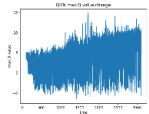
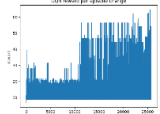
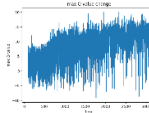
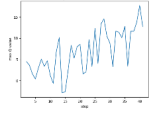
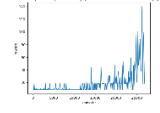
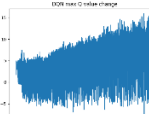
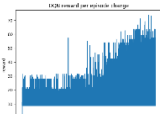
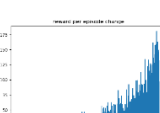


Breakout 实验结果

目前，已经完成和正在进行实验共有 7 组，实验参数如下：

实验编号	实验概述	干预概率 P	开始干预时间	是否保存原始数据	训练步数	最大 Q 值变化结果图	每轮游戏累计奖励变化图
01	原始 DQN 模型	0	无	否	3 百万		
02	加入人工干预	0.5	训练刚开始时就干预	否	3 百万		
03	延迟加入干预	0.5	训练 1 百万步后再开始干预	是	4 百万		
04	原始 DQN 模型	0	无	是	4 百万		
05	干预概率递减	干预概率在训练过程中从 0.1 逐渐递增至 0.5		是	4 百万		
06	加入人工干预	0.4	训练刚开始就干预	是	6 百万		
07	无模型，随机	0.5	训练刚开始就干预	是	6 百万		

- 人工干预是指根据小球的位置来移动挡板，相当于是一个“正确答案”，当干预概率是 1 的时候，挡板每次都可以接到小球，游戏可以无限地玩下去。
- 干预概率就是选择这个人造的“正确答案”的概率。
- 原始数据是指模型跑出来的那些结果数据，也就是用来画图的那些点的坐标。第一次和

第二次实验缺乏经验，所以是把画图的代码和训练的代码写到了一起，训练完就直接画图了，未保存原始的结果数据，无法二次处理数据和重新绘图（除非再花 3 天时间再跑一遍）。后面的实验都保存了原始的数据。

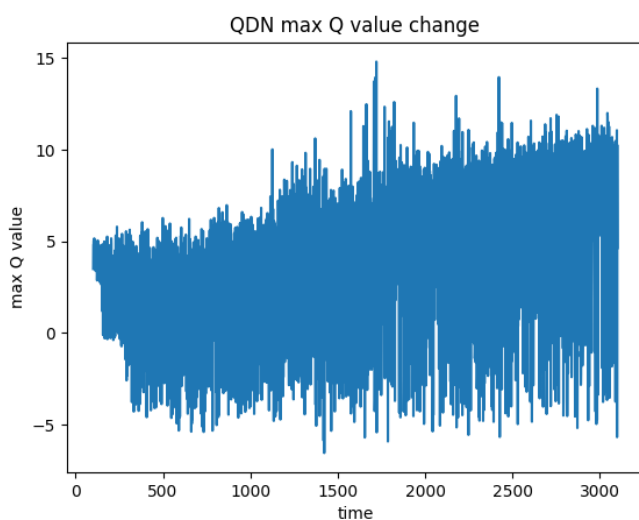
- 训练步数是模型一共进行训练的步数，前面是 3 百万次，但通过实验结果发现 3 百万次的时候模型的最大 Q 值没有出现明显的收敛，后面就把步数扩大到 4 百万步，但还是没有明显的收敛，现在扩大到了 6 百万步看看结果如何。

实验 1~5 的结果图如下（实验 6 和实验 7 还在跑）：

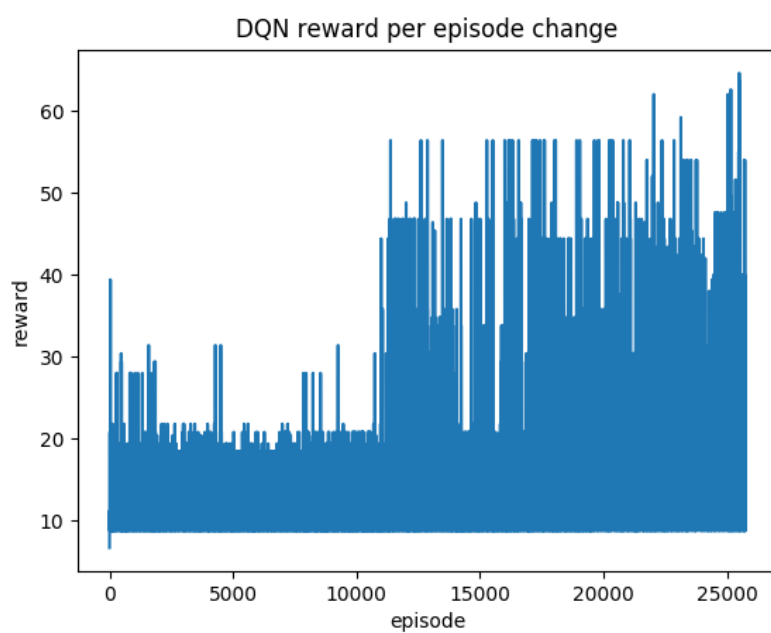
实验一：

原始 DQN 模型，无干预，训练 3 百万步，未保存原始数据。

最大 Q 值随训练步数(时间)的变化(每 100 步记录一次,所以实际上横坐标要大 100 倍)：



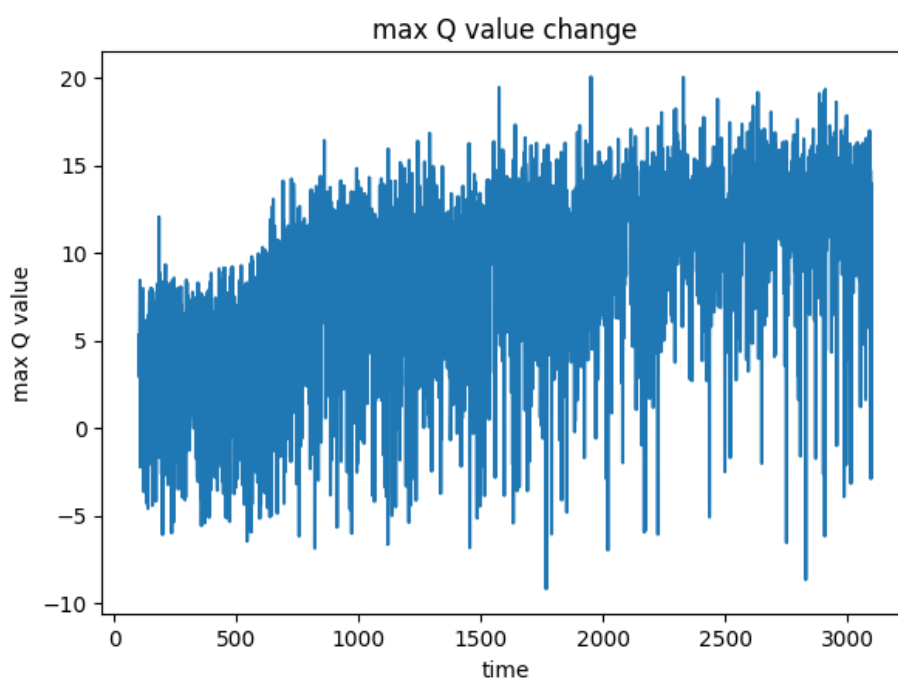
每轮游戏累计奖励的变化：



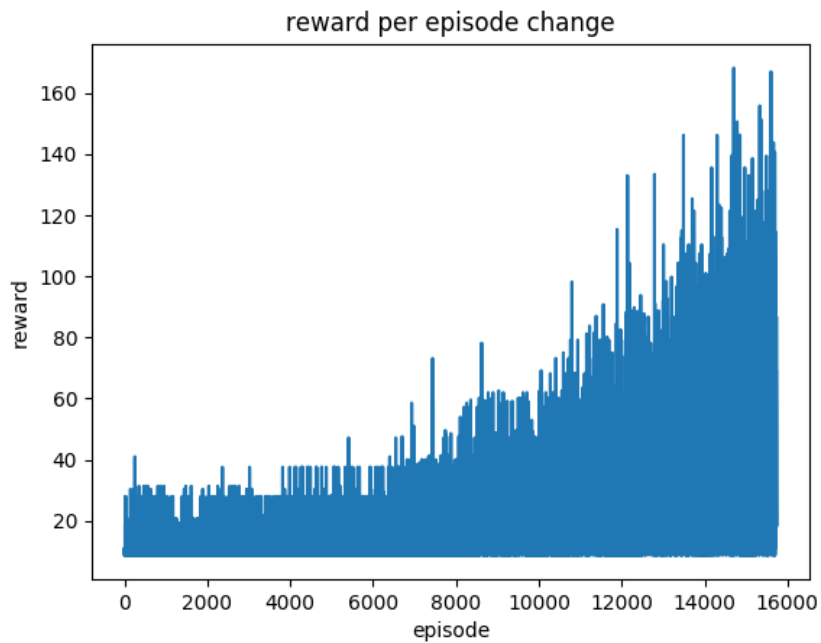
实验二：

从训练开始就加入概率为 0.5 的人工干预，训练 3 百万步，未保存原始数据。

最大 Q 值随训练步数（时间）的变化：



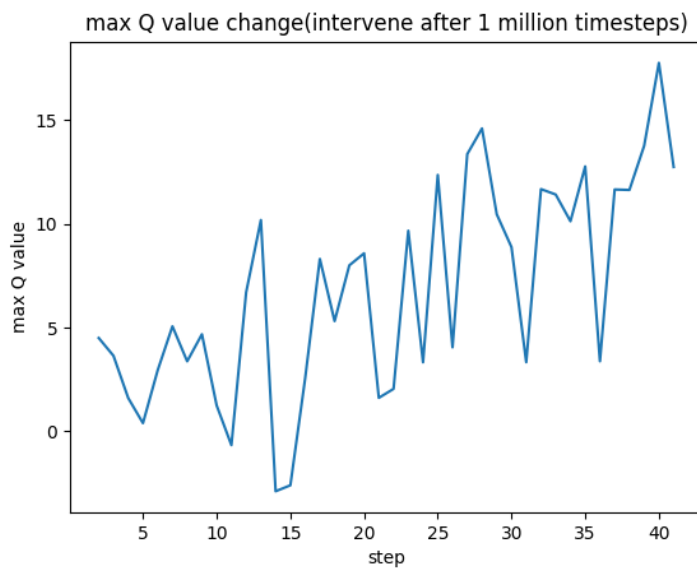
每轮游戏累计奖励的变化：



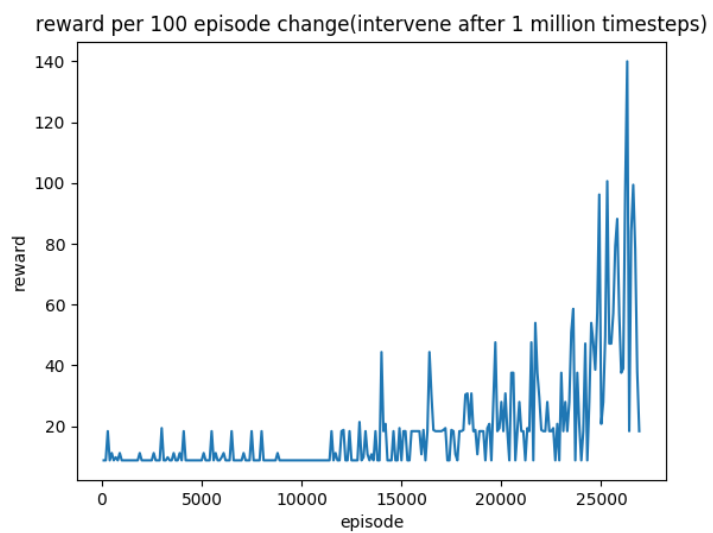
实验三：

以 0.5 的概率，在训练了 100 万次以后再干预，训练 400 万步，保存了原始数据。因为觉得之前的图太密集了，有点不好观察，所以把稀疏程度扩大了 100 倍，但事实证明不是疏密的问题。

最大 Q 值随训练步数（时间）的变化：



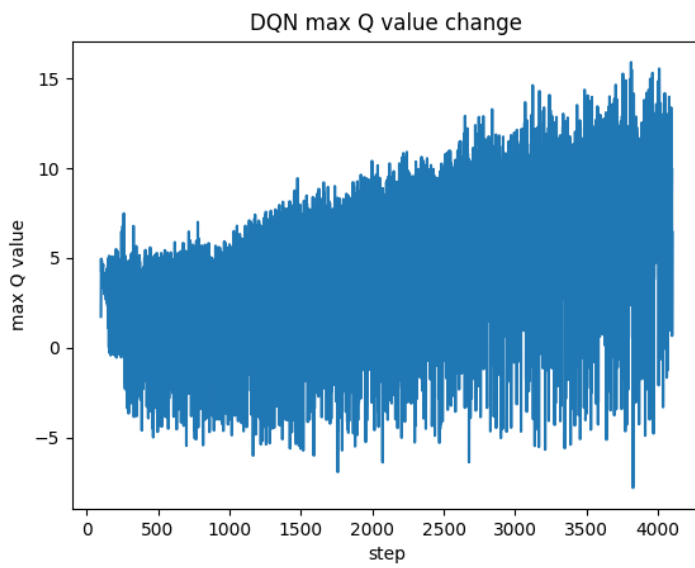
每轮游戏累计奖励的变化：



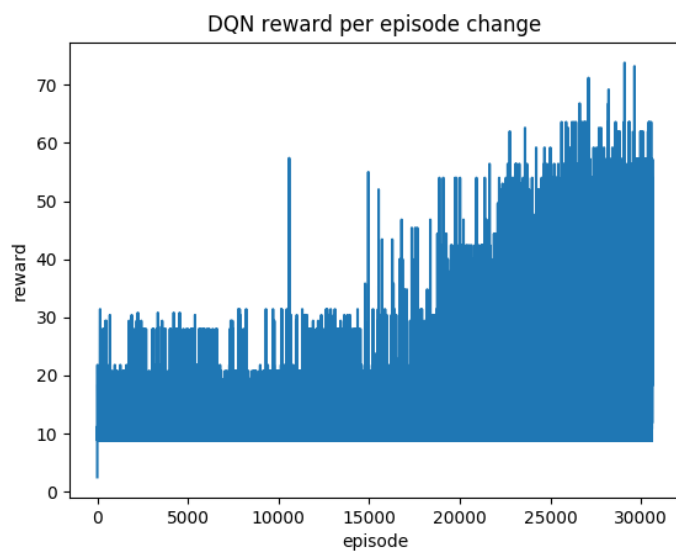
实验四：

把实验一重新做了一遍，以保存原始数据。

最大 Q 值随训练步数（时间）的变化：



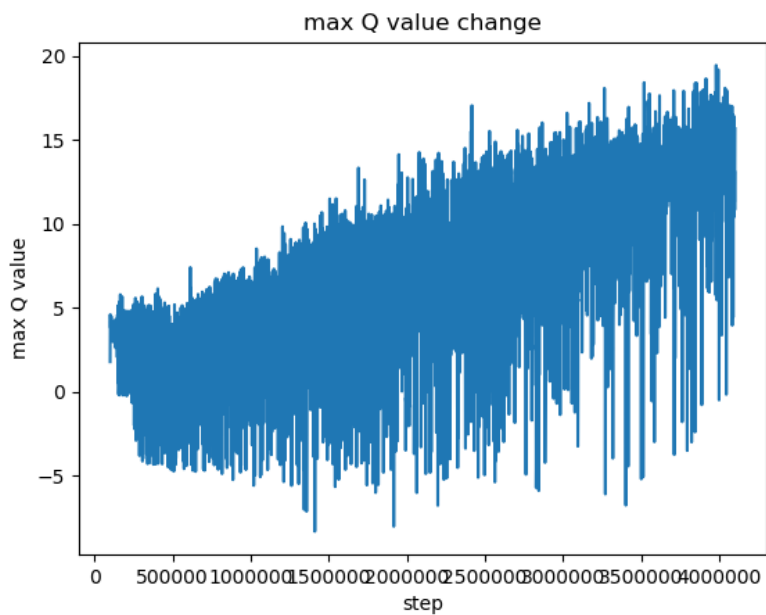
每轮游戏累计奖励的变化:



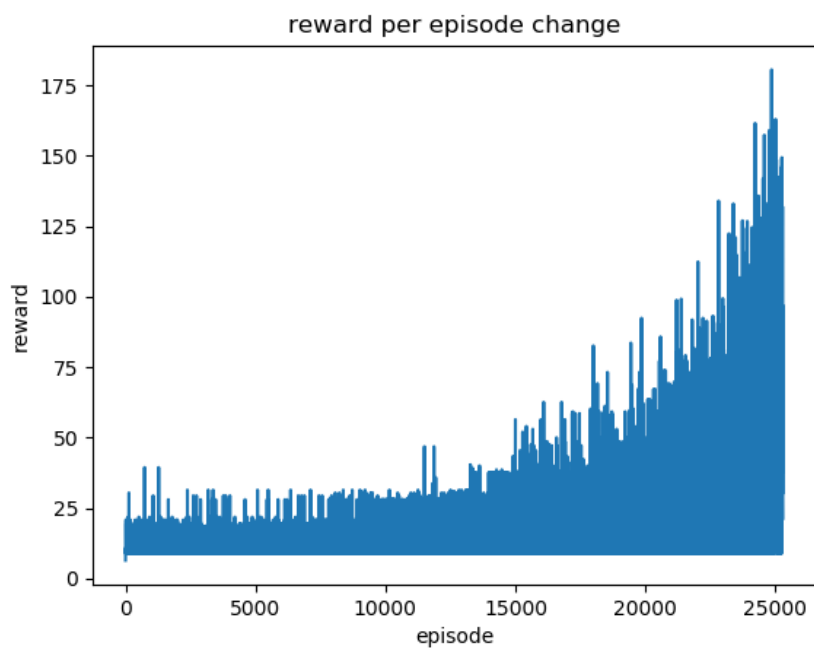
实验五：

干预概率在整个训练过程中从 0.1 逐渐递增至 0.5，共训练 400 万步，保存了原始数据。

最大 Q 值变化：



每轮游戏累计奖励变化：



总的来说，从结果图可以看出，加入规则的模型无论是最大 Q 值的变化还是每轮游戏的累计奖励都是要比原始的 DQN 模型要好一些的。

目前的问题有几点：

1. Q 值的变化是一直在波动的，不像 flappy bird 那样稳定，导致对比观察起来不是很直

观。具体原因，我觉得应该是因为游戏的原因。Flappy bird 游戏的反馈更即时点，也就是说，一个动作，其引起的结果的好坏（reward 大小变化）可以在接下来很短的时间内就可以显现出来，所以最大 Q 值的变化比较平稳。而 breakout 游戏的奖励存在很大的滞后性，小球的运动是需要时间的，在小球的运动期间挡板所做出的动作的好坏，要等到小球打到砖块或者死掉以后才能显现出来，动作与奖励之间的滞后较大，这就导致了 Q 值的变化在整个游戏过程中一直都是在波动的。但每个波动周期平均的最大 Q 值是不断增加的。波动周期应该与小球从挡板运动到砖块的时间有关。所以，我目前解决整个问题的想法就是看看能不能找到大致的波动周期，然后对原始数据进行处理，通过求出每个波动周期的平均 Q 值来重新绘图，结果应该会更加清楚更好比较点。这个在下周之前可以完成。

2. 另外就是每局奖励的问题，因为人为干预相当于是给了一个“正确答案”，虽然加入干预的模型比未加入的模型的每局的最好奖励要高一倍以上，但是到底是因为模型本身有提高的原因还是因为给的“正确答案”的原因，或者说除开“正确答案”对奖励的提高，模型本身对奖励的提高到底有多少？所以，这个问题我准备再做一个空白对照，不用模型，就以 0.5 的概率选择正确答案的动作，0.5 的概率随机选择动作，看一下结果怎么样。这个应该也能在下周之前完成。
3. 还有每局累计奖励波动的问题。即使训练到后期，可以打游戏打到很厉害，但有时候模型还是会因为一些失误而死掉，但和前期是有区别的。前期训练的时候，游戏死掉的时候挡板与小球的落点很远，但根据观察，后期训练到很好了以后，游戏死掉的时候小球与挡板的位置其实是很接近的，很多时候就在旁边一点点的位置，所以游戏到后期累计奖励还是会因为很接近成功的失误而有一些的波动，而这个失误其实是不能反映在游戏死掉之前的一系列的动作的好坏的（也是由于游戏延迟奖励的原因），因为按照给的奖

励来说，挡板离小球落点很近和挡板离小球落点很远是一样的，都没接到球。这就导致了游戏的奖励的波动和不稳定。影响最后结果的观察和比较。目前还没有想出比较好的方法来解决整个问题。