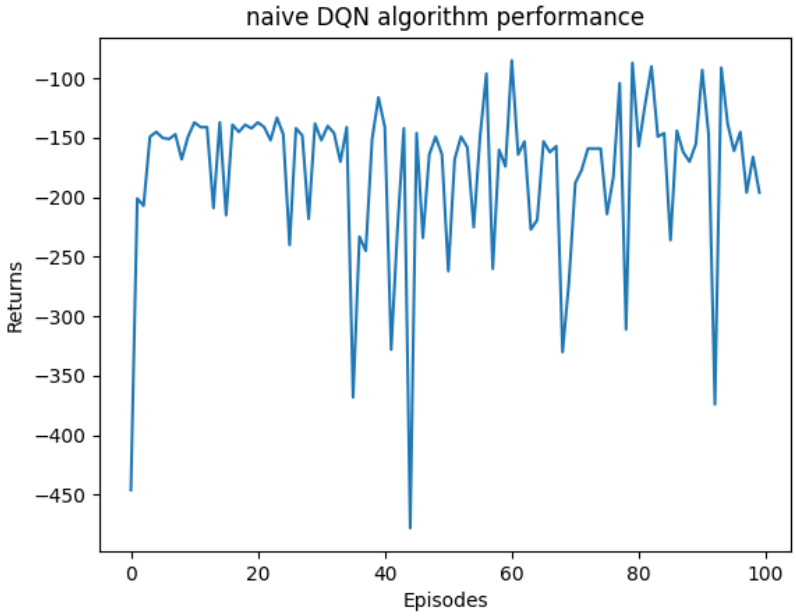
强化学习作业 A4 DQN

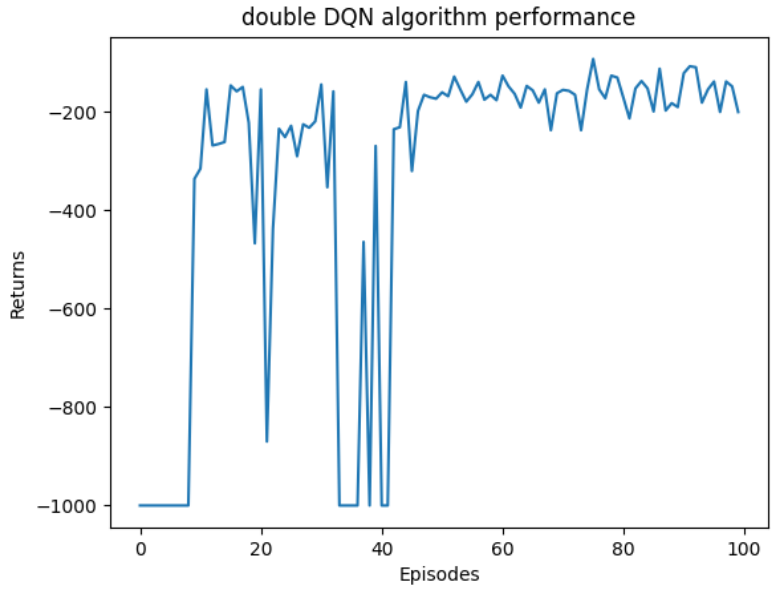
叶增渝 123033910090

本次采用了传统的 DQN 网络与 DoubleDQN 网络。

由于 MountainCar-v0 下，除非达到终点可以得到 reward 为 0，否则在每一个位置上得到的 reward 均为-1，不利于网络训练拟合，所以重新设计了 reward，根据小车实际到达的高度，当高度大于 0 时，会给予一个与高度相关的 reward，当达到终点时，会给予一个 1000 的高 reward 以保证训练的有效性，我们记此 reward 为 virtual_reward。

下图为 2 个 DQN 的训练的实际 reward（在环境中，真实的 reward 代表了步数，我们设置训练时的 maxstep 为 1000）：





由于任务简单容易学习，可以看到 DQN 网络可以在前几个 episode 就可以快速完成收

敛，然后就能顺利地完成任务，完成任务的所需步数在 100~200 步之间。

而由于我们设计的不够完美 reward 不够完美，可能导致 DQN 在左右来回晃荡但不到达终点，以此来刷分，如 DoubleDQN 中间几个 episode 所示，但是最终还是能稳定在 200 步以内。

我们不难发现 DoubleDQN 的训练过程比普通的 DQN 更加曲折，这是由于任务足够简单，导致两者没有什么分别，但 DoubleDQN 的机制使得其收敛速度比较慢，所以需要更长时间收敛。

最后我们对两个记录下来的 best_model 进行 10 次的测试（测试条件较为苛刻，需要在 200 步内完成任务），普通 DQN 的成绩为 4 次通过，平均实际 reward 为 -182.4；而 DoubleDQN 的成绩为全部通过，平均实际 reward 为 -124.5。

综上来看，DoubleDQN 的最终效果较好，但两者均有完成任务的能力。

代码与对应的模型均放在文件中，可以直接使用 ipynb 的最后一个 cell 进行测试（但要记得将前面的环境与函数定义运行一遍，训练过程可以不用运行）