Deep Reinforcement Learning Based Resource Allocation for V2V Communications

摘要

摘要：在本论文中基于深度强化学习开发了一种新的分布式的车对车 (V2V) 通信资源分配机制，该机制可以应用于单播和广播场景。根据分散的资源分配机制，自主的“代理”，在不需要或不必等待全局信息的情况下，自行决定寻找最佳的子频段和功率水平进行传输。由于所提出的方法是分布式的，它只产生有限的传输开销。从仿真结果来看，每个智能体都可以有效地学习满足 V2V 链路上严格的延迟约束，同时最大限度地减少对车与基础设施通信的干扰。

关键词：深度强化学习；V2V 通信；资源分配

1 引言

车对车 (V2V) 通信已经发展成为一项关键技术，通过支持近距离车辆之间的合作来增强交通和道路安全。出于安全考虑，对 V2V 通信的服务质量 (QoS) 要求非常严格，具有超低延迟和高可靠性。由于基于邻近的设备到设备 (device-to-device, D2D) 通信提供直接的本地消息传播，大大降低了延迟和能耗，第三代合作伙伴 (Third Generation Partnership, 3GPP) 支持基于 D2D 通信的 V2V 业务，以满足 V2V 应用的 QoS 要求。

为了管理 D2D 链路和蜂窝链路之间的相互干扰，需要有效的资源分配机制。在前人的工作中，提出了一种三步走的方法，其中控制发射功率并分配频谱，以最大限度地提高系统吞吐量，同时限制蜂窝和 D2D 链路的最小信噪比 (SINR)。在车对车通信网络中，高机动性车辆带来了新的挑战。由于高移动性导致无线信道的快速变化，传统的基于全信道状态信息 (CSI) 假设的 D2D 通信资源管理方法不再适用于 V2V 网络。

为了解决基于 D2D 的 V2V 通信中的新挑战，提出了资源分配方案。它们中的大多数以集中式的方式进行，其中 V2V 通信的资源管理在中央控制器中执行。为了做出更好的决策，每辆车必须向中央控制器报告本地信息，包括本地信道状态和干扰信息。在收集到车辆信息后，资源管理通常被表述为优化问题，其中 V2V 链路的 QoS 要求约束在优化约束中得到解决。然而优化问题通常是 NP-hard，并且通常很难找到最优解。而且由于解决资源分配优化问题需要向中央控制器报告车辆的信息，因此传输开销很大，并且随着网络规模的扩大而急剧增长，这使得这些方法无法扩展到大型网络。所以本文专注于分布式的资源分配方法，其中没有中央控制器收集网络的信息。此外，分布式资源管理方法也将更加自治和健壮，因为当支持基础设施中断或不可用时，它们仍然可以很好地运行。并且在以往的工作中，V2V 链

路的 QoS 仅包括 SINR 的可靠性，由于延迟约束难以直接表述为优化问题，因此没有对 V2V 链路的延迟约束进行深入的考虑。如今为了解决这些问题，本文使用深度强化学习来处理单播和广播车载通信中的资源分配问题。

在本文中，利用深度强化学习来寻找局部观测值 (包括局部 CSI 和干扰水平) 与资源分配和调度解决方案之间的映射。在单播场景中，将每条 V2V 链路视为一个 agent，根据对瞬时信道状况的观察和每个时隙与邻居共享的交换信息来选择频谱和发射功率。除了单播场景，基于深度强化学习的资源分配框架也可以扩展到广播场景。在这种情况下，每辆车都被视为一个代理，并根据学习到的策略选择频谱和消息。一般情况下，代理之间会自动平衡最大限度地减少 V2V 链路对 V2I 链路的干扰和满足 V2V 链路严格的延迟限制要求之间的关系 [1]。

2 相关工作

2.1 单播通信

2.1.1 集中式的 V2V 通信资源分配机制

由之前所述，集中式资源分配方案意味着其中 V2V 通信的资源管理在中央控制器中执行。在收集到车辆信息后，资源管理通常被表述为优化问题，其中 V2V 链路的 QoS 要求约束在优化约束中得到解决。然而，优化问题通常是 NP-hard 的，并且通常很难找到最优解。作为备选解决方案，通常将问题分为几个步骤，以便为每个步骤找到局部最优解和次最优解。例如有文献将 V2V 通信的可靠性和延迟需求转化为优化约束，仅使用大规模衰落信息即可计算，并提出了一种启发式方法来解决优化问题。还有文献仅基于信道缓慢变化的大尺度衰落信息开发了一种资源分配方案，在保证 V2V 可靠性的前提下优化了 V2I 遍历总容量。

2.1.2 分布式的 V2V 通信资源分配机制

近年来，一些分布式的 V2V 通信资源分配机制得到了发展。例如有文献提出了一种利用每辆车的位置信息进行 V2V 通信频谱分配的分布式方法。首先根据位置和负载相似度将 V2V 链路分组成簇。然后将资源块 (RBs) 分配给每个集群，并在每个集群内通过迭代交换两个 V2V 链路的频谱分配来改进分配。还有文献设计了一种基于二部匹配的 V2V 通信分布式算法来优化中断概率。

2.2 广播通信

在 V2V 通信的某些应用程序中，交换的消息没有特定的目的地。实际上，每个消息的兴趣区域包括周围的车辆，它们是目标目的地。使用广播方案来共享安全信息比使用单播方案更合适。但是，盲目广播消息会造成广播风暴问题，导致包碰撞。为了解决这个问题，前人研究了基于统计信息或拓扑信息的广播协议，已经提出了几种基于到最近发送方的距离的转发节点选择算法，其中 p-persistence 提供了最好的性能。

2.3 基于深度强化学习的分布式资源分配机制

近年来，深度学习在语音识别、图像识别、无线通信等领域取得了长足的进步。随着深度学习技术的发展，强化学习在许多应用中表现出了令人印象深刻的进步，它也被应用于各

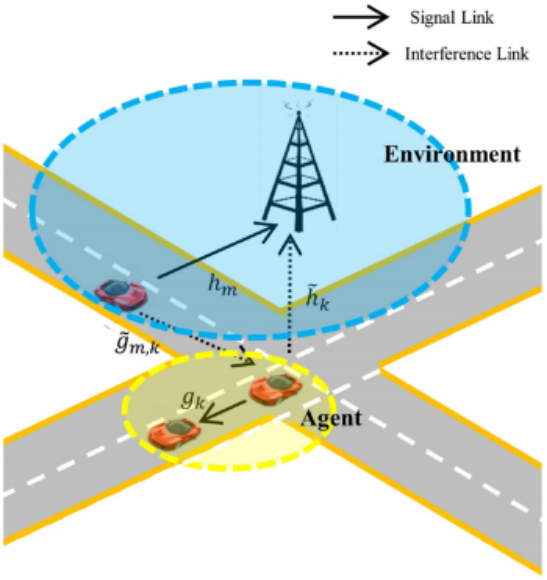

图 1. 单播车辆通信网络的说明性结构

个领域的资源配置。本文首次提出了一种基于深度强化学习的分布式资源分配机制，以解决 V2V 链路上的延迟限制。为了解决 V2V 资源分配问题，针对单播和广播场景，对动作空间、状态空间和奖励函数进行了细致的设计。

3 本文方法

3.1 单播通信的系统模型

如图 11所示，车载网络包含 M 个蜂窝用户 (CUE)，M =1,2，…，M 和 K 对 V2V 用户 (VUE)，表示 K =1,2，…K。CUE 需要 V2I 链路来支持与基站 (BS) 的高容量通信，而 VUE 需要 V2V 链路来共享交通安全管理的信息。

为了提高频谱利用效率，假设 V2I 链路正交分配的上行频谱由 V2V 链路共享。第 m 个 CUE 的 SINR 可以表示为:

$$\gamma^c[m] = \frac{P_m^c h_m}{\sigma^2 + \sum_{k\in K} \rho_k[m] P_k^v \widetilde{h}_k} \tag{1}$$

其中 p_m^c 和 p_k^v 分别表示第 m 个 CUE 和第 k 个 VUE 的发射功率，σ^2 为噪声功率，h_m 为第 m 个 CUE 对应的信道功率增益，$\widetilde{h}_k$ 为第 k 个 VUE 的干扰功率增益，$\rho_k[m]$ 为频谱分配指标，如果第 k 个 VUE 复用第 m 个 CUE 的频谱，则 $\rho_k[m] = 1$，否则为 $\rho_k[m]= 0$。第 m 个 CUE 的容量为:

$$C^c[m] = W \cdot \log(1 + \gamma^c[m]) \tag{2}$$

同理，第 k 个 VUE 的 SINR 可以表示为:

$$\gamma^v[k] = \frac{P_k^v g_k}{\sigma^2 + G_c + G_d} \tag{3}$$

其中,

$$G_c = \sum_{m\in M} \rho_k[m] P_m^c \widetilde{g}_{m,k} \tag{4}$$

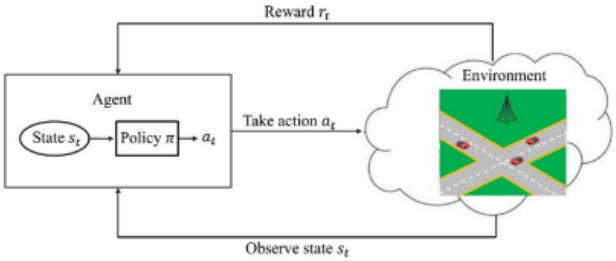

图 2. V2V 通信的深度强化学习

$$G_d = \sum_{m \in M} \sum_{k' \in K, K \notin k'} \rho_k[m] \rho_{k'}[m] P^v_{k'} \widetilde{g}^v_{k',k} \tag{5}$$

G_c 为 V2I 链路共用同一个 RB 的干扰功率，G_d 为共享同一 RB 的所有 V2V 链路的总干扰功率，g_k 为第 k 台 VUE 的干扰功率增益，$\widetilde{g}_{m,k}$ 为第 m 台 CUE 的干扰功率增益，$\widetilde{g}^v_{k',k}$ 第 k' 台 VUE 的干扰功率增益。

第 k 个 VUE 的容量可以表示为:

$$C^v[k] = W \cdot \log(1 + \gamma^v[k]) \tag{6}$$

由于 V2V 通信在车辆安全保护中的重要作用，对 V2V 链路有严格的延迟和可靠性要求，而数据速率相比并没有那么重要。在系统设计和考虑中，将 V2V 链路的时延和可靠性要求转换为中断概率。在深度强化学习中，这些约束被直接表述为奖励函数，当这些约束被违反时，就会给出负奖励。

而且 BS 中没有关于 V2V 链路的信息，所以整个 V2I 网络的资源分配过程应该独立于 V2V 链路的资源管理。然后在给定 V2I 链路的资源分配情况下，所提出的资源管理方案的目标是保证满足 V2V 链路的延迟约束，同时使 V2V 链路对 V2I 链路的干扰最小化。在分布式资源管理场景下，V2V 链路将根据本地观测值选择 RB 和传输功率。

3.2 单播资源分配的深度强化学习

强化学习的框架由 agent 和环境相互作用组成，在如图 22这个场景中，每个 V2V 链路都被视为一个代理，而特定 V2V 链路之外的一切都被视为环境，并且由于分布式设置下其他 V2V 链路的行为无法控制，因此每个 agent(单个 V2V 链路) 的行为都基于集体表现的环境条件，如频谱、传输功率等。

然后是本文比较重点的部分，奖励函数的设计，在文献中框架中，奖励函数由三部分组成，即 V2I 链路的容量、V2V 链路的容量和延迟条件。V2I 链路和 V2V 链路的总容量分别是用来度量对 V2I 和其他 V2V 链路的干扰，延迟条件表示为惩罚。因此奖励函数可以表示为:

为了获得长期的良好绩效，既要考虑眼前的奖励，也要考虑未来的奖励。因此，强化学习的主要目标是找到一个策略来最大化预期累积折扣奖励。

经典的 q -学习方法可以用来寻找状态-动作空间较小时的最优策略，在状态-动作空间中，可以维护一个查找表来更新每个项目的 q 值。然而，如果状态-动作空间变得巨大，就不能应用经典的 Q-learning，就像 V2V 通信的资源管理一样。这是因为大量的状态很少被访问，相应的 q 值也很少更新，导致 q 函数收敛的时间要长得多。为了解决这个问题，深度 Q-network 通过将深度神经网络 (DNN) 与 Q-learning 相结合来改进 Q-learning。

DNN 可以基于大量的训练数据来处理通道信息和期望输出之间的复杂映射，这些映射将用于确定 Q 值。Q-network 在每次迭代中更新它的权值 ，以最小化由相同的 Q-network 在数据集 D 上的旧权值得出的损失函数:

$$Loss(\theta) = \sum_{(s_t,a_t)\in D} (y - Q(s_t, a_t, \theta))^2 \tag{7}$$

其中,

$$y = r_t + \max_{a\in A} Q_{old}(s_t, a, \theta) \tag{8}$$

其中 r_t 为奖励函数,

$$r_t = \lambda_c \sum_{m\in M} C^c[m] + \lambda_d \sum_{k\in K} C^v[k] - \lambda_p(T_0 - U_t) \tag{9}$$

T_0 为最大可容忍延迟，λ_c,λ_d,λ_p 为三部分的权值，$(T_0 - U_t)$ 为传输所用时间。

3.3 在广播资源分配上的拓展

与单播场景类似，选择传输通道和消息的目标是在保证 VUE 延迟约束的情况下，最大限度地减少对 V2I 链路的干扰。为了达到这一目标，每辆车选择的频段和消息对所有 V2I 链路以及其他 VUE 的干扰都应该很小。它还需要满足延迟约束的要求。因此，与单播场景类似，奖励函数仍然由三部分组成:V2I 链路容量、V2V 链路容量和延迟情况。为了抑制冗余转播，只考虑未接收到消息的接收机的容量。因此，如果要重播的消息已被所有目标接收器接收，则不增加 V2V 链路的容量。如果消息没有被所有目标接收者接收，则延迟条件表示为惩罚，随着剩余时间 U_t 的减少，延迟条件线性增加。因此，奖励函数可以表示为:

$$r_t = \lambda_c \sum_{m\in M} C^c[m] + \lambda_d \sum_{k\in K, j\notin E\{k\}} C^v[k, j] - \lambda_p(T_0 - U_t) \tag{10}$$

其中，$E\{k\}$ 表示已接收到发送消息的目标接收者。广播的训练和测试算法与单播算法非常相似。

4 复现细节

4.1 与已有开源代码对比

作者分享的源码链接在 gitub 上目前消失了，但我从另外的搜索渠道找到了相关的源码引进。对于本论文的改进，我参考了别人的想法，由于 DDQN 算法可以视为 DQN 算法对连续型动作预测的一个扩展, 而且原论文没有提到使用 DDQN 算法的实验结果，因此我决定使用 DDQN 对本论文实验进行改进。通过在查阅资料了解深度强化学习，找到了 DDQN 的相关代码，并且加入到代码中，需要通过在初始定义 self.double_q = True 指令即可运行。

4.2 实验环境搭建

整个代码需要在 tensorflow 的环境下运行，因此通过查询资料成功安装了 Anaconda，并且搭建了 tensorflow 的环境。由于 tensorflow2 中无法运行源代码，因此这里整个环境采用的是 tensorflow1.14.0，并且安装对应版本的 7.4 的 CUDNN 以及 10 的 CUDA。并且安并装并导入对应所需要的库 *os time random numpy* = 1.13.1 *math matplotlib logging _pickle*

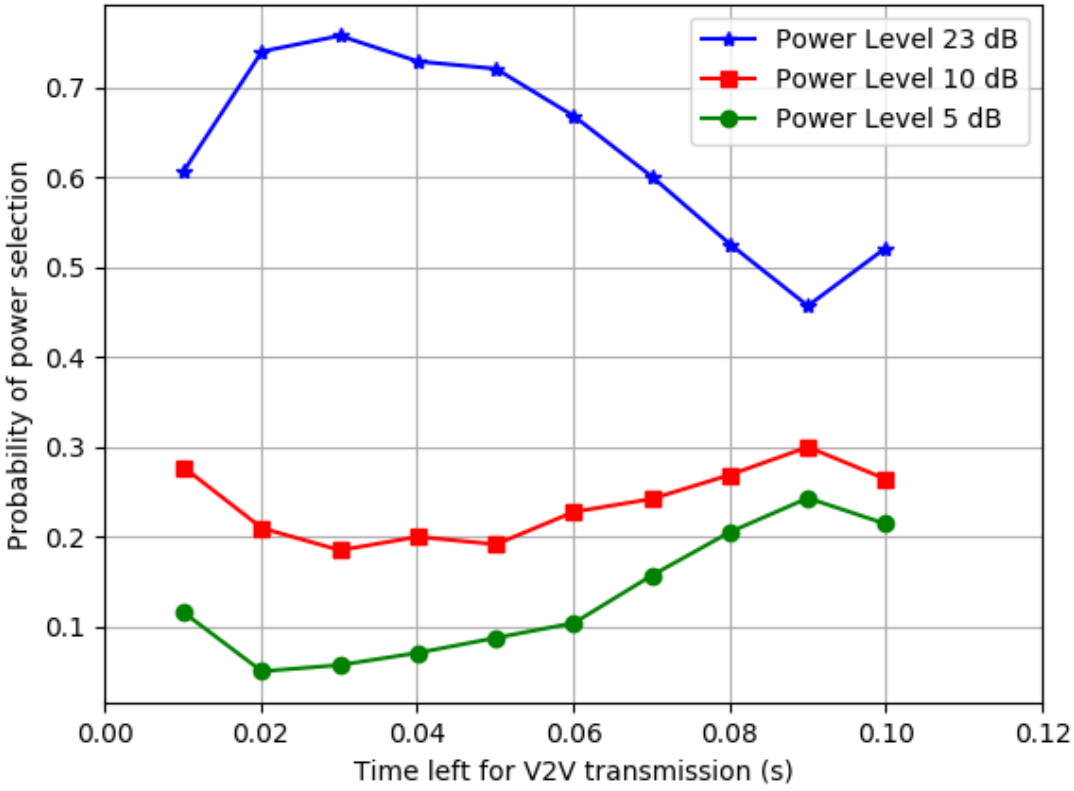

图 3. DQN

4.3 创新点

由于原论文采用的时 DQN 算法，有可能在当时的条件下这是当时基于深度强化学习比较基础的论文，论文中只使用了 DQN 算法。因此我对本本篇论文深度强化学习是采取更换了 DDQN 算法重复进行了实验，对比了 DDQN 算法以及 DQN 算法实验中产生的区别。

5 实验结果分析

整个一次实验需要大约一天至两天的时间，通过 DQN 以及 DDQN 两次实验分别对比，其中 DQN 实验与论文中的数据基本一致。

图 33，图 44分别显示了不同算法下 agent 在不同剩余时间下选择功率级别进行传输的概率。一般情况下，在传输时间充裕时，agent 选择最大功率的概率较低，而在传输时间较短时，为了保证满足 V2V 时延约束，agent 选择最大功率的概率较大。然而，当只剩下 10ms 时，选择最大功率水平的概率突然下降到 0.6 左右，因为 agent 了解到即使使用最大功率，也很有可能违反延迟约束，切换到更低的功率将通过减少对 V2I 和其他 V2V 链路的干扰而获得更多的奖励。

图 3 为通过 DQN 算法产生的结果，图 4 为通过 DQNN 算法产生的结果。通过对比可以发现差别主要在于 DQNN 当运用于 V2V 传输时的时间特别少时，选择一个最优通道的概率会更加降低，我认为是由于 DQNN 更能获取到较大奖励函数的值，当违反延迟约束的概率比

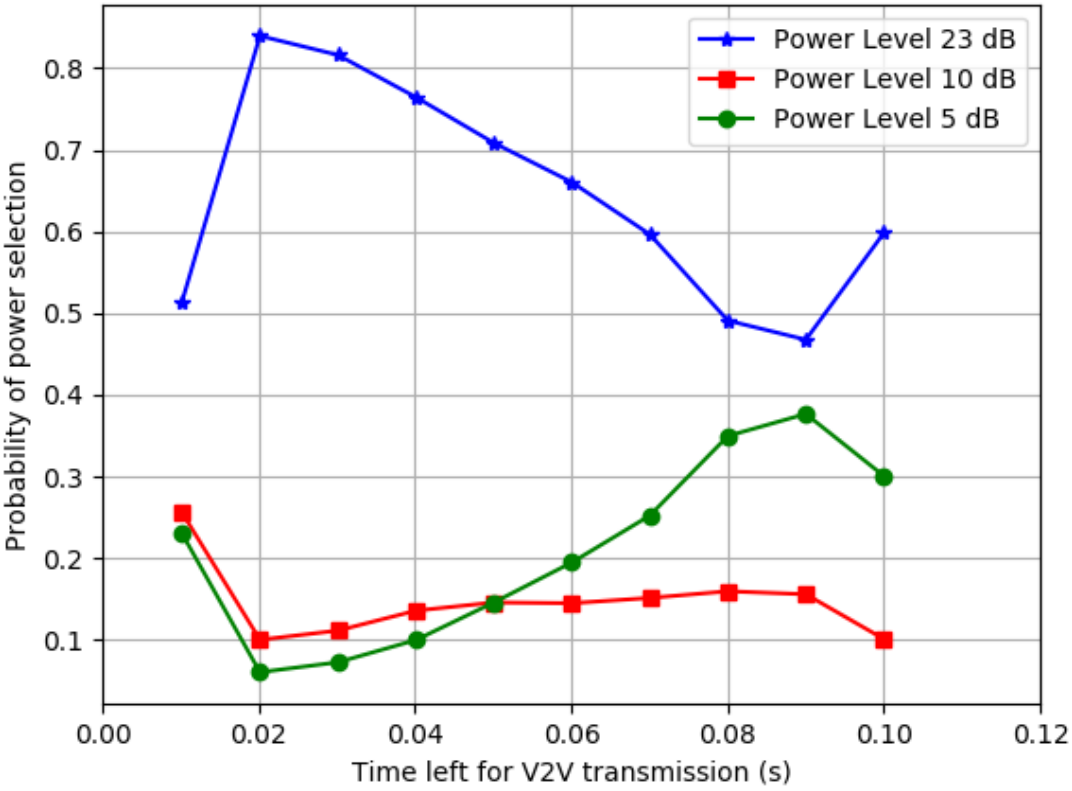
Power Level 23 dB
Power Level 10 dB
Power Level 5 dB
Probability of power selection
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.00
0.02
0.04
0.06
0.08
0.10
0.12
Time left for V2V transmission (s)

图 4. DDQN

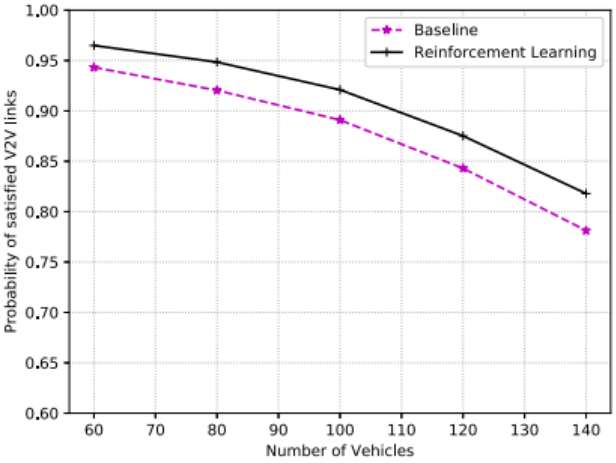

图 5. DQN 和 Baseline 满足 vue 的概率与车辆数量的关系

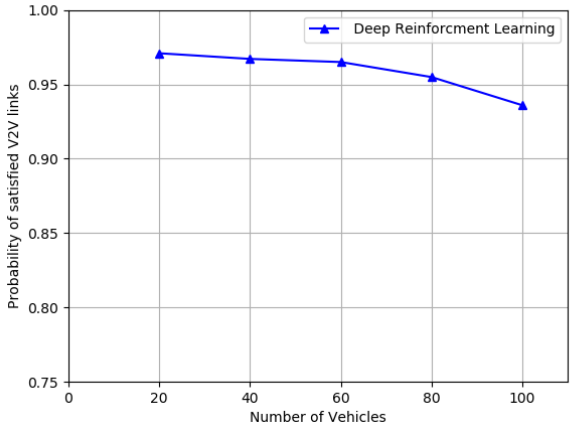

图 6. DDQN 满足 vue 的概率与车辆数量的关系

较大时，这里通过切换更低的功率来减少对 V2I 和其他 V2V 链路的干扰而获得更多的奖励。从当只剩下 20ms 时也可以看出，此时 DDQN 能选择最大功率的概率比 DQN 更高，能获取到的奖励也更高。因此在传输时间充裕的情况下，DDQN 的性能更加优越。

图 55显示了原论文中 DQN 和 Baseline random 下 vue 满足延迟约束的概率与车辆数量的关系，图 66显示了实验中 DDQN 算法下 vue 满足延迟约束的概率与车辆数量的关系，从结果图分析也可以看出 DDQN 相比 DQN 而言在满足延迟约束的概率上会更加优越。

6 总结与展望

论文中提出了一种基于深度强化学习的分布式 V2V 通信资源分配机制。单播和广播场景均可使用。由于所提出的方法是分散的，每个代理不需要全局信息来做出决策，因此传输开销很小。从仿真结果中，每个智能体可以学习如何在满足 V2V 约束的同时最小化对 V2I 通信的干扰。但在基于深度强化学习的机制上，V2V 算法还有很长的路可以走，仍然存在很高的性能提升空间。

参考文献

[1] Hao Ye, Geoffrey Ye Li, and Biing-Hwang Fred Juang. Deep reinforcement learning based resource allocation for v2v communications. *IEEE Transactions on Vehicular Technology*, 68(4):3163–3173, 2019.