1. Implement Advanced RL algorithm
   1. Choose one algorithm from REINFORCE with baseline、Q Actor-Critic、A2C, A3C or other advance RL algorithms and implement it.
   2. Please explain the difference between your implementation and Policy Gradient
   3. Please describe your implementation explicitly (If TAs can’t understand your description, we will check your code directly.
2. 我使用一般的Actor Critic 來提升我的performance
3. 與Policy Gradient 不同的是Actor Critic會多learn一個value network 來當作critc，評估當前的reward的好壞，也就是說Actor Critic 結合了Value-based (例如 Q learning) 和 Policy-based (例如 Policy Gradients) 這兩類的方法。
4. 使用MC的方式來更新Value network，玩完一回合遊戲，將每一步reward和value network predict出的value收集起來，定義loss為 actor的loss + critic的loss，actor的loss為，critic的loss為，每玩完一回合出現done，同時更新Actor network 與 Value network，這裡的reward都有做normalize，discounted reward 的gamma = 0.99，optimizer使用 Adam，lr =0.02，

self.affine = nn.Linear(8, 128)

self.action\_layer = nn.Linear(128, 4)

self.value\_layer = nn.Linear(128, 1)，action\_layer壓到4，因為共有四個動作，action layer 的 output會進softmax function，做normalize。