基于强化学习的光传送网路由波长优化

孔英会[1,2], 杨佳治[1*], 高会生[1,2], 胡正伟[1,2]

(1. 华北电力大学 电子与通信工程系, 河北 保定 071003;
2. 华北电力大学 河北省电力物联网技术重点实验室, 河北 保定 071003)

摘 要: 针对光传送网中动态业务的路由和波长问题,提出一种基于强化学习的深度路由波长分配算法 DeepRWA。算法基于软件定义网络架构,通过强化学习灵活地调整控制光传送网,实现光网络路由波长分配策略优化。针对路由选择问题,结合链路上的波长使用情况,使用 A3C 算法选择合适的路由,使得阻塞率最小;针对波长分配问题,使用首次命中算法选择波长。考虑阻塞率、资源利用率、策略熵、价值损失、运行时间及收敛速度等多个指标,利用 14 节点 NSFNET 网络拓扑仿真实验。结果表明:当信道中包含 18 个波长时,与传统 KSP-FF 算法相比,所提出的路由波长分配算法的阻塞率降低了 0.06,资源利用率提高了 0.02,但运行时间有增加;在波长数超过 45 以后,与传统 KSP-FF 算法相比,所提算法保持阻塞率和资源利用率的同时,运行时间开始降低;当信道中包含波长数为 58 时,与传统 KSP-FF 算法相比,所提算法运行时间减少了 0.07 ms。由此可见,提出的算法使路由选择和波长分配得到了优化。

关键词: 光传送网; 路由波长优化; 强化学习; 路由选择; 波长分配

中图分类号: TN929.1 **文献标志码:** A **DOI:** 10.3788/IRLA20220084

Optimization of routing and wavelength optimization algorithm for optical transport network based on reinforcement learning

Kong Yinghui[1,2], Yang Jiazhi[1*], Gao Huisheng[1,2], Hu Zhengwei[1,2]

(1. Department of Electronic and Communication Engineering, North China Electric Power University, Baoding 071003, China;
2. Hebei Key Laboratory of Power Internet of Things Technology, North China Electric Power University, Baoding 071003, China)

Abstract: Aiming at the routing and wavelength problems of dynamic services in optical transport network, a deep routing wavelength assignment algorithm based on reinforcement learning is proposed. The algorithm is based on a software defined network architecture, flexibly adjusts and controls the optical transport network through reinforcement learning, and realizes the optimization of the optical network routing wavelength assignment strategy. For the problem of routing selection, combined with the wavelength usage on the link, the A3C algorithm is used to select the appropriate route to minimize the blocking rate; for the problem of wavelength assignment, the first fit algorithm is used to select the wavelength. Considering multiple indicators such as blocking rate, resource utilization, policy entropy, value loss, execution time, and speed of algorithm convergence, the 14-node NSFNET network topology simulation experiment is implemented. The results show that when the channel contains 18 wavelengths, compared with the traditional KSP-FF algorithm, the blocking rate of this routing wavelength assignment algorithm is reduced by 0.06, and the resource utilization rate is

收稿日期: 2022−02−07; **修订日期:** 2022−07−23

基金项目: 国家自然科学基金 (52177083); 河北省省级科技计划 (SZX2020034); 国网山西省电力公司科学技术项目 (SGSXXT00JFJS2100106)

作者简介: 孔英会, 女, 教授, 博士, 主要从事机器学习、通信系统与通信网技术方面的教学和研究工作。

通讯作者: 杨佳治, 男, 硕士生, 主要从事光网络方面的研究。

increased by 0.02, but the execution time is increased. When the number of wavelengths exceeds 45, compared with KSP-FF, the proposed algorithm maintains the blocking rate and resource utilization, while the execution time begins to decrease. When the number of wavelengths is 58, compared with KSP-FF, the proposed algorithm's execution time is reduced by 0.07 ms. It can be seen that the proposed algorithm optimizes the routing and wavelength assignment.

Key words: optical transport network; routing and wavelength optimization; reinforcement learning; routing selection; wavelength assignment

0 引 言

网络业务的激增对骨干网传输带宽提出了更高的要求。如何在有限资源网络中为业务选择合适的路由和分配优化的波长对于提升网络资源的利用效果、优化管理和灵活控制都有较大的影响。所以路由与波长分配 (Routing and Wavelength Assignment, RWA) 成为光传送网中的核心问题之一[1−3]。

RWA 问题一般被分为路由问题和波长两个方面, 首先选取合适的路由作为链路, 然后为链路分配波长[4−6]。常用的路由算法有最短路径算法 (Shortest Pathes, SP)、K 条最短路径算法 (K Shortest Pathes, KSP); 其中 SP 根据源节点-目的节点计算最短路径, 当业务请求到来选择最短路径路由。这种方法计算复杂度低, 但会导致网络阻塞率高。KSP 是在 SP 的基础上, 在源节点-目的节点计算 K 条路径并且按照距离排序。当业务到达时, 可按照优先级顺序选择可用路由。常用波长分配算法有随机分配 (Random Assignment, RA)、首次命中 (First Fit, FF) 等。其中 RA 是在可用波长的集合随机选择一个波长传输资源, 该算法实现简单, 被使用波长的随机性较大。而 FF 按照优先级搜索可用波长, 使用首次找到可用的波长传输信息。FF 计算开销较小、阻塞率较低。参考文献 [7] 对解决 RWA 问题波长分配的常用方法做了对比实验, 仿真表明波长分配算法对 RWA 问题的解决影响较小, 因此 FF 是光网络中目前应用较多的典型算法。

由于目前的云计算、数据互联等新业务呈现动态特性, 上述路由方法由于缺乏灵活性不再适用, 需要根据业务特性快速按需部署网络资源, 并借助智能算法为光网络的路由选择提供灵活的优化管理与控制方案, 软件定义网络 (Software Defined Network, SDN) 与深度强化学习的思想可以为上述方案实施提供支持。

近几年, 基于机器学习的光网络路由和波长智能分配算法引起了学者的广泛关注。参考文献 [7] 提出一种遗传算法解决光网络中 RWA 问题, 较传统算法具有更低的的网络阻塞率。参考文献 [8] 提出一种蚁群 RWA 算法, 实现了卫星光网络的负载均衡, 但是所提出的蚁群算法易陷入局部最优, 收敛速度较慢。参考文献 [9]考虑卫星光网络传输延迟和波长连续性限制, 提出改进蚁群算法的 RWA 方法, 降低了计算复杂度, 但是也增加了阻塞率。参考文献 [10] 使用监督学习的机器学习方法解决 RWA 问题, 将 RWA 问题映射为分类问题, 该算法大大减少了生成 RWA 策略的计算时间, 但训练数据集难以得到。

自 2015 年开始, 深度强化学习已用于解决通信网络的优化问题[11]。参考文献 [12] 中, 作者针对空中移动无线通信节点与水面通信节点信息交互的应用场景, 提出基于深度强化学习的方法引导空中无线通信节点移动路径, 实现最小代价水面通信节点覆盖。参考文献 [13] 中, 作者针对通信组网对抗中干扰资源分配的优化问题, 提出了一种基于最大策略熵深度强化学习的干扰资源分配方法。参考文献 [14] 中, 作者提出一种基于深度强化学习算法用于解决光传送网中路由、调制、波长和端口分配问题, 目的是降低成本。但是该算法处理复杂的拓扑时, 由于可用的路由路径较多, 将导致模型训练时间较长。在参考文献 [15] 中, 作者提出一种基于深度强化学习的路由和资源分配算法, 该算法可以选择跳数最小的路径和数量最少的波长转换器, 降低光网络运维成本, 简单实验拓扑验证算法可以达到预期目标, 但实验所需的网络拓扑及评价指标需要进一步扩展。在参考文献 [16] 中, 作者针对弹性光网络提出一种基于深度强化学习的路由-模式-频谱分配算法, 根据业务需求动态分配带宽, 进而提高频谱利用率, 有效降低弹性光网络中业务请求的阻塞概率, 对光传送网中 RWA 问题的解决有一定借鉴意义。

针对光网络中的路由选择和波长分配问题，借鉴弹性光网络中频谱优化的思想，提出一种深度路由波长分配算法 (Deep Routing and Wavelength Assignment, DeepRWA)，该算法采用 SDN 框架灵活控制光传送网络的路由选择和波长分配，基于深度强化学习策略实现 RWA 的智能化处理。深度强化学习使用异步优势行动-评论算法 (Asynchronous Advantage Actor-Critic, A3C) 算法并考虑波长使用情况选择路由；在此基础上使用 FF 实现波长分配，使路由阻塞率最小，提升资源利用率。

1 光传送网系统模型及 RWA 问题描述

光传送网由光分叉复用器 (Optical Add/Drop Multiplexers, OADM) 和链路构成。光传送网连接模型可使用图结构 G(V, E, F) 表示，如图 1 所示，其中 V 和 E 表示拓扑的节点 (主要是光分插复用器 OADM) 和链路，$F=\{F_{e,f} \mid e,f\}$表示某条光纤链路 e 中的波长 f 使用情况。空闲波长分布情况用 $f=\left\{z^{1,j}_k, z^{2,j}_k\right\}$表示，$z^{1,j}_k$表示第 k 条路由第 j 段波长可用波长的数量；$z^{2,j}_k$表示第 k 条路由第 j 段波长第一个可用波长的索引。当光网络处于工作状态时，业务由主机发出，服务器对服务做出响应。把一个从源节点到目的节点的光路请求设为 $R_t(o, d, \tau)$，其中 o 和 d 表示光路请求的源节点和目的节点，τ 表示光路业务请求的持续时间。

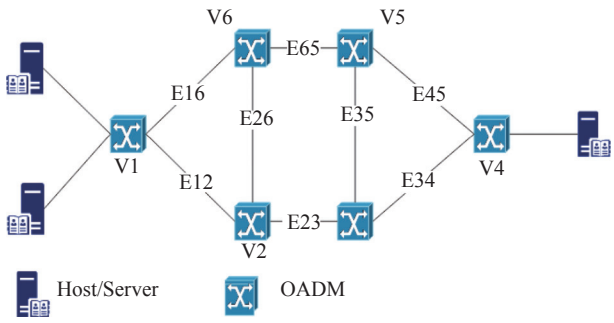


图 1 光传送网模型

Fig.1 Optical transport network model

RWA 包含路由选择和波长分配两个子问题，路由选择是在业务的源节点、目的节点之间选择一条合适的物理路径进行数据传输，以最小化阻塞率为目标；波长分配就是路由选择结果确定之后为业务请求分配波长，主要以提高波长利用率为目标，即提高已用波长数在总波长数的比例。在波长分配阶段，需要遵循两个原则：(1) 波长一致性限制，连接源-目的节点的光通道必须分配相同的波长。(2) 不同波长限制，所有使用同一条光纤链路的光通道都必须分配不同的波长。

根据应用场景的不同，RWA 工作方式主要分为两种：静态 RWA 与动态 RWA。静态 RWA 主要解决固定的业务请求下如何最小化网络成本和网络资源使用 (例如所需波长数最少) 的问题，即提高资源利用率。动态 RWA 主要解决在业务请求随机到达的情况下如何最小化业务的阻塞率，并提高网络资源利用率的问题。目前光传送网中动态业务较多，文中针对动态 RWA 问题进行研究。假设光网络拓扑的节点不具备把光信号的波长转换为一个不同的波长，即波长转换功能，因此光路请求需要服从波长一致性限制。

2 DeepRWA 算法设计与实现

2.1 基于 SDN 的 DeepRWA 总体结构设计

文中提出的基于强化学习的深度路由波长分配算法 DeepRWA 总体结构如图 2 所示，算法采用 SDN 架构，由光网络模块、控制器、强化学习模块三部分组成。SDN 本地控制器作为核心，首先根据数据平面获得网络状态和光路请求，利用深度强化学习算法生成 RWA 策略然后下发至数据平面。光网络模块主要由光分叉复用器作为节点，节点间相互连接形成拓扑；SDN 控制器是光网络模块和强化学习模块的桥梁，可以感知光网络信息、波长使用情况、业务请求信息，为光网络提供高效的路由与波长分配策略；强化学习模块利用控制器收集的信息进行训练，把生成包括路由路径和选用波长的动作发送至控制器模块。DeepRWA 工作过程是通过 SDN 智能控制光传送网的数据平面，DeepRWA 具体实现步骤如下：

(1) SDN 本地代理首先收到主机发出的光路请求 R_t；

(2) SDN 控制器获得网络拓扑信息和光路请求信息，调用特征处理模型产生 DeepRWA 的状态信息，其中包括可选路由和每条路由的波长使用信息；

(3) DeepRWA 强化学习模块通过神经网络读取状态 s_t，利用 A3C 算法迭代训练，将神经网络的输出作为 RWA 策略发送至 SDN 控制器；

(4) 控制器将 RWA 策略下发至本地代理，由本地

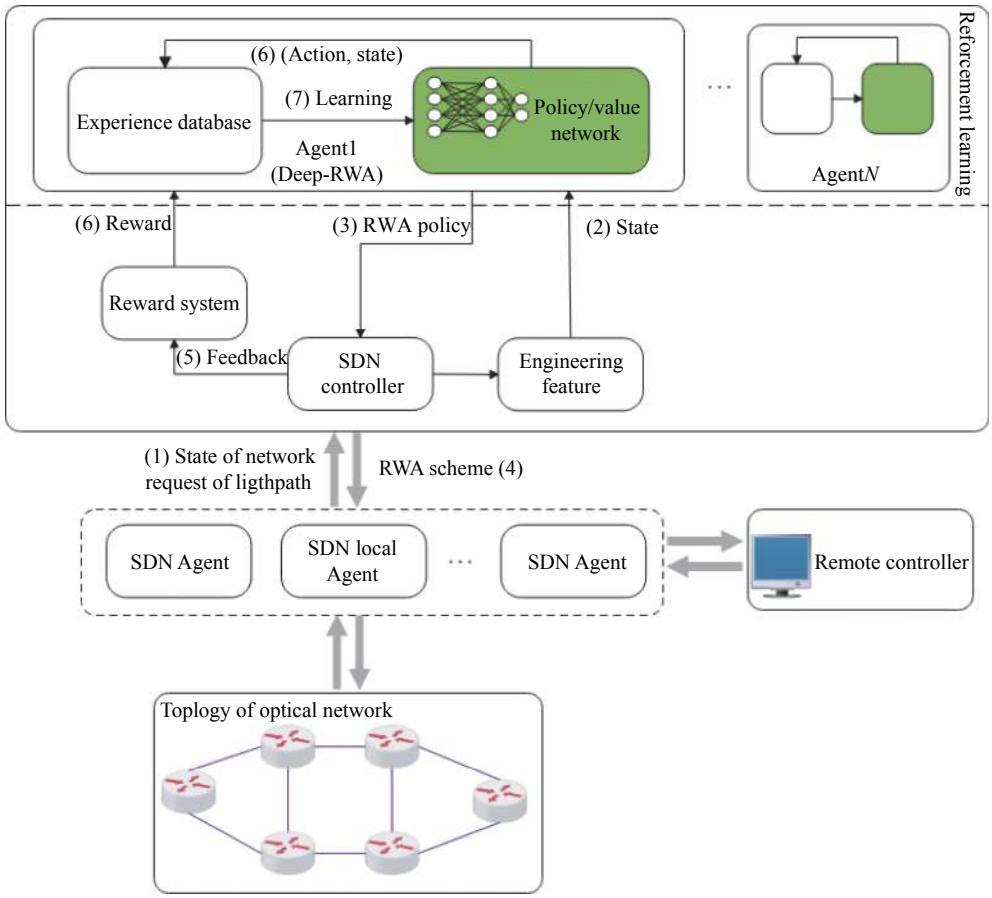


图 2 DeepRWA 结构图

Fig.2 Structure of DeepRWA

代理执行相应的路由和波长分配策略，即 DeepRWA 的动作；

(5) 奖励系统将获得以往的历史奖励作为反馈和生成 DeepRWA 的即时奖励 r_t；

(6) 存储即时奖励、动作和状态三个元素数据至经验池；

(7) 利用经验池的数据训练策略神经网络和价值神经网络，更新神经网络的参数，利用策略神经网络获取 RWA 策略。RWA 策略是选择策略神经网络中概率最大的动作。

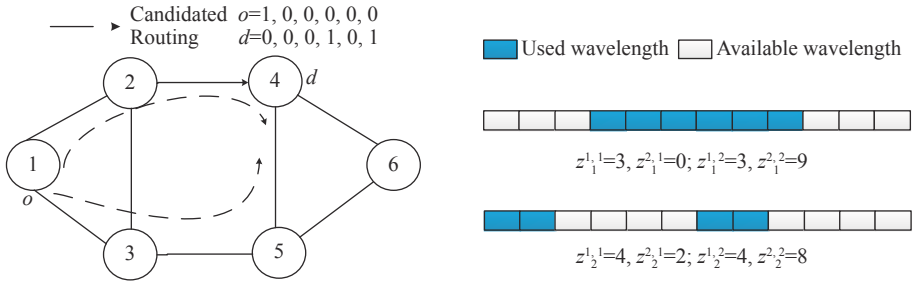
2.2 强化学习模块设计

文中提出 DeepRWA 的核心是基于强化学习的路由波长优化算法，而强化学习关键是设计状态、动作和奖励，文中具体设计如下：

(1) 状态：针对拓扑中的单业务请求，使用特定的矩阵表示状态 s_t 信息和候选路径的波长使用情况，文中状态设定如图 3 所示，其中图 3(a) 为光网络拓扑；图 3(b) 为光网络链路的波长使用示意图，定义状态为 1×(2 |V|+1+2JK) 大小的矩阵，如公式 (1) 所示：

$$s_t=\left\{o,d,\tau,\left\{z^{1,j}_k,z^{2,j}_k\right\}\right\} \tag{1}$$

式中：o、d 和 τ 表示业务请求的源节点、目的节点和服务时长；$k\in[1,2,\cdots,K]$，$j\in[1,2,\cdots,J]$，K 为业务请求的所有候选路由数 (如图 3(a) 虚线部分)，J 表示某条路由所有可用波长段 (单个可用波长段至少包含一个可用波长且波长段内的所有波长位置连续，如图 3(b) 中白色部分)，V 表示拓扑的节点数。候选路由数目的不同会导致状态的不同，因此需采用一个固定的候选路由数目。强化学习可以通过探索多条候选路由获得更大的收益，然而收敛时间也会更长。候选路由数太少，强化学习的收益较小；候选路由数太多，算法收敛速度慢，文中参考文献 [15] 的设置，即 K=5。$\left\{z^{1,j}_k,z^{2,j}_k\right\}$表示业务请求某条候选路由的未占用波长分布情况，$z^{1,j}_k$,表示第 k 条路由第 j 段波长可用波



(a) 光网络拓扑
(a) Optical network topology

(b) 波长使用情形
(b) Example of wavelength usage

图 3 状态示意图

Fig.3 Example of state

长的数量，$z^{2,j}_k$,表示第 k 条路由第 j 段波长第一个可用波长的索引。具体实例见图 3(a) 中，假设业务有两条路由可选，1-2-4 为路由 1 (k=1)，1-3-5-4 为路由 2 (k=2)。图 3(b) 中两条路由 1、2 的波长使用情形分上下两部分，上方路由 1 的第 1 段波长可用波长数量$z^{1,1}_1$ = 3，路由 1 的第 1 段波长可用波长集合的起始索引 $z^{2,1}_1 = 0$；路由 1 的第 2 段波长可用波长数量$z^{1,2}_1 = 3$，路由 1 的第 2 段波长可用波长集合的起始索引$z^{2,2}_1 = 9$；下方路由 2 第 1 段波长的可用波长数量$z^{1,1}_2 = 4$，路由 2 第 1 段波长的可用波长集合的起始索引$z^{2,1}_2 = 2$；路由 2 第 2 段波长的可用波长数量$z^{1,2}_2 = 4$，路由 2 第 2 段波长的可用波长集合的起始索引$z^{2,2}_2 = 8$。因此图 3 的状态表示为：$S_t = \left\{o, d, \tau, z^{1,1}_1 \sim z^{2,2}_1, \quad z^{1,1}_2 \sim z^{2,2}_2\right\}$。

(2) 动作：动作空间需要对光网络中业务的路由和波长进行部署，每一个动作就是代表一种路由与波长分配方案。对于一个业务请求，DeepRWA 算法从光网路拓扑的 K 条候选路径选择一个路由路径，根据状态计算可选波长数 I，从 I 个可选波长选择一个可用波长传输数据，因此动作空间包含 K*I 个动作。

(3) 奖励：奖励的设定对算法的收敛速度和动作决策都有很大影响。由于动态 RWA 问题的优化目标是最小化阻塞率，所以用业务的成功率衡量该动作的价值。即时奖励作为一个标记，记录当前业务请求是否成功。DeepRWA 算法如果成功地使一个业务通过光网络传输，即路由与波长分配成功，此时即时奖励 r_t=1，否则 r_t=−1。即时奖励为 1 时，表明代理在状态 S 下执行动作 a 策略会更好，有助于策略的收敛；即时奖励为−1 时，表明代理在状态 S 下执行动作 a 策略比较差，不利于策略的收敛。DeepRWA 中神经网络不断迭代使累积收益 Γ 最大化，累积收益如公式 (2) 所示：

$$\Gamma_t = \sum_{t' \in [t,\infty]} \gamma^{t'-t} \cdot r_{t'} \tag{2}$$

式中：$\gamma \in [0,1]$ 是未来收益的折扣因子；t'表示未来某个时刻；t 表示当前时刻。

3 仿真实验和结果分析

文中从阻塞率、资源利用率、策略熵、价值损失和运行时间及收敛速度五个方面评价 DeepRWA 算法的性能。并与现有的算法 KSP-FF 对比，验证所提出算法的有效性，借鉴参考文献 [16] 的参数设置，KSP-FF 算法的备选路径设为 5 条，即 K=5，下面实验中 KSP-FF 参数均采用上述设置。

3.1 实验环境和参数设置

仿真实验采用如图 4 所示的 14 节点 NSFNET 拓扑，链路间的数字代表节点间的距离 (单位：km)。实验采用动态的业务模型，业务服从泊松分布，服务时

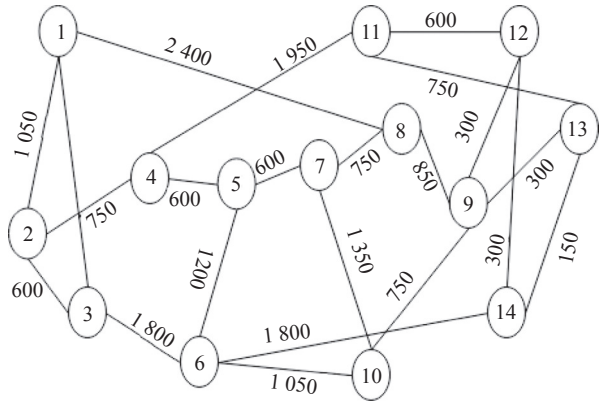


图 4 14 节点¬NSFNET 拓扑

Fig.4 14-node ¬NSFNET topology

间遵循指数分布，具体参数如表 1 所示，软件配置操作系统选用 Ubuntu16.04，GPU 为 NVIDIAGeforce 1080 Ti，CPU 为 i7-6700 k，程序设计采用 python3.7，tensorflow 版本为 1.14.0。

表 1 仿真实验业务参数

Tab.1 Serve parameters of simulation experiment

Parameters	Value
Average arrival time of dynamic service/s	1/12
Continuing times of dynamic service/s	13
Available wavelength of channel	18

3.2 实验结果与分析

(1) 阻塞率

动态 RWA 问题的目标是最小化阻塞率，对比了 DeepRWA 算法和 KSP-FF 算法阻塞率指标随业务数变化情况。图 5 给出了上述两种算法随着业务数增加的变化曲线。从图 5 可以看出，业务请求数较小时 DeepRWA 阻塞率较大，为 0.16 左右。这是因为业务请求数较少时，DeepRWA 训练次数较少，没有学习到理想的路由与波长分配策略。当业务请求数达到 370000 个后，经训练的 DeepRWA 路由和波长分配方案的阻塞率降到 0.01，比 KSP-FF 算法的阻塞概率近似低 0.06。原因是 KSP-FF 寻找路由只考虑最短路径，遇到无可用波长的链路就会发生阻塞；而 DeepRWA 在选路过程中考虑路径的跳数及波长占用情况，综合考虑选择一条路径，尽可能避开了拥塞链路。因此，DeepRWA 较 KSP-FF 算法具有更低的 (业务) 阻塞率。

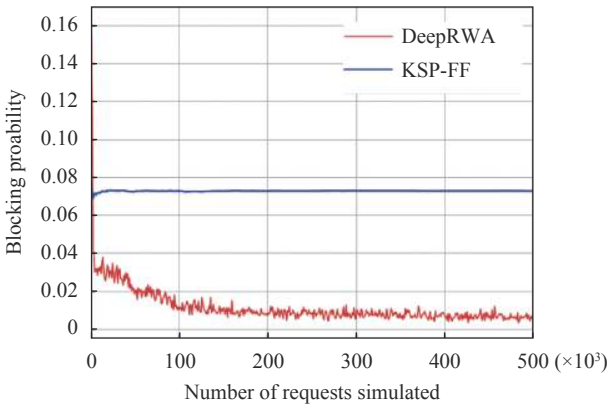


图 5 阻塞率变化曲线

Fig.5 Curve of blocking rate

(2) 资源利用率

资源利用率表示被业务利用的资源占总资源的比例，资源利用率越大，代表网络中被充分利用，算法性能会更好。图 6 比较了 DeepRWA 算法和 KSP-FF 算法资源利用率随业务数变化的性能曲线。从图 6 可以看出，DeepRWA 算法资源利用率在训练过程中逐步提高最后趋于稳定。起初资源利用率较低且波动较大，因为随着业务请求数增多，要想服务这些业务，就要不断地分配波长，网络的资源利用率必然呈上升趋势。在训练过程中，业务数较小时资源利用率为 0.51，当业务数达到 150000 后，DeepRWA 资源利用率上升至 0.59，表 1 中设定的 18 个波长平均有 10.6 个波长被利用；而在 KSP-FF 算法下，设定的 18 个波长中平均有 10.1 个波长被利用。可见 DeepRWA 的路由和波长策略使光网络中资源利用率超过了 KSP-FF 算法。图 7 表示实验环境的其它参数不变，资源

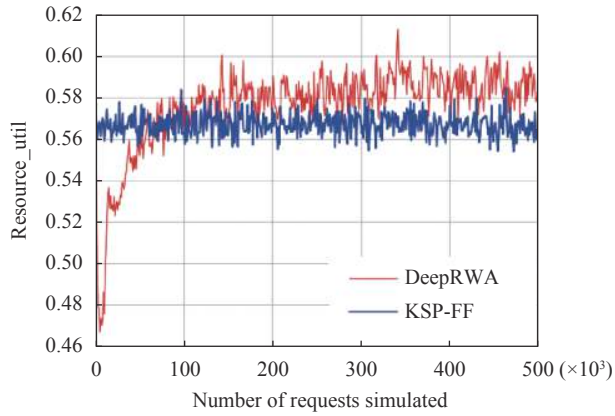


图 6 资源占用率曲线

Fig.6 Curve of resource occupancy

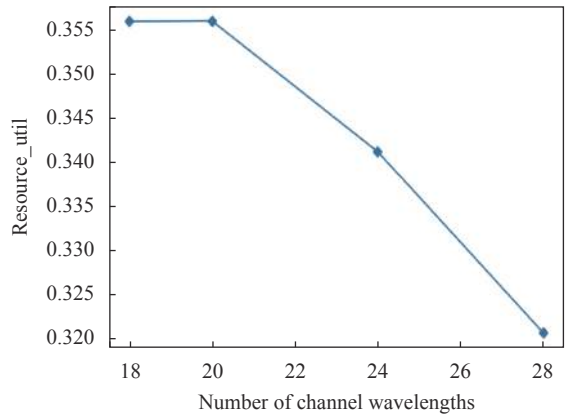


图 7 资源利用率随信道波长变化关系

Fig.7 Curve of resource utilization rate vs. channel wavelength

利用率随着信道波长数变化关系曲线图。随着信道波长数增加，资源利用率逐渐降低。由于业务强度一定，当信道波长数增加时，链路间空闲信道明显增加。根据资源利用率的定义，它会逐渐降低，使网络拥塞问题有所改善。

(3) 策略熵

图 8 描绘了 DeepRWA 算法随着业务强度增加策略熵的变化曲线，熵值的变化可以反映所提出的算法是否进行了有效的学习。熵值越大说明算法得到的路由和波长策略不确定性越强；策略熵收敛表明算法的随机性降到最低，模型已经训练完成。当业务强度较小时，熵值较大，生成的路由与波长决策随机性较大；随着业务请求数量的增加，当该算法通过 3 000 000 个业务请求后，策略熵收敛，熵值从 1.5 下降至 0.15，说明 DeepRWA 已经学习到一些规则并可以做出随机性相对较小的路由与波长决策。

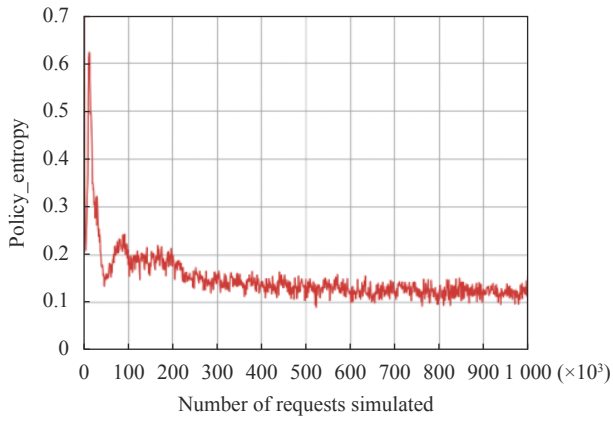


图 8 策略熵变化曲线图

Fig.8 Curve of policy entropy

(4) 价值损失

价值损失函数用于计算价值神经网络预估值和真实值的差距，从而指导下一步的训练向正确的方向进行。价值损失描述的是累积收益与价值函数预测值的接近程度，用于评价算法的优劣。图 9 描述了价值损失随着训练过程的变化曲线，训练过程中价值损失逐渐降低，训练逐渐稳定。当业务数较少时，选择的路由与波长策略的预测值与累积收益差距较大。随着业务数的增加，训练后算法的预测值逐渐接近于累积收益，当通过 450 000 个业务请求后训练算法的价值损失减少至 2.5 左右，预测值和累积收益比较接近，表明采取的路由与波长策略比较合适。

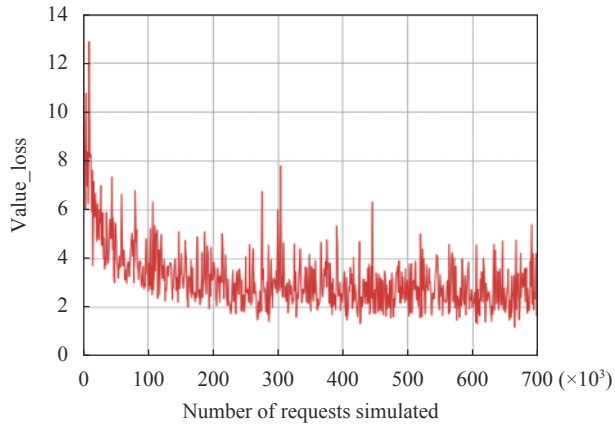


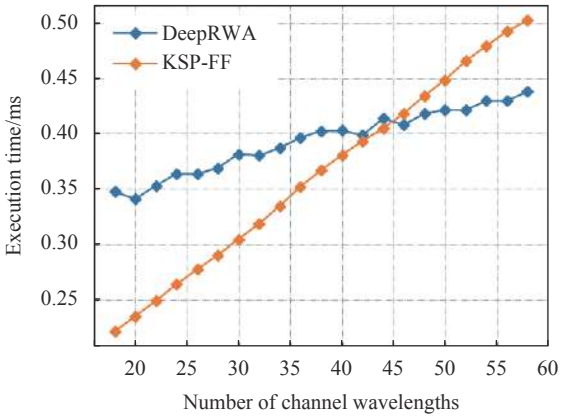
图 9 价值损失曲线

Fig.9 Curve of value Loss

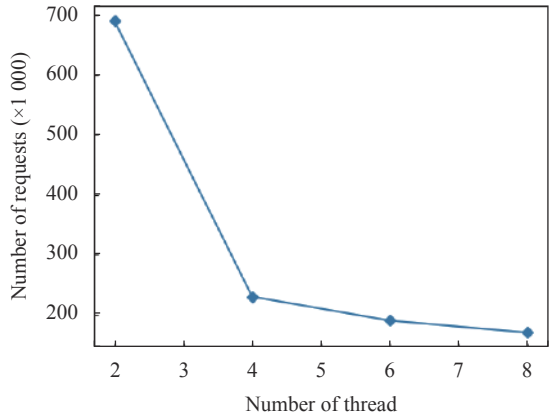
(5) 运行时间及收敛速度

运行时间是衡量算法时间复杂度的评价指标，运行时间越短表明算法的时间复杂度越小、适应性越好。针对单个业务请求，在业务服从泊松分布、到达率和服务时长的设置同表 1 时，对比了 DeepRWA 算法和 KSP-FF 算法运行时间随波长数变化情况，如图 10(a) 所示。由图 10(a) 可知，当信道波长数增加时，两种算法运行时间均明显增加。因为当信道波长数增加时，路由与波长分配策略复杂度增加。但是，随波长数增加，DeepRWA 算法运行时间增长的速度明显慢于 KSP-FF。当信道中包含 18 个波长时，与 KSP-FF 算法相比，DeepRWA 算法运行时间增加了 0.12 ms；当信道中包含的波长数增加至 45 时，两种算法的运行时间大致相同；之后，DeepRWA 算法运时间优势逐渐明显，当信道中包含的波长数增加至 58 时，与 KSP-FF 算法相比，DeepRWA 算法运行时间减少了 0.07 ms。因此，DeepRWA 较 KSP-FF 算法更适应于波长数较多的光网络。

算法的收敛速度反应了智能体使网络达到最低阻塞率需要的训练时长，也即算法收敛需要的业务数量。使用多线程训练神经网络，线程数的不同也会影响网络收敛速度。文中采用 A3C 算法，业务服从泊松分布，到达率、服务时长和信道波长数的设置同表 1，仿真得到算法收敛所需业务数随线程数变化曲线如图 10(b) 所示，可以看出当线程数增加时，算法训练至收敛需要的业务数呈指数式减少。当线程数为 2 时，算法需要 690 000 个业务进行训练；当线程数增加至 8 时，算法收敛所需的业务数已减少至 190 000。由此



(a) 运行时间与波长数变化关系
(a) Execution time vs channel wavelengths



(b) 业务数与线程数变化关系
(b) Requests vs thread

图 10 算法的运行时间及收敛速度

Fig.10 Algorithm's execution time and converage speed

可知，在 CPU/GPU 支持的最大线程范围内，增加线程数可以明显减少算法的训练时间。

4 结 论

为了适应大量动态业务的需求，针对光网络中的 RWA 问题进行研究，考虑阻塞率、资源利用率两个目标提出一种基于强化学习的路由波长分配算法 DeepRWA，采用 SDN 网络架构实现光网络灵活控制，通过强化学习实现路由选择和波长分配的优化。针对路由选择问题，结合链路上的波长使用情况，使用 A3C 算法选择合适的路由，使得阻塞率最小；针对波长分配问题，使用首次选中算法选择波长。利用 NSFNET 网络拓扑下进行了仿真实验，结果表明文中所提的 DeepRWA 算法阻塞率更低，改善了资源利用率，提升了网络的性能；当链路波长数较多时，与 KSP-FF 算法相比，文中所提 DeepRWA 算法运行时间更短，适应性更好。后续结合实际网络和业务进行进一步的研究和测试，为实际应用提供有力的支持。

参考文献：

[1] Nath P K, Venkatesh T. Lightpath routing and wavelength assignment for static demand in translucent optical networks [J]. ***Photonic Network Communications***, 2020, 39(7): 103-119.

[2] Yang Xiuqing, Chen Haiyan. Application of optical communication technique in the internet of things [J]. ***Chinese Optics***, 2014, 7(6): 889-896. (in Chinese)

[3] Li Haitao. Technical approach analysis and development prospects of optical communication technology in China deep space TT&C network [J]. ***Infrared and Laser Engineering***, 2020, 49(5): 20201003. (in Chinese)

[4] Yang Junbo, Yang Jiankun, Li Xiujian, et al. Choice and control of routes in crossover optical interconnection network [J]. ***Optics and Precision Engineering***, 2010, 18(6): 1249-1257. (in Chinese)

[5] Sun Zhaowei, Liu Xuekui, Wu Xiande, et al. Path planning based on ant colony and genetic fusion algorithm for communication supporting spacecraft [J]. ***Optics and Precision Engineering***, 2013, 21(12): 3308-3316. (in Chinese)

[6] Guo Xiuzhen, Hou Lixin, Yin Zhaotai, et al. All-optical routing control based on coherently induced high reflection band and high transmission band in a medium of cold atoms [J]. ***Chinese Optics***, 2011, 4(4): 355-362. (in Chinese)

[7] Zhang Min, Xu Bo, Cai Yi, et al. Routing and wavelength assignment based on genetic algorithm in large scale WDM network [J]. ***Optical Communication Technology***, 2018, 42(11): 1-4. (in Chinese)

[8] Wang Weilong, Li Yongjun, Zhao Shanghong, et al. Routing and wavelength assignment based on load balance for optical satellite network [J]. ***Laser & Optoelectronics Progress***, 2021, 58(7): 0706004. (in Chinese)

[9] Shi Xiaodong, Li Yongjun, Zhao Shanghong, et al. Ant colony optimization routing and wavelength technology for software-defined satellite optical networks [J]. ***Infrared and Laser Engineering***, 2021, 51(7): 20200125. (in Chinese)

[10] Martín I, Troia S, Hernández J A, et al. Machine learning based routing -and wavelength assignment in software-defined optical networks [J]. ***IEEE Transactions on Network and Service Management***, 2019, 16(3): 871-883.

[11] Mnih V, Kavukcuoglu K, Silver D, et al. Human level control through deep reinforcement learning. [J]. ***Nature***, 2015, 518(7540): 529-533.

[12] Li Zhongtao. Wireless communication node coverage optimization based double deep Q-learning [J]. ***Electronic Technology & Software Engineering***, 2021(14): 1-3. (in Chinese)

[13] Rao Ning, Xu Hua, Qi Zisen, et al. Communication interference resource allocation method of deep reinforcement learning based on maximum policy entropy [J]. ***Journal of Northwestern Polytechnical University***, 2021, 39(5): 1077-1086. (in Chinese)

[14] Zhao Zipiao, Zhao Yongli, Ma Haoli, et al. Cost-efficient routing, modulation, wavelength and port assignment using reinforcement learning in optical transport networks [J]. ***Optical Fiber Technology***, 2021, 64: 102571.

[15] Li Xin, Zhao Yongli, Li Yajie, et al. Multi-objective routing and resource allocation based on reinforcement learning in optical transport networks[C]//2020 Asia Communications and Photonics Conference (ACP) and International Conference on Information Photonics and Optical Communications (IPOC), 2020: 1-3.

[16] Chen Xiaoliang, Li Baojia, Proietti Roberto, et al. DeepRMSA: A deep reinforcement learning framework for routing, modulation and spectrum assignment in elastic optical networks [J]. ***Journal of Lightwave Technology***, 2019, 37(16): 4155-4163.