引用本文：尚晓凯，韩龙龙，翟慧鹏. 基于改进 DQN 强化学习算法的弹性光网络资源分配研究[J]. 光通信技术，2023，47(5)：12-15.

基于改进 DQN 强化学习算法的弹性光网络资源分配研究

尚晓凯，韩龙龙*，翟慧鹏
（国家计算机网络与信息安全管理中心河南分中心，郑州 450000）

摘要：针对光网络资源分配中频谱资源利用率不高的问题，提出了一种改进的深度 Q 网络（DQN）强化学习算法。该算法基于 ε-greedy 策略，根据动作价值函数和状态价值函数的差异来设定损失函数，并不断调整 ε 值，以改变代理的探索率。通过这种方式，实现了最优的动作值函数，并较好地解决了路由与频谱分配问题。此外，采用了不同的经验池取样方法，以提高迭代训练的收敛速度。仿真结果表明：改进 DQN 强化学习算法不仅能够使弹性光网络训练模型快速收敛，当业务量为 300 Erlang 时，比 DQN 算法频谱资源利用率提高了 10.09%，阻塞率降低了 12.41%，平均访问时延减少了 1.27 ms。

关键词：弹性光网络；改进深度 Q 网络强化学习算法；资源分配
中图分类号：TN929.1　**文献标志码**：A　**文章编号**：1002-5561(2023)05-0012-04
DOI：10.13921/j.cnki.issn1002-5561.2023.05.003　**开放科学(资源服务)标识码(OSID)**：

Research on elastic optical network resource allocation based on improved DQN reinforcement learning algorithm

SHANG Xiaokai, HAN Longlong*, ZHAI Huipeng
(National Computer Network and Information Security Management Center, Henan Branch, Zhengzhou 450000, China)

Abstract: Aiming at the low utilization of spectrum resources in optical network resource allocation, an improved deep Q network (DQN) reinforcement learning algorithm is proposed. Based on the ε-greedy strategy, the algorithm sets the loss function according to the difference between the action value function and the state value function, and constantly adjusts the ε value to change the exploration rate of the agent. In this way, the optimal action value function is realized, and the routing and spectrum allocation problems are solved well. In addition, different experience pool sampling methods are used to improve the convergence speed of iterative training. The simulation results show that the improved DQN reinforcement learning algorithm can not only make the elastic optical network training model converge quickly, but also improve the spectrum resource utilization by 10.09%, reduce the blocking rate by 12.41% and reduce the average access delay by 1.27 ms compared with DQN algorithm when the traffic volume is 300 Erlang.

Key words: elastic optical network, improved reinforcement learning algorithm for deep Q network, resource allocation

0 引言

随着数据中心、云平台和第五代移动通信（5G）网络等互联网业务的蓬勃发展，网络数据流量急剧增加，人们对基础网络传输提出了更高要求[1]，传统的波分复用光网络传输方法已难以满足网络业务需求[2]。基于正交频分复用的弹性光网络（EON）通过对频域需求的调控来提升频谱资源利用率，并针对不同业务需求提供灵活的配置方案[3]。路由与频谱分配（RSA）[4]是 EON 中至关重要的技术之一，一些研究学者已将深度 Q 网络（DQN）强化学习算法（下文简称 DQN 算法）应用于 EON 的 RSA 中[5-7]，但该算法在应用过程中仍然存在分配效果不佳、灵活性差、收敛速度慢等问题[8-9]。基于此，本文对 DQN 算法进行改进，提出改进 DQN 强化学习算法（下文简称改进 DQN 算法），以便提高 EON 的频谱利用率，降低网络阻塞率和平均访问时延。

收稿日期：2023-03-09。
基金项目：国家计算机网络与信息安全技术研究专项（242 研究计划）（2022Q66）资助；国家自然科学基金项目（批准号：61901159）资助。

作者简介：尚晓凯（1990—），男，硕士，工程师，从事信息系统集成建设与维护工作，主要研究方向为人工智能、网络传输技术、网络安全，参与主持了国家计算机网络与信息安全技术研究专项 2 项，公开发表论文 2 篇，授权国家发明专利 1 项。

* **通信作者**：韩龙龙（1988—），男，硕士，工程师，主要研究方向为信号传输及人工智能。

1 改进 DQN 算法

由于 DQN 算法[12]执行 ε-greedy 策略时会造成动作空间误差过大、动作选择不合理等问题，因此本文对 DQN 算法进行了改进。改进方法如下：在 ε-greedy 策略的基础上根据动作价值函数和状态价值函数的差异性分别确定损失函数，并不断调整 ε 值。在 DQN 训练的开始阶段和最后阶段，可以通过观察不同动作产生的状态价值函数的差异性来确定 ε 值，从而使动作选择更加合理。这种方法可以提升算法的迭代速度，并得到最优的频谱资源分配模型，从而更好地解决 RSA 问题。同时，为了提高样本利用率，本文采用不同经验回收池取样的方法，以提升迭代训练收敛速度，为复杂的 EON 网络环境提供解决方案。改进 DQN 算法流程图如图 1 所示。

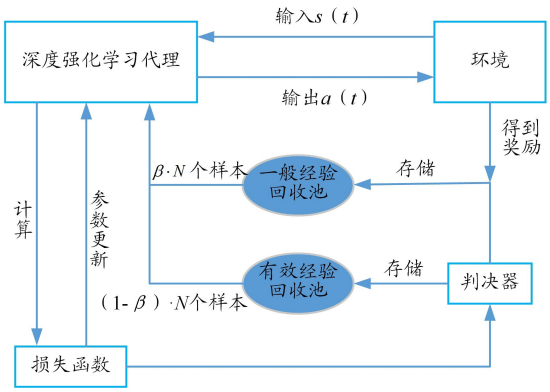

图 1 改进 DQN 算法流程图

改进 DQN 算法流程步骤如下：

1）初始化折扣因子、网络带宽、更新因子等参数，并使初始化状态函数值为 0。

2）通过 ε-greedy 策略生成状态函数。

3）计算动作函数值。

4）将获得的奖励函数通过不同经验回收池取样。

5）计算损失函数值，从而不断更新迭代参数。

6）重复步骤 2）~ 步骤 5），当训练次数大于等于最大迭代次数时终止算法。

由于动作价值函数 $Q^{\pi}(s(t),a(t))$ 须须在玻尔兹曼分布下满足一定条件[15]，因此可以将不同环境下采取动作的概率更新为 $h(s(t))$，即

$$h(s(t+1))=\delta\cdot f(s(t),a(t),\sigma)+(1-\sigma)h(s(t)) \quad (1)$$

$$f(s(t),a(t),\sigma)=\frac{1-e^{\frac{-|Q(s(t+1),a(t+1))-Q(s(t),a(t))|}{\sigma}}}{1+e^{\frac{-|Q(s(t+1),a(t+1))-Q(s(t),a(t))|}{\sigma}}} \quad (2)$$

其中，δ 表示动作价值函数对探索概率的影响因数，$\delta=1/|A(s)|$；σ 表示探索的灵敏度，σ 值越大，探索概率越小；$s(t)$、$a(t)$分别为状态函数和动作函数。

在 DQN 训练起步阶段，首先将 $h(s(0))$设置为 1，然后通过深度神经网络强化学习，取得 $h(s(t))$的近似值为

$$h(s(t))\approx\hat{h}(s(t)) \quad (3)$$

为了获得最小损失函数，用 $y(t)$表示状态价值函数的目标值，可以得到

$$y(t)=\delta\cdot f(s(t-1),a(t-1),\sigma)+(1-\sigma)\cdot\hat{h}(s(t-1)) \quad (4)$$

损失函数 $L(w)$可表示为

$$L(w)=E\left[(y(t)-\hat{h}(s(t),w))^2\right] \quad (5)$$

其中，$E[\cdot]$表示求期望值；w为权重，可通过随机梯度下降算法[16]对 w进行更新。若在经验回收池中抽取 N 个样本，则 $L(w)$的表达式变更为

$$L(w)=\frac{1}{N}\sum_{i=1}^{i}\left[y(i)-\hat{h}(s(i),w)\right]^2 \quad (6)$$

其中，i为网络节点数。

同时，为了防止过拟合，在进行样本抽样时，从不同的经验回收池中分别抽取 $(1-\beta)\cdot N$和 $\beta\cdot N$个样本进行训练，其中 β 为经验系数。

2 仿真结果与分析

为了评估所提出的改进 DQN 算法在光网络资源分配方案中的有效性和收敛性，本文分别以收敛效果、频谱资源利用率、阻塞率、平均访问时延等性能指标，对该算法进行仿真分析验证，并与常用的 Q-Learning（简称 QL）、DQN 算法进行对比分析。搭建的仿真实验平台为多核服务器，设备具有 16 个 2.5 GHz Intel Core i5 CPU 内核，2 个 GPU 内核和 32 GB RAM，以保证 EON 资源分配对计算资源和网络的较高要求。

以 CERNET 网络为对象进行实验，设定每条链路有 400 个频隙资源，实验过程中均采用最短路径建立路由，业务到达率满足泊松分布。本次仿真的具体参

表 1 仿真参数

仿真参数	值
业务达到率	2
折扣因子	0.99
频隙带宽/GHz	12.5
更新因子	0.01
学习效率	0.000 1

数设置如表 1 所示。

在仿真过程中，对不同算法均采用固定调制格式和相同的学习效率，得到 3 种算法收敛效果对比情况如图 2 所示。可以看出，在迭代次数达到约 50 次时，改进 DQN 算法开始收敛，而 DQN、QL 算法则分别在

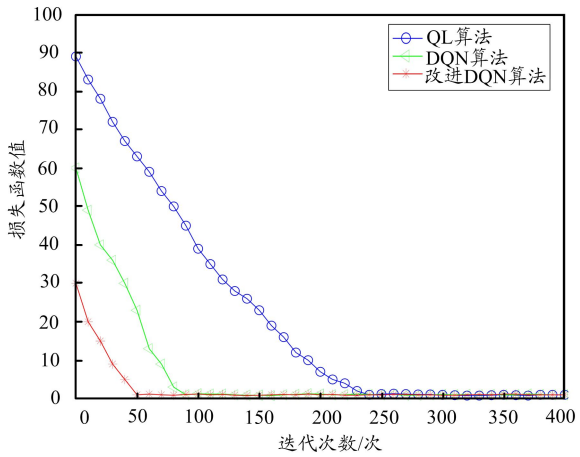

图 2 3 种算法收敛效果对比图

迭代次数约 100、250 次时开始收敛，这表明本文提出的改进 DQN 算法收敛速度更快。

在算法模型迭代训练趋于稳定后，本文验证了不同算法所获得的频谱效率，如图 3 所示。可以看出，本文提出的改进 DQN 算法的频谱效率整体高于 QL、

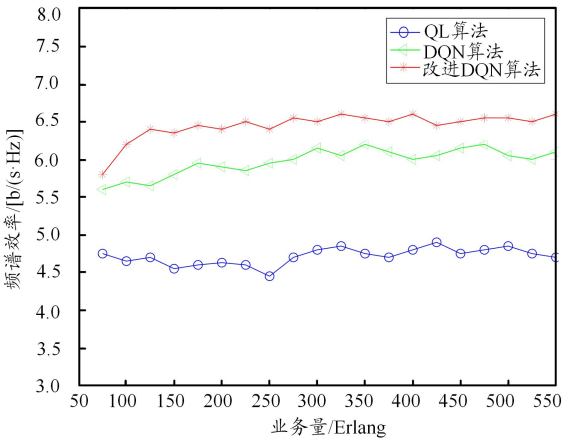

图 3 模型收敛后 3 种算法频谱效率的对比图

DQN 算法，这表明采用改进 DQN 算法的 EON 通信质量更有保障。

3 种算法在不同业务量下的频谱资源利用率、阻塞率和平均访问时延分别如图 4、图 5、图 6 所示。可以看出，当业务量在 300 Erlang 时，DQN 算法比 QL 算法频谱资源利用率提高了 11.03%，阻塞率降低了 5.64%，平均访问时延减少了 1.52 ms；改进 DQN 算法比 DQN 算法频谱资源利用率提高了 10.09%，阻塞率

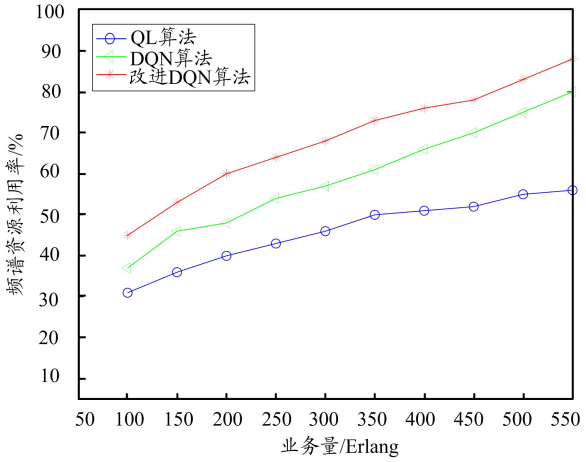

图 4 频谱资源利用率

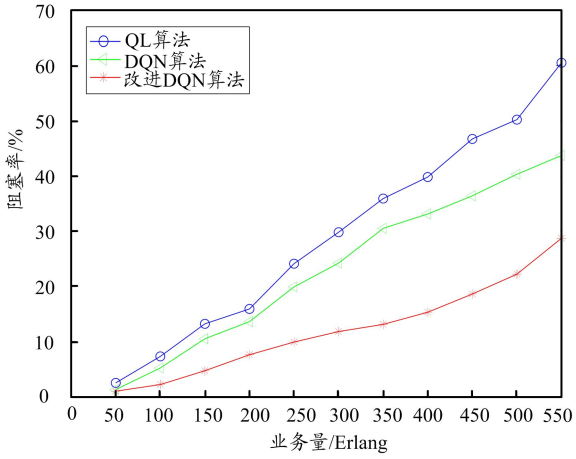

图 5 阻塞率

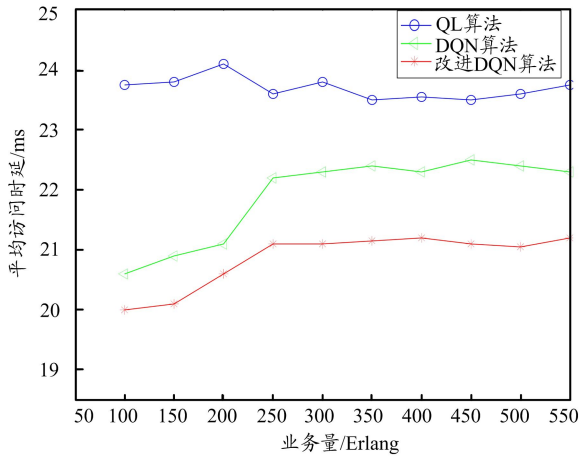

图 6 平均访问时延

降低了 12.41%，平均访问时延减少了 1.27 ms。上述结果表明，在 3 种算法中，改进 DQN 算法的频谱资源利用率最高，阻塞率和平均访问时延最低，改善了网络性能。

3 结束语

本文提出了一种改进的 DQN 算法，并获得了实时调整的 ε 值。该值有效地改善了 DQN 算法中的随机性和误差性，使其动作选择更加合理，并提高了算法的迭代速度，得到了最优的频谱资源分配模型。最后对不同算法在相同仿真环境下进行了收敛效果、频谱资源利用率、阻塞率和平均访问时延的仿真对比。仿真结果表明：本文提出的改进 DQN 算法不仅能够使训练模型快速收敛，同时在解决 RSA 问题时能更好地提升频谱资源利用率，并进一步降低了阻塞率，为下一代 EON 发展提供了思路。

参考文献：

[1] 李新伟，王敬勇. 互联网发展技术溢出与区域创新能力[J]. 科技管理研究，2020，40(22)：7–14.

[2] 姬卫玲. 波分复用技术的应用现状和发展前景[D]. 南京：南京邮电大学，2015.

[3] 张盛峰，何阿成，石鹏涛，等. EON 中资源节约型组播路由和频谱分配策略[J]. 光通信技术, 2018, 42(3)：52–55.

[4] 包博文，李娜娜，胡劲华，等. EON 中距离自适应调制技术综述[J]. 光通信技术，2018，42(9)：14–17.

[5] SHARMA D, KUMAR S. Evaluation of network blocking probability and network utilization efficiency on proposed elastic optical network using RSA algorithms[J]. Journal of Optical Communications, 2020, 41(4) : 403–409.

[6] 崔思恒. 光网络中基于强化学习的动态资源分配技术研究[D]. 北京：北京邮电大学，2021.

[7] 季晨阳，毕美华，周钊，等. 基于 DQN 的多租户 PON 在线带宽资源分配算法[J]. 光通信技术，2021，45(9)：36–39.

[8] 于佼良. 基于深度学习的 EON 路由和频谱分配策略的研究 [D]. 南京：南京邮电大学，2020.

[9] MUHAMMAD A, ZERVAS G, FORCHHEIMER R. Resource allocation for space-division multiplexing: optical white box versus optical black box networking[J]. Journal of Lightwave Technology, 2015, 33(23): 4928–4941.

[10] 尚晓凯，翟慧鹏，韩龙龙. 基于 DQN 的光网络资源分配方法研究[J]. 电子技术与软件工程，2022，(9)：1–4.

[11] 袁银龙. DQN 算法及应用研究[D]. 广州：华南理工大学，2019.

[12] GOSAVI A, PARULEKAR A. Solving Markov decision processes with downside risk adjustment [J]. International Journal of Automation and Computing, 2016,13(3): 235–245.

[13] RAZMINIA A, ASADIZADEHSHIRAZ M, TORRES D. Fractional order version of the Hamilton-Jacobi-Bellman equation[J]. Journal of Computational and Nonlinear Dynamics, 2019(1): 14–18.

[14] HUANG T, ZHANG R X, ZHOU C, et al. QARC: video quality aware rate control for real-time video streaming via deep reinforcement learning [C]//ACM. Proceedings of 2018 ACM Multimedia Conference. New York: ACM, 2018: 1208–1216.

[15] RIEDMILLER M. Neural fitted Q iteration-first experiences with a data efficient neural reinforcement learning method[C]// European Conference on Machine Learning, October 3–7, 2005, Berlin, Germany. Heidelberg: Spinger, 2005: 317–328.