Deep-Reinforcement Learning Multiple Access for Heterogeneous Wireless Networks

Yiding Yu, Student Member, IEEE,Taotao Wang, Member, IEEE,
and Soung Chang Liew, Fellow, IEEE

摘要

研究了一种基于深度强化学习的 MAC 协议，用于异构无线网络，称为深度强化学习多址接入 (DLMA)。考虑了多个网络运行不同的 MAC 协议，试图访问共同无线介质的时间槽的场景。假设 DLMA 网络不知道其他网络的 MAC 的操作原理——即 DLMA 不知道其他 MAC 如何决定何时传输和何时不传输。DLMA 的目标是能够学习一个最优的信道接入策略，以实现某个预设的全局目标。可能的目标包括最大化总吞吐量和最大化所有网络之间的阿尔法公平性。DLMA 的底层学习过程基于 DRL，通过在 DRL 中适当定义状态空间、动作空间和奖励，我们展示了 DLMA 可以通过精心选择某些时间槽来传输，从而轻松最大化总吞吐量。广泛的模拟结果表明，DLMA 可以在与 TDMA 和 ALOHA MAC 协议共存时，即使不知道它们是 TDMA 或 ALOHA，也能最大化总吞吐量或实现比例公平性。

关键词：异构网络；深度强化学习；多址接入

1 引言

随着智能设备（如手机、智能汽车和智能家居设备）和网络技术（如云计算和网络虚拟化）的快速发展，网络变的越来越异构和复杂，在异构无线网络中存在着多种网络，它们运行着不同的 MAC 协议，并且需要一个公共的无限介质。本文的研究背景是无线通信领域中的多址接入控制（MAC）协议设计。他是通讯系统给多个用户动态分配资源，使他们能同时进行数据传送，也就是确保多个设备在共享无线频谱时能够公平地获取带宽，避免冲突和碰撞，提高网络的效率和性能，设计一个基于深度强化学习（DRL）的 MAC 协议，称为深度强化学习多址接入（DLMA）。目的是通过一系列的观察和动作来学习一种最优的方案，并且不需要知道其他共存网络的 MAC 协议的运行机制。该设计利用了深度 Q 网络（DQN）算法 [1]，这是一种结合了深度神经网络和 Q 学习的 DRL 算法。DRL 代理从一系列状态-动作-奖励观察中收集的经验中学习，调整神经网络的权重。

2 相关工作

RL 是一种机器学习范式，其中代理学习成功的策略，这些策略从与环境的试错交互中产生最大的长期回报。最具代表性的 RL 算法是 Q-learning 算法。Q 学习可以通过更新状态-动

作值函数来学习好的策略，而无需环境的操作模型。当状态-动作空间较大且复杂时，可以使用深度神经网络来逼近 Q 函数，相应的算法称为 DRL。这项工作采用 DRL 来加快收敛速度并增加 DLMA 的鲁棒性。

2.1 强化学习概述

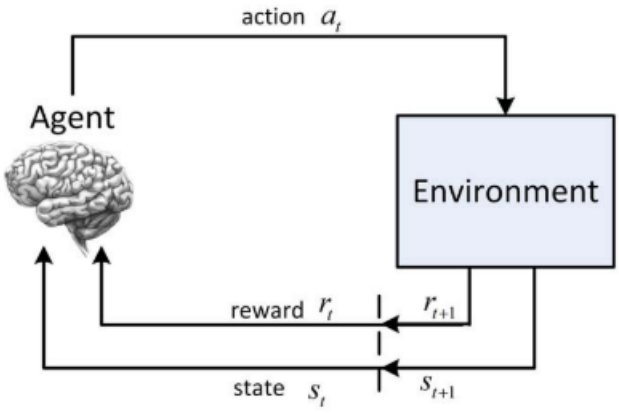


图 1. Agent 与环境交互过程

在强化学习（RL）中，一个智能体（agent）在一系列离散的时间点 $t = 0, 1, 2, \ldots$ 与环境进行交互，以完成一个任务，如图 1所示。在时间 t，智能体观察环境的状态 $s_t \in S$，其中 S 是可能状态的集合。然后它采取一个动作 $a_t \in A_{s_t}$，其中 A_{s_t} 是在状态 s_t 下可能的动作集合。由于状态-动作对 (s_t, a_t) 的结果，智能体收到一个奖励 r_{t+1}，并且环境在时间 $t+1$ 转移到新的状态 s_{t+1}。智能体的目标是通过其动作产生一系列奖励 $\{r_t\}_{t=1,2,\ldots}$ 来最大化某些性能标准。例如，在时间 t 要最大化的性能标准可能是 $R_t \triangleq \sum_{\tau=t}^{\infty} \gamma^{\tau-t} r_{\tau+1}$，其中 $\gamma \in (0, 1]$ 是用于加权未来奖励的折扣因子。通常，智能体根据某个决策策略 π 采取动作。强化学习方法指定了智能体如何根据其经验改变其策略。通过足够的经验，智能体可以学习到一个最优的决策策略 π^* 来最大化长期累积的奖励。

Q-learning 是最受欢迎的强化学习（RL）方法之一。Q-learning 强化学习智能体学习一个动作价值函数 $Q^{\pi}(s, a)$ ，它对应于在决策策略 π 下，当在环境状态 s 采取动作 a 时预期累积的奖励：

$$Q^{\pi}(s, a) \triangleq \mathbb{E}[R_t | s_t = s, a_t = a, \pi]$$

最优动作价值函数 $Q^*(s, a)$，定义为所有策略 π 下 $Q^{\pi}(s, a)$ 的最大值，遵循贝尔曼最优方程：

$$Q^*(s, a) = \mathbb{E}_{s'} \left[r_{t+1} + \gamma \max_{a'} Q^*(s', a') | s_t = s, a_t = a \right]$$

其中 s' 是状态-动作对 (s, a) 之后的新状态。Q-learning 的核心思想是，我们可以在每个状态-动作对 (s, a) 出现时迭代估计 $Q^*(s, a)$。设 $q(s, a)$ 为迭代过程中估计的状态-动作价值函数。在状态-动作对 (s_t, a_t) 和得到的奖励 r_{t+1} 之后，Q-learning 更新 $q(s_t, a_t)$ 如下：

$$q(s_t, a_t) \leftarrow q(s_t, a_t) + \alpha \left[r_{t+1} + \gamma \max_{a'} q(s_{t+1}, a') - q(s_t, a_t) \right]$$

其中 $\alpha \in (0,1)$ 是学习率。在系统更新 $q(s,a)$ 的同时，它也基于 $q(s,a)$ 做出决策。通常采用 ϵ-贪婪策略。对于 ϵ-贪婪策略，代理以概率 $1-\epsilon$ 选择贪婪动作 $a_t = \arg\max_a q(s_t,a)$，并以概率 ϵ 随机选择一个动作。随机选择动作的一个原因是为了避免陷入尚未收敛到 $Q^*(s,a)$ 的 $q(s,a)$ 函数。

2.2 异构网络结构

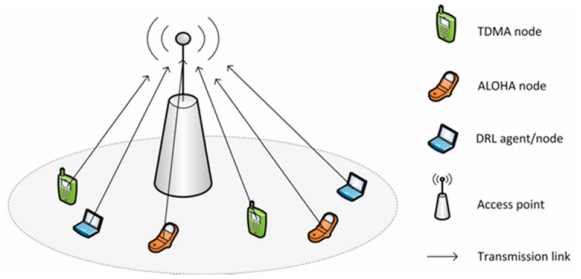


图 2. 异构无限网络结构

该系统（如图 2）中所有的节点总是有数据包要传输，一个节点只能在一个时隙的开始处开始传输，并且必须在该时隙内完成传输。节点会使用不同的 MAC 协议：TDMA、ALOHA，并且至少有一个节点使用 DLMA 协议。DRL 节点是一个 Agent 可以学习最优的 MAC 策略，从而与 TDMA 和 ALOHA 节点很好地协同使用网络资源。在该网络中，节点总是要发送数据包的。

TDMA：TDMA 节点以帧到帧的重复方式在 Y 个时隙的每个帧内的 X 个特定时隙中进行发送。

q-ALOHA：一个 q-ALOHA 节点在每个时隙以固定的概率 q 发送，且各个时隙之间相互独立同分布。

Fixed-window ALOHA：固定窗口 ALOHA（FW-ALOHA）节点在其在时隙中发送之后生成在 [0，W-1] 范围内的随机计数器值 w。然后，它在下一次传输之前等待 w 个时隙。参数 W 被称为窗口大小。

Exponential backoff ALOHA:：指数退避 ALOHA（EB-ALOHA）是基于窗口的 ALOHA 的变体，其中窗口大小不固定。具体地，EB-ALOHA 节点每次在其传输遇到冲突时使其窗口大小加倍，直到达到最大窗口大小 $2^m W$，其中 m 是“最大退避阶段”。在成功传输时，窗口大小恢复到初始窗口大小 W。

DRL agent/DRL node：DRL 代理或 DRL 节点是采用我们的 DLMA 协议的无线电节点。对于 DRL 节点，如果它发送，它将立即从 AP 获得 ACK，指示传输是否成功；如果它不发送，它将侦听信道并从环境中获得观察，指示其他节点的传输结果或信道空闲。

3 本文方法

3.1 使用深度强化学习的 DLMA

在时间 t 时，DRL 代理采取的动作是 $a_t \in \{\text{TRANSMIT,WAIT}\}$，TRANSMIT 表示代理进行传输，WAIT 表示代理不进行传输。我们用 $z_t \in \{\text{SUCCESSFUL,COLLIDED,IDLE}\}$ 表示采取动作 a_t 后的信道观察结果，其中 SUCCESSFUL 表示只有一个站点在信道上传输；COLLIDED 表示多个站点同时传输导致冲突；IDLE 表示没有站点进行传输。DRL 代理通过来自接入点的 ACK 信号和监听信道来确定 z_t。

将时间 $t+1$ 的信道状态定义为动作-观察对 $c_{t+1} \triangleq (a_t, z_t)$。$c_{t+1}$ 有五种可能的情况：$\{\text{TRANSMIT, SUCCESSFUL}\}, \{\text{TRANSMIT, COLLIDED}\}, \{\text{WAIT, SUCCESSFUL}\}, \{\text{WAIT, COLLIDED}\}$ 和 $\{\text{WAIT, IDLE}\}$。将时间 $t+1$ 的环境状态定义为 $s_{t+1} \triangleq [c_{t-M+2}, \ldots, c_t, c_{t+1}]$，其中参数 M 是代理需要跟踪的状态历史长度。

在采取动作 a_t 后，从状态 s_t 到 s_{t+1} 的转换会产生一个奖励 r_{t+1}，如果 $z_t = \text{SUCCESSFUL}$，则 $r_{t+1} = 1$；如果 $z_t = \text{COLLIDED}$ 或 IDLE，则 $r_{t+1} = 0$。这里的奖励定义对应于最大化总吞吐量的目标。

在深度强化学习中，使用深度神经网络来近似动作价值函数，$q(s,a;\theta) \approx Q^*(s,a)$，其中 $q(s,a;\theta)$ 是神经网络给出的近似值，而 θ 是包含神经网络中边权重的参数向量。神经网络的输入是一个状态 s，输出是不同动作的近似 q 值 $Q = \{q(s,a;\theta) \mid a \in A_s\}$。我们将整个神经网络称为 Q 神经网络（QNN），相应的强化学习算法称为 DRL。通过训练过程调整 QNN 中的 θ 来更新 $q(s,a;\theta)$。

特别地，QNN 通过最小化 $q(s,a;\theta)$ 的预测误差来训练。假设在时间 t，状态是 s_t 而 QNN 的权重是 θ。DRL 代理采取行动 $a_t = \arg\max_a q(s_t,a;\theta)$，其中对于不同的行动 a，$q(s_t,a;\theta)$ 由 QNN 的输出给出。假设结果的奖励是 r_{t+1} 并且状态转移到 s_{t+1}。然后，$(s_t, a_t, r_{t+1}, s_{t+1})$ 构成了一个“经验样本”，这将被用来训练 QNN。为了训练，我们定义 QNN 对于特定经验样本 $(s_t, a_t, r_{t+1}, s_{t+1})$ 的预测误差为

$$L_{s_t,a_t,r_{t+1},s_{t+1}}(\theta) = \left(y^{QNN}_{r_{t+1},s_{t+1}} - q(s_t,a_t;\theta)\right)^2,$$

其中 $q(s_t,a_t;\theta)$ 是 QNN 给出的近似值，$y^{QNN}_{r_{t+1},s_{t+1}} = r_{t+1} + \gamma \max_{a'} q(s_{t+1},a';\theta)$。注意 $y^{QNN}_{r_{t+1},s_{t+1}}$ 是一个基于当前奖励 r_{t+1} 加上 QNN 给出的预测折扣奖励 $\gamma \max_{a'} q(s_{t+1},a';\theta)$ 的精细目标输出。通过应用半梯度算法来训练 QNN，即更新 θ。θ 的迭代过程由下式给出：

$$\theta \leftarrow \theta + \rho \left[y^{QNN}_{r_{t+1},s_{t+1}} - q(s_t,a_t;\theta)\right] \nabla q(s_t,a_t;\theta),$$

其中 ρ 是每次调整中的步长。

3.2 ResNet 网络

模型的 Q 神经网络使用 Res Net，如图 3所示，拥有一个状态输入层，4 个容度为 64 的隐藏层，以及一个动作 Q 值的输出层。激活函数使用斜坡函数。隐藏层的前两层使用全连接层，中间两层是两个 Res Net 模块，由全连接层和一个加法器组成。输入是状态 s，输出是针对不同动作近似的 Q 值。同时采用残差网络 ResNet，相同的静态 ResNet 架构可以用于不同的无线网络场景的深度强化学习；而对于普通的 DNN，其最优神经网络深度因情况而异。

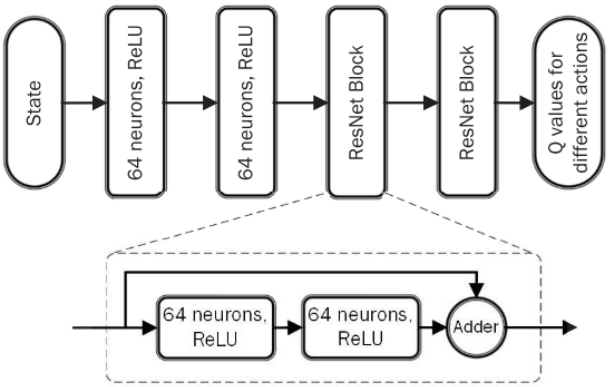


图 3. ResNet 网络结构

4 复现细节

4.1 与已有开源代码对比

根据伪代码如图 4进行复现，没有参考任何相关源代码。

Algorithm 1 DLMA With the Sum Throughput Objective

Initialize s_0, ε, γ, ρ, N_E, F
Initialize experience memory EM
Initialize the parameter of QNN as $\boldsymbol{\theta}$
Initialize the parameter of target QNN $\boldsymbol{\theta}^- = \boldsymbol{\theta}$
for $t = 0, 1, 2, \cdots$ in DLMA **do**
  Input s_t to QNN and output $\mathbb{Q} = \{q(s_t, a, \boldsymbol{\theta}) | a \in A_{s_t}\}$
  Generate action a_t from $\mathbb{Q}$ using ε-greedy algorithm
  Observe z_t, r_{t+1}
  Compute s_{t+1} from s_t, a_t and z_t
  Store $(s_t, a_t, r_{t+1}, s_{t+1})$ to EM
  if Remainder$(t/F == 0)$ **then** $I = 1$ **else** $I = 0$
  TRAINQNNγ, ρ, N_E, I, EM, $\boldsymbol{\theta}$, $\boldsymbol{\theta}^-$
end for
procedure TRAINQNN(γ, ρ, N_E, I, EM, $\boldsymbol{\theta}$, $\boldsymbol{\theta}^-$)
  Randomly sample N_E experience tuples from EM as E
  for each sample $e = (s, a, r, s')$ in E **do**
    Calculate $y_{r,s'}^{QNN} = r + \gamma \max_{a'} q(s', a'; \boldsymbol{\theta}^-)$
  end for
  Perform Gradient Descent to update $\boldsymbol{\theta}$ in QNN:

$$\text{Iterate } \boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \frac{\rho}{N_E} \sum_{e \in E} \left[y_{r,s'}^{QNN} - q(s, a; \boldsymbol{\theta}) \right] \nabla q(s, a; \boldsymbol{\theta})$$

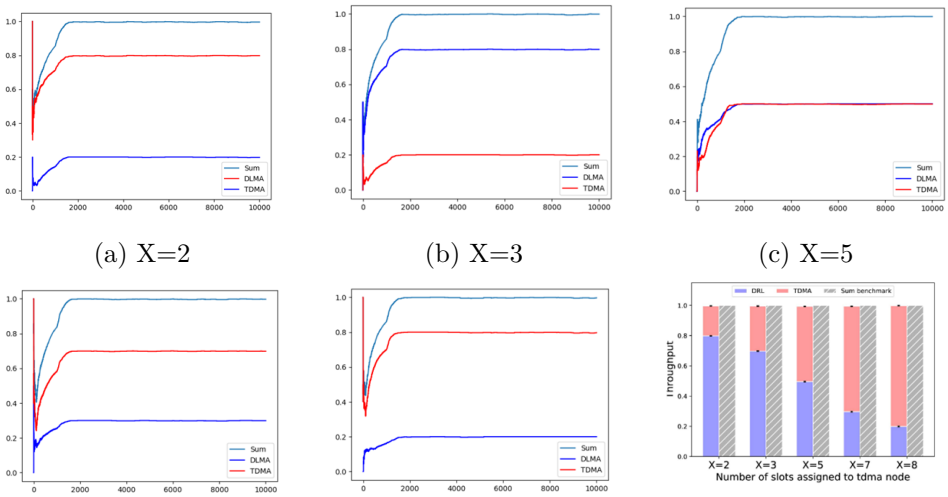
  if $I == 1$ **then**
    Update $\boldsymbol{\theta}^-$ in target QNN by setting $\boldsymbol{\theta}^- = \boldsymbol{\theta}$
  end if
end procedure

图 4. DLMA 算法

5 实验结果分析

首先展示一个 DRL 节点与一个 TDMA 节点共存的结果（如图 5）。TDMA 节点在每个包含 Y 个时隙的帧中，以重复的方式在 X 个特定的时隙内进行传输。为了进行基准测试，考

虑了一个 TDMA 感知节点，该节点完全了解 TDMA 节点使用的 X 个时隙。为了最大化整个系统的吞吐量，TDMA 感知节点将在 TDMA 节点未使用的所有 Y-X 个时隙中进行传输。最优的总吞吐量是每个时隙一个数据包。与 TDMA 感知节点不同，DRL 节点不知道另一个节点是 TDMA 节点，只是使用 DRL 算法来学习最优策略。图 4(a-e) 展示了当 TDMA 帧 Y=10，分配给 TDMA 节点的时隙数 X 为 2、3、5、7、8 时的吞吐量结果。从图 4(f) 中我们可以看到，不同 X 的总吞吐量都可以近似。这表明 DRL 节点能够在不知道 TDMA 节点采用的 TDMA 协议的情况下，捕捉到所有未使用的时隙。

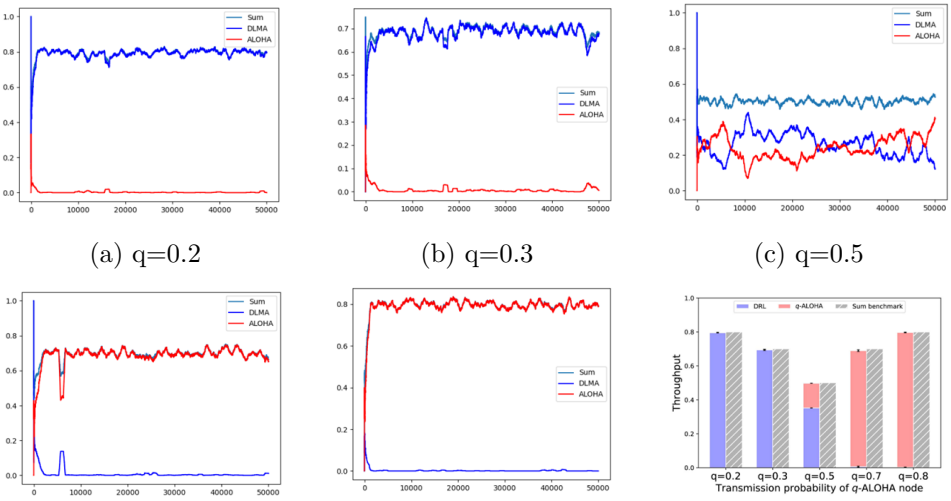


(a) X=2 (b) X=3 (c) X=5

(d) X=7 (e) X=8 (f) 平均值

图 5. TMDA 和 DLMA 共存

接下来，我们展示了一个 DRL 节点与一个 q-ALOHA 节点（如图 6）、一个 DRL 节点和 FW-ALOHA 节点（如图 7）和一个 DRL 节点和 EB-ALOHA 节点（如图 8）共存的结果。我们强调，即使其他协议不再是 TDMA，这里使用的仍然是与 A 部分完全相同的 DLMA 算法。为了进行基准测试，我们考虑了模型感知节点，这些节点运行着为三种 ALOHA 变体的操作机制量身定制的最优 MAC。

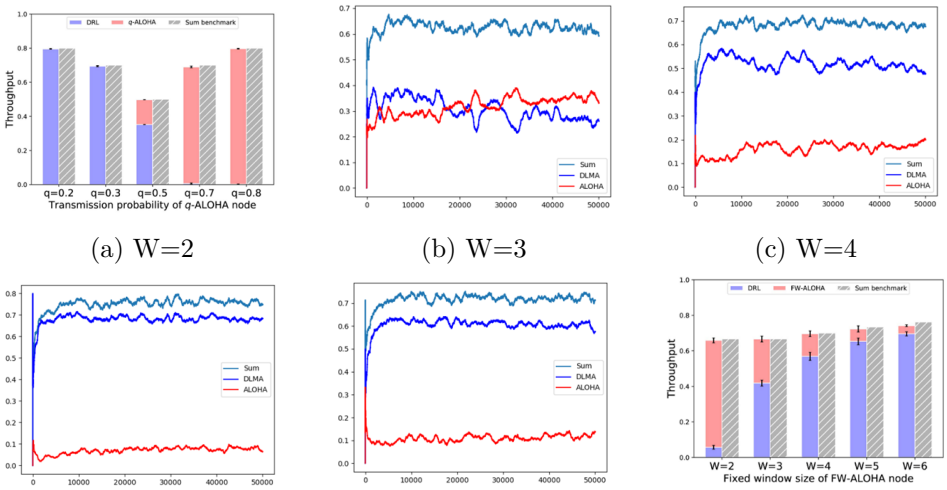


(a) q=0.2 (b) q=0.3 (c) q=0.5

(d) q=0.7 (e) q=0.8 (f) 平均值

图 6. DLMA 与 q-ALOHA 共存

DLMA 和 FW-ALOHA 共存（W 表示固定窗口的大小）

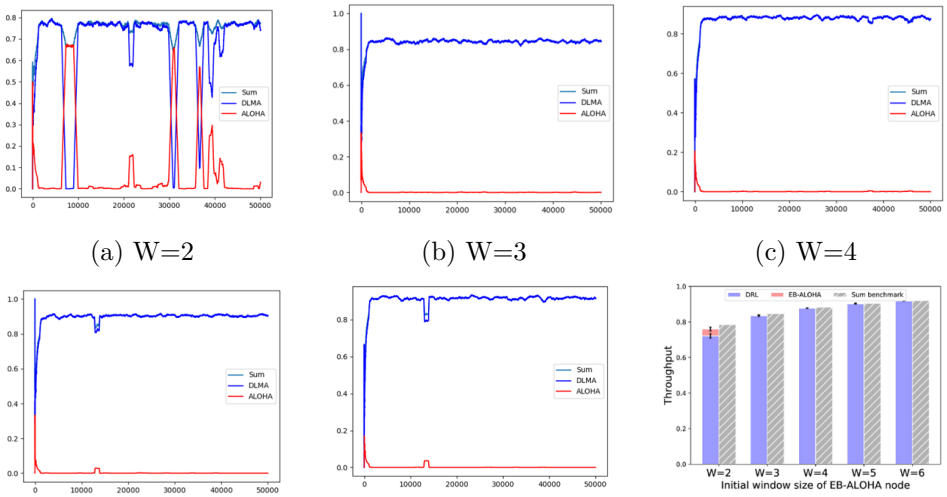


(a) W=2 (b) W=3 (c) W=4

(d) W=5 (e) W=6 (f) 平均值

图 7. DLMA 与 FW-ALOHA 共存

DLMA 和 EB-ALOHA 共存（W 表示窗口大小，最大回退值为 2）



(a) W=2 (b) W=3 (c) W=4

(d) W=5 (e) W=6 (f) 平均值

图 8. DLMA 与 EB-ALOHA 共存

6 总结与展望

按照文中伪代码的思路实现模型，并验证其性能和文中结论相符。本文提出了一种基于深度强化学习（DRL）的通用 MAC 协议，用于异构无线网络，称为 DLMA。DLMA 的一个显著特点是，它能够通过一系列状态-动作-奖励观察，在异构环境中学习实现整体目标。特别地，它能够在不了解共存 MAC 的详细操作机制的情况下，实现接近最优的性能。

本文还展示了在无线网络的强化学习中使用神经网络的优势。具体来说，与传统的强化学习（RL）相比，DRL 能够更快地获得接近最优的策略和性能，并且具有更强的鲁棒性，这

是在动态变化的无线环境中实际部署 MAC 协议的两个基本属性。本文专注于最大化异构环境中所有网络的总吞吐量。

参考文献

[1] Yiding Yu, Taotao Wang, and Soung Chang Liew. Deep-reinforcement learning multiple access for heterogeneous wireless networks. *IEEE Journal on Selected Areas in Communications*, 37(6):1277–1290, 2019.