Advances in Education 教育进展, 2024, 14(1), 603-608
Published Online January 2024 in Hans. https://www.hanspub.org/journal/ae
https://doi.org/10.12677/ae.2024.141091

Hans汉斯

强化学习实验教学现状与探究

——以新疆大学计算机科学与技术学院为例

余银峰[1]，祝美玲[2*]，汪烈军[1]

[1]新疆大学计算机科学与技术学院，新疆 乌鲁木齐
[2]乌鲁木齐市第五十九中学，新疆 乌鲁木齐

收稿日期：2023年12月13日；录用日期：2024年1月10日；发布日期：2024年1月18日

摘　要

强化学习是一门理论性和实践性都很强的课程，实验教学起着重要的作用。本文以新疆大学计算机专业的强化学习课程为例，针对该专业特点，提出以在线Python Notebook为平台，构建适合该专业学生的强化学习课程实验教学内容，并探讨了立体化教学、案例教学、“学研”结合和鼓励学生参加人工智能算法竞赛的实验课程教学方法和手段，对提升课程教学效能具有一定的参考价值。

关键词

实验教学，策略梯度，强化学习

Current Status and Exploration of Reinforcement Learning Experimental Teaching

—Taking Xinjiang University’s School of Computer Science and Technology as an Example

Yinfeng Yu[1], Meiling Zhu[2*], Liejun Wang[1]

[1]School of Computer Science and Technology, Xinjiang University, Urumqi Xinjiang
[2]Urumqi No. 59 Middle School, Urumqi Xinjiang

Received: Dec. 13th, 2023; accepted: Jan. 10th, 2024; published: Jan. 18th, 2024

*通讯作者。

文章引用: 余银峰, 祝美玲, 汪烈军. 强化学习实验教学现状与探究[J]. 教育进展, 2024, 14(1): 603-608.
DOI: 10.12677/ae.2024.141091

Abstract

Reinforcement learning is a course that is both theoretical and practical, with experimental teaching playing a crucial role. Taking the reinforcement learning course in the computer science program at Xinjiang University as an example, this article proposes the use of an online Python Notebook platform. It aims to build experimental teaching content suitable for students in this program, considering the characteristics of the major. The article explores three-dimensional teaching, case-based teaching, the integration of learning and research, and encourages students to participate in artificial intelligence algorithm competitions as methods and means for experimental course teaching. This approach has certain reference value for improving the effectiveness of course teaching.

Keywords

Experimental Teaching, Policy Gradients, Reinforcement Learning

1. 引言

人工智能是当今 21 世纪的前沿领域[1]，包括机器学习、深度学习、自然语言处理等多个方面。随着互联网、物联网等新兴技术的迅速发展，人工智能技术成为了社会发展的重要驱动力之一。我国政府寄予了人工智能技术极高的期望，并提出建设人工智能强国的战略目标，加大对人工智能领域的支持力度。在新疆地区，随着经济的快速发展和产业结构的转型升级，对人工智能领域人才的需求也不断增加。为了满足这一需求，2020 年 8 月，新疆大学成立了未来技术学院，开始了人工智能专业的探索和建设。这是一项十分重要的改革，也是响应国家战略发展需求的重要举措。作为新疆地区的重点高校之一，新疆大学将人工智能专业建设作为重点学科和重点发展方向。加大对师资队伍和实践机会的投入，完善课程设置十分必要。只有这样才能提高学生的实际操作能力和解决问题的能力，为他们未来的就业打下坚实的基础。

然而，强化学习作为一门理论性和实践性都极强的课程，要求学生在掌握基本理论和方法的同时，必须加强教学实验和实践。本文以在线 python notebook 为平台，采用 PyTorch 框架，对计算机科学与技术学院计算机专业的强化学习课程安排、实验教学内容、实验教学方法和手段等进行了探讨。

2. 教学目标和课程安排

强化学习的教学目标是让学生了解并掌握强化学习的基本原理和方法，学会利用强化学习解决人工智能领域的应用问题，更好地将强化学习服务于人工智能的相关领域。针对该专业的特点，需要制定符合实际需求的实验教学大纲、实验教学计划和实验大纲。这些计划需要确定适当的授课方式、实验比例、学分和课时[2]，同时也需要根据专业需要对课程内容进行必要的调整[3]、增删改、精简或拓宽应用[4]。在课时分配方面，需要合理分配理论课和实验课的时间[5]。本专业强化学习课程采取理论与实践相结合的教学方式，实验教学以课程相关内容为主线，与课堂教学交替进行[6]，旨在通过实验加深学生对强化学习的相关理论知识和算法的学习，并让学生熟悉主流强化学习算法的使用和实现等。建议根据以上要

求和本专业特点，在计算机专业三年级下学期开设强化学习课程。由于概率论和 Python 编程语言已经学习过，学生已经具备了强化学习所必需的专业知识和能力。同时，这些学生需要在四年级开始选择毕业研究课题，并有大约一年的时间来准备毕业设计，提前半年开设该课程可以为后续的毕业研究课题的开展打下良好的基础和做好充足的准备。根据教学计划，该课程共计 48 学时，其中理论授课 32 学时，上机实验 16 学时，这样的比例设置符合强化学习是理论性和实践性都很强的特点。

3. 实验教学内容

为了加强学生对强化学习理论知识的掌握，提高其消化吸收课堂讲授内容的能力，并锻炼其实际操作技能，建议选择难度适中、体系完整的强化学习实验配套教材。实验内容的设计应与理论知识的教学内容同步，同时也应与学生的知识背景紧密结合，最好以学生所要解决的应用问题为实验目标。基于以上考虑，选取《动手学强化学习》为实验教学参考书[7]，结合本专业特点，以在线 python notebook 为平台，以 PyTorch 为框架，探究各种算法在车杆、倒立摆等问题上的实现方法与性能表现。实验内容主要包括 DQN (deep Q-learning)算法、DQN 改进算法、REINFORCE 算法、Actor-Critic 算法、TRPO (信任区域策略优化[8]，trust region policy optimization)算法、PPO (近端策略优化[9] [10]，proximal policy optimization)算法、DDPG (深度确定性策略梯度[11] [12]，deep deterministic policy gradient)算法和 SAC (soft actor-critic)算法。

4. 实验教学内容设计

对于计算机专业的学生来说，只有将强化学习的理论知识与实际应用相结合，并将 DQN 算法、DQN 改进算法、REINFORCE 算法、Actor-Critic 算法、TRPO 算法、PPO 算法、DDPG 算法或者 SAC 算法等应用于本学科具体场景中，才能真正理解和掌握这门课程。因此，建立强化学习的实践教学系统对于提高学生的实践操作和应用能力具有至关重要的意义。下面是实验内容的具体设计：

1) DQN 算法。主要目的是让学生掌握 DQN 算法的使用和实现。主要内容包括 DQN 算法的基本原理、在车杆环境中的实验以及如何使用神经网络表示最优策略函数，同时介绍了经验回放和目标网络等模块的引入，以提高 DQN 算法在解决连续状态下离散动作问题时的训练效果和稳定性。

2) DQN 改进算法。主要目的是让学生掌握两种 DQN 改进算法(Double DQN 和 Dueling DQN)的使用和实现。主要内容包括 Double DQN 和 Dueling DQN 算法的基本原理，并通过倒立摆实验验证了它们能够改善 DQN 算法的效果。特别是 Dueling DQN，在动作空间较大的环境下表现非常出色。研究深度强化学习的重点在于如何将深度学习和强化学习有效结合。Double DQN 算法可以解决过高估计 Q 值的问题，而 Dueling DQN 则通过设计高效的网络结构来学习状态值函数和动作优势函数[13] [14]。

3) REINFORCE 算法。主要目的是让学生掌握 REINFORCE 算法的使用和实现。主要内容包括 REINFORCE 算法的基本原理和在车杆环境中的实验。强化学习包括基于值函数和基于策略的方法[15]。DQN 及其改进算法是基于值函数的方法，而基于策略的方法则是直接学习目标策略。REINFORCE 算法是策略梯度方法的代表[16]，通过采样得到的轨迹数据直接计算出策略参数的梯度来更新策略，使其向最大化策略期望回报的目标靠近。REINFORCE 算法[17]的优点是可以得到无偏的梯度估计，并且理论上能保证局部最优。

4) Actor-Critic 算法。主要目的是让学生掌握 Actor-Critic 算法的基本原理与代码实现。主要内容包括 Actor-Critic 算法的基本原理和在车杆环境上的实验，旨在解决 REINFORCE 算法的梯度估计方差过大问题[18]。Actor 采样数据，Critic 学习分辨好坏动作，指导 Actor 更新策略。Critic 需要适应数据分布的变化，给出好的判别。

5) TRPO 算法。主要目的是让学生掌握 TRPO 算法的基本原理与代码实现。主要内容是 TRPO 算法的基本原理，并分别在离散动作和连续动作交互的环境中进行了实验。TRPO 算法是在线策略学习方法，只使用上一轮采样的数据进行训练。该算法是基于策略的深度强化学习算法中的代表性工作之一。TRPO 通过限制策略学习区域，保证策略学习的稳定性和有效性[19]。

6) PPO 算法。主要目的是让学生掌握 PPO 算法的基本原理与代码实现。主要内容是介绍 PPO 的基本原理，并在车杆和倒立摆这两个环境中测试 PPO 算法。PPO 是 TRPO 的一种改进算法，它在实现上简化了 TRPO 中的复杂计算，并且它在实验中的性能大多数情况下会比 TRPO 更好，因此目前常被用作一种常用的基准算法。

7) DDPG 算法。主要目的是让学生掌握 DDPG 算法的基本原理与代码实现。主要内容是介绍 DDPG 算法的基本原理，并以倒立摆为例，编程实现 DDPG 算法。DDPG 算法构造一个确定性策略，用梯度上升的方法来最大化 Q 值。

8) SAC 算法。主要目的是让学生掌握 SAC 算法的基本原理与代码实现。主要内容是介绍 SAC 算法的基本原理，首先用倒立摆进行实验，然后再尝试将 SAC 应用到以离散动作与环境交互的车杆问题上。

按照以上内容设计，实验教学课程体系如表 1 所示。

Table 1. Experimental teaching content schedule
表 1. 实验教学内容安排表

实验项目	实验内容摘要	实验类别	学时数
DQN 算法	DQN 算法的基本原理与代码实现	综合	2
DQN 改进算法	Double DQN 和 Dueling DQN 算法的基本原理与代码实现	综合	2
REINFORCE 算法	REINFORCE 算法的基本原理与代码实现	综合	2
Actor-Critic 算法	Actor-Critic 算法的基本原理与代码实现	综合	2
TRPO 算法	TRPO 算法的基本原理与代码实现	综合	2
PPO 算法	PPO 算法的基本原理与代码实现	综合	2
DDPG 算法	DDPG 算法的基本原理与代码实现	综合	2
SAC 算法	SAC 算法的基本原理与代码实现	综合	2

5. 探索实验教学方法和手段

1) 使用高效的实时在线教学，以实现立体化教学。算法实践教学讲解宜采用在线 python notebook 的形式进行，主要用于实验内容、要点、难点的讲解，可以将原理讲解部分(包括配图和公式)与对应的代码耦合在一起，这样能使学生在学习完一个原理知识点后立即以代码的形式学习其实现方式。更重要的是，这样的代码块可以在线直接运行和修改，以致于可以在一个 notebook 里完成对一个强化学习算法的原理学习和实验。这样的学习方式能帮助学生更好地对应上理论知识点和实践能力点，也能帮助老师更高效地授课、布置和批改作业。实时在线教学可以做到条理清晰、利于更新、便于分享。与此同时，在线 python notebook 的形式还可以为学生提供代码小作业。讲、练、改都十分方便。

2) 实行案例教学法，提高教学效果。由于车杆的状态值就是连续的，动作值是离散的，所以 DQN 算法采用车杆为例；由于要演示 DQN 算法存在对 Q 值估计过高的缺陷，DQN 改进算法采用倒立摆为例。案例教学的方式使得枯燥的理论知识变得形象直观，并提高了学生的学习主动性，同时也使学生熟悉采用强化学习解决实际问题的基本套路和方法，从而提高学生的综合应用能力。在学生练习的过程中，鼓励学生自己动手写代码，而不仅仅是简单的复制粘贴代码，运行代码，贴出实验结果图。

3) “学研”结合，培养学生的创新能力。教师带领学生参与科研，从课题选择、方案制定、算法实现到论文撰写的全过程中，让学生全程参与。通过这种全过程的科研参与，可以提高学生应用理论知识和解决问题的能力，促进他们分析和解决问题的能力培养。

4) 激励学生参加人工智能算法竞赛，提高学生综合技能。参加人工智能算法竞赛可以提高学生的综合素质和能力，包括专业知识、交流表达、合作、问题解决、改革创新、自我学习等方面。例如，多目标导航竞赛(比赛网址：http://multion-challenge.cs.sfu.ca/)提供基本的代码框架，让参赛者快速上手。该竞赛要求代理使用 RGB 图像、深度图像和相对位置信息来自主学习导航策略，并按指定顺序导航到目标对象。这需要学生具备计算机视觉、强化学习和调参能力，参加比赛可以大大提高这些综合技能。

6. 结束语

综上所述，强化学习课程对数学特别是概率论的要求比较高，同时对学生的编程能力的要求也比较高。探讨强化学习的实验教学内容和教学方法，推动强化学习的理论学习与实践的有效衔接，不仅有助于学生掌握强化学习的基础知识，将其与专业背景知识融合，还有助于学生深刻体会到强化学习在人工智能应用中的作用，掌握应用强化学习算法来解决专业相关问题的方法，这符合计算机专业发展的总体趋势。

基金项目

中国新疆维吾尔自治区天山卓越计划项目(2022TSYCLJ0036)；中央引导地方科技发展基金项目(ZYYD2022C19)；新疆维吾尔自治区自然科学基金项目(2022D01C58、2020D01C026 和 2015211C288)。

参考文献

[1] 蔡红娟. 新工科背景下人工智能人才培养模式探索与实践[J]. 教育教学论坛, 2022(40): 107-110.

[2] 贾泽露. 非 GIS 专业地理信息系统课程教学思考[J]. 测绘科学, 2008(5): 230-232.

[3] 钱敏. 城市规划专业 GIS 课程教学改革探讨[J]. 科教文汇(中旬刊), 2014(9): 61-62.

[4] 僧德文, 王红霞. 基于 SuperMap 的地理信息系统课程教学设计[J]. 浙江水利水电专科学校学报, 2009, 21(3): 79-81.

[5] 刘桂萍, 陈川, 杨焱青, 等. 资源勘查工程专业 GIS 实验教学改革与探讨[J]. 教育教学论坛, 2018(11): 77-79.

[6] 张应武, 刘素君. 基于研究性学习的本科计量经济学教学策略研究[J]. 佳木斯教育学院学报, 2014(4): 132-133.

[7] 张伟楠, 沈键, 俞勇. 动手学强化学习[M]. 北京: 人民邮电出版社, 2022.

[8] 陈红名, 刘全, 闫岩, 等. 基于经验指导的深度确定性多行动者——评论家算法[J]. 计算机研究与发展, 2019, 56(8): 1708-1720.

[9] 张建行, 刘全. 基于情节经验回放的深度确定性策略梯度方法[J]. 计算机科学, 2021, 48(10): 37-43.

[10] 王鸿涛. 基于强化学习的机械臂自学习控制[D]: [硕士学位论文]. 哈尔滨: 哈尔滨工业大学, 2019.

[11] 张大胜. 基于深度强化学习的智能体决策与控制研究[D]: [硕士学位论文]. 南京: 南京师范大学, 2021.

[12] 申怡, 刘全. 基于自指导动作选择的近端策略优化算法[J]. 计算机科学, 2021, 48(12): 297-303.

[13] 苏畅. 基于强化学习的雷达辐射源识别技术研究与应用[D]: [硕士学位论文]. 北京: 北京邮电大学, 2021.

[14] 郁洲, 毕敬, 苑海涛. 基于改进 DQN 算法的复杂海战场路径规划方法[J]. 智能科学与技术学报, 2022, 4(3): 418-425.

[15] 梁宏斌. 基于 openAI Gym 和 DRL 的移动机器人路径规划算法研究[D]: [硕士学位论文]. 重庆: 重庆理工大学, 2021.

[16] 韩国亮. 基于强化学习的末制导导引律设计[D]: [硕士学位论文]. 哈尔滨: 哈尔滨工业大学, 2019.

[17] 刘开宇. 基于强化学习的物体抓取方法研究[D]: [硕士学位论文]. 哈尔滨: 哈尔滨工业大学, 2020.

[18] 朱文文, 金玉净, 伏玉琛, 等. 连续空间的递归最小二乘行动者——评论家算法[J]. 计算机应用研究, 2014, 31(7): 1994-1997+2000.

[19] 黄俊宁. 基于有界动作策略的强化学习探索方法[D]: [硕士学位论文]. 广州: 广东工业大学, 2018.