强化学习 A3C 算法在电梯调度中的建模及应用

刘 宇[1,2]，张 聪[2+]，李 涛[3]

（1. 武汉大学 计算机学院，湖北 武汉 430000；2. 武汉轻工大学 数学与计算机学院，湖北 武汉 430000；3. 国网湖北省电力有限公司荆州供电公司 发展策划部，湖北 荆州 434000）

摘 要：为让电梯调度算法在电梯电力能耗、用户乘梯体验和算法适应性方面具备更好表现，在目前主流的电梯调度算法基础之上，提出对调度环境、电梯行为和调度目标 3 个方面进行统一建模的基于强化学习 A3C 的电梯智能调度算法。让调度电梯在不断地和环境交互学习过程中逐渐学习得到最优电梯调度策略，与基于具体环境建模的相关电梯调度算法进行对比实验，基于 A3C 的调度算法具有建模简单规范、适应性强和控制目标多样的优势，对比 A3C 算法与部分强化学习算法在电梯调度中的优劣，实验结果表明，A3C 算法具备较好的调度性能。

关键词：智能调度；电梯调度算法；电梯节能；强化学习；A3C

中图法分类号：TP391 **文献标识号**：A **文章编号**：1000-7024（2022）01-0196-07

doi：10.16208/j.issn1000-7024.2022.01.026

Modeling and application of reinforcement learning A3C in elevator scheduling algorithm

LIU Yu[1,2], ZHANG Cong[2+], LI Tao[3]

(1. School of Computer Science, Wuhan University, Wuhan 430000, China; 2. School of Mathematics and Computer Science, Wuhan Polytechnic University, Wuhan 430000, China; 3. Development Planning Department, Jingzhou Power Supply Company of State Grid Hubei Electric Power Limited Company, Jingzhou 434000, China)

Abstract: To make the elevator scheduling algorithm have better performance in the aspects of elevator power consumption, user experience and algorithm adaptability, based on some elevator scheduling algorithms, the elevator intelligent scheduling algorithm based on reinforcement learning A3C was constructed, which unified the three aspects of scheduling environment, elevator behavior and scheduling objectives. The elevator learned the optimal scheduling strategy in the process of continuous interactive learning with the environment. Compared with some elevator scheduling algorithms, the scheduling algorithm of reinforcement learning modeling has the advantages of simple modeling and high scheduling efficiency. At the same time, the experiment explores that A3C algorithm has better scheduling performance than some reinforcement learning algorithms.

Key words: intelligent scheduling; elevator scheduling algorithm; elevator energy saving; reinforcement learning; A3C

0 引 言

楼宇电梯调度[1]是一个复杂的过程，调度算法的设计复杂性一般取决于调度环境的复杂性，当调度电梯数量越多楼宇楼层越高时，从众多调度策略中选择最优调度策略这本质上类似一个 NP 完全问题。当前随着认知智能技术的发展，探索更加智能的调度算法，让调度算法更具多环境适应性、调度高效性和低能耗性成为新的研究热点。

目前电梯调度算法往往基于特定需求方面进行建

收稿日期：2020-07-29；**修订日期**：2021-07-16

基金项目：湖北省重大科技专项基金项目（2018ABA099）；国家自然科学基金面上基金项目（61272278）；湖北省自然科学基金重点基金项目（2015CFA061）；湖北省自然科学基金青年基金项目（2018CFB408）；2020 年国网湖北省电力科技基金项目（5215J0200012）

作者简介：刘宇（1994-），男，重庆人，博士研究生，CCF 学生会员，研究方向为人工智能技术及其运用；+通讯作者：张聪（1968-），男，上海人，博士，教授，研究方向为多媒体信息处理与网络通信；李涛（1986-），男，湖北荆州人，硕士，高级工程师，研究方向为电力系统及其自动化。E-mail：hb_wh_zc@163.com

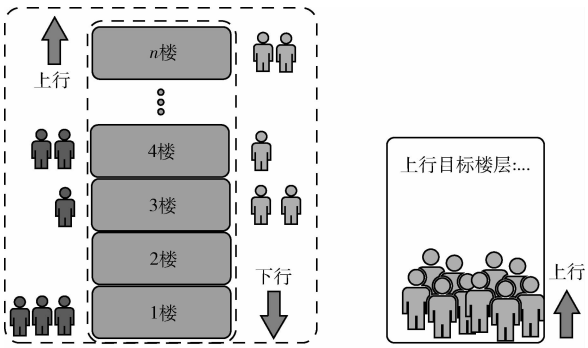
模，例如基于最短等待时间的算法（先来先服务算法、扫描算法和 LOOK 算法等），这类算法往往以满足各楼层用户需求为目的，但在多梯复杂环境下建模往往困难且缺乏适应性，特别是在人流高峰阶段的复杂环境之下，整个电梯群控效率较低[1]。同时基于算法进行最优解搜索也是研究点之一，如刘桂雄等通过引入统计方法提出了以节能优先的电梯调度算法[2]，仲惠琳根据客流量变化同时引入启发式算法提出了基于神经网络的电梯调度重规划算法[3]，郎曼等基于现有算法控制参数多和计算较复杂特性提出了采用人工蜂群算法的电梯群控系统[4]。刘清等针对传统的电梯群控系统的问题，提出一种模糊控制结合神经网络算法的电梯群控系统来进行优化调度[5]。刘剑等提出了基于 Fast R-CNN 的多轿厢电梯调度算法[6]。

电梯调度算法设计需要考虑用户体验、电梯能耗和空载情况等多个目标，上述文献的调度算法往往基于特定环境建模，存在很难平衡多个调度目标的问题。因此，本文提出了基于 A3C 的电梯调度算法，利用强化学习自主智能学习的优点，让调度算法在电梯和环境地不断交互中学习得到使调度目标收益最大化的调度策略[7]。同时对调度环境、电梯行为和调度目标 3 个方面进行统一建模，这极大方便模型在新调度环境和新目标下的低成本部署。本文基于开源 LiftSim 电梯调度仿真环境[8]进行了相关实验，实验结果表明 A3C 算法运用于群组电梯环境下进行电梯调度具有建模便捷、多环境移植适应性强、调度效率高和有效平衡多个调度目标的优势。

1　电梯调度环境

基于强化学习的电梯调度算法建模需要将不同的电梯调度环境归纳为统一的表达形式。近年来相关研究者针对电梯调度环境进行研究，开发了众多的电梯运行环境实例[8,9]，电梯调度环境可用图 1 进行表示。本文将电梯调度环境分为楼宇环境和电梯环境，楼宇环境主要反映楼宇基本信息和楼宇各楼层用户乘梯需求信息，其中楼宇信息包括楼层和电梯桥箱数等信息，电梯调度同一时间内乘梯需求主要包含上下行需求和用户所在楼层。楼宇环境作为电梯调度算法需要考虑的主要环境，是当前调度算法的主要考虑因素。电梯环境主要指具体电梯桥箱内环境，这往往包括电梯桥箱内用户的意图楼层、载重量、桥箱所在楼层和桥箱开关门情况等信息。

高层楼宇的发展让电梯调度算法面对的调度环境越来越复杂多样[10]，越来越多的楼宇具备更高楼层的同时也具备了更多的电梯数量，同时不同楼层不同时段的人流往往呈现较大波动。将电梯调度环境以楼宇环境和电梯环境进行表示，一方面能较完整反映瞬时电梯调度环境，另一方面为强化学习中环境状态建模提供依据。



(a) 楼宇环境　　(b) 电梯环境

图 1　电梯调度环境

2　强化学习

强化学习（reinforcement learning，RL）学习方式类似人类的学习模式，通过在实践环境中执行相应动作以期望获得环境给予的最大奖励值，从而依据奖励不断地进行试错学习来不断地修正自己在具体环境中的动作策略。强化学习典型学习模式如图 2 所示，图示中智能体在和环境交互过程中会根据环境状态 S_t 选择执行相应的动作策略 A_t，执行动作 A_t 后环境的状态值 S_t 会变为新的状态 S_{t+1}，同时会得到在环境 S_t 下执行动作 A_t 的奖励 R_t，然后智能体对于新的状态 S_{t+1} 又会选择执行相应的动作策略 A_{t+1}，以此往复不断地在和环境交互过程中进行学习以期获得更多奖励[11,12]。

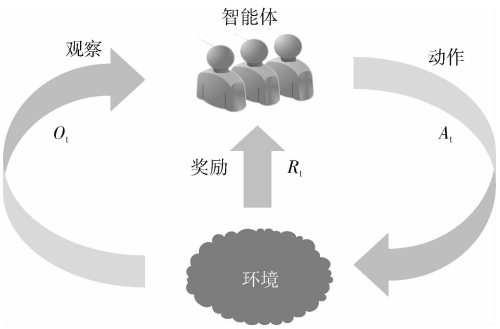


图 2　强化学习框架

强化学习模式和电梯调度模式二者建模之间存在一定的相似性，基于强化学习的电梯调度算法建模中可将电梯调度环境和电梯分别视为强化学习中环境状态（State）和智能体（Agent），同时电梯调度目标和电梯上下行动作可视为强化学习中的奖励函数（Reward）和动作策略（Action）。电梯调度算法中引入强化学习方法，让电梯智能体在不断地和环境交互中自主学会最优调度策略，这对目前设计电梯调度算法具有一定的实用价值。近年随着强化学习的发展也逐渐涌现出众多的强化学习算法[13,14]，例如 DQN 和 MADDPG 算法，探索强化学习算法在电梯调度领

域的运用潜力具有一定研究意义。

3 基于强化学习 A3C 的电梯调度算法建模

基于强化学习的电梯调度模型通过强化学习实现电梯的自主调度策略学习，这大大减少了不同调度环境目标下的建模成本，实现了电梯的智能高效调度。本文基于强化学习 A3C 的电梯调度算法建模如图 3 所示，图示中首先需要对调度环境、奖励函数和电梯动作行为 3 个方面进行定义并进行编码，然后基于 Actor 和 Critic 角色构建融合网络实现 A3C 算法的自主策略输出以及模型参数更新，最后通过多线程异步训练的方式对模型进行训练直至收敛。

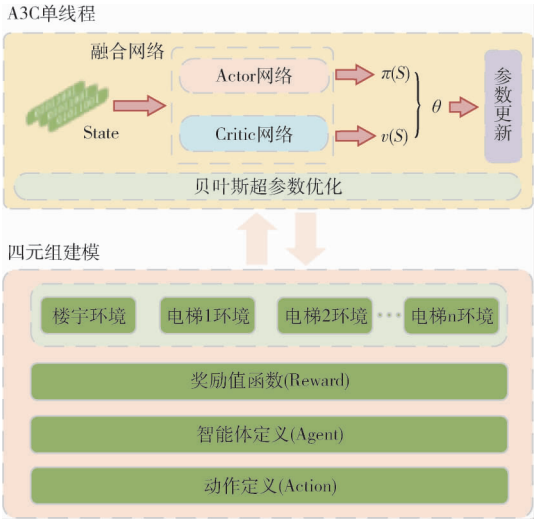


图 3 基于强化学习 A3C 的电梯调度算法模型框架

3.1 强化学习四元组建模

基于强化学习算法的电梯调度算法建模依赖马尔科夫决策过程，需要对其四元组 $<S,A,R,f>$ 进行定义，其分别是状态 S、动作 A 和奖励值 R，而状态转移函数 f 则由强化学习中深度神经网络进行表现。实际中电梯调度环境可以视作楼宇环境和电梯自身环境两者的组合，本文所提强化学习算法中电梯自身环境建模状态 S 所需定义参数项见表 1，其中针对多电梯情况，只需要将多部单电梯自身环境进行顺序排列组合。强化学习运用于电梯调度算法中的智能体可视整体电梯为一个智能体，也可针对每个电梯桥箱进行多智能体建模，即每个调度电梯作为一个智能体进行强化学习建模，如强化学习中 MADDPG 算法就是多个智能体在同一环境下进行策略学习[15]。

复杂环境下电梯调度算法的调度目标是在提升用户体验的情况下，使电梯的能耗尽可能低，因此建模过程中奖励函数的设计需要充分考虑用户体验和电梯能耗两方面。本文强化学习的奖励函数 R 设计如式（1）所示，式子中 $time$ 表示所有乘客在一个时间周期内的等待时长总和，$energy$ 表示一个时间周期内电梯消耗的电力，$person$ 表示一个时间内放弃的人数，实验仿真环境中默认排队等待一定时间后自动放弃等待。式子中调节参数 α、β 和 γ 分别调节各自比例，本文 α、β 和 γ 调节因子取值分别为 0、0.01 和 100

表 1 电梯调度环境建模参数项

序号	环境	参数项
1	楼宇环境	向上需求楼层、向下需求楼层、电梯状态和禁停楼层等。
2	电梯环境	当前楼层、门开比例、箱内按键楼层、最高楼层、调度楼层、是否超载、当前速度、调度方向、电梯门正在打开、最大速度、承载重量、电梯门正在关闭、运行方向和最大载重等。

$$R=\alpha\times time+\beta\times energy+\gamma\times person \tag{1}$$

电梯调度算法中电梯作为智能体，需要在强化学习中定义其动作空间 A，电梯实际调度过程调度动作由目标楼层和方向构成，则其动作空间 A 可由二元组 $<DispatchTarget,DispatchTargetDirection>$ 组成，其中 $DispatchTarget$ 表示是相应电梯调度分配的目标楼层，$DispatchTargetDirection$ 表电梯抵达目标楼层后的后继方向，1 表向上运行，−1 表向下运行，0 则为无方向[8]。

3.2 贝叶斯优化下 A3C 算法建模

强化学习 A3C 算法来源于 Actor-Critic 算法，经典 A3C 算法基于 Actor-Critic 算法在网络结构、Critic 评估点和异步训练框架 3 个方面进行了优化，这一定程度上克服了 Actor-Critic 算法收敛性问题[16]。但在实际建模训练过程中很多时候依然存在收敛慢和调参困难的问题，基于 A3C 算法依然存在性能提升的可能性，本文基于经典 A3C 算法引入贝叶斯优化算法对超参数进行学习，同时借鉴著名算法 Alpha Zero 中使用的融合网络结构设计 Actor 和 Critic 网络结构对整个 A3C 算法进行建模。

A3C 是 Actor-Critic 算法的改进升级，Actor-Critic 算法[16]是一种策略（Policy Based）和价值（Value Based）相结合的强化学习方法。如图 3 所示 A3C 算法建模主体亦包括演员（Actor）和评论家（Critic）两部分，A3C 算法中 Actor 负责生成动作和环境进行交互，Critic 则对 Actor 的行动进行评估，指导 Actor 的动作。Actor 网络输入环境状态 S，输出为当前环境下的动作策略 $\pi(S)$，Critic 网络输入环境的状态 S，输出为对状态 S 的评估值 $v(S)$，$v(S)$ 表示状态下 S 的平均期望价值[17]。A3C 算法中采用优势函数（动作价值函数和状态价值函数的差值）作为 Critic 评估标准，同时结合 N 步采样加速收敛，A3C 中优势函数表达式如式（2）所示，式中 $A(S,t)$ 是优势函数，其表当前状态 S 的价值，γ 是衰减因子

$$A(S,t)=R_t+\gamma R_{t+1}+\cdots\gamma^{n-1}R_{t+n-1}+\gamma^n v(S')-v(S) \tag{2}$$

Critic 网络的评估值直接作用于 Actor 网络的参数更

新，由于策略函数损失函数中加入了策略的熵项，Actor 网络参数 θ 更新公式如式（3）所示

$$\theta = \theta + \alpha \nabla_\theta \log\pi(s_t, a_t)\mathrm{A}(S,t) + c \nabla_\theta \mathrm{H}(\pi(S_t, \theta)) \quad (3)$$

A3C 算法中 Critic 网络则通过计算 TD 误差值 δ，使用均方差作为损失函数对自身网络参数 w 进行参数更新，相关计算式如式（4）所示

$$\delta = R + \gamma \mathrm{v}(S') - \mathrm{v}(S)$$
$$loss = \sum (R + \gamma \mathrm{v}(S') - \mathrm{v}(S, w))^2 \quad (4)$$

强化学习中超参数的设置对于结果收敛性起到一定程度的作用，A3C 算法建模过程中需要设置部分的参数项，例如需要设置全局共享迭代轮数 T、相应网络参数更新步长、熵系数值和衰减因子等，本文为了 A3C 算法性能有效提升引入经典贝叶斯优化方法对部分超参数进行搜寻，贝叶斯优化作为最受欢迎的方法，其充分利用相关历史信息来指导进行最优解的搜索。贝叶斯优化中假设待优化参数为 $X = (x_1, x_2, \cdots, x_n)$，同时存在一个与 X 相关的损失函数 $\mathrm{f}(x)$，贝叶斯优化目标函数即是寻找 X 得到最小损失函数 $\mathrm{f}(x)$。本文基于 A3C 算法的电梯调度算法建模中采用 Hyperopt 工具来实现贝叶斯优化算法。

A3C 算法中 Actor 网络和 Critic 网络一般针对实际任务可采取融合网络和分离网络两种网络形式，本文基于强化学习 A3C 算法的电梯调度算法建模中 Actor 网络和 Critic 网络采用融合网络，融合网络即是让 Actor 网络和 Critic 网络为同一个神经网络，融合网络的输入状态为 S，输出为状态的价值 V 和对应的策略，强化学习著名算法 Alpha Zero 中也使用融合网络结构[18]。

3.3　多线程异步训练

A3C 算法相比较 Actor-Critic 算法更容易收敛，同时 A3C 强化学习算法借用 DQN 经验回放的技巧，利用多线程训练的方法。本文 A3C 算法多线程异步训练部署框架如图 4 所示，训练过程中多个线程分别和环境进行交互学习，同时将结果汇总保存到一个全局线程中进行智能体学习和参数更新，而各个线程智能体则定期从全局线程获取参数指导后面和环境的交互。通过多线程交互 A3C 算法避免了经验回放相关性强的问题，同时多线程训练能够确保电梯智能体能在最短时间学习得到最优调度策略。

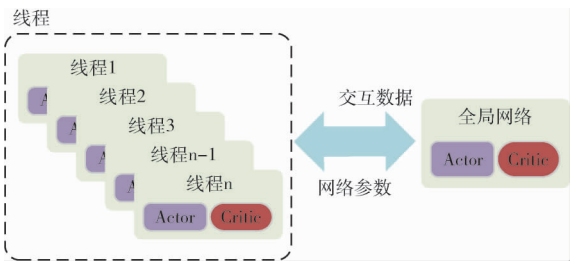


图 4　A3C 算法多线程训练部署框架

4　实验仿真及结果分析

为了充分验证强化学习 A3C 算法运用于楼宇电梯调度的可行性，本文进行了相关的仿真实验，实验环境基于百度飞桨和百度 AI Studio 平台进行，实验 GPU 采用英伟达 Tesla V100。电梯仿真环境采用 LiftSim 进行，LiftSim 是一款类似 gym 的仿真环境[8]，同时基于 LiftSim 开源代码能够方便进行环境的自定义。本文通过在单电梯环境和多电梯复杂环境下对强化学习 A3C 调度算法和部分常用的传统电梯调度算法进行了对比实验，同时在多电梯且具有分时人流的复杂环境下引入多种强化学习算法和 A3C 算法进行对比实验。

4.1　单梯环境下电梯调度

单电梯调度环境下假设整栋楼宇只存在一部电梯，为了验证在单电梯调度环境下强化学习 A3C 算法和传统调度算法的有效性，本文基于 LiftSim 进行了仿真实验，传统调度算法中选用 FCFS、SCAN、SCAN-EDF 和 LOOK 算法作为对比算法。其中 FCFS 算法是一种先来先服务算法，它根据用户楼层的请求电梯先后次序进行依次调度。SCAN 算法即扫描算法，是一种按照楼层顺序依次服务请求，它让电梯在最底层至最高层之间往返运行。SCAN-EDF 算法是一种实时调度算法，结合了调度算法 SCAN 和 EDF 的优势，SCAN-EDF 首先按照 EDF 算法选择请求队列中下一个服务对象，而对于相同时限的请求，则按照 SCAN 算法选择下一个服务对象。LOOK 算法调度电梯在最底层和最顶层之间运行，当电梯前进方向未有用户请求时则立即改变运行方向运行。

本实验设计强化学习 A3C 算法和传统算法得分评价标准为对 10 层大楼的 4 小时人流进行调度，综合得分计算公式如式（3）所示，所得得分值为负，分值越大代表电梯能耗和用户等待时长越短，电梯调度效率越高。调度环境大楼人流产生是一个均匀随机过程，与时间无关。强化学习 A3C 算法超参数学习率设置为 0.001，训练线程数设置 25 线程，A3C 训练过程每一万次迭代进行评测一次。整个 A3C 算法训练过程情况如图 5 所示，图示为 A3C 算法 60 万次迭代训练过程图。

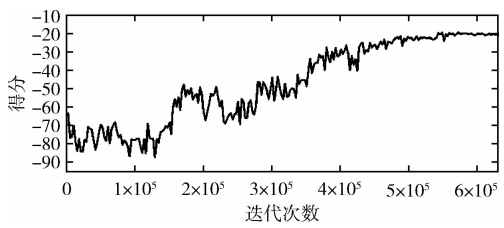


图 5　单梯环境 A3C 算法训练过程

由图 5 强化学习 A3C 算法在不断地和环境进行交互过

程中，可以明显看出 A3C 算法获得的奖励值逐渐增加，调度综合得分逐渐增大这说明电梯的能耗和人员等待时间逐渐降低，智能体逐步学到了最优调度策略。图 5 中大约在 50 万次迭代后得分达到－22 分左右，大约 60 万次迭代后得分达到－21 分左右。利用收敛后的算法和传统相关算法进行对比实验，对比实验调度得分如图 6 所示，图示中 A3C_50 算法采用约 50 万次迭代训练后的模型，A3C_60 算法采用约 60 万次迭代后的模型。

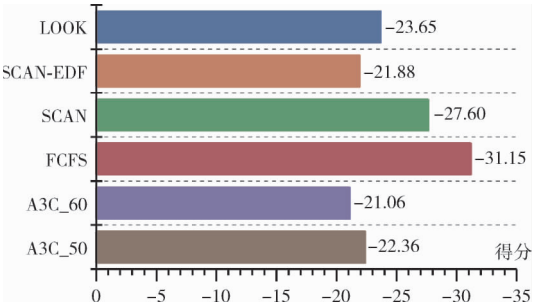


图 6 单梯环境下各调度算法性能对比

由图 6 可知 A3C 强化学习算法 60 万次迭代后调度效果最好达到－21.06 分，调度效果超过其它传统电梯调度算法，传统调度算法中 SCAN-EDF 算法具有较好调度效率，最终得分为－21.88 分。A3C 算法 50 万次迭代后调度得分－22.36 分，相比较 SCAN-EDF 算法存在不足。在单电梯环境中 A3C 调度算法随着迭代数上升相比较传统算法存在一定优势，但迭代数的增加一定程度上也将消耗一定的 GPU 资源。综合来说单电梯环境中传统调度算法具备一定的运用价值，但强化学习 A3C 算法在一定的迭代训练之后相比较传统调度算法也存在一定优势。

4.2 多梯复杂环境下电梯调度

多梯复杂环境下电梯调度算法建模更加复杂，需要考虑调度电梯的运行情况和其它电梯的调度情况，而强化学习中，通过在多电梯环境下增加环境状态 S 维度就容易达到算法建模的要求。为了验证本文所提调度算法的有效性，本文针对 A3C 算法和传统 SCAN-EDF 算法进行了对比实验，基于 SCAN-EDF 算法在多部电梯情况之下采用最近派遣法则，即电梯距离需求楼层最近则派遣此电梯。同时本文为了验证不同强化学习算法的性能，引入了调度算法 DDQN 和 MADDPG 进行了相关对比实验，DDQN 是经典的强化学习算法[19]，而 MADDPG 则是 DDPG 算法运用于多智能体的升级版算法。对比实验是在固定 4 小时时长情况下对大楼 4 部电梯进行联合调度，同时设定 4 小时人流分布呈现一定非随机性，在初期 1 小时左右引入早高峰上行，在后期 3 小时左右引入晚高峰下行，通过引入早晚高峰上下行人流可以模拟当前高层楼宇写字楼的调度环境。

由图 7 可知在模型训练初期 MADDPG 算法能够比较好的到达分值区间（－50，－60），但是随后在训练过程中 MADDPG 却没有表现出更高的分值区间，相反针对 DDQN 算法和 A3C 算法在训练初期阶段其得分情况波动较大，大约在 60 万次训练之后 A3C 算法表现出了综合得分逐步上升的趋势，最终抵达得分区间大约为（－30，－35），而 DDQN 算法则达到了得分区间（－45，－55），这说明 A3C 算法运用多梯电梯调度环境具有一定的优势。同时本实验待强化学习模型训练较收敛后和传统调度算法 SCAN-EDF 算法进行了对比实验，最终实验调度得分情况如图 8 所示。

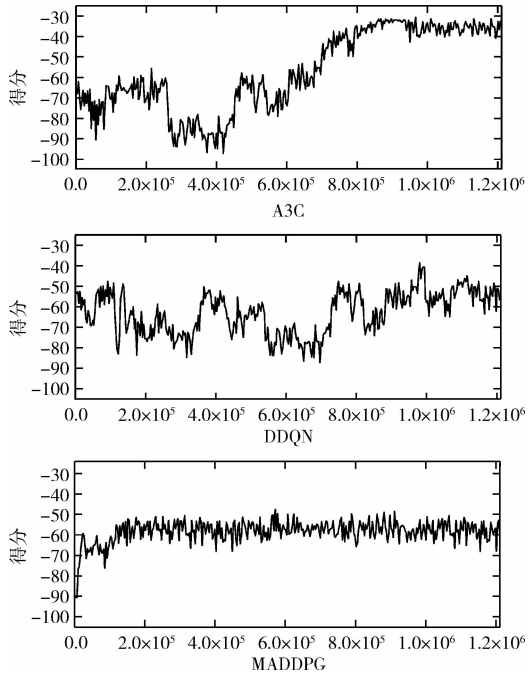


图 7 多梯复杂环境下强化学习算法训练过程

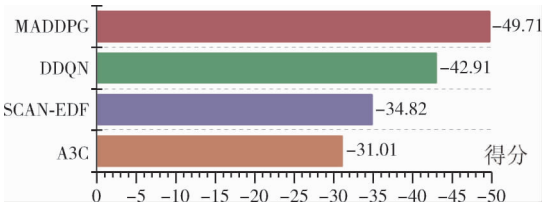


图 8 多梯复杂环境下各调度算法性能对比

由图 8 各调度算法调度综合得分可知，在复杂多电梯环境下由于传统电梯调度算法建模困难，很难做到用户体验和电梯能耗的平衡，而强化学习算法中建模相对简单，由于长时间的模型训练让模型逐渐从和环境的交互过程中学习得到了最佳的调度策略，这说明强化学习运用于复杂环境下电梯调度算法具有可行性，同时对比相关强化学习算法调度分值发现其中 A3C 算法运用于电梯电梯调度具有一定的优势，更容易收敛。

5 结束语

本文基于强化学习 A3C 算法构建电梯调度算法，通过对调度环境、电梯行为和调度目标 3 个方面进行统一建模，让电梯在不断地和调度环境进行交互过程中学习最优调度策略，一方面克服了传统调度算法的建模复杂性，让调度算法在新环境下的也能方便移植，另一方面利用强化学习设计满足用户需求和能耗的奖励函数能有效平衡用户体验和电梯能耗等多方面调度目标。通过进行单电梯和多电梯复杂环境下电梯调度对比实验，实验结果表明强化学习 A3C 算法经过一定迭代次数的训练之后相比较传统调度算法具有一定的优势，能有效平衡用户用梯和电梯节能需求。同时本文通过对比 A3C、DDQN 和 MADDPG 强化学习算法运用于复杂环境电梯调度中的效果，实验结果表明 A3C 算法具有更好的调度性能，更易收敛。本文所提基于强化学习的调度方法同时也在国网湖北省电力公司科技项目的湖北电网运营监控实用化关键技术研究项目中有进行相关研究，强化学习在诸多调度领域存在运用可能。在以后的进一步研究中，可以结合强化学习和相关深度学习方法进行电梯调度的建模和算法优化，让电梯调度模型更快收敛。

参考文献：

[1] Smarter Buildings. Smarter buildings study [EB/OL]. [2020-07-23]. https://www.stepjockey.com/media.ashx/smarter-buildings-stepjockey-svma.pdf.

[2] LIU Guixiong, LIN Jia, CHEN Guoyu, et al. Elevator scheduling algorithm for energy-saving based on statistical analysis [J]. China Measurement & Test, 2015, 41 (7): 85-89 (in Chinese). [刘桂雄，林佳，陈国宇，等. 基于统计分析的节能优先电梯调度算法 [J]. 中国测试，2015，41 (7)：85-89.]

[3] ZHONG Huilin. The elevator dispatching heavy planning based on neural network [J]. China Academic Journal Electronic Publishing House, 2016 (7): 63-64 (in Chinese). [仲惠琳. 基于神经网络的电梯调度重规划 [J]. 数字技术与应用，2016 (7)：63-64.]

[4] LANG Man, LI Guoyong, XU Chenchen. Elevator group control system for energy scheduling optimization simulation [J]. Computer Simulation, 2017, 34 (2): 375-379 (in Chinese). [郎曼，李国勇，徐晨晨. 电梯群控系统的节能调度优化仿真 [J]. 计算机仿真，2017，34 (2)：375-379.]

[5] LIU Qing, GUAN Yujun. Energy saving optimal dispatching control of elevator group control system [J]. Computer Simulation, 2018, 35 (10): 350-354 (in Chinese). [刘清，关榆君. 电梯群控系统节能优化调度控制 [J]. 计算机仿真，2018，35 (10)：350-354.]

[6] LIU Jian, ZHAO Yue, XU Meng, et al. Multi-car elevator systems using dynamic zoning based on Fast R-CNN [J]. Control Engineering, 2019, 26 (2): 32-38 (in Chinese). [刘剑，赵悦，徐萌，等. 基于 Fast R-CNN 的动态分区多轿厢电梯调度研究 [J]. 控制工程，2019，26 (2)：32-38.]

[7] LIANG Xingxing, FENG Yanghe, MA Yang, et al. A survey on deep reinforcement learning [J/OL]. Acta Automatica Sinica, 2020, 46 (12): 2537-2557. [2019-09-26]. https://doi.org/10.16383/j.aas.c180372 (in Chinese). [梁星星，冯旸赫，马扬，等. 多 Agent 深度强化学习综述 [J/OL]. 自动化学报，2020，46 (12)：2537-2557. [2019-09-26]. https://doi.org/10.16383/j.aas.c180372.]

[8] Baidu. LiftSim [EB/OL]. [2019-09-23]. https://github.com/PaddlePaddle/RLSchool/tree/master/rlschool/liftsim.

[9] SUN Zhewei, BI Chao. Development and application of group elevator running simulation program [J]. Computer Engineering and Design, 2020, 41 (5): 1472-1480 (in Chinese). [孙哲伟，毕超. 群组电梯运行仿真程序的开发及应用 [J]. 计算机工程与设计，2020，41 (5)：1472-1480.]

[10] WANG Xudong, SUN Weixiang, XU Xiaozhuo, et al. Research on scheduling control strategy of direct drive multi-car elevator system [J]. Engineering Journal of Wuhan University, 2019, 52 (8): 716-721 (in Chinese). [汪旭东，孙伟翔，许孝卓，等. 直驱多轿厢电梯系统的调度控制策略 [J]. 武汉大学学报（工学版），2019，52 (8)：716-721.]

[11] WANG Yunpeng, GUO Ge. Signal priority control for trams using deep reinforcement learning [J/OL]. Acta Automatica Sinica, 2019, 45 (12): 2366-2377. [2019-09-26]. http://kns.cnki.net/kcms/detail/11.2109.TP.20190917.1556.005.html (in Chinese). [王云鹏，郭戈. 基于深度强化学习的有轨电车信号优先控制 [J/OL]. 自动化学报，2019，45 (12)：2366-2377. [2019-09-26]. http://kns.cnki.net/kcms/detail/11.2109.TP.20190917.1556.005.html.]

[12] FENG Chao. Essentials of reinforcement learning: Core algorithm and TensorFlow implementation [M]. 1st ed. Beijing: Electronic Industry Press, 2018: 2-11 (in Chinese). [冯超.《强化学习精要：核心算法与 TensorFlow 实现》[M]. 1 版. 北京：电子工业出版社，2018：2-11.]

[13] LIU Quan, ZHAI Jianwei, ZHANG Zongchang, et al. A survey on deep reinforcement learning [J]. Chinese Journal of Computers, 2018, 40 (1): 1-27 (in Chinese). [刘全，翟建伟，章宗长，等. 深度强化学习综述 [J]. 计算机学报，2018，40 (1)：1-27.]

[14] XU Xijian, WANG Zilei, XI Hongsheng. Session scheduling strategy for streaming media edge cloud based on deep reinforcement learning [J]. Computer Engineering, 2019, 45 (5): 243-248 (in Chinese). [徐西建，王子磊，奚宏生. 基于深度强化学习的流媒体边缘云会话调度策略 [J]. 计算机工程，2019，45 (5)：243-248.]

[15] Lowe R, Wu Y, Tamar A, et al. Multi-agent actor-critic for mixed cooperative-competitive environments [C] //Neural

Information Processing Systems，2017：6379-6390.

[16] Sutton R，Barto A. Reinforcement learning：An introduction [M]. Cambrige：MIT Press，2017：274-276.

[17] WANG Yeli. Research on game agent with prediction based on A3C model [D]. Harbin：Harbin Institute of Technology，2018：31-33 (in Chinese). [王耶利. 基于 A3C 模型的带预判游戏智能体研究 [D]. 哈尔滨：哈尔滨工业大学，2018：31-33.]

[18] Silver D，Schrittwieser J，Simonyan K，et al. Mastering the game of Go without human knowledge [J]. Nature，2017，550 (7676)：354-359.

[19] Wang Z，Schaul T，Hessel M，et al. Dueling network architectures for deep reinforcement learning [C] //International Conference on Machine Learning，2016：1995-2003.