《强化学习应用实践》教学大纲

《强化学习应用实践01》

一、课程基本信息

课程名称/英文名称	强化学习应用实践/Project Practice for Reinforcement Learning	课程代码	CS290T
课程层次	研究生课程	学分/学时	3/48
先修课程	无	授课语言	双语
开课单位	信息科学与技术学院	课程负责人	田政

二、课程简介

本课程为强化学习工程实践课程，旨在通过项目实践帮助学生掌握强化学习的核心技术与应用。学生将围绕智能决策问题，设计并实施相关项目，涵盖的主题包括动态规划、强化学习、基于模型的强化学习、深度强化学习、多智能体系统等。项目将应用于游戏智能体开发、机器人控制等实际场景，深入探索强化学习在这些领域中的关键作用。

三、课程教学目标

- 知识认知能力：掌握强化学习的核心概念以及常见算法的基本实现方法。
- 工程实践技能：学习使用Python和PyTorch构建强化学习模型，并创建训练、评估和策略优化工具。
- 实际问题解决：通过将强化学习算法应用于实际场景中的问题，训练学生分析和处理复杂环境数据的能力，启发学生对强化学习算法性能的深入理解，引导学生提出针对不同应用场景的优化方案。
- 沟通与展示:通过小组讨论、小组演示等形式，帮助学生提高其沟通与团队合作能力。

四、课程教学方法

课堂教学：强化学习知识点基本以课堂教学为主，在讲解基本知识点和各课题的基础上，关注课程重点难点内容的讲授，采用启发式教学方法，引导学生对问题展开思考和讨论，使学生对强化学习的基本概念与方法等有清晰的认知。
项目实践：理论与实践相结合，学生将根据老师给出的实际项目，应用课堂教学中的理论展开项目设计与实施。

五、课程教学内容与安排

以章节名称方式安排教学内容

章节名称	主要教学内容（主要知识点）	教学周	学时安排	教学方法（仅列名称）
1	强化学习基础（马尔可夫决策过程、贝尔曼等式、值函数估计、策略学习等）	1	3	课堂教学
2	实践1：动态规划实践	2	6	课堂教学

	(针对强化学习中的经典toy example，利用策略迭代、值迭代、以及基于模型的方法进行动态规划算法的实践) 授课内容（1周）： • 动态规划基础概念回顾：介绍动态规划的基本思想及其在强化学习中的应用。 • 经典强化学习toy example：选择具体的强化学习环境，如迷宫问题或网格世界，说明任务背景及目标。 • 常见问题与调试技巧：讨论在实现动态规划算法时可能遇到的问题及解决方案。 • Python中Gym库的使用：教学如何使用Gym库创建和操作强化学习环境，包括环境的初始化、状态空间和动作空间的理解，以及如何与环境交互进行训练。 实验内容（1周）： • 环境搭建：设置强化学习环境，确保所有必要库和工具安装完成。 • 实现策略迭代：编写代码实现策略迭代算法，对选定的toy example进行训练，并观察策略的变化过程。 • 实现值迭代：编写代码实现值迭代算法，分析收敛速度与最终策略的表现。 • 比较策略与值函数：对比策略迭代和值迭代在相同环境下的效果，观察优缺点。 • 基于模型的方法实践：创建模型并利用模型进行决策，评估其性能。			实践指导

3	实践2：深度强化学习实践
	（学习经典深度强化学习算法，如DQN、PPO、DDPG等，并针对深度强化学习中的经典环境：OpenAI Gym、Atari、Mujoco开展实践）
	授课内容（1周）：
	• 深度强化学习概述：介绍深度强化学习的基本概念及其发展背景。
	• 经典算法介绍：
	  ◦ DQN（Deep Q-Network）：讲解DQN的原理、架构和实现，重点介绍经验回放和目标网络。
	  ◦ PPO（Proximal Policy Optimization）：讨论PPO的优势及其策略优化的机制。
	  ◦ DDPG（Deep Deterministic Policy Gradient）：介绍DDPG的原理，适用于连续动作空间的特点。
	• 环境选择：阐述Atari和Mujoco的特点和适用场景。
	• 模型评估与调优：介绍如何评估深度强化学习模型的性能，并进行超参数调优。
	• 常见问题与调试技巧：讨论在实现深度强化学习算法时可能遇到的问题及解决方案。
		4	12	课堂教学
				实践指导

	实验内容（3周）： • 环境搭建：设置深度强化学习环境，确保安装必要的库（如TensorFlow/PyTorch和Gym）。 • 实现DQN：编写代码实现DQN算法，在OpenAI Gym的经典环境中进行训练，并观察学习曲线。 • 实现PPO：编写代码实现PPO算法，针对Atari游戏进行训练和测试，分析模型的表现。 • 实现DDPG：编写代码实现DDPG算法，应用于Mujoco环境，评估其在连续动作空间的效果。 • 模型评估：对训练后的模型进行评估，比较不同算法在相同环境下的表现。 • （拓展）超参数调优：调整学习率、折扣因子等超参数，观察对模型性能的影响。 • （拓展）可视化分析：使用图表可视化训练过程，包括奖励曲线和策略演变。			

4	实践3：多智能体强化学习实践
	（学习多智能体强化学习算法，如IQL、MAPPO、Double Oracle,并针对多智能体系统中的经典环境：Multi-Agent Particle Environment、Kuhn Poker、谷歌足球开展实践）
	授课内容（1周）：
	• 多智能体强化学习概述：介绍多智能体强化学习的基本概念及其在现实世界中的应用。
	• 经典算法介绍：
	  ◦ IQL（Independent Q-Learning）：讲解IQL的原理及其在多智能体环境中的实现。
	  ◦ MAPPO（Multi-Agent Proximal Policy Optimization）：介绍MAPPO的特点及其在多智能体协作中的应用。
	  ◦ Double Oracle：讨论Double Oracle算法的工作机制及其在博弈论中的应用。
	  ◦ PSRO（Policy Space Response Oracles）：阐述PSRO的原理和流程，强调其在动态多智能体环境中的有效性和灵活性。
	• 随机博弈（Stochastic Games）：介绍随机博弈的基本概念及其与多智能体强化学习的关系，讨论如何在不确定环境下进行决策。
	• 环境选择：阐述Multi-Agent Particle Environment、Kuhn Poker和谷歌足球的特点及适用场景。
	• 策略评估与合作机制：介绍如何评估多智能体策略的性能，以及不同智能体之间的合作与竞争机制。
		4	12	课堂教学
				实践指导

	实验内容（3周）： • 环境搭建：设置多智能体强化学习环境，确保安装必要的库（如Gym和相关环境包）。 • 实现IQL：编写代码实现IQL算法，在Kuhn Poker中进行训练。 • 实现MAPPO：编写代码实现MAPPO算法，应用于Multi-Agent Particle Environment游戏，观察智能体之间的互动。 • 实现PSRO：编写代码实现PSRO算法，在谷歌足球中探索如何利用该算法生成有效的策略，并评估智能体在动态环境中的表现。 • 模型评估：对训练后的模型进行评估，比较不同算法在相同环境下的效果。			

5	实践4：快思考与慢思考双系统实践
	（学习蒙特卡洛树搜索、Alpha Zero算法，了解树搜索与神经网络结合的迭代学习模式，并选择适合场景进行实践）
	授课内容（1周）：
	• 快思考与慢思考概述：介绍丹尼尔·卡尼曼的双系统理论，区分快思考（直觉、快速决策）与慢思考（分析、深思熟虑）。
	• 蒙特卡洛树搜索（MCTS）：讲解MCTS的基本原理、步骤及其在决策中的应用，强调探索与利用的平衡。
	• AlphaZero算法：介绍AlphaZero的架构，包括如何结合MCTS与深度神经网络进行决策，重点讨论其训练过程和迭代学习模式。
	• 树搜索与神经网络的结合：分析树搜索与神经网络结合的优势，讨论如何通过这种结合提高决策质量和效率。
	• 适合场景选择：探讨适合应用MCTS和AlphaZero的具体场景，如棋类游戏、围棋等复杂策略游戏。
	实验内容（3周）：
	• 环境搭建：设置实验环境，确保安装必要的库（如Gym和相关游戏环境）。
	• 实现简化的MCTS：
	  ◦ 填空实现核心功能：在开源代码的基础上，填充MCTS的关键部分，如选择（Selection）、扩展（Expansion）、模拟（Simulation）和回传（Backpropagation）的具体实现。
	  ◦ 测试和调试：运行代码并测试每个部分的功能，确保实现正确。
	• 实现简单的策略网络：
	  ◦ 使用开源的神经网络结构：在现有的代码中，找到并填充构建和训练策略网络的关键部分。
		4	12	课堂教学
				实践指导

	◦ 训练与验证：使用生成的游戏数据来训练神经网络，并验证模型的表现。 • 结合MCTS与策略网络： ◦ 修改代码以整合两者：在开源代码中，将训练好的策略网络集成到MCTS中，用于指导选择动作。 ◦ 观察效果：运行实验，比较使用策略网络与随机选择的效果。 • 模型评估：对训练后的模型进行评估，记录不同策略的胜率和表现。 • （拓展）价值网络实现：在现有代码中，添加价值网络的实现，用于评估状态的价值。 • （拓展）价值网络与MCTS的结合：探索如何将价值网络与MCTS结合，以提高决策的准确性和效率。			
6	课程结题：课程项目汇报	1	3	考试

六、考核方式和成绩评定方法

1. 实践1：20%
2. 实践2：20%
3. 实践3：20%
4. 实践4：35%
5. 出勤：5%

七、教材和参考书目

(一)、推荐教材

无

(二)、参考书目

无

八、学术诚信教育

本课程高度重视学术诚信，严禁抄袭、作弊等行为。

“在学习、科研、实习实践等活动中，学生应恪守学术道德，坚守学术诚信，保护知识产权，坚持勇于创新、求真务实的科学精神，努力培养自己严谨求实、诚实自律、真诚协作的科学态度，成为良好学术风气的维护者、严谨治学的力行者、优良学术道德的传承者。”

九、其他说明(可选)

无

《Project Practice for Reinforcement Learning》Syllabus

1.Basic course information

Course name	Project Practice for Reinforcement Learning	Course code	CS290T
Course Level		Credit/Contact Hour	3/48
Prerequisite	Null	Teaching Language	

School/Institute	School of Information Science and Technology	Instructor	田政

2.Course Introduction

Null

3.Learning Goal

Null

4.Instructional Pedagogy

Null

5. Course Content and Schedule

Null

6.Grading Policy

Null

7. Textbook & Recommended Reading

(1) Textbook

Null

(2) Recommended Reading

Null

8.Academic Integrity

Null

9.Other Information (Optional)

Null