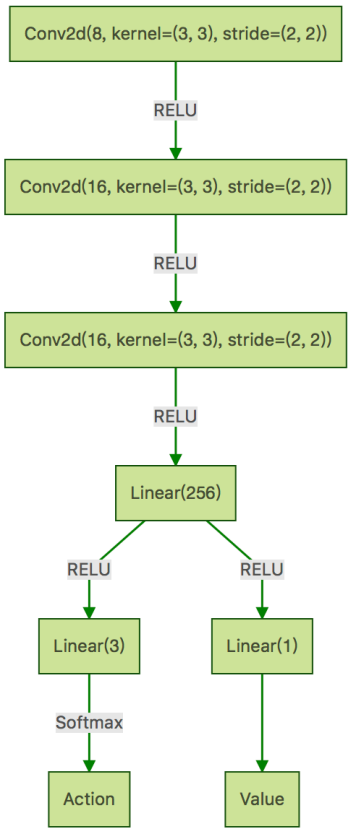
ADL HW3 Report

B04902004 王佑安

Model Description:

Policy Gradient



Actor-Crtic:

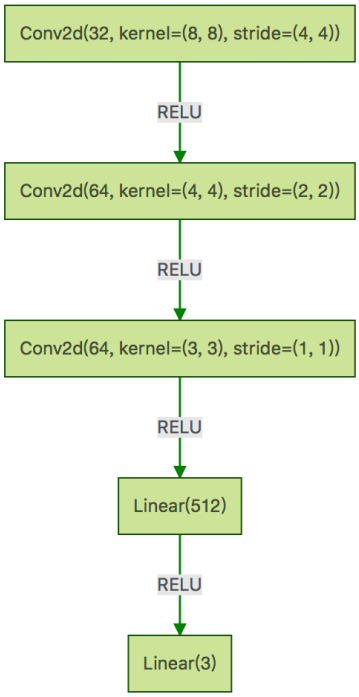
Optimizer: Adam

Learning rate: 1e-4

Gamma: 0.99

Initializer: Xavier initializer

DQN



Double DQN

Optimizer: RMSprop

Learning rate: 1e-4

Gamma: 0.99

Memory size: 10000

Batch size: 32

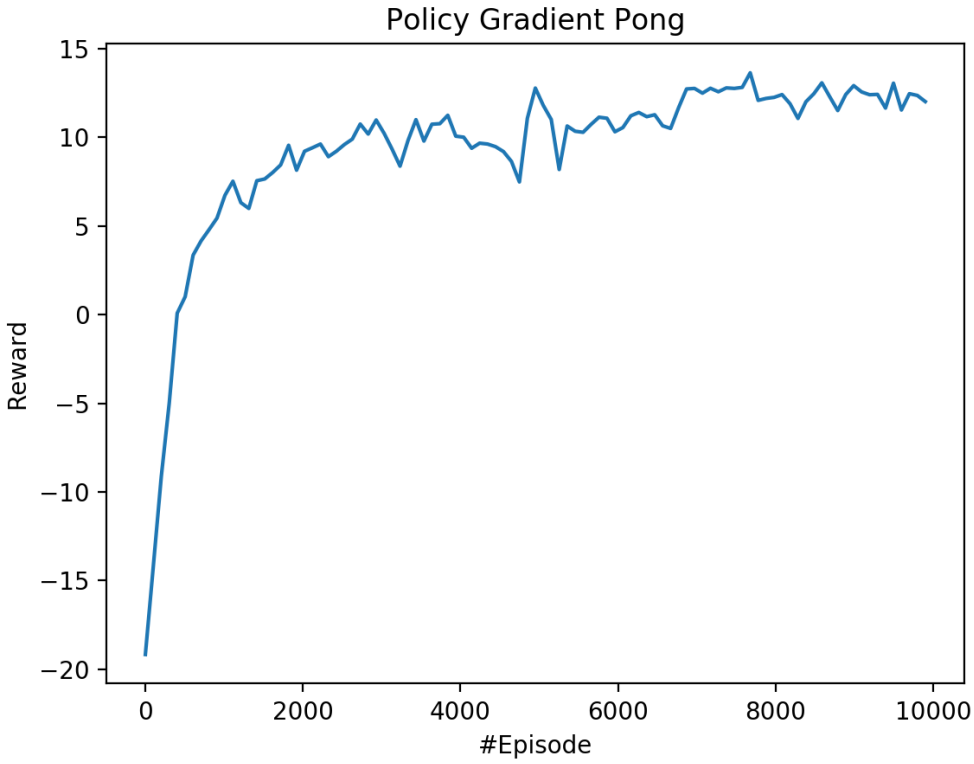
Target update freq: 1000

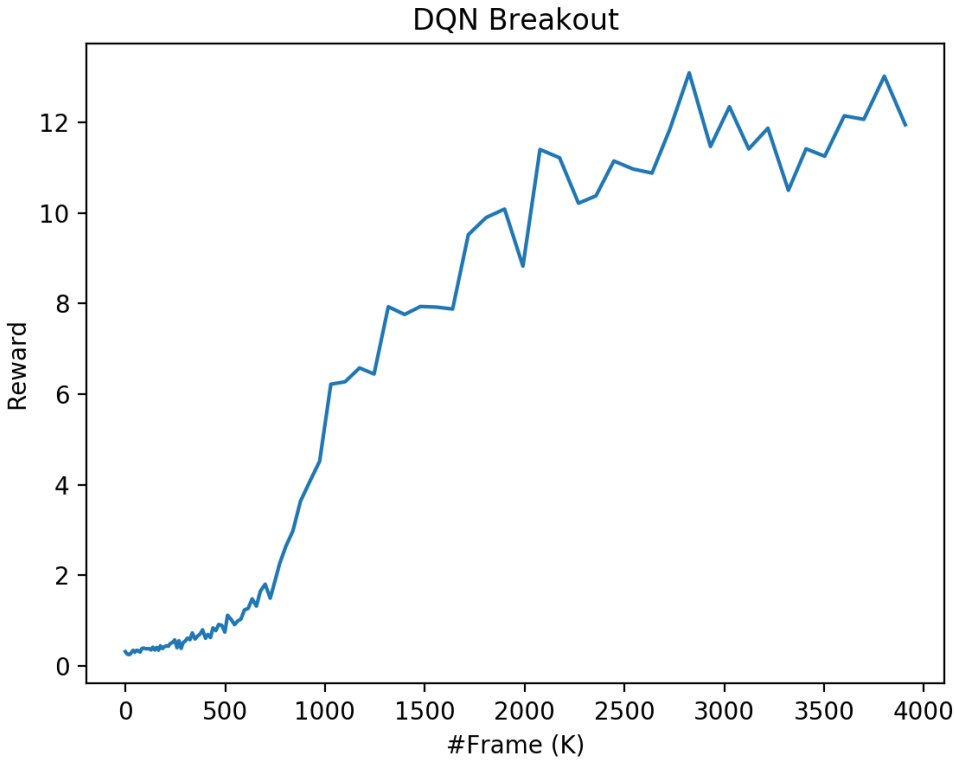
Online update freq: 4

Exploration: 1 to 0.05 in first 100000 frames

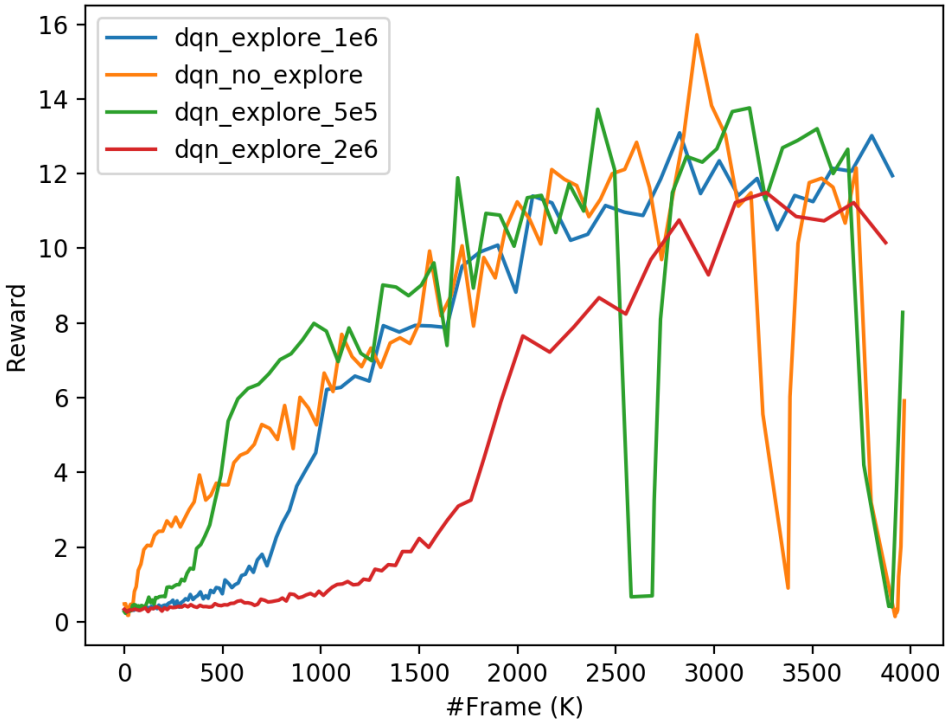
Initializer: Xavier initializer

Performance:





DQN Hyperparameters: Exploration rate



觀察 training 時 exploration rate 的下降對 reward 的上升感覺看不出什麼顯著的差異，因此決定測試不同 exploration rate 對最後整體表現的影響。這裡測試了 explore 前 500k、1000k、2000k 個 frame 以及 no explore 四種 setting。

從結果出 explore 2000k 收斂得最慢，而 500k、1000k、no explore 都在差不多的 frame 數 reward 超過 10。推測是因為 explore 太久，model 學到太多隨機出來的東西導致一開始的 weight 偏掉太多，後面得花更多時間學回來。

而另外觀察到 explore 較少的 500k 跟 no explore 的表現很不穩，認為可能是因爲前面 explore 到的 penalty 太少，導致 train 到後面還是會選擇錯的 step 才再學回來。

Improvement of DQN: Double DQN & Dueling DQN

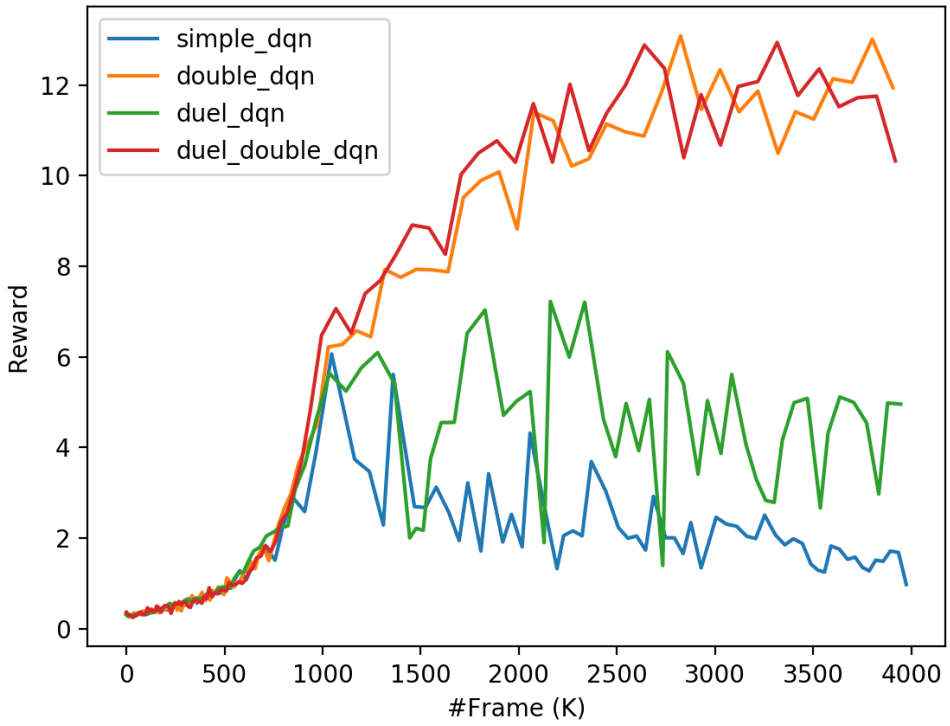
Double DQN:

原本的 DQN 的 target network output 直接選擇 Q value 最大的 action 去 optimize，而 Double DQN 用 online network output 的 action 決定要 optimize target network 的哪個 output，由於選擇的 Q value 都是 online network 實際選擇的，因此會跟實際的 reward 比較接近，可以解決 Q value overestimate 的問題。

Dueling DQN:

Dueling DQN 將 Q value 分成兩個部分，對 state 的估計 V(s)跟 action 的 advantage A(s,a)，而最後要 optimize 的 Q = V(s) – A(s,a)。Dueling DQN 多了一個對 state 的估計，network 可以得到的資訊更多，就可以收斂的更快。

Performance:



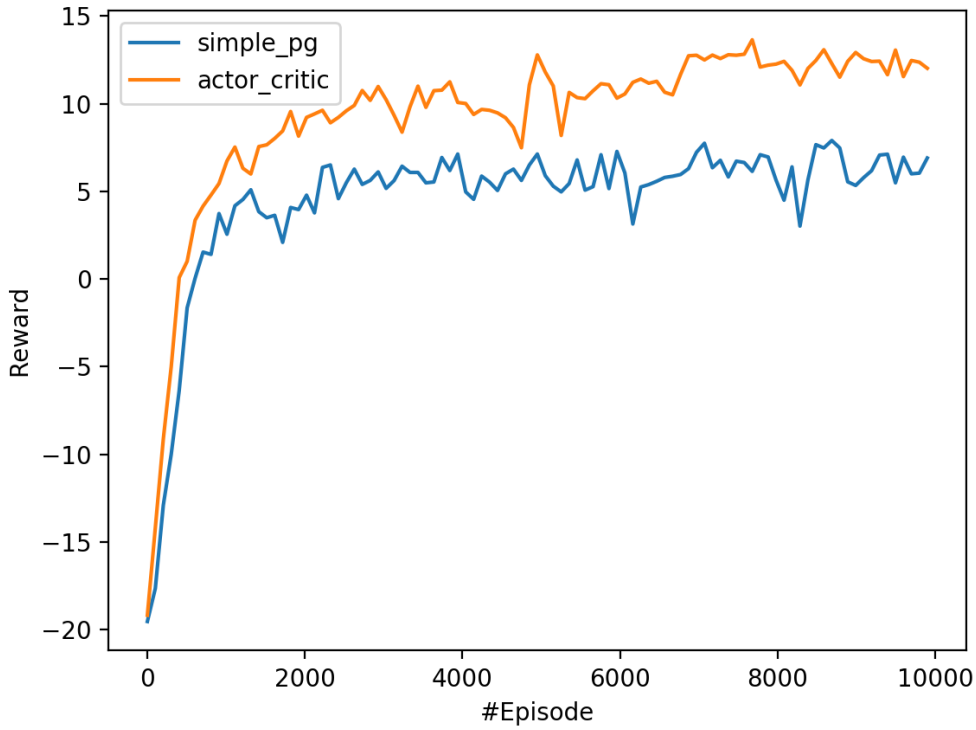
可以看出 Dueling DQN 跟 Double DQN 的表現確實明顯比原本的 DQN 好，尤其 Double DQN 進步的幅度比較明顯。而 Dueling DQN + Double DQN 收斂的速度稍微比 Double DQN 又快了一點，也證實了 Dueling DQN 收斂較快的性質。

Other RL Method Implementation: Actor-Critic

Actor-Critic 是將 Value based 跟 Policy based 結合的做法，跟原本的 Policy Gradient 的差異在於 network 的 output 多了一個 critic network 當作 function V，由 model 本身對這個 state 評分來決定 gradient 要 ascent 多少。而最後 gradient 的算式如下：

$$\nabla_\theta J(\theta) \approx \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\nabla_\theta \log \pi_\theta(\mathbf{a}_{i,t}|\mathbf{s}_{i,t})\left(Q(\mathbf{s}_{i,t},\mathbf{a}_{i,t}) - V(\mathbf{s}_{i,t})\right)$$

Perfomance on Pong:



可以看出在 pong 的表現上，actor-critic 如預期的比原本的 Policy gradient 明顯來得好。