Dopamine DQN采用的是Off-Policy的策略，可以用之前所有的经验数据，所以不像On-Policy需要基于当前的策略去进行产生数据。Off-Policy只需要从Buffer的经验数据中进行重要性采样就行。需要做的就是去设置一个Buffer的大小。

并行的好处在于可以让没个Agent都去产生数据，然后把数据汇总起来去寻找策略梯度或者去估算Value Function

现在要做的就是如何去加reward或者去给agent一个惩罚，惩罚它一直做一个动作，这个是很不合理的。

另外就是要提升一些探索性，感觉现在很容易陷入一个动作的局部最优