# 课程大作业：强化学习

## 大作业概述

本大作业要求同学们实现一个马里奥游戏的智能体：在给定的 baseline 的基础上，通过对“观测空间（observation space）”、“动作空间（action space）”与“奖励空间（reward space）”的特征处理，提升智能体的性能。

大作业并不会专注于强化学习算法的改进。对强化学习的要求程度为入门级别，了解强化学习的基本概念与基础的 DQN（Deep Q-Network）算法即可。

相反，一些特征工程才是本作业的重点。

## 任务介绍

（1）阅读所附的强化学习相关的资料和论文，进行文献综述和算法总结；

（2）依据所给出的 baseline 代码，实现一个超级马里奥的游戏 AI；

（3）按照给出的提示与特征空间处理代码，了解强化学习特征空间处理的方式与着手点，并学会分析智能体性能好坏及其原因。

*baseline 仓库地址：https://github.com/opendilab/DI-adventure（可点击打开链接），其中 main 分支包含所使用的代码，analysis 分支包含分析示例，results 分支包含供参考的运行结果。*

# 具体要求

仔细阅读下述大作业的具体要求，并遵照要求完成大作业。

## （1）强化学习基本概念

了解强化学习的基础知识，完成相关要求。

### 1.1【基本要求】10 分

（1） 给出“强化学习”的概念与要解决的问题，说明它和“机器学习”、“深度学习”这两个概念的关系；

（2） 给出强化学习中一些概念的定义：包括智能体（agent）、奖励（reward）和回报（return）、状态转移函数、马尔可夫决策过程（MDP）、价值函数（Value function）、策略（Policy）、时序差分算法（TD），以及观测空间、动作空间、奖励空间。

*关于强化学习的入门教程不一而足，这里提供一个可供参考的资料：https://lilianweng.github.io/posts/2018-02-19-rl-overview/*

## （2）强化学习文献阅读与综述

阅读所提供的强化学习（Reinforcement Learning, RL）主题相关的论文（见下文 2.1 中的“A 论文列表”），对给定论文进行文献综述与算法总结，并完成以下任务。

### 2.1【基本要求】10 分

A. 论文列表（可点击论文题目，访问原始论文）

i. Human-level control through deep reinforcement learning

ii. Action Space Shaping in Deep Reinforcement Learning

iii. Playing FPS Games with Deep Reinforcement Learning

iv. MinAtar: An Atari-Inspired Testbed for Thorough and Reproducible Reinforcement Learning Experiments

B. 完成以下任务

（1） 给出 DQN 优化的伪代码（请根据自己的理解完成，而非完全拷贝论文细节）；

（2） 画出上述第一篇 DQN 的论文中，针对 Atari 的 DQN 的网络结构图；

（3） DQN 相较于 Q-Learning，使用神经网络来替代 Q-table 的好处是什么？

（4） 使用 Target Network 的好处是什么？

（5） 使用 Replay Buffer 的好处是什么？Replay Buffer 的增删改查都以什么样的规则进行？Replay Buffer 的 Buffer Size 应该怎么设置？放大和缩小 Buffer Size 可能会有什么影响？

（6） 总结四篇文章里面所有的对于 observation、reward、action 空间处理的操作。

### 2.2【加分项：强化学习算法进阶】≤10 分

对 DQN 的一些改进进行文献综述与算法总结，完成以下任务。

A. 论文列表（可点击论文题目，访问原始论文）

i. 改进论文一：Dueling DQN

ii. 改进论文二：Revisiting Fundamentals of Experience Replay

B. 完成以下全部或部分任务

（1） 指出“改进论文一”相关算法提出时，拟解决的问题或改进的出发点（Motivation）；

（2） “改进论文一”是如何解决这些问题的（Methodology）；

（3） “改进论文一”最后的实验是如何设计以验证其方法有效性的（Experiment），为什么这么设计？

（4） 尝试用 3-4 句话总结“改进论文一”的全文（Abstract）；

（5） “改进论文二”通过实验，得到了哪些结论？

（6） “改进论文二”中，上述结论是通过怎样的实验验证得到的？

## （3）智能体设计

### 3.1【熟悉环境】5 分

熟悉所给 马里奥环境（https://github.com/Kautenja/gym-super-mario-bros）（可点击链接访问）。*本次实验使用 v0 环境，安装过程可能会由于 gym 版本过高而遇到问题，推荐使用 gym==0.25.1 版本（如遇到问题请在群里提问）。*完成以下任务。

（1） 尝试用随机动作的智能体挑战 1-1 关卡，渲染出来游戏过程看看效果；

（2） 尝试通过键盘操控马里奥闯关（*有键盘输入接口，具体可以参考 nes-py 里面的任天堂红白机操作说明*），体会人是怎么进行决策的；

（3） 通过环境给定的接口，来保存某一局游戏录像（示例代码如下）；

（4） 分析特征空间构成*（提示：打印 state、reward、action 或环境的 action_space、observation_space、reward_range 等属性）*:

i. 观测空间（observation space）是怎样的？（维度是多少？内容是什么含义？）

ii. 动作空间（action space）是怎样的？

iii. 奖励空间（reward space）是如何构成的？

```python
Python
'''
保存录像的代码示例，使用 gym.wrappers.RecordVideo 类
'''
import gym
import time
from nes_py.wrappers import JoypadSpace
import gym_super_mario_bros
from gym_super_mario_bros.actions import SIMPLE_MOVEMENT

video_dir_path = 'mario_videos'
env = gym_super_mario_bros.make('SuperMarioBros-v0')
env = JoypadSpace(env, SIMPLE_MOVEMENT)
env = gym.wrappers.RecordVideo(
    env,
    video_folder=video_dir_path,
    episode_trigger=lambda episode_id: True,
    name_prefix='mario-video-{}'.format(time.ctime())
)

# run 1 episode
env.reset()
while True:
    state, reward, done, info = env.step(env.action_space.sample())
    if done or info['time'] < 250:
        break
print("Your mario video is saved in {}".format(video_dir_path))
try:
    del env
except Exception:
    pass
```

## 3.2【baseline 跑通】15 分

跑通所给定的 baseline 代码:

（1） 学习深度学习框架（PyTorch）的基本使用方法;

（2） 配置 baseline 环境;

（3） 训练出能够通关简单级别关卡（1-1）的智能体;

（4） 评估指标：对于训练好的智能体模型，设置多个 seed，在每个 seed 下运行多个 episode，然后把所有 seed 的所有 episode 的分数取均值；

*在实验课上会讲 seed 对于环境的重要性；大体而言需要在更多样性的环境下取得好的效果，以减轻过拟合与随机性的影响。*

（5） 查看训练和测试结果：

i. 训练：

a) 查看损失曲线，判断是否收敛；

b) 和监督学习的损失曲线相比，强化学习的损失曲线有什么特点？可能原因是什么？

c) 怎样的 Q 值变化表明当前训练正常？

ii. 测试：

a) 算法在多少 env_step 后收敛？

b) episode return 能达到多少？

c) 通关用时如何？

d) （附加）你的智能体吃到了多少 coin、蘑菇、花朵？怎么让智能体学会去吃这些道具，而不是只考虑通关？*如果有这方面的分析和实现可以加分。*

（6） 查看得到的智能体的回放，分析智能体遇到了哪些问题？

## 3.3【特征空间处理尝试】20 分

特征空间处理在强化学习中包含对于观测空间（observation space）、动作空间（action space）与奖励空间（reward space）的设计。特征空间处理方案对智能体的能力至关重要。

本部分会给出一些特征空间处理的方案以及对应的代码实现。请尝试在代码中整合不同的特征空间处理方案，并完成给定的任务。

A. 不同特征空间的一些典型的处理方案例子如下：

i. Observation space：多帧堆叠、放缩观测图像、跳帧等；

ii. Action space：降低动作空间复杂度，比如只有“向右”、“向右同时跳跃”两个动作，可以降低训练难度；

iii. Reward space：为了告诉智能体需要一直向右边走才能通关，设置一个向右边位移即可获得的奖励。

B. 完成以下任务：

（1） 阅读所提供的代码查看实现，了解所给出的特征空间处理方案的细节；

（2） 分别验证所给定的每个特征空间处理方案的效果；

（3） 将经过自己验证的有效特征空间处理方案组合起来，跑通代码，构成新的智能体；

（4） 上述所构建的智能体，达到什么样的水平？

### 3.4【结果分析】20 分

在上一步工作的基础上，阅读所提供的智能体性能的标准化分析过程的案例，并完成以下任务：

（1） 模仿所提供的标准化分析过程的案例，尝试对给定的方案进行效果分析；*注意：不需要每一组参数都分析，可以选择有代表性或你想要分析的参数与 wrapper 组合，从 TensorBoard 结果曲线、评估视频与 CAM 激活图三个方面出发进行分析；另外，由于视频无法放入实验报告与海报，可以对你认为有意思的部分进行截图放入到实验报告或海报中即可；*

（2） 分析加上所提供的特征处理方案后，当前的智能体仍然存在哪些问题？如何改进？

*如需更好了解智能体性能的标准化分析过程，感兴趣的同学可以看看这篇论文：DRLIVE*

### 3.5【加分项：特征空间处理深入】≤10 分

（1） 根据进一步的提示，尝试自己修改、加强或添加特征处理方案；

（2） 按照前面提供的分析方法，分析自己实现的特征处理方案的效果。

### 3.6【加分项：算法深入】≤10 分

（1） 根据所给出的算法文章，或者自己调研的文章，从算法角度做出改进。

## （4）实验报告与海报展示

### 4.1【实验报告】10 分

（评价实验报告撰写是否规范、内容是否全面丰富、逻辑是否清晰、重点是否突出）

（1） 实验报告可以以中文撰写、也可以以英文撰写。要求重点突出、逻辑清晰。

（2） 实验报告的格式参考正式的 paper，建议包括：

报告题目：基于强化学习的超级马里奥兄弟游戏 AI 设计

个人信息：包括小组成员的姓名及学号；具体专业方向（不能只是电子信息）；电子邮箱

中文摘要及关键词

英文摘要及关键词

引言

1. 强化学习基础知识（这里 1 为建议编号，下同）（可细分为子章节，下同）

2. 文献综述

3. 算法总结

4. 马里奥环境

5. 智能体训练

6. 实验结果及分析

7. 结论

8. 所完成的加分项（以表格方式给出所完成的加分项，并给出实验报告中的对应子章节索引）

9. 成员分工及贡献比（以表格方式给出，可以按照具体要求中的项目划分，也可更加细分）

10. 心得体会

参考文献

附录：给出包括实验报告在内的大作业相关文件清单及相应说明 （即上传到网络学堂的文件内容；代码可放于一个目录，并对该目录作说明）

（3） 实验报告的表格、图片等要给出相应的表题、图题，并顺序编号，并在正文中相应地方给出引用；参考文献应在正文中给出相应的引用。

## 4.2【海报展示】10 分

（老师/助教/同学互评的加权成绩：包括海报的美观度、工作亮点总结、汇报展示的效果等）

（1） 请每个小组准备一张海报，应当包括报告题目、小组成员姓名、学号、具体专业方向（不能只是电子信息）、电子邮箱等;

（2） 海报内容：除了基本算法/模型的介绍之外，应突出自己工作的亮点部分：可以是模型的亮点、实验结果的亮点、除了基本要求之外完成的加分项的亮点、甚至是实验报告撰写的亮点、实验结果呈现形式的亮点、心得体会的亮点等等，总之能够凸显自己工作特色的所有东西都可以作为亮点给出来。

（3） 完成大作业后，会花一次课程的时间，让大家在课堂上分组展示和介绍自己的海报（需打印海报）。

（4） 海报电子版需使用 pptx 格式准备，设置为 A0 大小。

（5） 海报电子版需在规定时间（具体时间请等待通知）之前上传到网络学堂。

## （5）在截止日期前上传实验结果

将实验报告、海报电子版 pptx 文件、所复现代码以及相关说明文件（如代码运行环境需求说明、代码运行方法说明等），打包成一个 zip 文件上传到网络学堂。

请在截止日期前上传实验结果。否则将按以下公式扣分：

$$S' = S \times \min(0.85, 0.95^D)$$

其中，$S'$是迟交作业的评分，$S$是作业的原始得分，$D$是向上取整的迟交天数（超过 deadline 后即记为迟交一天）。例如：作业的 deadline 是 10 月 11 日，10 月 12 日补交的作业评分为原始作业得分的 85%，10 月 18 日补交的作业评分将被折合为原始作业得分的 69.8%。

# 其他大家关心的问题

Q1：训练所需的计算资源和时间

A1：对于涉及到的关卡，有一块普通 GPU 的情况下，都能在数个小时内训练收敛。

Q2：训出通过关卡的智能体有什么具体限制吗

A2：只要在时间限制内（非常宽裕的时间）通过游戏关卡即可，其他游戏属性（比如金币收集，马里奥成长都可以不考虑）

Q3：是否可以使用已经训好的智能体或者人类玩这个游戏的相关数据

A3：不可以，本次实验主要是理解强化学习如何从零开始，在与环境的交互探索和利用中在线学习，不能使用离线强化学习或者模仿学习的方法。

Q4：可以使用人工设计的规则吗

A4：可以探索人工规则和强化学习的结合方法，但纯规则的代码是不可以的。

Q5：我觉得所提供的 baseline 代码不行，我可以完全自己实现来完成任务吗？

A5：可以，如果自己能力够强，可以不使用所提供的 baseline 来完成任务，最后是按照任务完成情况给分的（但请不要提交 GitHub 上的其它开源代码，会有查重过程）；

Q6：我遇到了问题怎么办？

A6：请放松随意地在课程群里提问，会有助教进行回答。

Q7：我能做完全部的加分项吗？

A7：非常鼓励有兴趣的同学自行尝试加分项的内容，但加分项最多只能累计 20 分。

Q8：如何获取训练所需的 GPU 资源？

A8：获取 GPU 资源三种可行渠道：

（1）自己电脑的 GPU 资源。

（2）实验室服务器/独显机器：推荐。开发环境熟悉，GPU 资源使用时长可控。

（3）学校提供的计算资源：学校提供小组账户(一组一号，可同时登陆/跑作业)，每组提供配置为 2 核 Intel6348CPU，2*V100 GPU(23GB)，100G 存储的机器，可用时长为 54h。