September 2019, 1-6. https://doi.org/10.1109/VTCFall.2019.8891331

[9] Jeon, Y., Kuk, S. and Kim, H. (2018) Reducing Message Collisions in Sensing-Based Semi-Persistent Scheduling (SPS) by Using Reselection Lookaheads in Cellular V2X. *Sensors*, **18**, Article No. 4388. https://doi.org/10.3390/s18124388

[10] Honnaiah, P.J., Maturo, N. and Chatzinotas, S. (2020) Foreseeing Semi-Persistent Scheduling in Mode-4 for 5G Enhanced V2X Communication. 2020 *IEEE* 17*th Annual Consumer Communications & Networking Conference* (*CCNC*), Las Vegas, 10-13 January 2020, 1-2. https://doi.org/10.1109/CCNC46108.2020.9045276

[11] Heo, S., Yoo, W., Jang, H. and Chung, J.-M. (2021) H-V2X Mode 4 Adaptive Semipersistent Scheduling Control for Cooperative Internet of Vehicles. *IEEE Internet of Things Journal*, **8**, 10678-10692. https://doi.org/10.1109/JIOT.2020.3048993

[12] Bonjorn, N., Foukalas, F. and Pop, P. (2018) Enhanced 5G V2X Services Using Sidelink Device-to-Device Communications. 2018 17*th Annual Mediterranean Ad Hoc Networking Workshop* (*Med-Hoc-Net*), Capri, 20-22 June 2018, 1-7. https://doi.org/10.23919/MedHocNet.2018.8407085

[13] 余翔, 陈晓东, 王政, 石雪琴. 基于 LTE-V2X 的车联网资源分配算法[J]. 计算机工程, 2021, 47(2): 188-193.

[14] 金久一, 邱恭安. C-V2X 通信中资源分配与功率控制联合优化[J]. 计算机工程, 2020, 47(10): 147-152.

[15] Liang, L., Ye, H. and Li, G.Y. (2019) Spectrum Sharing in Vehicular Networks Based on Multi-Agent Reinforcement Learning. *IEEE Journal on Selected Areas in Communications*, **37**, 2282-2292. https://doi.org/10.1109/JSAC.2019.2933962

[16] Gupta, J.K., Egorov, M. and Kochenderfer, M. (2017) Cooperative Multi-Agent Control Using Deep Reinforcement Learning. In: Sukthankar, G. and Rodriguez-Aguilar, J., Eds., *Autonomous Agents and Multiagent Systems*. *AAMAS* 2017. *Lecture Notes in Computer Science*, Vol. 10642, Springer, Cham, 66-83. https://doi.org/10.1007/978-3-319-71682-4_5

[17] Bazzi, A., Cecchini, G., Menarini, M., Masini, B.M. and Zanella, A. (2019) Survey and Perspectives of Vehicular Wi-Fi versus Sidelink Cellular-V2X in the 5G Era. *Future Internet*, **11**, Article No. 122. https://doi.org/10.3390/fi11060122

[18] (2011) 3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Physical Layer Procedures. http://www.arib.or.jp/english/html/overview/doc/STD-T104v1_30/5_Appendix/Rel10/36/36213-a60.pdf

[19] Zhang, K., Yang, Z. and Başar, T. (2021) Multi-Agent Reinforcement Learning: A Selective Overview of Theories and Algorithms. In: Vamvoudakis, K.G., Wan, Y., Lewis, F.L. and Cansever, D., Eds., *Handbook of Reinforcement Learning and Control*. *Studies in Systems*, *Decision and Control*, Vol. 325, Springer, Cham, 321-384. https://doi.org/10.1007/978-3-030-60990-0_12

[20] Hernandez-Leal, P., Kaisers, M., Baarslag, T. and de Cote, E.M. (2017) A Survey of Learning in Multiagent Environments: Dealing with Non-Stationarity. ArXiv Preprint ArXiv: 1707.09183.

[21] Naparstek, O. and Cohen, K. (2018) Deep Multi-User Reinforcement Learning for Distributed Dynamic Spectrum Access. *IEEE Transactions on Wireless Communications*, **18**, 310-323. https://doi.org/10.1109/TWC.2018.2879433

[22] Schroeder de Witt, C., Foerster, J., Farquhar, G., *et al*. (2019) Multi-Agent Common Knowledge Reinforcement Learning. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. and Garnett, R., Eds., *Advances in Neural Information Processing Systems* 32 (*NeurIPS* 2019), Curran Associates, Inc., Red Hook.

[23] Van Hasselt, H., Guez, A. and Silver, D. (2016) Deep Reinforcement Learning with Double Q-Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, **30**, 2094-2100. https://doi.org/10.1609/aaai.v30i1.10295

[24] Foerster, J., Assael, I.A., De Freitas, N. and Whiteson, S. (2016) Learning to Communicate with Deep Multi-Agent Reinforcement Learning. In: Lee, D., Sugiyama, M., Luxburg, U., Guyon, I. and Garnett, R., Eds., *Advances in Neural Information Processing Systems* 29 (*NIPS* 2016), Curran Associates, Inc., Red Hook.

[25] Li, M., Lu, F., Zhang, H. and Chen, J. (2020) Predicting Future Locations of Moving Objects with Deep Fuzzy-LSTM Networks. *Transportmetrica A*: *Transport Science*, **16**, 119-136. https://doi.org/10.1080/23249935.2018.1552334

[26] Feng, J., Li, Y., Zhang, C., *et al*. (2018) Deepmove: Predicting Human Mobility with Attentional Recurrent Networks. *Proceedings of the* 2018 *World Wide Web Conference*, Lyon, 23-27 April 2018, 1459-1468. https://doi.org/10.1145/3178876.3186058

[27] Sukhbaatar, S. and Fergus, R. (2016) Learning Multiagent Communication with Backpropagation. In: Lee, D., Sugiyama, M., Luxburg, U., Guyon, I. and Garnett, R., Eds., *Advances in Neural Information Processing Systems* 29 (*NIPS* 2016), Curran Associates, Inc., Red Hook.