

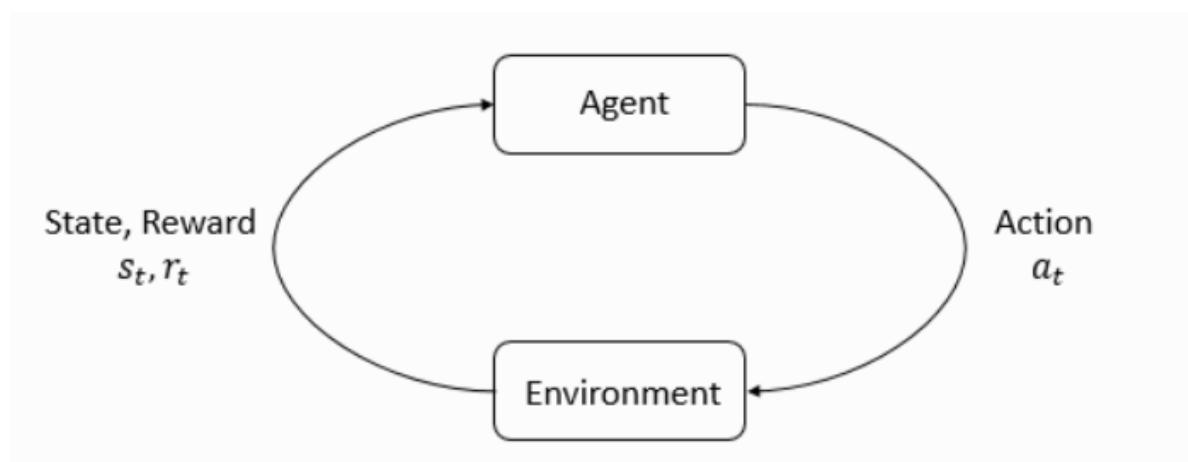
坦克大战作业Tutorial

看了之前的文章，相信大家已经对神经网络和强化学习有了一定的了解。在坦克大战这个小项目中，我们将结合强化学习与神经网络，训练一个Agent（智能体）去玩坦克大战这个游戏。

这个作业的目的**不是**让大家去掌握强化学习的学习算法或者学习如何搭建、训练神经网络等技术上的细节，**而是**希望大家可以从宏观的角度了解神经网络是怎么和学习算法结合起来的，以及如何利用自己对游戏、对技术的理解去让你训练的坦克更智能。

Part 1. Recall：强化学习&神经网络

首先让我们再一次简单回忆一下强化学习这个概念，回想一下什么才是强化学习的最终目的--



上图是强化学习的一张示意图，其只涉及到两个主体 -- Agent & Environment。而我们的最终目的是让Agent学会如何在环境中行动，以最大化他累计的Reward。更加具体一点的说，强化学习的最终目标是让这个名为Agent的玩意儿学习到一个**映射（or 函数）**，这个映射的输入是State，输出是Action，这个映射需要最大化**一段时间内**（如 一局游戏）**累计的Reward**（有些小伙伴在这里可能会有小问号：为啥是累计的Reward而不是Reward？**一段时间**有点含糊，啥意思？这个问题我会在最后说明）。





之前的文章已经提过了，学习映射，尤其是复杂的映射，就可以用神经网络。所以，这里的Agent它就是一个神经网络！强化学习中有很多学习算法去操作这个神经网络，让他能更好的学习由state到Action的**映射**，以最大化累计的Reward。

Part 2. 你需要做的

虽然学习算法在作业中是给定的，**但是State和Reward是需要你自己设计的**，而如何设计他们很重要！举个例子，Agent就像是学生，他的学习方式（学习算法）是固定的，但是他的考试题是由State决定的，他的考试目标是由Reward决定的。比如你希望他就考60分，多一分少一分都不行，那么当他考60时的Reward应该大于他考其他分数时的Reward。如果你设计的State很好，就像给Agent出了一张只考一个知识点的卷子，他很容易就精确控分了。

但是实际情况中，Agent考试的目标往往不是那么明确的，他的目标很有可能是考哈哈哈佛，但是你只能根据Agent的考试分数设置奖励。因此，根据我们丰富的知识，我们知道，哈哈哈佛是世界顶级大学，大概率收分高的不收分低的，因此我们聪明滴给Agent设置这样一个Reward：你考多少分，我就给你多少Reward，反正你往高了考就对了！这个Reward就很清晰，配上你给他精心设计的考卷，他就考上了。

Part 3. Environment -- 坦克大战

Environment通常是给定的，在这个项目中，Environment就是坦克大战这个游戏。坦克大战这个游戏在运行过程中，可以接受6种动作：开火，, , , , 以及啥也不干（是的，啥也不干也是一种动作--）。坦克大战这个游戏限制死了Agent可以采取的动作，但是坦克大战游戏本身可没有State和Reward的概念。那，他们是啥？

这里我们先定义一个简单的概念：‘步’（也就是图中各个标记的下标 t）。在每一步，你训练的Agent会采取一次动作，Environment 反馈State和Reward给Agent。在本项目中，一步约等于真实游戏中的八帧，这不重要。重要的是，你要注意到，当Agent做出一步动作后，整个游戏状态就变了，变成了八帧后的状态。所以这里的状态，你可以定义的很灵活，如坦克的位置，游戏进行时间等。至于奖励，比如你希望坦克别原地不动就行，那你的Reward可以是他动了就reward=+1，开火或者啥也不干就reward=-1。

最后说明

1) 为啥是累计的Reward而不是Reward？

答：根据Reward的定义，他是当前动作的一个即时的奖励，如果我们只注重当前的Reward，Agent可能为了眼前的利益放弃长久的利益，即竭泽而渔。

2) “一段时间”是如何定义的？

答：这个就要看任务了。比如围棋中，一段时间可以定义成一局游戏。在养鱼场景中，如果老板命令你3天必须3t鱼，以后不管了。这时，一段时间就是3天，为了3天的最大利益，Agent选择了竭泽而渔。换一个老板，他希望你可以在未来100年总收益最高，这时，一段时间就是100年，为了100年的收益，Agent选择了填湖造房。