Computer Science and Application 计算机科学与应用, 2022, 12(12), 2925-2936
Published Online December 2022 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2022.1212297

Hans汉斯

基于强化学习的多信道车联网频谱聚合共享

唐嘉程[1]，王辛果[2]

[1]成都信息工程大学，四川 成都
[2]中国航空工业无线电电子研究所，上海

收稿日期：2022年11月22日；录用日期：2022年12月21日；发布日期：2022年12月29日

摘　要

针对车联网需求日益增多以及频谱资源的短缺问题，本文结合认知无线电的频谱聚合功能以及多智能体强化学习方法，提出了基于强化学习的多信道车联网频谱聚合共享模型。每一条车辆到车辆链路作为一个智能体，共同与通信环境交互。各链路独立获得观测结果，同时获得共同的奖励。用这样的设置来促进多个智能体进行合作来训练Q网络，达到改善频谱聚合位置选取和功率分配这一智能体动作的目的。仿真结果表明，通过适当的奖励设计和训练机制，多个智能体能成功学会以分布式方式合作。在不损失车辆到基础设施链路传输总带宽的前提下，本模型能大幅度提高车辆到车辆链路的负载交付率。

关键词

车联网，多智能体强化学习，认知无线电，DQN，多信道

Reinforcement Learning-Based Aggregated Spectrum Sharing for Multi-Channel Vehicular Networking

Jiacheng Tang[1], Xinguo Wang[2]

[1]School of Computer Science, Chengdu University of Information Technology, Chengdu Sichuan
[2]Chinese Aeronautical Radio Electronics Research Institute, Shanghai

Received: Nov. 22nd, 2022; accepted: Dec. 21st, 2022; published: Dec. 29th, 2022

Abstract

In response to the increasing demand of vehicular networks and the shortage of spectrum resources, this paper proposes a reinforcement learning-based spectrum aggregation and sharing model for

文章引用: 唐嘉程, 王辛果. 基于强化学习的多信道车联网频谱聚合共享[J]. 计算机科学与应用, 2022, 12(12): 2925-2936. DOI: 10.12677/csa.2022.1212297

multi-channel vehicular networks by combining the spectrum aggregation function of cognitive radio and a multi-agent reinforcement learning. Each vehicle-to-vehicle link, as an agent, interacts with the communication environment together. Each link obtains observations independently while receiving a common reward. Such a setup is used to facilitate cooperation among multiple agents to train the Q-network for the purpose of improving spectrum aggregation location picking and power allocation as an agent action. Simulation results show that multiple agents can successfully learn to cooperate in a distributed manner through appropriate reward design and training mechanisms. Without losing the total bandwidth of vehicle-to-infrastructure link transmission, this model can substantially improve the load delivery rate of vehicle-to-vehicle links.

Keywords

The Vehicular Network, Multi-Agent Reinforcement Learning, Cognitive Radio, Deep Q Network, Multi-Channal

1. 引言

车联网 V2X (vehicle-to-Everything)主要由 V2I 链路(Vehicle-to-Infrastructure)和 V2V 链路(Vehicle-to-Vehicle)组成。车载娱乐服务要求通常需要更高带宽的 V2I 链路，将每辆车连接到 BS (Base Station)进而连接到互联网，因此判断 V2I 性能的 QoS 优劣看重的指标是链路的传输总带宽大小。在尽可能保证 V2I 传输总带宽的条件下，高级驾驶功能需要通过 V2V 链路在相邻车辆之间定期相互通信，交换单位数据包的安全消息来获取 CSI (Channel State Information)，CSI 通常包括车辆位置、速度、行驶方向等信息，用来提高所有车辆对实时驾驶环境的“合作意识”，所以 V2V 链路的 QoS (Quality of Service)要求尽可能低的延迟和高度的可靠性。在频谱资源日益紧缺的当下，共享信道是提高 V2X 链路 QoS 的一个不错的方向。

针对如何使用共享信道提高 V2X 链路 QoS 的问题，本文结合了认知无线电中的频谱聚合功能[1]。频谱聚合是指次要用户可以通过正交频分多路复用方法[2]同时访问多个未被主要用户有效利用的离散频谱空洞并将空洞聚合成足够宽的频带以成功地完成传输任务。为了充分利用频谱资源，我们对 V2I 链路进行了简化，使其以正交方式占据频谱，并用固定功率进行传输。因此，频谱聚合优化的方向将聚焦于 V2V 链路将采取何种设计策略与 V2I 链路共享频谱进行数据传输，包括频谱子带的聚合范围选择和发射功率的控制。本模型虽然增加了 V2V 链路在有限频谱上获得共享信道的机会，但也使系统的干扰设计更加复杂，解决此问题这也是本研究的重点。

然而，车辆的高速运动带来了 CSI 变化的极大不确定性，但 RL (Reinforcement Learning)提供了一种稳定且有效的方法来处理车联网的环境变化并执行一系列动作。文献[3]针对车载环境中信道快速变化所带来的挑战，提出了一种基于设备到设备的空间频谱复用方案，缓解了对全局 CSI 的要求。在[4]中，通过对 V2X 资源进行合理分配可最大化 V2I 链路的吞吐量，以适应缓慢变化的大规模信道衰落，从而降低网络信令开销。除了传统的优化方法外，最近的几个研究中还提出了基于 RL 的方法来解决 V2X 网络中的频谱分配问题[5] [6]。在[7]中研究了提高多智能体 V2V 链路传输交付成功率问题，该研究中对于多智能体的处理思路值得参考。因此，我们将在研究中使用 MARL (Multi-Agent Reinforcement Learning)来解

决多信道聚合接入中的 V2X 频谱访问控制问题。

本文研究了多信道车联网的频谱聚合共享问题，多个 V2V 链路通过多信道频谱聚合共享 V2I 链路使用的频谱。为了实现车联网的多样化需求，我们设计了 V2V 链路的频谱聚合和功率分配方案。该方案最大限度地提高 V2V 链路的安全信息负载交付率，同时实现 V2I 链路的高带宽内容交付。在本研究中至少在两个方面与现有研究有所不同。首先，本研究结合频谱聚合的方法，在提高 V2V 负载交付率的同时，还扩展了 V2I 链路的传输总带宽。其次，本研究提出了一种基于 MARL 的频谱聚合共享方法，让 V2V 链路作为 agent 以合作的方式根据本地 CSI 分布式地决策出频谱聚合接入策略。

2. 问题建模

如图 1 所示，本研究基于蜂窝网对车联网进行建模，如 3GPP 第 15 版中针对蜂窝 V2X 网络补充中所讨论的一样[8]。本模型包含 M 条 V2I 链路和 K 条 V2V 链路，在为车载高速娱乐提供同时支持的同时，为高级驾驶服务提供稳定的定期安全消息共享。V2I 链路利用蜂窝网的 Uu 接口将车辆连接到 BS 以提供高速率内容服务，V2V 链路则通过具有本地化 D2D 通信的侧链 PC5 接口传输周期生成的安全消息。我们假设所有收发器都使用单个天线，车联网中 V2I 和 V2V 的聚合信道集合可表示为 $M=\left\{M_1^1,M_2^1,\cdots,M_h^1,M_1^2,\cdots,M_h^m\right\}$ 和 $K=\left\{K_1^1,K_2^1,\cdots,K_g^1,K_1^2,\cdots,M_g^k\right\}$，其中 h 表示每条 V2I 链路聚合的信道数，g 表示每条 V2V 链路聚合的信道数。

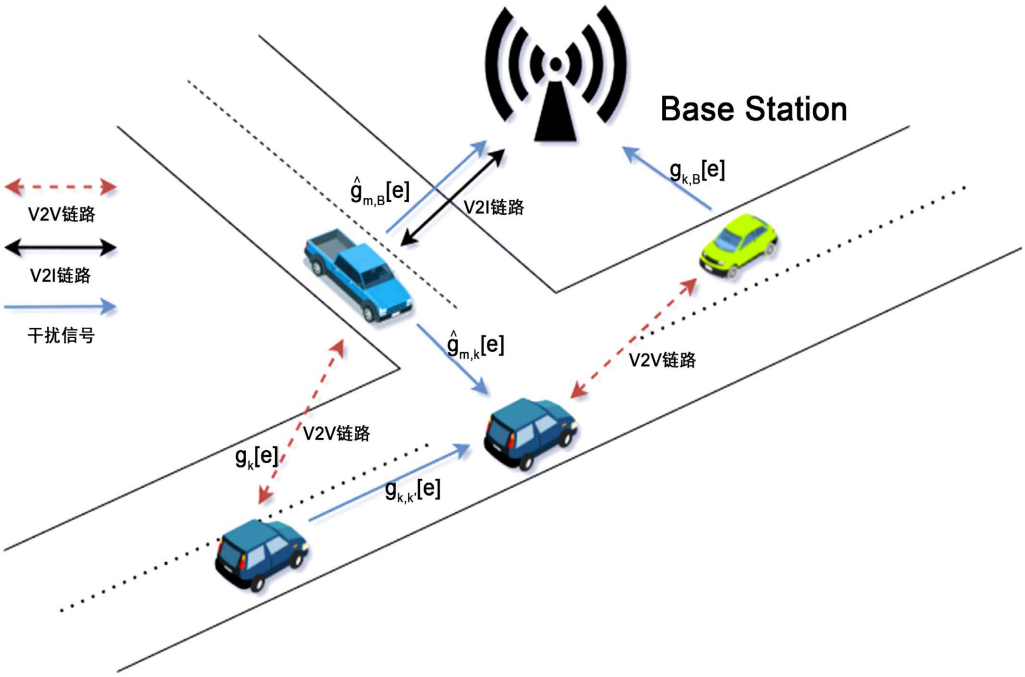

Figure 1. Vehicular network scenario diagram
图 1. 车联网场景示意图

在蜂窝 V2X 架构的模式 4 中，车辆有一个无线电资源池可以自主选择是否用于 V2V 通信[9]。如果有合适的干扰管理设计，那么该资源池就能与 V2I 链路的资源池重叠，提高频谱利用率。我们在考虑上行链路的情况下，进一步假设 m 条 V2I 链路已经提前分配完了聚合范围内具有固定传输功率的正交频谱子带，即第 m 条 V2I 链路使用了第 $h\times(m-1)+1$ 到 $h\times m$ 共 h 条频谱子带。因此，设计的主要任务是如何为 V2V 链路设计一种有效的频谱聚合共享方案，使 V2I 和 V2V 链路在车辆复杂高机动性环境下以最小

的信令开销实现目标。

我们将几个连续的子载波组成一个频谱子带，假设在同一个子带内信道衰落相同，并且不同子带之间独立。第 m 条 V2I 链路和第 k 条 V2V 链路在第 e 条子带上的接收信噪比表示为

$$\gamma_m^i[e]=\frac{P_m^i[e]\hat{g}_{m,B}[e]}{\sigma^2+\sum_k \rho_k[e]P_k^g[e]g_{k,B}[e]}, \tag{1}$$

和

$$\gamma_k^j[e]=\frac{P_k^j[e]g_k[e]}{\sigma^2+I_k[e]}, \tag{2}$$

其中 $P_m^i[e]$ 和 $P_k^j[e]$ 分别表示第 m 条 V2I 链路和第 k 条 V2V 链路在第 e 条子带上的发射功率，i 和 j 分别表示 V2I 链路和 V2V 链路在聚合范围内的子带序号。在第 e 条子带上，第 k 条 V2V 链路功率增益为 $g_k[e]$，第 k'条 V2V 链路对第 k 条 V2V 链路的干扰衰减为 $g_{k',k}[e]$，第 k 条 V2V 链路对 BS 的干扰衰减为 $g_{k,B}[e]$，第 m 条 V2I 链路对 BS 的干扰为 $\hat{g}_{m,B}[e]$，第 m 条 V2I 链路到第 k 条 V2V 链路的干扰为 $\hat{g}_{m,k}[e]$，σ^2 是噪声功率，干扰功率表示为

$$I_k[e]=P_m^i\hat{g}_{m,k}[e]+\sum_{k'\neq k}\rho_{k'}[e]P_{k'}^i[e]g_{k',k}[e], \tag{3}$$

其中 $\rho_k[e]$ 是二元频谱分配标志，$\rho_k[e]=1$ 表示 V2V 链路在使用第 e 条子带，否则标志为 0。我们假设每条 V2V 链路聚合 g 条子带，即 $\sum_g \rho_k[e]\le g$。第 m 条 V2I 链路和第 k 条 V2V 链路在第 e 条子带上的带宽为

$$C_m^i[e]=W\log\left(1+\gamma_m^i[e]\right), \tag{4}$$

和

$$C_k^j[e]=W\log\left(1+\gamma_k^j[e]\right), \tag{5}$$

其中 W 是每条频谱子带的带宽。

V2I 链路目的是提供移动环境下的高速率服务，因此其设计目标是最大化传输总带宽 $\sum_m\sum_h C_m^i[e]$。同时，V2V 链路负责周期性传递高级驾驶服务的关键安全消息。对于这样的需求，V2V 链路在固定时间约束 T 内交付率为

$$\Pr\left\{\sum_{t=1}^{T}\sum_{k=1}^{K}\sum_g \rho_k[e]C_k^j[e,t]\ge B/\Delta T\right\}, \tag{6}$$

式中 B 表示周期性生成的 V2V 有效负载的大小，ΔT 是信道相干时间，$C_k^j[e,t]$ 中的时隙 t 表示第 k 条 V2V 链路传输时不相干时隙的间隔。

因此，本文研究的车联网多信道聚合频谱共享问题可以表述为：设计一个合理的 V2V 频谱聚合共享策略，用二进制变量 $\rho_k[e]$ 和 V2V 传输功率 $P_k^j[e]$ 同时最大化 V2I 链路的总带宽 $\sum_M\sum_h C_m^i[e]$ 和(6)中定义的 V2V 链路的数据交付率。

3. 基于 MARL 的多信道频谱聚合分配

车辆高移动性带来的复杂环境变化使得集中处理全局 CSI 不切实际，因此采用分布式来进行决策更加合理。在本研究中，多条 V2V 链路尝试聚合共享 V2I 链路占用的有限频谱可以建模为 MARL 问题。每条 V2V 链路作为一个 agent 并与未知环境交互获得经验，然后将其用于更新自己的决策网络。多个 agent

共同探索环境，并根据自己对环境状态的观察来改善频谱聚合和功率控制策略。本文通过对所有 agent 设置相同的奖励，将原本的竞争博弈变为了合作博弈。提出的基于 MARL 的方法分为训练和应用两个阶段，本文将聚焦于集中式训练和分布式应用的设置。在训练阶段，每个 agent 都可以单独获得以系统性能为导向的奖励，然后通过更新 DQN (Deep Q-Network)来使动作决策调整为最优策略。在应用阶段，每个 agent 基于其本身对于环境的观察，使用与小规模信道衰落相当的时间尺度上根据训练出来的 DQN 进行动作选择。

3.1. 状态与观测空间

在频谱聚合共享问题的 MARL 模型中，在时间约束 T 内，每条 V2V 链路作为一个 agent，按时间步 t 为周期同时进行未知环境探索[10] [11]。在数学以按照部分可观察马尔可夫决策过程进行建模。如图 2 所示，在每个相干时间的时间步 t 内，给定当前环境状态 S_t，每个 agent 观测的环境 $Z(k)_t$ 由观测函数 O 确定 $Z(k)_t = O(S_t, k)$，然后采取一个动作 $A(k)_t$，形成一个联合动作 A_t。之后 agent 收到奖励 R_t，环境以概率 $P(s', r \mid s, a)$ 进入下一个状态 S_{t+1}，每个 agent 也能收到新的观测值 $Z(k)_{t+1}$，此时所有的 agent 共享同一个奖励，以促进他们的合作意识。

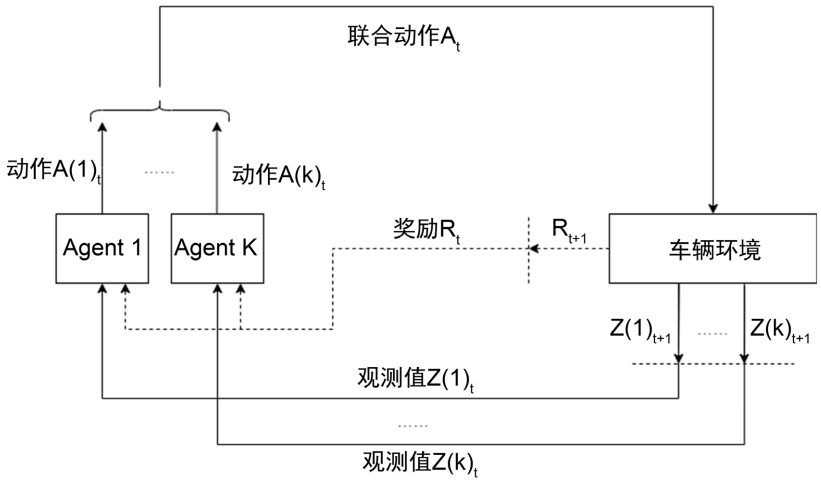

Figure 2. The interaction between agent and vehicular network
图 2. Agent 与车联网交互

环境状态 S_t 包含全局 CSI 和所有 agent 的动作，但对于单独的 agent 来说 S_t 是未知的，每个 agent 只能通过观测函数来获取局部 CSI。局部 CSI 包含 agent 自身对所有信道的干扰 $g_k[e]$，其在信道 e 上受到其他 agent 的衰减 $g_{k',k}[e]$，以及此 agent 受到 V2I 链路的衰减 $\hat{g}_{m,k}[e]$。这些信道信息在每个时隙 t 开始时由 agent 提前计算完毕，我们假设这些信息可以被 agent 即时获取[12]。而 agent 对 BS 的干扰 $g_{k,B}[e]$ 在每个时隙 t 开始时在 BS 处计算，以很小的开销广播给其覆盖范围内所有车辆。(4)中的信道上的干扰功率 $I_k[e]$ 可以在 agent 处获得。此外，观测空间还包括表示 agent 传输状态的 V2V 剩余有效载荷 L_K 和剩余时间预算 T_k。因此，agent 的观测函数表示为

$$O(S_t, k) = \left\{L_k, T_k, \left\{I_k[e]\right\}, \left\{G_k[e]\right\}\right\}, \tag{7}$$

其中 $G_k[e] = \left\{g_k[e], g_{k',k}[e], g_{k,B}[e], \hat{g}_{m,k}[e]\right\}$。

通常在 MARL 问题中[13]，每个 agent 将其他 agent 视为环境的一部分，基于自己的动作和观测进行分布式学习。在 DQN 训练时，经验回放是必要的步骤，但每个 agent 面临都是非稳态环境，同时其他 agent

也在训练并调整他们的动作，这会导致经验回放取样的经验和当前环境相关性不高，DQN 训练将很难收敛。为了解决这个问题，在[11]中提出了一个低维特征方案来跟踪其他 agent 的策略变化。其思想是，在每个 agent 的观测空间里增加其他 agent 策略的低维特征观测项来促进模型收敛。同时 agent 的策略由一个高维 DQN 组成，低维特征代替高维策略可以避免模型复杂度的提升。进一步分析发现，Q-Learning 常用的 ε -greedy 策略中，agent 的决策变化与训练迭代次数 e 及其探索速率 ε 高度相关。因此 agent 的观测函数表示为

$$Z(k)_t = \{O(S_t,k),e,\varepsilon\} \tag{8}$$

3.2. 动作空间

车联网的多信道频谱聚合共享设计的动作可视为 V2V 链路的频谱子带聚合位置选择和发射功率控制。如图 3 所示，频谱分成 $h\times B$ 条不相交的子带，由 B 条 V2I 链路分别占据 h 条连续信道，K 条 V2V 链路分别独立地选择 g 条连续信道进行数据传输，每条 V2V 链路有 $h\times B-(K-1)$ 种聚合位置选择。出于训练和实际算力的限制，我们将 V2V 功率控制限制为[18, −100] dBm 两个级别，低于 V2I 链路的 23 dBm 传输功率。在功率选择中，−100 dBm 实际上意味着 V2V 传输功率为零。同时，对于 h 和 g 两个频谱聚合宽度都选取为 4 个信道。因此动作空间的维度为 $2\times\left[4\times B-(K-1)\right]$，每个动作对应一种特定的频谱子带聚合位置选择和功率选择的组合。

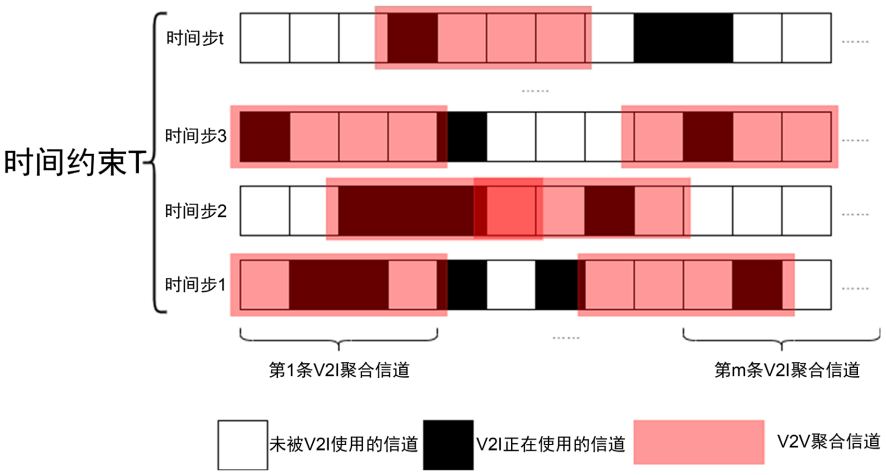

Figure 3. Agent aggregation channel action
图 3. Agent 聚合信道动作

3.3. 奖励设计

在第 2 节中说明了本研究有两个实现目标：在时间约束 T 内，最大化 V2I 传输总速率的同时，尽可能提高 V2V 有效负载的交付成功率。针对第一个目标，在每个时间步 t 中，我们简单地将奖励设为如(4)中定义所有 V2I 链路的即时总带宽 $\sum_{m\in M}C_m^i[e,t]$。为了实现二个目标，对每个 agent 我们都将奖励 L_k 设置为 V2V 传输速率减去未传输负载 B_n，直到有效负载交付完成，之后奖励设置为常数 β，其值大于最大的 V2V 传输速率。因此，每个时间步 t 的 V2V 相关奖励设置为

$$L_k(t)=\begin{cases}\sum_{k=1}^{K}\sum_g \rho_k[e]C_k^j[e,t]-B_n, & \text{if } B_k>0,\\ \beta, & \text{others}\end{cases} \tag{9}$$

训练目的是找到一个从状态集合 S 到动作集合 A 概率映射的最优策略 π^*，它使任意初始状态 s 的预期回报最大化，含折扣率 γ 的累积折扣回报 G_t 表示为

$$G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}, \quad 0 \le \gamma \le 1. \tag{10}$$

若将折扣率 γ 设置为 1，V2V 链路的累积奖励越大，传输数据量就越大，直到有效载荷交付完成。因此，当剩余有效载荷 $B_k > 0$ 时，最大化累积奖励将会使 V2V 链路传输更多数据。此外，训练中获得越多 β 奖励，V2V 有效负载的成功传输率就越高。

在训练中，理想情况下 β 的值要小于调整经验中 V2V 传输速率最大值的两倍。若将每个时间步 t 的奖励设置为 0，直到 V2V 有效负载传输完成后奖励变为 1，这样 agent 在每一回合开始时很难得到有效的反馈，模型很难收敛。所以在训练前将先验知识传授给奖励，这样有助于提高 V2V 有效载荷的成功交付率。当环境干扰过强时，agent 会采取加大传输速率的动作，这会造成环境干扰继续加强的负反馈循环。因此，我们提出了(9)中奖励设计，在未完成负载传输时将奖励与传输速率和剩余负载挂钩来避免上述极端的奖励设计。最终每个时间步 t 的奖励设置为

$$R_{t+1} = \lambda_m \sum_m C_m^i [e,t] + \lambda_k \sum_k L_k(t), \tag{11}$$

其中 λ_m 和 λ_k 用来调整 V2I 和 V2V 在设计中的权重。

3.4. 学习算法

在本研究场景中，每个回合开始会初始化包含所有 V2V 链路初始传输功率和 CSI 的环境状态，并开始传输大小为 B 的 V2V 负载直到时间约束 T 结束，若负载提前传输完成传输速率会提前降为 0。期间小规模信道衰落的变化会改变环境状态，并让每个 agent 有针对地调整其动作。

1) 集中式训练：我们同时对多个 agent 采用深度 Q-Learning 和经验回放[14]方法来训练其频谱聚合共享策略。Q-Learning [15]基于策略 π 的动作价值函数 $q_\pi(s,a)$，它被定义为从状态 s 开始，遵循策略 π 来进行动作 a，表示为

$$q_\pi(s,a) = \mathbb{E}_\pi [G_t \mid S_t = s, A_t = a], \tag{12}$$

其中 G_t 在(10)中定义。一旦获得动作价值函数，最优策略 $q^*(s,a)$ 也能推导出来[16]。中表明，随着学习率的随机近似条件变化和所有的状态－动作对不断更新，Q-Learning 中训练完成的动作价值函数必将收敛到最优 q^*。在深度 Q-Learning [14]中，用 θ 参数化的深度神经网络 DQN 表示动作价值函数。

每个 agent 都有一个专用的 DQN，其将当前观测 $Z(k)_t$ 作为输入，输出所有动作对应的价值。我们进行多次回合来确保 DQN 训练成功，训练中 agent 使用 ε-greedy 策略探索状态动作空间，这意味着为了避免收敛到局部最优，除了选出最大估值的动作外还可能以 ε 的概率来随机选取一个动作。因为信道环境一直在变化，所以每个 agent 都会收集当前时刻转换元组 $\left(Z(k)_t, A(k)_t, R_{t+1}, Z(k)_{t+1}\right)$ 并存储在回放内存中。在经验回放中，每一回合从内存中均匀地抽取一小批转换元组 D 通过随机梯度下降方法更新 θ：

$$\sum_D \left[R_{t+1} + \gamma \max_{a'} Q(Z_{t+1}, a'; \theta^-) - Q(Z_t, A_t; \theta) \right]^2, \tag{13}$$

式中 θ^- 是目标 Q 网络的参数集，到达一定回合数后固定从训练 Q 网络参数集 θ 中复制。训练过程在算法 1 中总结：

算法 1 基于 MARL 的频谱聚合分配学习算法

1) 启动模拟环境，生成车辆，建立链路
2) 随机初始化所有 agent 的 Q 网络
3) **For** 每一回合 **do**
4) 更新车辆位置与大尺度衰落
5) 重置所有 agent 的传输负载 B_k 和传输时间 T_k
6) **For** 每个时间步 t**do**
7) **For** 每个 agent**do**
8) 观测环境 $Z(k)_t$
9) 在 $Z(k)_t$ 中根据 ε-greedy 策略选择动作 $A(k)_t$
10) **Endfor**
11) 所有 agent 获得采取行动后的奖励 R_{t+1}
12) 更新小尺度衰落
13) **For** 每个 agent**do**
14) 观测环境 $Z(k)_{t+1}$
15) 将$(Z(k)_t, A(k)_t, R_{t+1}, Z(k)_{t+1})$储存在回放内存 D 中
16) **End for**
17) **Endfor**
18) **For** 每个 agent**do**
19) 从 D 中均匀取样小批次样本
20) 如(13)所示，用随机梯度下降方法优化 Q 网络
21) **Endfor**
22) **Endfor**

2) 分布式实现：在每个时间步 t 开始时，每个 agent 根据 ε 和(8)得出最新的环境观测值 $Z(k)_t$，然后通过观测值和训练完的 Q 网络选出估值最大的动作 $A(k)_t$。所有 agent 按照 $A(k)_t$ 进行发射功率和聚合频谱子带选取并进行传输。

在算法 1 中，集中式训练可在不同信道环境下离线执行多回合，而低消耗的实现过程可在线执行以进行网络部署。所有 agent 训练的 DQN 只需要在环境发生重大变化时进行更新，具体时间取决于环境变化和网络性能要求。

4. 仿真分析

我们按照 3GPP TR 36.885 [17]附录 A 中定义的城市案例评估方法搭建模拟场景，该方法详细描述了车辆衰落模型、密度、速度、移动方向、车辆通道、V2V 数据流量等。表 1 列出了主要仿真参数，本文所有参数均设置均依照表 1。

Table 1. Simulation parameters
表 1. 仿真参数

参数	数值
V2I 链路数 M	3
V2V 链路数 K	3
V2V 链路聚合信道数量	4
车辆衰落和移动模型	[17]中 A.1.2 的城市模型
V2I 传输功率	23 dBm
V2V 传输功率	[18, −100] dBm

Continued

V2V 负载传输时间约束 T	100 ms
V2V 负载传输时间步 t	1 ms
V2V 负载大小 B	[1, 2, …] × 1060 btypes

每个 agent 的 DQN 由 3 个全连接的隐藏层组成，分别包含 500 个、250 个、120 个神经元。校正线性单元 ReLU 用作激活函数 $f(x)=\max(0,x)$。RMSProp 优化器[18]以 0.001 的学习率更新神经网络参数。每个 agent 的 Q 网络训练 3000 回合，在前 2400 回合探索率 ϵ 从 1 线性退火降到 0.02 后保持不变。此外，训练阶段将 V2V 有效负载大小为 2 × 1060 btypes，但测试阶段会改变负载以验证本方法的鲁棒性。

图 4 是训练时模型的累积奖励，每回合累积奖励随着训练迭代次数不断增加直到趋于稳定，证明所研究的 MARL 达到收敛状态。从图中可以看出，当训练回合数达到 2600 左右时，尽管 CSI 由于车辆移动引起的信道衰落而出现一些性能的波动，但总体趋于收敛。基于对该图的观察，可以推断对 Q 网络进行 3000 回合的训练，能保证 DQN 达到评估 V2I 和 V2V 链路性能的标准。

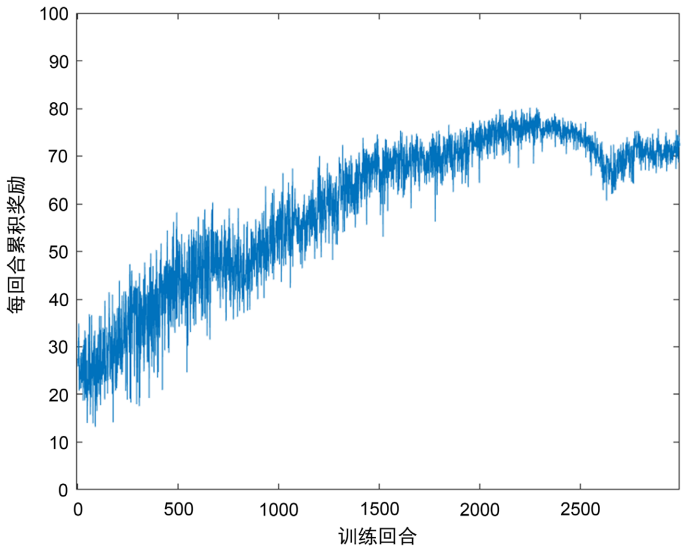

Figure 4. Training accumulative reward
图 4. 训练累积奖励

我们将以下两种分布式执行的方法作为 baseline，并在图 5 和图 6 中与 MARL 进行比较。一种是基于单代理 RL 的算法 SARL [19]，该方法中所有 agent 共享一个 DQN，每一时间步 t 只有一个 agent 基于训练的 DQN 更新其动作，而其他 agent 的行为保持不变。另外一种方法是随机 baseline，在每个时间步 t 开始时以随机方式为每个 agent 选择频谱聚合位置和传输功率。

图 5 显示了不同方法在不同 V2V 传输负载下 V2I 的传输性能。在图中可以看出，随着 V2V 有效负载的增加，所有方案的性能都会下降。V2V 有效负载增加会延长 V2V 传输持续时间和输出更高的 V2V 传输功率，这将在较长时间内对 V2I 链路造成更强的干扰。虽然它以 2 × 1060 字节的固定大小负载进行训练，但这足以证明其对 V2V 有效负载变化的鲁棒性。MARL 方法的性能与 baseline 比较，传输损失都只在 5 Mbps 内，性能差距几乎可忽略不计。

图 6 展现的是不同频谱聚合共享方案在有效负载增大时 V2V 传输成功率的变化。从图中可以看出，随着 V2V 有效载荷的增大，MARL 方法的传输成功率处于下降趋势，两个 baseline 方法则一直处于 20%到 40%

的传输成功率区间。与两种 baselines 分布式方法相比，本文所提出的 MARL 方法性能具有显著优势。对于小于 2 × 1060 字节的传输负载，MARL 方法的 V2V 负载传输成功率能高于 90%。并且对于 3 × 1060 字节的传输负载，MARL 的传输成功率也能高于 75%，而此时的传输负载大小已远远高于 V2V 所需的传输需求。

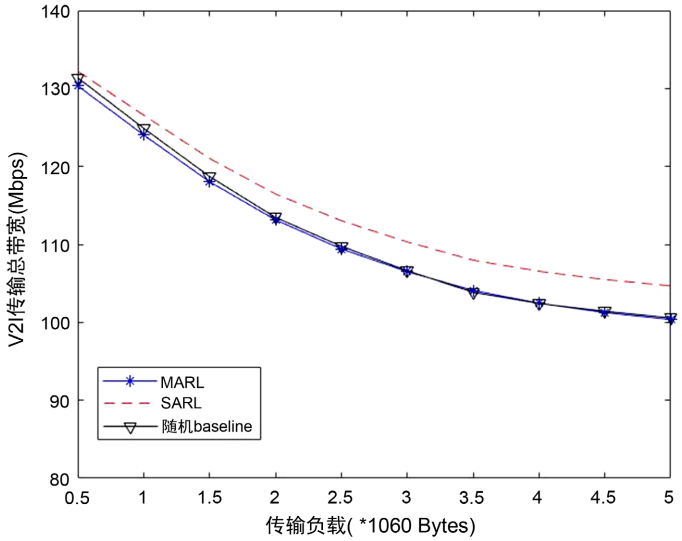

Figure 5. Total bandwidth of V2I transmission
图 5. V2I 传输总带宽

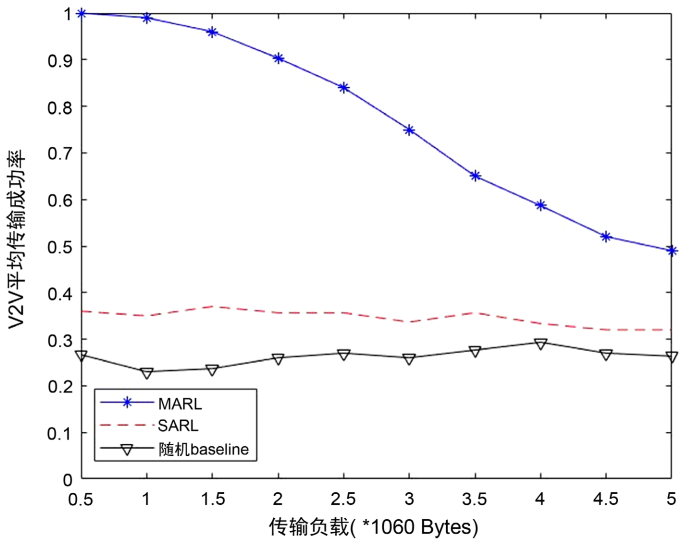

Figure 6. Average transmission success rate of V2V
图 6. V2V 的平均传输成功率

结合图 5 和图 6 来看，MARL 方法在基本没有牺牲 V2I 传输带宽的情况下，大幅度的提升了 V2V 的传输成功率。我们可以得出结论，MARL 方法在传输负载合适的区域内，经过训练的 DQN 表现良好，但如果 V2V 传输负债超出适应范围，则 DQN 需要重新训练。对于当前设置，我们可以得出结论，当数据包大小不大于 2 × 1060 字节时，不会发现明显的性能损失，即使 V2V 负载大小突增到 3 × 1060 字节，V2V 的传输成功率也保持在 75%以上。即使如此，我们仍然可确定所提出的 MARL 频谱聚合接入设计的优势，因为即使在未经训练的高负载情况下，它也优于其他两个分布式 baseline。

5. 结论

在本文中，我们研究了一种基于 MARL 的，用于具有多个 V2V 链路的车联网与其 V2I 链路共享其频谱。我们所提出的资源共享方案可以有效地鼓励 V2V 链路之间的合作，在不损失 V2I 传输性能的情况下，大幅度提高 V2V 负载的传输成功率。我们未来的工作将包括深入分析和比较 SARL 和 MARL 算法的鲁棒性，更好地理解何时需要更新 Q 网络以及如何有效地执行此类更新，以及尝试采用新的强化学习方法看是否可以提高传输效率。

参考文献

[1] Li, Y., Zhang, W., Wang, C.-X., Sun, J. and Liu, Y. (2020) Deep Reinforcement Learning for Dynamic Spectrum Sensing and Aggregation in Multi-Channel Wireless Networks. *IEEE Transactions on Cognitive Communications and Networking*, **6**, 464-475. https://doi.org/10.1109/TCCN.2020.2982895

[2] Poston, J.D. and Horne, W.D. (2005) Discontiguous OFDM Considerations for Dynamic Spectrum Access in Idle TV Channels. *First IEEE International Symposium on New Frontiers in Dynamic Spectrum Access Networks*, 2005, Baltimore, 8-11 November 2005, 607-610. https://doi.org/10.1109/DYSPAN.2005.1542679

[3] Botsov, M., Klügel, M., Kellerer, W. and Fertl, P. (2014) Location Dependent Resource Allocation for Mobile Device-to-Device Communications. 2014 *IEEE Wireless Communications and Networking Conference* (*WCNC*), Istanbul, 6-9 April 2014, 1679-1684. https://doi.org/10.1109/WCNC.2014.6952482

[4] Sun, W., Ström, E.G., Brännström, F., Sou, K.C. and Sui, Y. (2016) Radio Resource Management for D2D-Based V2V Communication. *IEEE Transactions on Vehicular Technology*, **65**, 6636-6650. https://doi.org/10.1109/TVT.2015.2479248

[5] Ye, H., Liang, L., Li, G.Y., Kim, J., Lu, L. and Wu, M. (2018) Machine Learning for Vehicular Networks: Recent Advances and Application Examples. *IEEE Vehicular Technology Magazine*, **13**, 94-101. https://doi.org/10.1109/MVT.2018.2811185

[6] Liang, L., Ye, H. and Li, G.Y. (2019) Toward Intelligent Vehicular Networks: A Machine Learning Framework. *IEEE Internet of Things Journal*, **6**, 124-135. https://doi.org/10.1109/JIOT.2018.2872122

[7] 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

[8] (2017) Technical Specification Group Radio Access Network. Study Enhancement 3GPP Support for 5G V2X Services. Release 15, Document 3GPP TR 22.886 V15.1.0, 3rd Generation Partnership Project.

[9] Molina-Masegosa, R. and Gozalvez, J. (2017) LTE-V for Sidelink 5G V2X Vehicular Communications: A New 5G Technology for Short-Range Vehicle-to-Everything Communications. *IEEE Vehicular Technology Magazine*, **12**, 30-39. https://doi.org/10.1109/MVT.2017.2752798

[10] Omidshafiei, S., Pazis, J., Amato, C., How, J.P. and Vian, J. (2017) Deep Decentralized Multi-Task Multi-Agent Reinforcement Learning under Par Tial Observability. *Proceedings of the* 34*th International Conference on Machine Learning*, (ICML), Sydney, 6-11 August 2017, 2681-2690.

[11] Foerster, J., *et al*. (2017) Stabilising Experience Replay for Deep Multi-Agent Reinforcement Learning. *Proceedings of the* 34*th International Conference on Machine Learning*, Sydney, 6-11 August 2017, 1146-1155.

[12] Nasir, Y.S. and Guo, D. (2018) Deep Reinforcement Learning for Distributed Dynamic Power Allocation in Wireless Networks. ArXiv: 1808.00490. https://arxiv.org/abs/1808.00490

[13] Tan, M. (1993) Multi-Agent Reinforcement Learning: Independent vs. Cooperative Agents. *Proceedings of the* 10*th International Conference*, *University of Massachusetts*, Amherst, 27-29 June 1993, 330-337. https://doi.org/10.1016/B978-1-55860-307-3.50049-6

[14] Mnih, V., et al. (2015) Human-Level Control through Deep Reinforcement Learning. *Nature*, **518**, 529-533. https://doi.org/10.1038/nature14236

[15] Watkins, C.J.C.H. and Dayan, P. (1992) Q-Learning. *Machine Learning*, **8**, 279-292. https://doi.org/10.1007/BF00992698

[16] Sutton, R.S. and Barto, A.G. (1998) Reinforcement Learning: An Introduction. MIT Press, Cambridge.

[17] (2016) Technical Specification Group Radio Access Network. Study LTE-Based V2X Services, Release 14, Document 3GPP TR 36.885 V14.0.0, 3rd Generation Partnership Project.

[18] Ruder, S. (2016) An Overview of Gradient Descent Optimization Algorithms. ArXiv: 1609.04747. https://arxiv.org/abs/1609.04747

[19] Ye, H., Li, G.Y. and Juang, B.-H.F. (2019) Deep Reinforcement Learning Based Resource Allocation for V2V Communications. *IEEE Transactions on Vehicular Technology*, **68**, 3163-3173. https://doi.org/10.1109/TVT.2019.2897134