基于深度强化学习的无人机数据采集和路径规划研究

牟治宇[1,2]，张煜[1,2]，范典[3]，刘君[2,4]，高飞飞[1,2]

（1. 清华大学自动化系，北京 100084；2. 清华大学北京信息科学与技术国家研究中心，北京 100084；
3. 中国信息通信研究院泰尔终端实验室，北京 100191；4. 清华大学网络科学与网络空间研究院，北京 100084）

摘 要：物联网时代需要实现海量的节点覆盖和连接，对于一些偏远地区，物联网通信技术存在无法及时采集数据的问题。而无人机具有灵活性和机动性等特点，因此，可用于物联网中的无线传感器网络的数据采集。所提方案着重对无人机辅助传感器网络数据采集时的路径规划问题进行了研究，同时满足无人机自身因电池容量有限而产生的充电需求。具体地，利用时间抽象分层强化学习思想，基于离散动作深度强化学习架构，提出了一种新颖的 option-DQN（option-deep Q-learning）算法，实现了高效的无人机数据采集和路径规划，同时控制无人机及时进行充电，保证其正常飞行。仿真结果表明，相比于传统 DQN（deep Q-learning）算法，所提算法在训练时的周期奖励上升速度更快，最终达到的周期奖励水平更高，并且无人机在执行任务时的轨迹更清晰、合理，所提算法可以判断无人机何时应进行充电，从而保证无人机的电量始终充足。

关键词：无人机；路径规划；数据采集；充电

中图分类号：TP92

文献标识码：A

doi: 10.11959/j.issn.2096−3750.2020.00177

Research on the UAV-aided data collection and trajectory design based on the deep reinforcement learning

MOU Zhiyu[1,2], ZHANG Yu[1,2], FAN Dian[3], LIU Jun[2,4], GAO Feifei[1,2]

1. Department of Automation, Tsinghua University, Beijing 100084, China
2. Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing 100084, China
3. Department of Strategic Planning & Research of CTTL-Terminals, China Academy of Information Communications Technology, Beijing 100191, China
4. Institute for Network Sciences and Cyberspace, Tsinghua University, Beijing 100084, China

Abstract: The Internet of things (IoT) era needs to realize the wide coverage and connections for the IoT nodes. However, the IoT communication technology cannot collect data timely in the remote area. UAV has been widely used in the IoT wireless sensor network for the data collection due to its flexibility and mobility. The trajectory design of the UAV assisted sensor network data acquisition was discussed in the proposed scheme, as well as the UAV charging demand in the data collection process was met. Specifically, based on the hierarchical reinforcement learning with the temporal abstraction, a novel option-DQN (option-deep Q-learning) algorithm targeted for the discrete action was proposed to improve the performance of the data collection and trajectory design, and control the UAV to recharge in time to ensure its normal flight. The simulation results show that the training rewards and speed of the proposed method are much better than the conventional DQN (deep Q-learning) algorithm. Besides, the proposed algorithm can guarantee the sufficient power supply of UAV by controlling it to recharge timely.

Key words: UAV, trajectory design, data collection, charging

收稿日期：2020−05−18；**修回日期：**2020−07−02

通信作者：刘君，juneliu@tsinghua.edu.cn

基金项目：国家重点研发计划（No.2018AAA0102401）；中国信息通信研究院 2020 年青年课题；清华大学自主科研项目（No.2019Z08QCX19）；国家自然科学基金资助项目（No.61902214）；北京市自然科学基金资助项目（No.4182030, No.L182042）

Foundation Items: The National Key R&D Program of China (No.2018AAA0102401), The China Academy of Information and Communications Technology Youth Project 2020, The Tsinghua University Independent Research Project (No.2019Z08QCX19), The National Natural Science Foundation of China (No.61902214), The Beijing Natural Science Foundation (No.4182030, No.L182042)

1　引言

随着 5G 通信技术的发展，无人机在无线通信中的应用越来越广泛[1-2]。以无人机的机动性、高灵活性和视距传输为主的信道模型提高了无线通信网络的覆盖率和吞吐量[3]。在无线传感器网络中，大量分布的传感器往往被用来监测环境，如温度、湿度、压强等。传感器需要把采集的数据通过多跳传输方式传送到融合中心。此时，每个传感器不仅需要传输自身的数据，也需要中转其他传感器的数据，造成传感器电量消耗过快，并且多跳的通信连接具有较强的不稳定性。若采用无人机进行数据采集，可以将传感器的数据直接发送给邻近的无人机[4]，大幅度提高了传输效率。

在无人机辅助无线通信的场景中，无人机的路径规划问题是关键的研究内容。文献[5-6]利用深度强化学习方法对无人机辅助路由进行路径规划。文献[7]提出了基于矩阵填充的方法优化无人机飞行路径。为了使数据采集速率最大化，文献[8-9]设计了联合算法优化传感器数据传输顺序、电量分配以及无人机飞行路径，并且通过交替优化的方法把构造的混合整数非凸优化问题转化为可解的问题。然而，当传感器的数量很多时，无人机针对每个传感器的数据采集会消耗大量时间。鉴于此，文献[10]提出了传感器聚类算法，在每一类中，传感器只需要把数据传递给头节点，而无人机只需要采集头节点的数据即可。在文献[10]中，采用遗传算法进行无人机的路径规划。然而，无人机的电量有限，若传感器的数量过多，则无人机的飞行路径会被拉长，那么在机体电量耗尽前，无人机很可能无法完成所有传感器节点的数据采集工作。因此，在执行任务过程中，让无人机按照电量阈值自行充电是一种有效的解决方法。文献[11]提出了无人机联合路径规划和循环充电分配的优化问题，通过交替迭代的方法来解决这一复杂的非凸问题。然而，这些传统的数学优化算法无法预先确定无人机何时需要充电以及充电次数。

为了实现一种动态、智能的无人机自充电方式，本文提出了基于深度强化学习的路径规划和充电分配策略。首先对无人机采集传感器数据的场景进行建模，并考虑动态充电形成优化问题；进而基于分层强化学习思想以及数据采集场景的特点设计了 option-DQN 算法进行路径规划；最后利用 option-DQN 算法对本文所考虑的场景进行仿真，并与传统 DQN 算法进行对比，证实了本文所提算法的优越性。

2　问题描述与系统模型

本文考虑一种无人机采集传感器数据的场景，无人机采集数据场景模型如图 1 所示。其中，黄色椭圆形表示待考察区域 Ψ，区域内随机分布有 N 个位置固定的传感器，每个传感器可以采集其周围地面环境的信号，传感器的通信模式为先接受无人机发射的激活信号，再通过反向散射（backscatter）方式向无人机发送存储在传感器内部的数据[12]。假设每个传感器携带的数据量不同，因此，传感器被采集的时间需求也不同。并且假设每个传感器的反射功率不同，因此，各传感器有一定的被采集范围，无人机只有在传感器的被采集范围内才能采集数据；一次数据采集周期是指具有一定初始电量的无人机从基地（拥有充电站）出发后，对所有传感器依次采集数据并返回基地，即完整的数据采集周期包括数据采集和返航充电两个阶段。

假设传感器之间距离较远，传感器的被采集范围较小，不同传感器的被采集范围没有重叠。在数据采集阶段，首先无人机需要进入传感器采集范围内并悬停一段时间采集数据，假设无人机采集各传感器数据的速率均为常数 C，而无人机在传感器之间飞行时不能同时进行数据采集。假设无人机始终在同一高度飞行与悬停，其高度由 H 表示。为了节省时间和能源，无人机由一个传感器飞行至另一个传感器时应按照两个传感器的连线进行直线飞行。此外，假设无人机在飞行过程中的飞行速度为常数 v_0。

无人机的电量会随着飞行动作和数据采集动作的增多而逐步降低。一种合理的设计是当电量低于一定阈值时，无人机需中断数据采集阶段并进入返航充电阶段。假设无人机只能在进入基地后进行充电，则从某一个传感器上方飞回充电站时，无人机也应该直线飞行，飞行速度仍假设为 v_0，无人机在充满电后离开基地并返回数据采集阶段。

在一次数据采集周期中，数据采集阶段和返航充电阶段交替进行，交替频率由无人机的初始电量、最大电量以及采集顺序等因素共同决定，因此，需要从全局设计考虑无人机电量的动态采集路径。

2.1　信道模型

考虑无人机到地面的信道模型为空—地链路

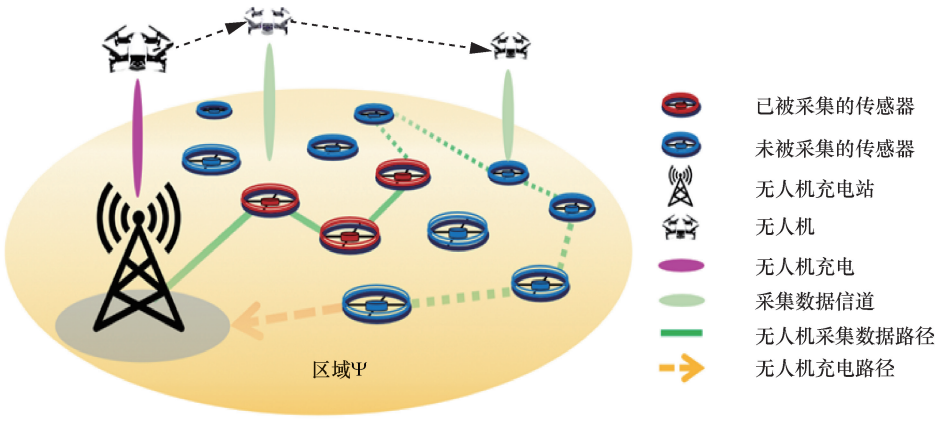


图 1 无人机采集数据场景模型

以及信道服从自由空间路径衰落，若周围存在建筑物、树木等，则会在自由空间路径损耗的基础上带来额外的损失，总体信道衰落可以表示为[13]

$$\Lambda_{\zeta} = \Lambda_{\text{free}} + \eta_{\zeta} \tag{1}$$

其中，$\Lambda_{\text{free}} = 20\log d + 20\log f_c + 20\log\dfrac{4\pi}{c}$ 为自由空间路径损耗，d 是无人机到地面传感器的距离，f_c 和 c 分别是载波频率和光的传播速度，η_{ζ} 是平均额外路径损耗，ζ 与衰落环境有关，可能是视距衰落（LoS, line of sight），也可能是非视距衰落（NLoS, non-line of sight）。假设 Pr_{LoS} 是 LoS 出现的概率，Pr_{NLoS} 是 NLoS 出现的概率，则空—地链路的平均路径损耗可以写成基于概率的形式为

$$\overline{\Lambda} = \Lambda_{\text{LoS}}\text{Pr}_{\text{LoS}} + \Lambda_{\text{NLoS}}\text{Pr}_{\text{NLoS}} \tag{2}$$

同时，用一个简单的 sigmoid 函数表示 LoS 信道的概率为[14]

$$\text{Pr}_{\text{LoS}} = \frac{1}{1 + \text{a}\mathrm{e}^{-\text{b}\left(\sin^{-1}\left(\frac{H}{d}\right)-\text{a}\right)}} \tag{3}$$

其中，a 和 b 是与环境有关的常数，H 是无人机到地面的垂直高度。

2.2 问题建模

本场景模型的研究目标是让无人机利用尽可能短的时间采集完地面所有传感器的数据，并保证在飞行途中无人机的剩余电量始终高于零。用 SN_i 表示第 i 个传感器，$i \in \mathcal{N}, \mathcal{N} = \{1,2,\cdots,N\}$，$d_i$ 表示 SN_i 存储的数据量，τ_i 表示采集 d_i 所需的单位时间。设无人机的总体数据采集量为 D_{total}，总体飞行时间为 T 个单位时间，无人机在 $t \in \mathcal{T} = \{1,2,\cdots,T\}$ 时刻是否采集 SN_i 由 $x_{t,i}$ 表示，$x_{t,i} \in \{0,1\}$，其中 1 代表采集，0 代表未采集。由于无人机需要采集完传感器中的所有数据，且每个传感器只采集 τ_i 个单位时间，可得

$$\sum_{i=1}^{N} d_i = D_{\text{total}} \tag{4}$$

$$\sum_{t=1}^{T} x_{t,i} = \tau_i, \forall i \in \mathcal{N} \tag{5}$$

其中，式(5)的具体含义为：每个传感器 i 需要 τ_i 个单位时间进行数据采集，$x_{t,i}$ 为在 t 时刻无人机是否采集第 i 个传感器的数据，对 t 求和即可得到整个飞行过程中无人机对传感器 i 数据采集的总时间为 $\sum_{t=1}^{T} x_{t,i}$。为了采集完每个传感器 i 的数据，$\sum_{t=1}^{T} x_{t,i}$ 应等于传感器 i 数据需要被采集的时间 τ_i，即 $\sum_{t=1}^{T} x_{t,i} = \tau_i, \forall i \in \mathcal{N}$。

设无人机飞行策略为 μ，无人机在时刻 t 的剩余电量为 e_t，本文所考虑的问题可以表述为受限的优化问题为

$$\min_{\mu} T \tag{6}$$

$$\text{s.t.}(4),(5) \tag{7}$$

$$x_{t,i} \in \{0,1\}, \forall i \in \mathcal{N}, t \in \mathcal{T} \tag{8}$$

$$e_t > 0, \forall t \in \mathcal{T} \tag{9}$$

其中，式(9)表示限制无人机在飞行过程中任意时刻的电量大于零。

3 分层深度强化学习

强化学习是人工智能领域的重要分支，可以用来解决许多大规模的决策问题，如机器人控制、自

动驾驶等。强化学习的核心是通过智能体与环境的交互，观察动作输出后状态的变化和环境反馈的奖励来调整自身的行为，从而使得积累奖励最大化。强化学习的算法可以分为三大类：基于策略梯度的强化学习[15]、基于值函数的强化学习[16]和基于演员—评论家（actor-critic）模型的强化学习[17]。此外，还可以分层次对行为动作进行学习，形成分层强化学习算法[18]，本文所提算法主要利用基于值函数的强化学习和基于选项（option）的分层强化学习结合的算法架构，下文将着重介绍该架构的相关理论。

3.1　马尔可夫决策过程

在强化学习中，智能体与环境的交互过程一般用马尔可夫决策过程（MDP, Markov decision process）描述。一个 MDP 是一个五元组 $<S,A,\boldsymbol{P},R,\gamma>$[19]，其中 S 表示状态空间，A 表示动作空间，$\boldsymbol{P}:S\times A\to\Delta(S)$ 表示状态转移概率矩阵，$R:S\times A\times S\to\mathbb{R}$ 表示即时奖励函数，$\gamma\in[0,1]$ 表示折扣因子。在 MDP 中的任意时刻 t，智能体观测到的当前环境状态为 $\boldsymbol{s}_t(\boldsymbol{s}_t\in\boldsymbol{S})$，并根据策略 μ 选择动作为 $\boldsymbol{a}_t(\boldsymbol{a}_t\in A)$，环境反馈给智能体相应的奖励为 $r_t(r_t\in R)$，并根据环境的状态转移矩阵 $\boldsymbol{P}$ 进入下一个状态 $\boldsymbol{s}_{t+1}(\boldsymbol{s}_{t+1}\in S)$，反复执行上述操作直至结束。

3.2　DQN

Q 学习是一种常用的、基于值函数的强化学习算法，但当强化学习场景中的动作和状态空间维度很大时，一般的 Q 学习很难完成这样复杂的任务。因此，文献[20]提出了采用神经网络估计 Q 函数，即 DQN。DQN 算法架构如图 2 所示，DQN 算法包含两个神经网络，即估计值网络和目标值网络。DQN 学习的目标是保证估计值网络输出的 Q 估计值和目标值网络输出的目标 Q 值越相近越好，该过程可以通过损失函数表示为

$$\mathrm{Loss}(\theta)=\mathrm{E}\left[\left(Q_{\mathrm{target}}-Q(\boldsymbol{s}_t,\boldsymbol{a}_t;\theta)\right)^2\right] \tag{10}$$

其中，$Q(\boldsymbol{s}_t,\boldsymbol{a}_t;\theta)$ 是当前状态的 Q 估计值，而 Q_{target} 是目标值，可表示为

$$Q_{\mathrm{target}}=r+\gamma\max_{\boldsymbol{a}_{t+1}}Q(\boldsymbol{s}_{t+1},\boldsymbol{a}_{t+1};\theta) \tag{11}$$

估计值网络的参数更新需要通过求损失函数的梯度得到，而目标值网络则通过每 N 步复制一次估计值网络的参数进行更新。为了避免强化学习的状态之间存在相关性，DQN 算法采用回放记忆单元来存放状态，动作和奖励组成的元组为 $(\boldsymbol{s}_t,\boldsymbol{a}_t,r,\boldsymbol{s}_{t+1})$。在训练时，从中随机取一些样本来训练，这样可以打破样本之间的相关性，从而提高学习效率。

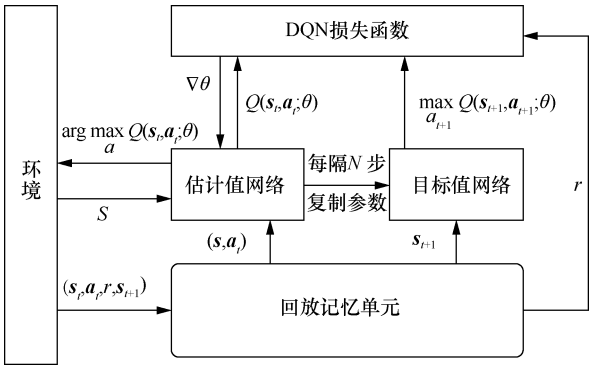


图 2　DQN 算法架构

3.3　option 与 semi-MDP

文献[21]提出的基于 option 的强化学习是分层强化学习的一种，其中 option 可以看成一系列有共同目的动作的整体抽象，定义如下。

定义 1　option 是一个三元组 $<I,\pi,\beta>$。其中，I 是可以执行该 option 的状态集合，π 是内部采取的策略，β 是终止条件。

如“开门”是一个 option，在执行“开门”的过程中按照某一个策略 π（如先旋转门把手再推开门等）执行具体的动作，然后根据是否开门来判断是否终止这个 option。

在一个 MDP 上，如果定义了一个由 option 构成的集合 O 就会成为半马尔可夫决策过程（semi-MDP），具体如下。

定理 1　(MDP+option=semi-MDP)，对于任何 MDP 以及任何定义在此 MDP 上的 option 的集合 O，仅从该集合中选择 option 并执行这个 option 直至结束的决策过程是 semi-MDP。

semi-MDP 可以用一个六元组表示，即 $<S,A,O,\boldsymbol{P},R,\gamma>$，其中 O 表示 option 集合，剩余元素的含义与 MDP 五元组中相应元素的含义相同。在 semi-MDP 中，智能体一般需要分层学习选择 option 的“高层”策略 μ 以及每个 option 内部的“底层”策略 π。

4　option-DQN 路径规划算法

在本文所考虑的场景中，无人机动作的数量与传感器的个数成正比，动作空间较复杂，且一系列

动作可以合成“高层级”的抽象动作，如一系列飞向某个传感器动作与采集动作的组合可以合成飞向该传感器并采集的抽象动作。因此，基于分层强化学习思想提出 option-DQN 算法来解决本文场景中的问题，具体过程如下。

4.1 无人机采集场景的 semi-MDP 模型

本文利用 semi-MDP 对文中描述的无人机采集任务进行建模，该模型的六元组可表示为 $<S,A,O,\boldsymbol{P},R,\gamma>$，其中智能体、状态、动作、option 集合以及即时奖励等基本要素建模如下。

1) 单智能体：无人机可以看作是一个单智能体，其选择 option 的策略为 μ。在任务开始时，无人机根据自身的初始状态 $\boldsymbol{s}_0$ 选择一个 option，记为 o_0，并按照 o_0 的内部策略 π_0 执行一系列动作，在 o_0 的结束时刻 ξ，无人机得到该 option 的总体即时奖励 r_0，再次观测自身的状态信息 $\boldsymbol{s}_\xi$，并依据 μ 选择下一个要执行的 option，记为 o_ξ，如此循环直至无人机任务结束。

2) 状态：假设无人机可以完全观测自身的状态，即在每一个时刻 t 的状态 $\boldsymbol{s}_t=(\mathbf{cr}_t,\boldsymbol{p}_t,e_t)$，$\boldsymbol{s}_t\in S$，其中，$\mathbf{cr}_t$ 为 t 时刻每个传感器被采集的百分比组成的向量，即 $\mathbf{cr}_t=(\mathrm{cr}_t^1,\mathrm{cr}_t^2,\cdots,\mathrm{cr}_t^N)$，$\mathrm{cr}_t^i\in[0,1]$，$\forall i\in\mathcal{N}$；$\boldsymbol{p}_t$ 为无人机 t 时刻所在的位置向量，即 $\boldsymbol{p}_t=(x_t,y_t,z_t)$，因为无人机的高度为一常数 H，则 $z_t=\mathrm{H}$；e_t 为无人机的剩余电量。

3) 动作：无人机的基本动作集合 A 中包含 3 种基本动作，即向某一目标直线飞行动作 $\boldsymbol{a}_\mathrm{f}$、充电动作 $\boldsymbol{a}_\mathrm{c}$ 以及采集数据动作 $\boldsymbol{a}_\mathrm{cr}$。在每一个时刻 t，执行动作 $\boldsymbol{a}_\mathrm{f}$ 使得无人机向指定目标方向以速度 v_0 直线飞行一个单位时间，执行动作 $\boldsymbol{a}_\mathrm{c}$ 使得无人机充满电量，执行动作 $\boldsymbol{a}_\mathrm{cr}$ 使得无人机采集数据量 C。

4) option 集合：本文中无人机执行的 option 集合 O 包含采集某个传感器 SN_i 的数据 $o_{\mathrm{s},i}(i\in\mathcal{N})$、返航充电 o_c 以及结束任务 o_p 等 3 种 option，即 $O=\{o_{\mathrm{s},1},o_{\mathrm{s},2},\cdots,o_{\mathrm{s},N},o_\mathrm{c},o_\mathrm{p}\}$。其中，每个 option 均为一个三元组 $<I_o,\pi_o,\beta_o>$，option 集合 O 包含元素个数为 $|O|=N+2$。无人机在任意一个状态可选择的 option 集合为整个 option 集合 O，即 $I_o=S,\forall o\in O$。在本文中每个 option 内部的策略 π_o 均设定为已知的固定策略，每个 option 的终止条件 β_o 均为执行完所有动作。具体地，对于采集传感器的 option $o_{\mathrm{s},i}$，其策略为由当前位置直线飞向传感器 SN_i 并采集该传感器的数据 d_i，直到采集完毕后才退出当前 option；对于返航充电 o_c，其策略为直线飞向充电站并进行充电，直到充满电量后才退出 o_c；对于结束任务 o_p，其策略为无人机直线飞向终点，告知环境任务结束并退出 o_p。在仿真时，每个 option 内部的策略不用训练。

5) 即时奖励：在本文中，智能体只在每个选项结束时刻 ξ 得到该 option 的总体即时奖励 r_ξ。r_ξ 是 option 初始状态 s_ξ 和选项动作 o_ξ 的函数，即 $r_\xi=r_\xi(s_\xi,o_\xi)$。设定一个 option 的总体即时奖励分为电量奖励 r_ξ^e、采集奖励 r_ξ^c 和路径奖励 r_ξ^l 3 个部分，其中，电量奖励主要用于惩罚无人机在执行该 option 过程中电量不足的情况，即

$$r_\xi^\mathrm{e}=\begin{cases}\mathrm{N_e},e_\xi\leqslant 0\\ 0,e_\xi>0\end{cases}\tag{12}$$

采集奖励用于惩罚无人机重复选择已经完成采集的传感器 option，即

$$r_\xi^\mathrm{c}=\begin{cases}\mathrm{N_c},\mathrm{cr}_\xi^{o_\xi}=1\\ 0,\mathrm{cr}_\xi^{o_\xi}<1\end{cases}\tag{13}$$

而路径奖励 r_ξ^l 与无人机在该 option 内飞过的距离 l_ξ 成反比，用于引导无人机学习飞行尽量短的路径来采集传感器数据，可以表示为

$$r_\xi^\mathrm{l}=\frac{\mathrm{N_l}}{l_\xi}\tag{14}$$

式(12)～式(14)中，$\mathrm{N_e}$、$\mathrm{N_c}$ 和 $\mathrm{N_l}$ 均为负常数。最终，无人机经历一个 option 得到的即时奖励 r_ξ 为上述 3 种奖励的和，表示为

$$r_\xi=r_\xi^\mathrm{e}+r_\xi^\mathrm{c}+r_\xi^l\tag{15}$$

semi-MDP 模型中状态转移矩阵 $\boldsymbol{P}$ 描述了本场景中的环境的动力学规则，其定义了无人机在任意状态 $\boldsymbol{s}$ 下采用动作 $\boldsymbol{a}$ 时，其下一个状态 $\boldsymbol{s}'$ 的概率分布为

$$\boldsymbol{P}\triangleq\mathrm{P}\left(\boldsymbol{s}'|\boldsymbol{s},\boldsymbol{a}\right),\forall\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}'\tag{16}$$

类似于 MDP 中状态动作值函数 $Q^\pi(\boldsymbol{s},\boldsymbol{a})$，在 semi-MDP 模型中，定义状态-option 值函数 $Q_O^\mu(\boldsymbol{s},o)$ 来表示在策略 μ 和 option 集合 O 下 $(\boldsymbol{s},o)$ 的状态-option 元组，可以得到累积回报的期望为

$$\begin{aligned} &Q_O^{\mu}(\boldsymbol{s},o)=\mathrm{E}_{\boldsymbol{s}'\sim\mu,\boldsymbol{P}}\left[r+\gamma r'+\gamma^2 r''+\cdots\right]= \\ &r+\gamma\mathrm{E}_{\boldsymbol{s}'\sim\mu,\boldsymbol{P}}\left[r'+\gamma r''+\cdots\right]= \\ &r+\gamma\mathrm{E}_{\boldsymbol{s}'\sim\boldsymbol{P}}\left[\mathrm{E}_{o'\sim\mu}Q_O^{\mu}(\boldsymbol{s}',o')\right]= \\ &r+\gamma\sum_{\boldsymbol{s}'}\mathrm{P}(\boldsymbol{s}'\mid\boldsymbol{s},o)\sum_{o'}\mu(o'|\boldsymbol{s}')Q_O^{\mu}(\boldsymbol{s}',o') \end{aligned} \tag{17}$$

其中，$\gamma\in[0,1]$ 表示累积即时奖励的折扣因子，式(17)也即 $Q_O^{\mu}(\boldsymbol{s},o)$ 满足的贝尔曼期望方程（Bellman expectation equation）。为了求解 semi-MDP 模型，需要找到最优状态-option 值函数 $Q_O^{*}(\boldsymbol{s},o)$，其应满足贝尔曼最优方程（Bellman optimality equation），如式(18)所示。为了求解贝尔曼最优方程，通常可以利用 Monte Carlo 算法、时序差分（TD, temporal difference）算法等进行求解。

$$Q_O^{*}(\boldsymbol{s},o)=r+\sum_{\boldsymbol{s}'}\mathrm{P}(\boldsymbol{s}'|\boldsymbol{s},o)\left[\max_{o'}Q_O^{*}(\boldsymbol{s}',o')\right] \tag{18}$$

4.2　option-DQN 算法

本文提出基于 option 与 DQN 算法相结合的 option-DQN 算法来解决无人机数据采集中的路径规划问题。option-DQN 算法执行架构如图 3 所示，图 3 中展示了算法执行时的推理过程，无人机与环境的交互过程中从上一个 option $o_{\xi-1}$ 退出，获得该 option 的整体即时奖励为 $r_{\xi-1}$ 以及下一步的状态信息为 $\boldsymbol{s}_{\xi}$，$\boldsymbol{s}_{\xi}$ 包括当前所有传感器节点已经采集的数据百分比向量 $\mathbf{cr}_{\xi}$、无人机的位置信息 $\boldsymbol{p}_{\xi}$ 以及无人机剩余的电量 e_{ξ}。进而将当前 $\boldsymbol{s}_{\xi}$ 输入到 option-DQN 算法中的值函数神经网络 Q_{op} 中，该网络由输入层、隐藏层和输出层组成，其中，隐藏层由两个全连接层构成，第一个全连接网络包含 1 024 个神经元，其激活函数采用线性整流函数（ReLU, rectified linear unit）[22]，第一层网络的输出可以表示为

$$\boldsymbol{X}_1=\mathrm{ReLU}(\boldsymbol{W}_1^{\mathrm{T}}\boldsymbol{s}_{\xi}+\boldsymbol{b}_1) \tag{19}$$

其中，$\boldsymbol{W}_1$ 是第一层神经网络的权重参数，$\boldsymbol{b}_1$ 是其偏差参数。第二层隐藏层的输入是第一层隐藏层的输出，第二层隐藏层由 300 个神经元构成，其激活函数与上一层相同，也采用 ReLU 函数，这一层的输出可以表示为

$$\boldsymbol{X}_2=\mathrm{ReLU}(\boldsymbol{W}_2^{\mathrm{T}}\boldsymbol{X}_1+\boldsymbol{b}_2) \tag{20}$$

其中，$\boldsymbol{W}_2$ 和 $\boldsymbol{b}_2$ 分别为第二层网络的权重参数和偏差参数。输出层接受第二层网络的输出 $\boldsymbol{X}_1$，并利用 softmax 激活函数输出 $|O|$ 维向量 $\boldsymbol{q}$ 为

$$\boldsymbol{q}=\mathrm{softmax}(\boldsymbol{W}_3^{\mathrm{T}}\boldsymbol{X}_2+\boldsymbol{b}_3) \tag{21}$$

其中，$\boldsymbol{W}_3$ 和 $\boldsymbol{b}_3$ 分别是输出层的权重参数和偏差参数，softmax 是归一化指数函数。

最后，值函数神经网络 Q_{op} 的输出 $\boldsymbol{q}=(q_1,q_2,\cdots,q_{|O|})$ 是选择每一个 option 的概率，即 $\sum_{j=1}^{|O|}q_j=1$。通过使用 $\varepsilon-$greedy 算法[23]来找到最优的 option。ε

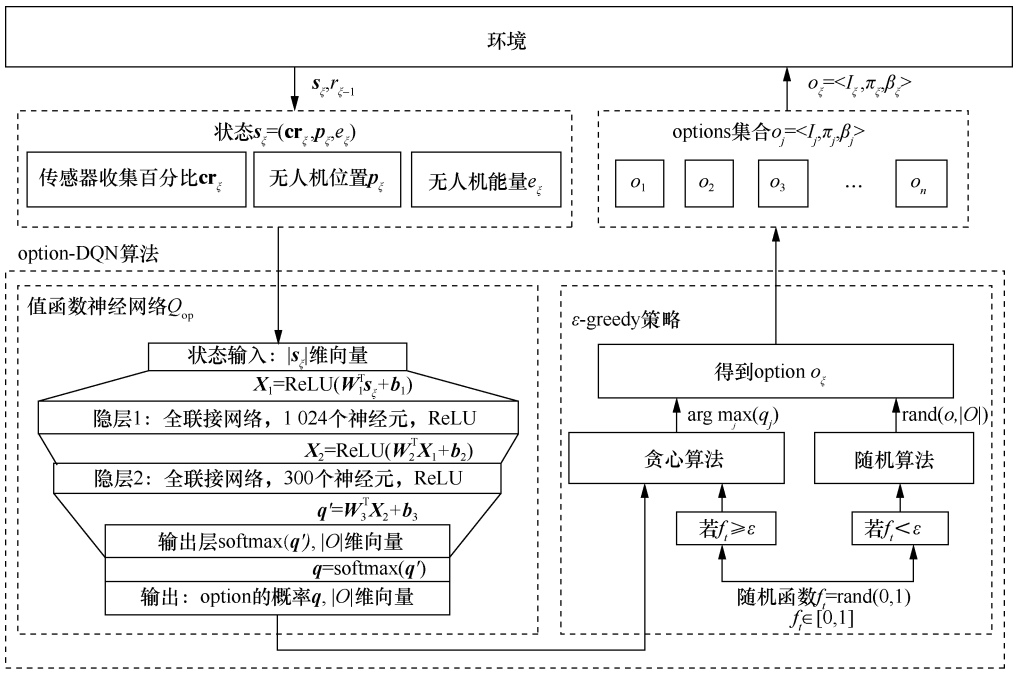


图 3　Option-DQN 算法执行架构

是 0～1 的一个较小的值，每次以 ε 的概率进行随机选择，以 $1-\varepsilon$ 的概率用贪心算法进行选择，即选择 $\boldsymbol{q}$ 中最大数值的索引作为要选择的 option o_ξ，贪心算法可以表示为

$$o_\xi = \arg\max_j (q_j) \tag{22}$$

最后在 option 集合 O 中选择 o_ξ 对应的策略 π_ξ 和终止条件 β_ξ 并输出相应的动作，继续与环境进行交互。

4.3 option-DQN 训练

在训练 options-DQN 算法时，设置无人机经验存储器 $\mathcal{D}=\{\boldsymbol{e}_k \mid k \in D_c\}$，其中，$\boldsymbol{e}_k$ 为第 k 条经验向量，即 $\boldsymbol{e}_k=(\tilde{\boldsymbol{s}}_k,\tilde{o}_k,\tilde{r}_k,\tilde{\boldsymbol{s}}'_k)$，$\tilde{\boldsymbol{s}}_k$ 表示当前状态，$\tilde{o}_k$ 表示根据当前 option-DQN 算法得到的 option 动作，$\tilde{r}_k$ 为 $(\boldsymbol{s}'_k,o'_k)$ 得到的总体即时反馈，$\tilde{\boldsymbol{s}}'_k$ 表示经过与环境交互后无人机转移的下一个状态，而 D_c 表示存储器最大存储量。采用经验回放和经验随机抽取的方式训练值函数神经网络 Q_{op}，可以打破数据之间的关联性，使得数据尽量满足独立同分布的特性，从而增强训练的稳定性。算法中值函数神经网络 Q_{op} 又称为评估网络，此外，设置目标网络 $Q_{\text{op}}^{\text{target}}$ 用于近似表示最优评估网络 Q_{op}^{*}，即 $Q_{\text{op}}^{\text{target}} \approx Q_{\text{op}}^{*}$。评估网络的损失函数可以表示为

$$\mathcal{L}(\theta)=\mathrm{E}_{(\tilde{\boldsymbol{s}}_k,\tilde{o}_k,\tilde{r}_k,\tilde{\boldsymbol{s}}'_k)\sim\mathcal{D}}\left[\left(\gamma\arg\max Q_{\text{op}}^{\text{target}}\left(\tilde{\boldsymbol{s}}'_k\right)+\tilde{r}_k-Q_{\text{op}}\left(\tilde{\boldsymbol{s}}_k,\tilde{o}_k;\theta\right)\right)^2\right] \tag{23}$$

在式(23)中，θ 表示值函数神经网络 Q_{op} 中的所有参数，其更新规则为

$$\theta_{\text{new}}=\theta_{\text{old}}-\alpha\nabla_\theta\mathcal{L}(\theta) \tag{24}$$

其中，α 为学习速率，θ_{new} 和 θ_{old} 分别表示评估网络更新后的参数和更新前的参数，损失函数的梯度 $\nabla_\theta\mathcal{L}(\theta)$ 可以表示为

$$\nabla_\theta\mathcal{L}(\theta)=\mathrm{E}_{(\tilde{\boldsymbol{s}}_k,\tilde{o}_k,\tilde{r}_k,\tilde{\boldsymbol{s}}'_k)\sim\mathcal{D}}\left[2\left(\arg\max Q_{\text{op}}^{\text{target}}\left(\tilde{\boldsymbol{s}}_k\right)+\tilde{r}_k-Q_{\text{op}}\left(\tilde{\boldsymbol{s}}_k,\tilde{o}_k;\theta\right)\right)\times\nabla_\theta Q_{\text{op}}\left(\tilde{\boldsymbol{s}}_k,\tilde{o}_k;\theta\right)\right] \tag{25}$$

为了得到的总体即时反馈，表示经过与环境交互采取“软更新”的方式对目标网络进行更新，即每隔一定的周期后，利用原有目标网络和当前估计网络的参数共同对目标网络进行更新，其更新规则为

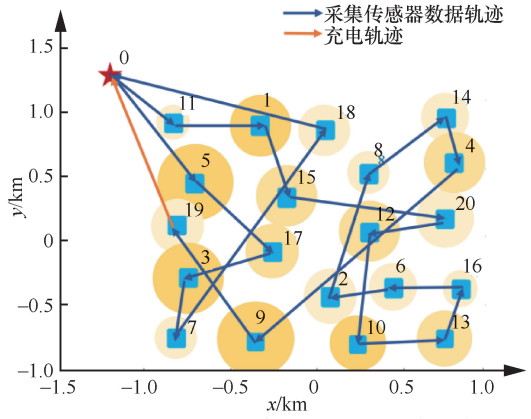
$$\theta_{\text{target,new}}=\beta\theta_{\text{target,old}}+(1-\beta)\theta \tag{26}$$

其中，β 为更新速率，并且 $\beta\in[0,1]$，$\theta_{\text{target,new}}$ 和 $\theta_{\text{target,old}}$ 分别表示目标网络 $Q_{\text{op}}^{\text{target}}$ 更新后的参数和更新前的参数，对目标网络采用“软更新”的方式可以增加神经网络训练的稳健性[20]。

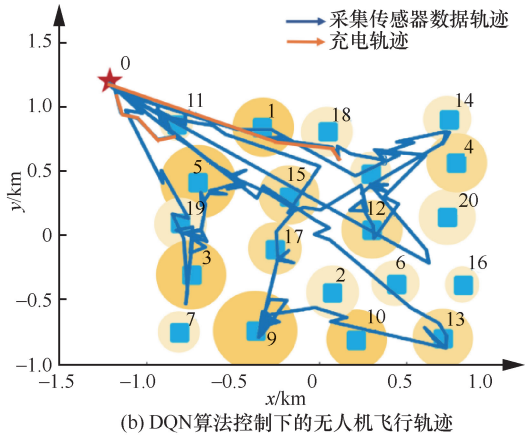
5 仿真分析

本节将对所提算法进行仿真验证和分析，并与传统的 DQN 算法[20]进行对比，验证了所提算法的有效性。

考虑一个 2.5 km×2.5 km 的区域，两种算法控制下的无人机飞行轨迹对比如图 4 所示。



(a) option-DQN算法控制下的无人机飞行轨迹



(b) DQN算法控制下的无人机飞行轨迹

图 4 两种算法控制下的无人机飞行轨迹对比

在该区域上随机分布有 20 个位置固定的地面传感器，每个传感器存储的数据量不同导致数据采集时间不同，在图 4 中将所需采集时间用不同的黄色圆盘表示，圆盘的面积越大则表示其所需的采集时间越长。无人机的出发点、充电站以及终点均在（−1.2 km, 1.2 km）位置，即图 4 中红色五角星位置处，无人机的飞行高度固定为 50 m，飞行速度固定为 1 m/s。无人机的初始电量为 100 个单位电量，

每飞行或采集 1 个单位时间将消耗 1 个单位电量，一个单位时间为 1 s，仿真参数设置如表 1 所示[24]。

表 1　　仿真参数设置

参数	描述	值
f_c	载波频率	2.5 GHz
$(a,b,\eta_{\mathrm{LoS}},\eta_{\mathrm{NLoS}})$	郊区环境	(4.88, 0.43, 0.1, 21)
	城市环境	(9.61, 0.16, 1, 20)
	密集城市环境	(12.08, 0.11, 1.6, 23)
	高度密集城市环境	(27.23, 0.08, 2.3, 34)
γ	奖励折扣因子	0.95

分别采用本文所提的 option-DQN 算法和传统的 DQN 算法对该场景中的无人机进行路径规划训练，利用 PyTorch 深度学习架构实现两种算法的神经网络部分，计算平台使用的 CPU 为英特尔酷睿 i7-8700k，内存为 32 GB，GPU 为 NVIDIA GeForce RTX 2080。每个算法均训练 1 200 个周期，从无人机任务开始直至任务结束为一个周期，两种算法的训练回报对比如图 5 所示。

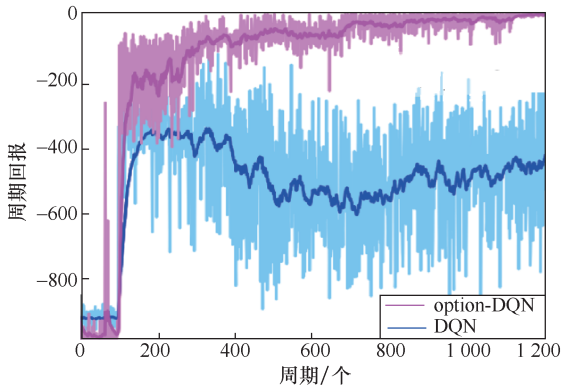


图 5　两种算法的训练回报对比

从图 5 中可以看出，所提 option-DQN 算法和 DQN 算法在前 100 个周期自探索阶段的周期回报大致相同；在开始训练后，两个算法的周期回报均迅速增长。然而，option-DQN 算法的周期回报增长更快；在后面的训练中，option-DQN 算法的周期回报持续增长并最终收敛，而 DQN 算法的周期回报则震荡明显且方差较大，其最终的周期回报明显低于 option-DQN 算法。上述比较说明了相比传统 DQN 算法，所提 option-DQN 算法利用直接学习高层次策略的方式能更快地掌握场景的含义，因此，所提算法更有效；而传统 DQN 算法每次只选择基本的动作，缺乏整体考虑，如其经常在采集一个传感器数据的途中转而采集另一个传感器数据，造成数据采集效率较低。

接下来，对已训练的 option-DQN 算法和 DQN 算法的应用效果进行对比，图 4 展示了两种算法控制下的无人机飞行轨迹对比，图 6 为两种算法控制下的无人机剩余电量对比，两种算法执行时的指标数据如表 2 所示。

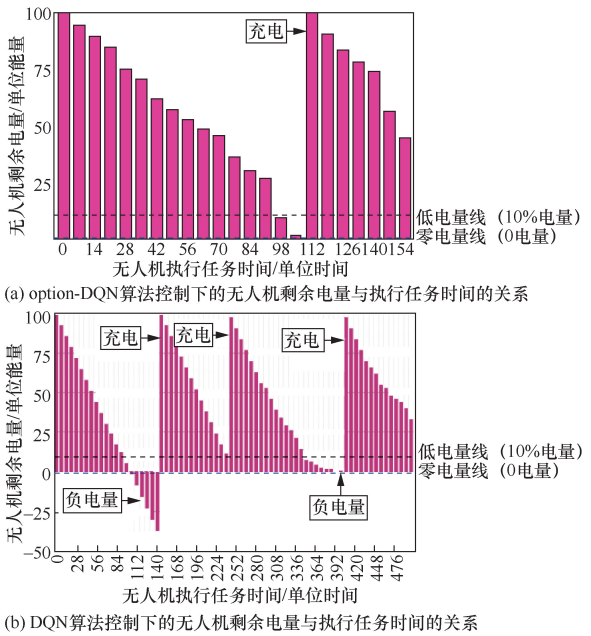


图 6　两种算法控制下的无人机剩余电量对比

表 2　　两种算法执行时的指标数据

指标	option-DQN 算法	DQN 算法
飞行总时长/s	162	519
采集节点数	20	15
是否采集完所有节点	是	否
充电次数/次	1	3
途中是否存在低电量状态	是	否
途中是否存在负电量状态	否	否

从图 4(a)可见，无人机由起始点出发遍历每个传感器一次，最后返回终点，途中返回充电桩充电一次，在整个飞行轨迹中，无人机共执行 22 次 option，用时 162 s；从图 4(b)中可见，使用 DQN 算法进行采集任务的无人机轨迹比用 option-DQN 算法的无人机轨迹更杂乱，无人机只采集了 15 个传感器数据且存在重复采集，在剩余 5 个传感器数据未被采集的状态下就返回终点结束了任务，整个任务周期用时 519 s。从图 6(a)中可见，使用

option-DQN 算法的无人机在采集前 15 个传感器数据 option 时的电量是逐次下降的，并逐渐低于低电量线（10%电量）；在第 16 个 option 选择充电后，电量恢复至 100 个单位电量，之后无人机继续采集剩余传感器数据，并在结束前保持电量充足，无人机在整个飞行途中存在低电量状态，但是电量始终大于 0，没有负电量时刻，说明该路径规划有效；而图 6(b)中，使用 DQN 算法的无人机在整个周期中进行 3 次充电，且存在负电量时刻，说明若要无人机完成整个规划路径，则需要在途中额外补充电量，否则该规划无效。由无人机轨迹及其电量变化可以看出，在本仿真环境中，option-DQN 算法直接学习抽象动作 option 后成功规避了无人机重复遍历同一个传感器的错误，学到了在电量低时返回充电站进行充电，避免无人机处于负电量状态，并尽量保持总路程最短。事实上，对于不同的传感器分布网络来说，option-DQN 算法基本可以使无人机智能地实现遍历所有传感器的路径最短和及时充电的方式，从而较好地解决了 2.2 节中式(6)～式(9)表述的路径优化问题。

6 结束语

本文研究了电量有限且可充电的无人机采集传感器网络数据的场景，并对该场景进行了通信建模。提出了基于分层强化学习的 option-DQN 算法来联合优化无人机的飞行路径和充电过程，从而保证完成数据采集任务的时间最短。仿真结果表明，相对于传统 DQN 算法，在所提算法控制下的无人机完成任务时间更短，路径更清晰，并且可以不重复采集传感器数据。

参考文献：

[1] ZHAO N, LU W D, SHENG M, et al. UAV-assisted emergency networks in disasters[J]. IEEE Wireless Communications, 2019, 26(1): 45-51.

[2] CHENG F, ZHANG S, LI Z, et al. UAV trajectory optimization for data offloading at the edge of multiple cells[J]. IEEE Transactions on Vehicular Technology, 2018, 67(7): 6732-6736.

[3] YOU C S, ZHANG R. 3D trajectory optimization in Rician fading for UAV-enabled data harvesting[J]. IEEE Transactions on Wireless Communications, 2019, 18(6): 3192-3207.

[4] ZHAN C, ZENG Y, ZHANG R. Energy-efficient data collection in UAV enabled wireless sensor network[J]. IEEE Wireless Communications Letters, 2018, 7(3): 328-331.

[5] SHAMSOSHOARA A, KHALEDI M, AFGHAH F, et al. Distributed cooperative spectrum sharing in UAV networks using multi-agent reinforcement learning[C]//2019 16th IEEE Annual Consumer Communications & Networking Conference (CCNC). IEEE, 2019: 1-6.

[6] YANG Q, JANG S J, YOO S J. *Q*-learning-based fuzzy logic for multi-objective routing algorithm in flying Ad Hoc networks[J]. Wireless Personal Communications, 2020, 113(1): 115-138.

[7] LIU X, LIU Y X, ZHANG N, et al. Optimizing trajectory of unmanned aerial vehicles for efficient data acquisition: a matrix completion approach[J]. IEEE Internet of Things Journal, 2019, 6(2): 1829-1840.

[8] ZHANG J, ZENG Y, ZHANG R. Multi-antenna UAV data harvesting: joint trajectory and communication optimization[J]. Journal of Communications and Information Networks, 2020, 5(1): 86-99.

[9] ZHAN C, ZENG Y, ZHANG R. Trajectory design for distributed estimation in UAV-enabled wireless sensor network[J]. IEEE Transactions on Vehicular Technology, 2018, 67(10): 10155-10159.

[10] ALFATTANI S, JAAFAR W, YANIKOMEROGLU H, et al. Multi-UAV data collection framework for wireless sensor networks[C]//2019 IEEE Global Communications Conference (GLOBECOM). IEEE, 2019.

[11] LI X W, YAO H P, WANG J J, et al. Joint node assignment and trajectory optimization for rechargeable multi-UAV aided IoT systems[C]//2019 11th International Conference on Wireless Communications and Signal Processing (WCSP). IEEE, 2019: 1-6.

[12] ZHANG Y, LI B, GAO F F, et al. A robust design for ultra reliable ambient backscatter communication systems[J]. IEEE Internet of Things Journal, 2019, 6(5): 8989-8999.

[13] CUI M, ZHANG G C, WU Q Q, et al. Robust trajectory and transmit power design for secure UAV communications[J]. IEEE Transactions on Vehicular Technology, 2018, 67(9): 9042-9046.

[14] AL-HOURANI A, KANDEEPAN S, LARDNER S. Optimal LAP altitude for maximum coverage[J]. IEEE Wireless Communications Letters, 2014, 3(6): 569-572.

[15] SCHULMAN J, WOLSKI F, DHARIWAL P, et al. Proximal policy optimization algorithms[J]. arXiv: 1707.06347, 2017.

[16] SCHAUL T, QUAN J, ANTONOGLOU I, et al. Prioritized experience replay[J]. arXiv: 1511.05952, 2015.

[17] MNIH V, BADIA A P, MIRZA M, et al. Asynchronous methods for deep reinforcement learning[C]//International Conference on Machine Learning. 2016: 1928-1937.

[18] KULKARNI T D, NARASIMHAN K, SAEEDI A, et al. Hierarchical deep reinforcement learning: integrating temporal abstraction and intrinsic motivation[C]//Advances in Neural Information Processing Systems. 2016: 3675-3683.

[19] 丁瑞金, 高飞飞, 邢玲. 基于深度强化学习的物联网智能路由策略[J]. 物联网学报, 2019, 3(2): 56-63.
DING R J, GAO F F, XING L. Intelligent routing strategy in the Internet of things based on deep reinforcement learning[J]. Chinese Journal on Internet of Things, 2019, 3(2): 56-63.

[20] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518(7540): 529-533.

[21] SUTTON R S, PRECUP D, SINGH S. Between MDPs and semi-MDPs: a framework for temporal abstraction in reinforcement learning[J]. Artificial Intelligence, 1999, 112(1-2): 181-211.

[22] 蒋昂波, 王维维. ReLU 激活函数优化研究[J]. 传感器与微系统, 2018, 37(2): 50-52.
JIANG A B, WANG W W. Research on optimization of ReLU activation function[J]. Transducer and Microsystem Technology, 2018, 37(2): 50-52.

[23] TOKIC M, PALM G. Value-difference based exploration: adaptive control between epsilon-greedy and softmax[C]//Annual Conference on Artificial Intelligence. Springer, 2011: 335-346.

[24] BOR-YALINIZ R I, EL-KEYI A, YANIKOMEROGLU H. Efficient 3-D placement of an aerial base station in next generation cellular networks[C]//2016 IEEE International Conference on Communications (ICC). IEEE, 2016: 1-5.

[作者简介]

牟治宇（1997− ），男，河北石家庄人，清华大学硕士生，主要研究方向为基于深度强化学习的无人机路径规划。

张煜（1993− ），女，河南郑州人，清华大学博士生，主要研究方向为物联网通信理论、基于强化学习的无人机路径规划。

范典（1992− ），男，山东菏泽人，中国信息通信研究院泰尔终端实验室战略规划与研究部工程师，主要研究方向为毫米波大规模多天线通信理论、阵列信号处理和无人机通信理论。

刘君（1982− ），女，山东济南人，博士，清华大学助理研究员，主要研究方向为天地一体化网络、无人机组网。

高飞飞（1980− ），男，陕西西安人，博士，清华大学副教授、博士生导师，主要研究方向为多天线通信和智能信号处理技术。