基于参数化强化学习的车联网内容缓存和功率分配联合优化

雒江涛*② 杨和平① 冉泳屹②
①(重庆邮电大学通信与信息工程学院 重庆 400065)
②(重庆邮电大学电子信息与网络工程研究院 重庆 400065)

摘 要：车联网场景下的业务内容具有海量和高度动态的特性，使得传统缓存机制无法较好感知内容动态变化，且巨量接入设备与边缘缓存设备的有限资源之间的矛盾会引起系统时延性能差的问题。针对上述问题，该文提出一种基于学习的联合内容缓存和功率分配算法。首先，考虑联合优化内容缓存和功率分配，建立最小化系统整体时延的优化模型。其次，将该优化问题建模为马尔可夫决策过程(MDP)，并进一步将内容缓存和内容提供者的选择映射为离散动作集，并将功率分配映射为与离散动作相对应的连续参数。最后，借助参数化深度Q-Networks (P-DQN)算法求解这个具有离散-连续混合动作空间的问题。仿真结果表明，相较对比算法，该文所提算法能提高本地缓存命中率并降低系统传输时延。

关键词：车联网；内容缓存；功率分配；深度强化学习
中图分类号：TN929.5 **文献标识码**：A
DOI: 10.11999/JEIT220857

Joint Optimization of Content Caching and Power Distribution for Internet of Vehicles Based on Parametric Reinforcement Learning

LUO Jiangtao② YANG Heping① RAN Yongyi②
①(*School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing* 400065, *China*)
②(*Electronic Information and Networking Research Institute, Chongqing University of Posts and Telecommunications, Chongqing* 400065, *China*)

Abstract: The service content in the Internet of Vehicles scenario is massive and highly dynamic, which makes the traditional caching mechanism unable to better perceive the dynamic changes of the content, and the contradiction between the huge number of access devices and the limited resources of edge cache devices will cause the problem of poor system latency performance. In view of the above problems, a learning-based joint content caching and power allocation algorithm is proposed. First, considering the joint optimization of content caching and power allocation, an optimization model is established to minimize the overall system delay. Second, this optimization problem is modeled as a Markov Decision Process (MDP), and the selection of content caches and content providers is further mapped as discrete action sets, and power allocation is mapped as continuous parameters corresponding to discrete actions. Finally, this problem with a discrete-continuous mixed action space is solved with the aid of the Parametric Deep Q-Networks (P-DQN) algorithm. The simulation results show that the proposed algorithm can improve the local cache hit rate and reduce the system transmission delay compared with the comparison algorithms.

Key words: Internet of vehicles; Content caching; Power distribution; Deep reinforcement learning

1 引言

随着汽车技术和车载网络的发展，诞生了大量以提高驾驶安全性、旅行舒适性和车内娱乐性为目的的车载应用，这些应用常对计算、通信和存储资源有极大需求，并对服务质量(Quality of Service, QoS)有特定的要求(如传输延迟和响应时间)，这给仅依靠蜂窝网络从云端数据中心获取数据的车载网络带来巨大的压力[1,2]。车载边缘缓存技术通过将云缓存部分迁移至诸如路边单元(RoadSide Unit,

收稿日期：2022-06-27；改回日期：2022-11-16
*通信作者： 雒江涛 Luojt@cqupt.edu.cn
基金项目：国家自然科学基金(62171072, 62172064, 62003067)
Foundation Items: The National Natural Science Foundation of China (62171072, 62172064, 62003067)

RSU)的边缘缓存设备，以满足此类QoS要求[3]。但RSU的缓存容量和通信资源有限，需设计一个协同内容缓存和功率分配的联合优化策略。

除RSU的存储和功率资源有限之外，联合优化策略还需考虑如下3个问题：(1) 车联网场景中内容的流行度具有时变性，进行内容缓存时应充分考虑有限存储资源和内容流行度的动态变化。(2) 大量存在的联网车辆和内容带来了“维度灾难”问题。(3) 内容缓存和功率分配联合优化问题可被表述为混合整数非线性规划(Mixed Integer NonLinear Programming, MINLP)问题，该非凸问题的复杂度极高。

一些基于传统优化算法的方案研究了车载网络中内容缓存和资源分配问题。文献[4]提出了一种基于流行度和社会相似性的缓存策略，并利用块坐标下降法算法为计算和通信资源分配方案。文献[5]研究了在通信、计算和缓存资源约束下服务延迟的优化问题，将MINLP问题转化为线性规划问题，并提出了一种基于交替方向乘子法的新迭代算法以求解目标问题。文献[6]提出了一种缓存辅助延迟更新和交付方案，以平衡车载网络中的内容新鲜度和服务时延。文献[7]将协同内容缓存问题表达为无线资源和计算资源的联合优化，并使用蚁群优化算法解决该优化问题。文献[8]将雾无线接入网络的联合缓存和无线资源优化问题建模为云资源管理器和雾接入点之间的Stackelberg博弈，并提出雾接入点的分布式集群形成算法。然而，这些解决方案所建立的待解问题往往是NP-hard的，其高计算复杂度使得解决方案只能推导出接近最优的次优解，且这些解决方案大多也无法较好地捕捉车载网络拓扑结构和内容流行度的动态变化。

为此，部分研究方案提出了基于学习的策略。文献[9]提出了一种基于点对点联合学习的主动缓存方案，以提高缓存和延迟性能。文献[10]利用Hawkes过程适应内容流行度的动态变化，提出了一种基于深度强化学习(Deep Reinforcement Learning, DRL)的协同内容缓存方案。文献[11,12]使用深度确定性策略梯度(Deep Deterministic Policy Gradient, DDPG)算法求解车载网络计算和网络中的内容放置和内容交付问题。文献[13]为边缘缓存问题构建了一个基于多智能体框架的协作缓存解决方案。这些方案往往通过离散化连续动作或松弛离散动作来处理联合优化问题中的离散-连续混合动作空间问题，但松弛化离散动作空间会导致问题复杂度增加，离散化连续动作空间则会增加问题复杂性并降低准确率。

针对上述问题，本文提出一种基于P-DQN算法的内容缓存和功率分配联合优化算法。首先，将响应内容请求的过程划分为内容缓存和内容交付阶段，建立以请求为驱动的最小化系统整体时延的优化模型。其次，将该优化问题构建为马尔可夫决策过程，并进一步将内容缓存和内容提供者的选择映射为离散动作集，将功率分配作为连续参数与离散动作进行关联。最后，借助P-DQN算法求解这个具有离散-连续混合动作空间的问题。

2 系统模型与问题建模

2.1 无线通信模型

如图1所示，系统模型由一个宏基站(Macro Base Station, MBS)、一系列路边单元$L=\{1,2,\cdots,L^N\}$和一组车辆$K=\{1,2,\cdots,K^N\}$构成。MBS可通过云端数据中心(Cloud Data Center, CDC)获取到任意流行内容。假设各RSU的信号覆盖范围有限，且各RSU覆盖范围之间无重叠区域。本地RSU是相对于请求车辆而言，车辆在某RSU的信号覆盖范围内发起内容请求，则该RSU即为请求车辆的本地RSU。本地RSU与其邻居RSU通过诸如光纤之类的有线连接方式进行通信，且二者均可从MBS或其邻居RSU获取流行内容并缓存至本地。RSU和MBS之间、RSU与车辆之间以及MBS与车辆之间均可建立基于C-V2X标准的无线通信连接[14]。RSU收集状态信息并发送至MBS，这些状态信息包括内容请求状态、内容缓存状态和信道状态。部署于MBS上的智能体根据状态信息做出决策，并将决策信息下发至各RSU, RSU据此执行相应动作。决策信息包括内容缓存和功率分配：一方面，若决策信息要求本地RSU将请求内容缓存至本地，则在邻居RSU已缓存请求内容的情况下，本地

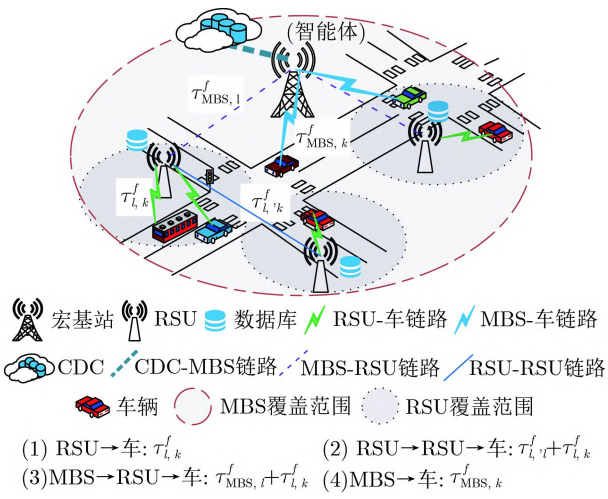

图 1 车联网内容缓存和交付模型

RSU从邻居RSU处获取该内容并缓存至本地，而在邻居RSU并未缓存该内容的情况下，本地RSU则从MBS处获取该内容并缓存至本地，该过程称为内容缓存阶段。另一方面，决策信息要求RSU或MBS以适当发射功率将请求内容交付给请求车辆，该过程称为内容交付阶段。系统以时隙为基础运行，将时间轴划分为T个持续时长为Δ的时隙，$t\in\{1,2,\cdots,T\}$为时隙索引。

2.2 内容模型

定义内容库集合$F=\{1,2,\cdots,F^N\}$，内容$f\in F$的大小定义为s_f。假设各时隙持续时间极短，车辆k在一个时隙内以概率$p\in(0,1)$发出最多一个内容请求，则k在t时隙发起内容请求的概率为$P(n_k(t)=1)=p$，其中$n_k(t)=1$表示车辆发起了内容请求，$n_k(t)=0$则表示未发起。已知车辆k在时隙t发起了内容请求的条件下，该车对内容k发起请求的概率为$P(n_f(t)=1|n_k(t)=1)=u_f(t)$，其服从Zipf分布[12]，其中$n_f(t)\in\{0,1\}$，$n_f(t)=1$表示内容$f$被请求，$n_f(t)=0$则表示未被请求。定义内容$f$的流行度$u_f(t)$为

$$u_f(t)=\frac{V(f)^{-\kappa}}{\sum_{i=1}^{F^N}V(i)^{-\kappa}} \tag{1}$$

其中，$\kappa\in[0.6,1.2]$为表征Zipf分布的偏度参数，$V(f)$为内容f的流行度排名。定义内容f的内容热度为π_f，内容f每被请求1次则π_f增加1。将内容热度集合$\{\pi_1,\pi_2,\cdots,\pi_{F^N}\}$按降序排列可得到序列$\Pi$，则具有最高内容热度的内容排列在序列$\Pi$的第1个位置，并以此类推。那么，将内容$f$在序列$\Pi$中的序号定义为该内容的流行度排名$V(f)$。车辆$k$在$t$时隙对内容$f$发起请求的概率为

$$\begin{aligned}&P(n_{k,f}(t)=1)\\&\quad=P(n_f(t)=1|n_k(t)=1)P(n_k(t)=1)\\&\quad=u_f(t)p\end{aligned} \tag{2}$$

其中，$n_{k,f}(t)\in\{0,1\}$为车辆在时隙t的请求状态，$n_{k,f}(t)=1$表示车辆k发起了对内容f的请求，$n_{k,f}(t)=0$则表示未发起。

定义缓存命中率z为

$$z=\frac{\sum_{l=1}^{L^N}\sum_{j=1}^{J_l}H_l(j)}{\sum_{l=1}^{L^N}J_l} \tag{3}$$

其中，J_l表示RSU$_l$覆盖范围下的车辆发起内容请求的总次数，$H_l(j)$为示性函数

$$H_l(j)=\begin{cases}1, & \text{命中}\\0, & \text{未命中}\end{cases} \tag{4}$$

其中，$H_l(j)=1$表示RSU$_l$覆盖范围内的车辆在第j次请求内容时，RSU$_l$已缓存了请求内容。

2.3 时延模型

MBS上的智能体根据RSU收集的状态信息产生并下发决策动作至RSU, RSU据此执行缓存和交付动作。鉴于实际业务大多为应用、视频下载等服务，与传输请求内容的时延开销相比，传输信令的时延开销可忽略不计，故本文主要考虑下行链路传输时延和排队时延。

2.3.1 传输时延

假设RSU$_l$的覆盖范围下的车辆k对内容f发起了请求，定义在时隙t时将f从其所在位置(MBS或RSU)发送到车辆k的总下行链路传输时延为$\tau_{k,f}^{\text{tran}}(t)$，则$\tau_{k,f}^{\text{tran}}(t)$有如下4种情况(图1)：

(1) 情况1。若RSU$_l$已缓存内容f，则RSU$_l$以最大可达传输速率k将内容f发送给车辆k，其中RSU$_l$为RSU$_l$与k之间的信号干扰噪声比，且RSU$_l$满足

$$\text{SINR}_{l,k}=\frac{h_{l,k}p_{l,k}}{\sum_{i\neq l,i\in L}h_{i,k}p_{i,k}+\sigma^2}\geq\gamma_{\min} \tag{5}$$

其中，$\gamma_{\min}$和σ^2分别为信号干扰噪声比阈值和噪声功率，$h_{l,k}$为RSU$_l$与车辆k之间的信道增益，$p_{l,k}$为RSU$_l$将内容f发送给车辆k时的发射功率。此种情况下，$\tau_{k,f}^{\text{tran}}(t)$等于RSU$_l$将$f$发送给车辆$k$的传输时延$\tau_{l,k}^{f}$，即$\tau_{k,f}^{\text{tran}}(t)=\tau_{l,k}^{f}=s_f/R_{l,k}$。

(2) 情况2。若决策动作要求RSU$_l$将内容f缓存至本地，且RSU$_l$的邻居RSU$_{l'}$已缓存f。由于RSU$_l$与RSU$_{l'}$之间通过光纤进行通信，为简化分析，设置二者间传输速率$R_{l',l}$为固定值[15]。那么，RSU$_l$先从RSU$_{l'}$处获取f，其传输时延为$\tau_{l,l'}^{f}=s_f/R_{l',l}$，而后RSU$_l$将$f$发送给请求车辆$k$，其传输时延为$\tau_{l,k}^{f}$，则有$\tau_{k,f}^{\text{tran}}(t)=\tau_{l,l'}^{f}+\tau_{l,k}^{f}$。

(3) 情况3。若决策动作要求RSU$_l$将内容f缓存至本地，但其邻居RSU$_{l'}$未缓存f，则RSU$_l$先从MBS处获取f，其传输时延为$\tau_{\text{MBS},l}^{f}=s_f/R_{\text{MBS},l}$，其中$R_{\text{MBS},l}$为MBS将$f$发送给RSU$_l$的传输速率，然后RSU$_l$将$f$发送给请求车辆$k$，那么$\tau_{k,f}^{\text{tran}}(t)=\tau_{\text{MBS},l}^{f}+\tau_{l,k}^{f}$。

(4) 情况4。若决策动作不要求RSU_l将内容f缓存至本地，则MBS将f发送至请求车辆k，此时$\tau_{k,f}^{\text{tran}}(t)=\tau_{\text{MBS},k}^{f}$，且$\tau_{\text{MBS},k}^{f}=s_f/R_{\text{MBS},k}$为MBS以传输速率$R_{\text{MBS},k}$将$f$发送给车辆$k$的传输时延。

综上所述，$\tau_{k,f}^{\mathrm{tran}}(t)$可表示为

$$\tau_{k,f}^{\mathrm{tran}}(t)=\begin{cases}\tau_{l,k}^{f}, & \text{RSU}\rightarrow\text{车}\\ \tau_{l',l}^{f}+\tau_{l,k}^{f}, & \text{RSU}\rightarrow\text{RSU}\rightarrow\text{车}\\ \tau_{\mathrm{MBS},l}^{f}+\tau_{l,k}^{f}, & \text{MBS}\rightarrow\text{RSU}\rightarrow\text{车}\\ \tau_{\mathrm{MBS},k}^{f}, & \text{MBS}\rightarrow\text{车}\end{cases} \tag{6}$$

2.3.2 排队时延

考虑到相邻RSU间通过具有极大容量的光纤传输内容，故认为相邻RSU间传输内容时不存在拥塞[15]。定义MBS发送内容至RSU_l和车辆的虚拟数据传输队列分别为Q_l^{MBS}和$Q_{\mathrm{V}}^{\mathrm{MBS}}$，$RSU_l$发送内容至车辆的虚拟数据传输队列为$Q_l$。不失一般性，假设某一内容在时刻$t_0$进入数据队列$Q$，则该内容的预期排队时延$T_{\exp}$可计算为

$$T_{\exp}(t_0)=\sum_{i=1}^{q(t_0)}\frac{s_i}{R_i} \tag{7}$$

其中，$q(t_0)$为数据队列Q中的文件数量，s_i和R_i分别为该队列中第i个内容的大小和传输内容i的传输速率。假设t_s为时隙t的起始时刻，则队列Q在时隙t的预期平均排队时延$T_Q(t)$为[16]

$$T_Q(t)=\frac{1}{\Delta}\int_{t_s}^{t_s+\Delta}T_{\exp}(\tau)\mathrm{d}\tau \tag{8}$$

与2.3.1节中讨论的传输时延的4种情况相对应：在情况1中，RSU_l在时隙t处将内容ff发送给车辆k，则此种情况有排队时延$D=\{0,1,2\}$；在情况2中，由邻居RSU将f发送至本地RSU，再由本地RSU将f发送至车辆k，由于忽略邻居RSU到本地RSU的排队时延，故有情况2的排队时延f；在情况3中，MBS将f发送至本地RSU，其排队时延为f，再由本地RSU将f发送给车辆，其排队时延为$T_{Q_l^{\mathrm{MBS}}}(t+\tau_{\mathrm{MBS},k}^{f}+T_{Q_l^{\mathrm{MBS}}}(t))$，则情况3的排队时延为$Q(s_t,d_t,p_d(s;\theta);\varphi)\approx\sup_{p_d\in P_d}Q(s,d,p_d;\varphi)$。在情况4中，由MBS将内容发送给车辆，此时的排队时延为φ。

综合上述对传输时延和排队时延的讨论，从车辆k发起对内容f请求到车辆k接收完成请求内容f之间的时延$\tau_{k,f}(t)$为

$$\tau_{k,f}(t)=\begin{cases}\tau_{l,k}^{f}+\tau_1^{\mathrm{que}}, & \text{RSU}\rightarrow\text{车}\\ \tau_{l',l}^{f}+\tau_{l,k}^{f}+\tau_2^{\mathrm{que}}, & \text{RSU}\rightarrow\text{RSU}\rightarrow\text{车}\\ \tau_{\mathrm{MBS},l}^{f}+\tau_{l,k}^{f}+\tau_3^{\mathrm{que}}, & \text{MBS}\rightarrow\text{RSU}\rightarrow\text{车}\\ \tau_{\mathrm{MBS},k}^{f}+\tau_4^{\mathrm{que}}, & \text{MBS}\rightarrow\text{车}\end{cases} \tag{9}$$

2.4 问题表述

本文提出一种车联网场景下的联合内容缓存和功率分配的优化模型，旨在最小化系统整体时延，建立模型为

$$\left.\begin{aligned}&\min_{c_{l,f},p_k}\sum_{t=1}^{T}\sum_{k=1}^{K^N}\sum_{f=1}^{F^N}(\tau_{k,f}(t)n_{k,f}(t))\\ &\text{s.t.C1:}\sum_{f=1}^{F^N}(c_{l,f}s_f)\le G_l,\forall l\\ &\quad\text{C2:}\sum_{k-1}^{K^N}p_{l,k}\le P_l^{\max},\forall l\\ &\quad\text{C3:}\sum_{k=1}^{K^N}p_{\mathrm{MBS},k}\le P_{\mathrm{MBS}}^{\max}\\ &\quad\text{C4}:\mathrm{SINR}_{l,k}\ge\gamma_{\min},\forall l,\forall k\end{aligned}\right\} \tag{10}$$

其中，$c_{l,f}\in\{0,1\}$为内容f在RSU_l中的缓存状态，$c_{l,f}=1$表示$\mathrm{RSU}_l\mathrm{RSU}_1$已缓存$f$，$c_{l,f}=0$则表示未缓存。$p_k\in\{p_{l,k},p_{\mathrm{MBS},k}\}$表示发射功率，$p_{l,k}$为$\mathrm{RSU}_l$将$f$发送给车辆$k$的发射功率，$p_{\mathrm{MBS},k}$为MBS将$f$发送给车辆$k$的发射功率。$G_l$为$\mathrm{RSU}_l$的存储容量。$P_l^{\max}$和$P_{\mathrm{MBS}}^{\max}$分别为$\mathrm{RSU}_l$和MBS的总功率。约束C1描述了RSU可以存储数量有限的内容，约束C2和C3则描述了RSU和MBS的总功率是有限的，约束C4描述了信号干扰噪声比值的下限以保证QoS。

3 基于P-DQN的算法设计

鉴于优化问题式(10)是一个MINLP问题，用常规方法不易解决，为此，本文提出了一种基于参数化DRL的联合优化内容缓存和功率分配算法。常用DRL算法包括深度Q-network(DQN)和DDPG，其中DQN使用神经网络逼近Q-learning的值函数，而DDPG则可看作DQN对连续动作预测的扩展。但DQN和DDPG分别适合处理具有离散动作和具有连续动作的问题，二者均无法单独直接处理具有混合动作空间的问题。因此，本文使用P-DQN算法解决目标优化问题，从而无需对混合动作空间进行离散化或松弛化处理。

3.1 DRL模型

3.1.1 状态空间

智能体根据状态空间做出决策动作。智能体需感知内容流行度变化，以动态决策请求内容的缓存，故而应考虑将内容流行度、各RSU中内容的缓存状况和车辆对各内容的请求状况作为状态空间组成部分。此外，在RSU或MBS交付内容时，智能体需能做出合理的功率分配决策，故而亦将请求内容的大小和信道增益作为状态空间组成部分。那么，将状态空间S定义为

$$S=(C,U,M,H) \tag{11}$$

其中，$C=\{c_{l,f}|c_{l,f}\in\{0,1\},\forall l,\forall f\}$表示内容缓存状态。$U=\{u_f\in(0,1),\forall f\in F\}$表示内容流行度，且$u_f$由式(1)计算。$M=\{f,s_f\}$表示请求内容信息，且$f$和$s_f$分别为请求内容的索引和该内容的大小。$H=\{h_{l,k},h_{\mathrm{MBS},k}\}$表示信道增益，且$h_{l,k}$为$RSU_l$和车辆$k$之间的信道增益，$h_{\mathrm{MBS},k}$为MBS和车辆$k$之间的信道增益。

3.1.2 动作空间

系统以车辆请求事件为驱动，即针对车辆发起的某个内容请求事件，智能体根据状态空间做出决策动作，该决策动作指示本地RSU是否将本次请求的内容缓存到本地，并指出应由哪一内容提供者(RSU或MBS)将内容交付给请求车辆，同时还指定交付该请求内容时发射功率的大小。定义离散动作d决定本地RSU是否将该请求内容缓存至本地，以及由MBS还是本地RSU将内容发送给请求车辆，并定义连续动作p决定发射功率大小，则决策动作可定义为$a=(d,p)$。进一步将离散动作与连续动作进行关联，即离散动作d对应的连续参数为p_d，即$a=(d,p_d)$。定义动作空间A为

$$A=\{(d,p_d)\,|d\in D,p_d\in P_d\} \tag{12}$$

其中，$D=\{0,1,2\}$为离散动作集。

$(d=0,p_d)$表示请求车辆所在的本地RSU已缓存请求内容f，并由本地RSU以发射功率p_d将f交付于请求车辆。$(d=1,p_d)$表示本地RSU未缓存内容f，本地RSU需先从邻居RSU或MBS处获取并缓存f至本地，本地RSU再以发射功率p_d将f交付于请求车辆，若邻居RSU缓存有内容f，本地RSU则从邻居RSU处获取f，否则，从MBS处获取f。$(d=2,p_d)$表示本地RSU未缓存请求内容f，且f在RSU覆盖范围内也不具有成为流行内容的可能性，则由MBS以发射功率p_d将f交付于请求车辆。

3.1.3 奖励函数

奖励函数与优化问题(10)密切相关，本文优化目标旨在最小化系统整体时延，故而设置奖励函数与时延$\tau_{k,f}(t)$呈负相关。此外，较高的缓存命中率反映出大部分内容请求由本地RSU服务，可极大缓解回程网络压力，故而设置奖励与缓存命中率z间呈正相关。综上所述，奖励函数$r(S_t,A_t)$表示为

$$r\left(S_t,A_t\right)=\begin{cases}\left(\tau^{\mathrm{tol}}-\tau_{k,f}(t)\right)(1+z), & d=0\\ \tau^{\mathrm{tol}}-\tau_{k,f}(t), & d=1\\ 0, & d=2\end{cases} \tag{13}$$

其中，S_t和A_t分别为t时隙的状态空间和动作空间，$\tau^{\mathrm{tol}}\tau^{\mathrm{toler}}$为最大容忍时延。

3.2 基于P-DQN的内容缓存和功率分配算法

P-DQN算法的整体流程图如图2所示。首先，使用确定性策略网络$p(\theta)$根据状态S生成连续动作值，其中θ为$p(\theta)$的网络权重。接着将连续动作值连同状态S输入至深度Q网络$Q(\varphi)$中，其中φ为$Q(\varphi)$的网络权重。最后，选择出Q值最大的离散动作及其对应的连续动作。

具体地，将动作值函数表示为$Q(s,a)=Q(s,d,p_d)$，其中$s\in S$。假设智能体在t时刻选择离散动作d_t，p_{d_t}为离散动作d_t相对应的连续参数值，则可将贝尔曼方程表示为

$$Q(s_t,d_t,p_{d_t})=\underset{r_t,s_{t+1}}{\mathbb{E}}\left[r_t+\gamma\max_{d\in D}Q\left(s_{t+1},d,p_d^Q(s_{t+1})\right)|s_t=s\right] \tag{14}$$

其中，$\gamma\in[0,1]$为折扣因子，r_t为智能体在t时刻所获即时奖励。

利用深度神经网络$Q(s,d,p_d;\varphi)$逼近$Q(s,d,p_d)$，并利用确定性策略网络$p_d(s;\theta):S\to P_d$逼近$p_d^Q(s)$。换言之，在固定网络权重φ时，欲寻得θ使得下式成立

$$Q(s_t,d_t,p_d(s;\theta);\varphi)\approx\sup_{p_d\in P_d}Q(s,d,p_d;\varphi) \tag{15}$$

结合n-step算法，对于固定的$n\geq1$，将n-step目标值y_t定义为

$$\begin{aligned}y_t=&\gamma^n\max_{d\in D}Q\left(s_{t+n},d,p_d(s_{t+n};\theta_t);\varphi_t\right)\\&+\sum_{i=0}^{n-1}\left(\gamma^i r_{t+i}\right)\end{aligned} \tag{16}$$

与DQN类似，对φ使用最小二乘损失函数。此外，为了在φ固定时找到使$Q(s,d,p_d(s;\theta);\varphi)$最大化的$\theta$，设置$\varphi$和$\theta$的损失函数为

$$\begin{aligned}\ell_t^Q(\varphi)&=\frac{1}{2}[Q\left(s_t,d_t,p_{d_t};\varphi\right)-y_t]^2\\\ell_t^\Theta(\theta)&=-\sum_{d\in D}Q\left(s_t,d,p_d(s_t;\theta);\varphi_t\right)\end{aligned} \tag{17}$$

在每一轮训练之后，φ和θ可通过下式更新

$$\begin{aligned}\varphi_{t+1}&\leftarrow\varphi_t-\alpha_t\nabla_\varphi\ell_t^Q(\varphi_t)\\\theta_{t+1}&\leftarrow\theta_t-\beta_t\nabla_\theta\ell_t^\Theta(\theta_t)\end{aligned} \tag{18}$$

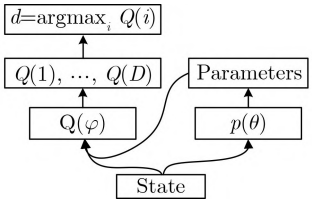

图 2 P-DQN算法流程

其中，α和β分别为更新φ和θ时的学习率。

算法1为基于P-DQN的联合优化算法流程。首先将状态s_t输入至网络$p_d(s_t;\theta_t)$中以生成连续动作p_d，接着将p_d连同状态s_t输入至网络$Q(s_t,d,p_d;\varphi_t)$并选择出Q值最大的离散动作d_t。为避免模型陷入局部最优，在获得最优d_t之后使用ε-贪心策略来增加动作探索概率。在执行混合动作(d_t,p_{d_t})之后评估时延和缓存命中率，并将状态更新为s_{t+1}，根据式(13)计算即时奖励r_t。接着，将四元组(s_t,a_t,r_t,s_{t+1})存储在经验回放池Γ中，并从经验回放池Γ中采样得到的mini-batch集。最后利用式(16)计算得到y_t，据式(18)更新网络参数φ和θ。

4 仿真结果与分析

4.1 实验设置

本文利用Python3.7.0和Pytorch1.9.0搭建仿真平台，并在平台上进行模拟实验以验证所提算法的可行性和有效性，系统主要的仿真参数由表1给出。

4.2 实验结果

图3展示了基于P-DQN和基于DDPG的内容缓存和功率分配方案收敛过程。从图中可知，虽然基于DDPG的方案较基于P-DQN的方案更快收敛，但前者的收敛所得奖励却不如后者，这是由于基于DDPG的方案对离散动作进行松弛处理导致问题复杂度增加，从而陷入局部最优。

图4展示了所提方案在不同学习率和mini-batch大小为64的条件下的性能。在学习率为0.0001的条件下，由于学习率过小，系统经历了一个缓慢的学

算法1 基于P-DQN的联合优化算法

初始化：设置最大训练轮数T、学习率$\{\alpha,\beta\}$、探索参数ε、概率分布参数ξ、mini-batch大小为B、经验回放池Γ、网络权重φ_1和θ_1

for $t=1$ **to** T **do**
  for $k=1$ **to** K **do**
    计算动作参数$p_d \leftarrow p_d(s_t;\theta_t)$。
    使用ε-greedy策略选择动作$a_t=(d_t,p_{d_t})$，其中
    $d_t=\arg\max_{d\in D}Q(s_t,d,p_d;\varphi_t)$
    $a_t=\begin{cases}\text{以概率}\xi\text{采样}, & \varepsilon\\ (d_t,p_{d_t}), & 1-\varepsilon\end{cases}$
    执行a_t，并获取时延和命中率，观测奖励r_t和下一状态s_{t+1}
    将$[s_t,a_t,r_t,s_{t+1}]$存入Γ
    从Γ中采集B个$\{s_b,a_b,r_b,s_{b+1}\}_{b\in[B]}$样本
    $y_b=r_b+\max_{d\in D}\varphi Q(s_{b+1},d,p_d(s_{b+1};\theta_t);\varphi_t)$
    使用$\{y_b,s_b,a_b\}_{b\in[B]}$计算$\nabla_\varphi \ell_t^Q(\varphi)$和$\nabla_\theta \ell_t^\Theta(\theta)$
    计算$\varphi_{t+1}\leftarrow\varphi_t-\alpha_t\nabla_\varphi\ell_t^Q(\varphi)$和$\theta_{t+1}\leftarrow\theta_t-\beta_t\nabla_\theta\ell_t^\Theta(\theta)$
  end for
end for

表 1 仿真参数

参数	数值
RSU覆盖半径 (m)	250
RSU数量	4
RSU存储容量 (GB)	16
RSU总功率 (dBm)	40
内容大小 (MB)	[8,12]
车辆数量	100
带宽 (MHz)	10
噪声功率 (dBm)	–60
SINR门限 (dB)	20
路径损耗模型	128.1+37.61lg(d)
pP和Q网络的隐藏层	128×64
mini-batch大小	64
经验回放池容量	5000
学习率$\alpha=\beta\alpha=\beta$	0.001
折扣因子$\gamma\gamma$	0.95

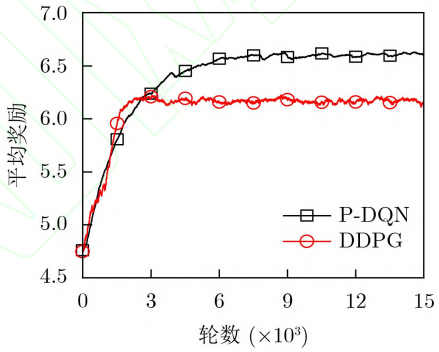

图 3 基于P-DQN和基于DDPG方案的收敛过程

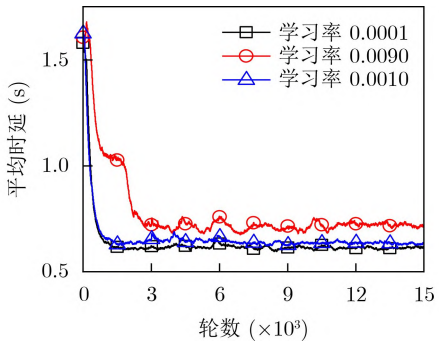

图 4 不同学习率下的平均时延

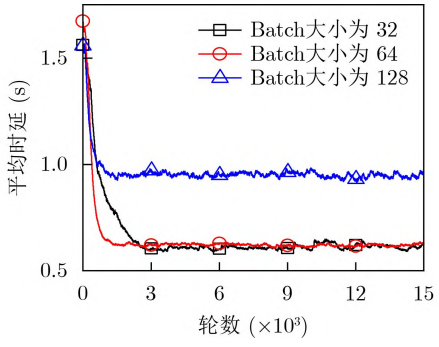

图 5 不同mini-batch大小下的系统性能

习过程，而当学习率为0.09时，系统的时延增加了，这是由于过大的学习率使得算法陷入局部最优。因此本文在后续实验中采用0.001的学习率。

图5展示了所提方案在不同mini-batch大小和学习率为0.001的条件下的性能。从图中可知，过小的mini-batch使得梯度表现出非常粗略的近似，系统需要很长时间才能找到最优策略。而过大的mini-batch使得梯度计算更准确，但学习过程可能会陷入局部最优。与大小为32和128相比，当mini-batch大小为64时，系统实现了更好的性能，具有更早的收敛和更低的延迟，因此本文在后续实验中采用大小为64的mini-batch。

图6展示了本文提出的基于P-DQN的方案在不同时隙处，系统将请求内容从内容源发送给车辆时不同途径的占比变化。随着时间推移，由RSU直接将内容交付给车辆的事件比例逐渐增大并保持稳定，其他交付事件则逐渐降低至稳定，这是由于RSU动态捕捉流行内容并将其缓存至本地使得命中率增加。

图7展示了基于P-DQN的方案、基于DDPG的方案和基于最近最少使用(Least Recently Used, LRU)的方案在缓存命中率方面的性能对比。在RSU缓存容量从8 GB增加至32 GB的过程中，所有方案的缓存命中率都提高，这是由于更大的缓存容量意味着系统可以缓存数量更多的流行内容。较基于DDPG和基于LFU的方案而言，所提方案可更好感知内容流行度的动态变化，使得平均缓存命中率分别提高了5%和15%。

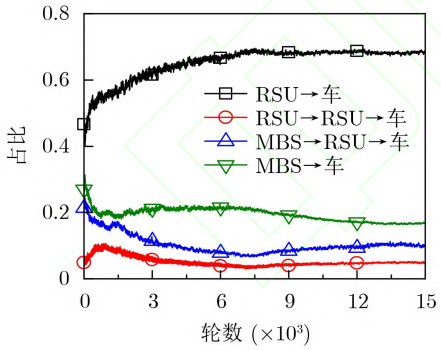

图 6 提出方案中内容交付途径的比例变化

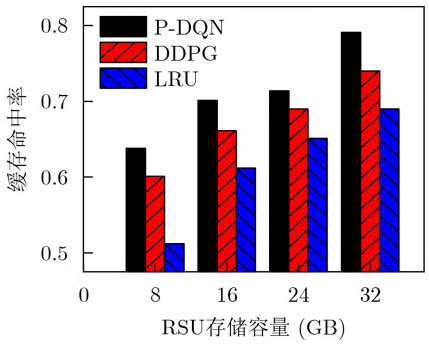

图 7 存储容量对缓存命中率的影响

图8展示了RSU最大可分配功率对平均系统时延的影响。本文对比了基于P-DQN的方案、基于DDPG的方案和随机方案在平均系统时延方面的性能。在RSU最大可分配功率从35 dBm增加至50 dBm的过程中，所有方案的平均系统时延都有所降低，这是由于最大可分配功率的增加会提高系统的平均传输速率。此外，由于所提方案无需对混合动作进行离散化或松弛化，使得其在动态和复杂的无线环境下能更加精准地分配适当大小的功率，因此所提方案较基于DDPG方案和随机方案而言，平均系统时延分别降低了8%和21%。

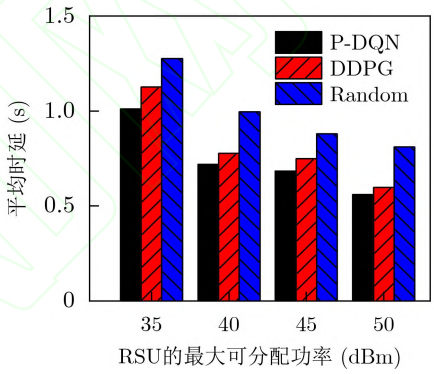

图 8 最大可分配功率对时延的影响

5 结束语

本文研究了车联网场景中内容缓存和传输功率分配的联合优化问题，旨在最小化系统整体时延。为实现该目标，本文将该联合优化问题建模为MDP，并将缓存内容和选择内容提供者的动作映射为离散动作集，将功率分配动作映射为与离散动作相对应的连续参数。同时，为避免将离散和连续变量处理为同一类型变量会带来的维度或复杂性的增加的问题，利用P-DQN算法实现混合动作空间的参数化。实验结果表明，相较于基准方案，本文提出方案实现了更低的系统时延和更高的缓存命中率。

参 考 文 献

[1] YOUSEFPOUR A, ISHIGAKI G, GOUR R, *et al*. On reducing IoT service delay via fog offloading[J]. *IEEE Internet of Things Journal*, 2018, 5(2): 998–1010. doi: 10.1109/JIOT.2017.2788802.

[2] HE Ying, ZHAO Nan, and YIN Hongxi. Integrated networking, caching, and computing for connected vehicles: A deep reinforcement learning approach[J]. *IEEE Transactions on Vehicular Technology*, 2018, 67(1): 44–55.

doi: 10.1109/TVT.2017.2760281.

[3] TANG Fengxiao, MAO Bomin, KATO N, *et al*. Comprehensive survey on machine learning in vehicular network: Technology, applications and challenges[J]. *IEEE Communications Surveys & Tutorials*, 2021, 23(3): 2027–2057. doi: 10.1109/COMST.2021.3089688.

[4] XU Lianming, YANG Zexuan, WU Huaqing, *et al*. Socially driven joint optimization of communication, caching, and computing resources in vehicular networks[J]. *IEEE Transactions on Wireless Communications*, 2022, 21(1): 461–476. doi: 10.1109/TWC.2021.3096881.

[5] KAZMI S M A, DANG T N, YAQOOB I, *et al*. Infotainment enabled smart cars: A joint communication, caching, and computation approach[J]. *IEEE Transactions on Vehicular Technology*, 2019, 68(9): 8408–8420. doi: 10.1109/TVT.2019.2930601.

[6] ZHANG Shan, LI Junjie, LUO Hongbin, *et al*. Towards fresh and low-latency content delivery in vehicular networks: An edge caching aspect[C]. 2018 10th International Conference on Wireless Communications and Signal Processing, Hangzhou, China, 2018: 1–6. doi: 10.1109/WCSP.2018.8555643.

[7] CHEN Jiayin, WU Huaqing, YANG Peng, *et al*. Cooperative edge caching with location-based and popular contents for vehicular networks[J]. *IEEE Transactions on Vehicular Technology*, 2020, 69(9): 10291–10305. doi: 10.1109/TVT.2020.3004720.

[8] SUN Yaohua, PENG Mugen, and MAO Shiwen. A game-theoretic approach to cache and radio resource management in fog radio access networks[J]. *IEEE Transactions on Vehicular Technology*, 2019, 68(10): 10145–10159. doi: 10.1109/TVT.2019.2935098.

[9] YU Zhengxin, HU Jia, MIN Geyong, *et al*. Proactive content caching for internet-of-vehicles based on peer-to-peer federated learning[C]. 2020 IEEE 26th International Conference on Parallel and Distributed Systems, Hong Kong, China, 2020: 601–608. doi: 10.1109/ICPADS51040.2020.00083.

[10] XING Yuping, SUN Yanhua, QIAO Lan, *et al*. Deep reinforcement learning for cooperative edge caching in vehicular networks[C]. 2021 13th International Conference on Communication Software and Networks, Chongqing, China, 2021: 144–149. doi: 10.1109/ICCSN52437.2021.9463666.

[11] QIAO Guanhua, LENG Supeng, MAHARJAN S, *et al*. Deep reinforcement learning for cooperative content caching in vehicular edge computing and networks[J]. *IEEE Internet of Things Journal*, 2020, 7(1): 247–257. doi: 10.1109/JIOT.2019.2945640.

[12] DAI Yueyue, XU Du, LU Yunlong, *et al*. Deep reinforcement learning for edge caching and content delivery in internet of vehicles[C]. 2019 IEEE/CIC International Conference on Communications in China, Changchun, China, 2019: 134–139. doi: 10.1109/ICCChina.2019.8855951.

[13] CHEN Shuangwu, YAO Zhen, JIANG Xiaofeng, *et al*. Multi-agent deep reinforcement learning-based cooperative edge caching for ultra-dense next-generation networks[J]. *IEEE Transactions on Communications*, 2021, 69(4): 2441–2456. doi: 10.1109/TCOMM.2020.3044298.

[14] CHETLUR V V and DHILLON H S. Coverage and rate analysis of downlink cellular vehicle-to-everything (C-V2X) communication[J]. *IEEE Transactions on Wireless Communications*, 2020, 19(3): 1738–1753. doi: 10.1109/TWC.2019.2957222.

[15] ASHERALIEVA A and NIYATO D. Game theory and lyapunov optimization for cloud-based content delivery networks with device-to-device and UAV-enabled caching[J]. *IEEE Transactions on Vehicular Technology*, 2019, 68(10): 10094–10110. doi: 10.1109/TVT.2019.2934027.

[16] HAO Linchun, REN Pinyi, and DU Qinghe. Satellite QoS routing algorithm based on energy aware and load balancing[C]. 2020 International Conference on Wireless Communications and Signal Processing, Nanjing, China, 2020: 685–690. doi: 10.1109/WCSP49889.2020.9299827.

雒江涛：男，教授，博士生导师，研究方向为新一代网络技术、通信网络测试与优化、移动大数据等.

杨和平：男，硕士生，研究方向为车联网.

冉泳屹：男，讲师，硕士生导师，研究方向为绿色数据中心、卫星互联网等.

责任编辑：马秀强