隊長：R06922002 姚嘉昇　組員：R06942054 潘仁傑、R06522620 王仁蔚

0. 分工表 (0%)

	HW4-1	HW4-2	HW4-3	Report 4-1	Report 4-2	Report 4-3
姚嘉昇 R06922002						
王仁蔚 R06522620						
潘仁傑 R06942054						

1. HW4-1 Policy Gradient (5%)

- **Describe your Policy Gradient model (1%)**

兩層 Convolutional Layer，Filter 數分別為 32, 64，再經過 128 個 unit 的 FC Layer 後接 2 個 unit 的 Output Layer，因此 action size 為 2(up and down)，其他參數設定如下：optimizer = tf.train.AdamOptimizer (lr=1e-4, epsilon=1e-5)，Batch Size = 32。

表 1.1 Policy Gradient 的 Model 架構

Input：image (80, 80, 1)
conv2d(filters=32, kernel_size=[8, 8], strides=[4, 4], padding='same') + relu
conv2d(filters=32, kernel_size=[8, 8], strides=[4, 4], padding='same') + relu
Flatten()
Dense(units=128) + relu
Dense(units=2) + softmax

- **Plot the learning curve to show the performance of your Policy Gradient on Pong (1%)**

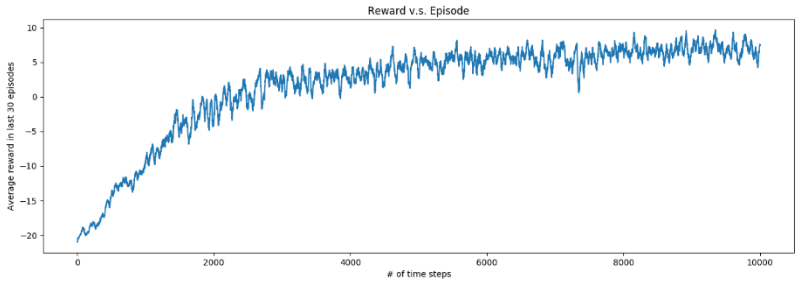


圖 1.1 Learning Curve of Original Policy Gradient

- **Implement 1 improvement method on page 8**
  - **Describe your tips for improvement (1%)**

    我們使用 Policy Gradient with Proximal Policy Optimization 2 (PPO2)增進 PG 的效果。並設定下面參數：clip_value=0.2, c_2=0.01，loss = loss_clip + c_2 * entropy。
  - **Learning curve (1%)**

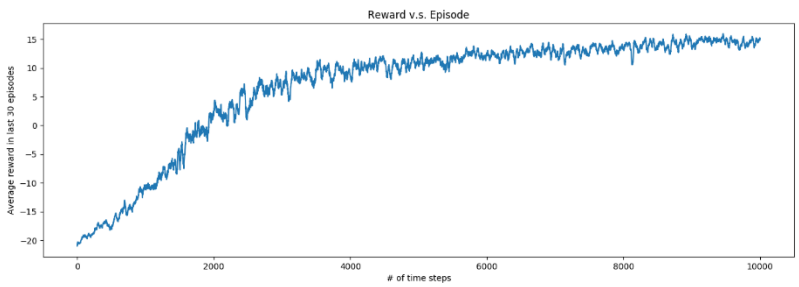


圖 1.2 Learning Curve of Policy Gradient with PPO2

- **Compare to the vallina policy gradient (1%)**

  根據下圖我們可以發現使用 PPO2 後，可以使 model 最後的分數從 5 左右進步至 15 左右。

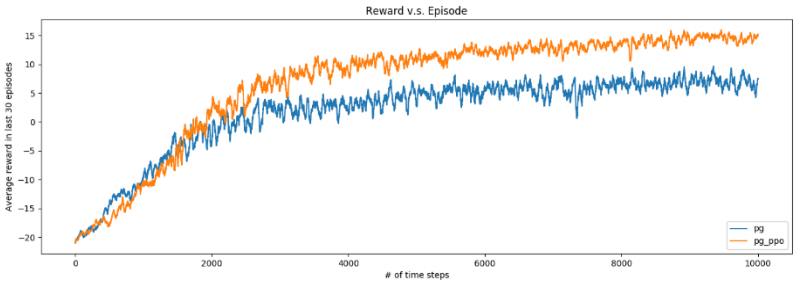


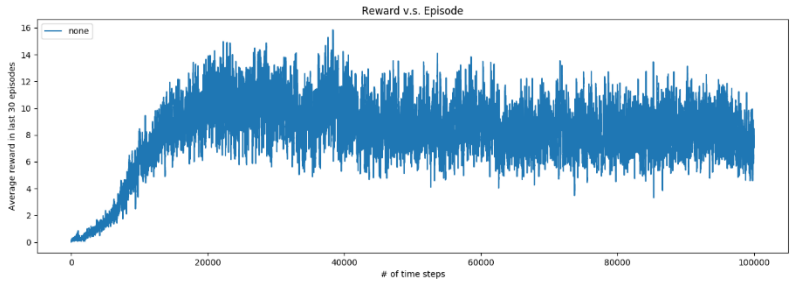
圖 1.3 Comparison of Original PG and PG with PPO

2. HW4-2Deep Q Learning (5%)

- **Describe your DQN model (1%)**

current network 與 target network 都有相同的架構，皆有三層 Convolutional Layer，Filter 數分別為 32, 64, 64，再經過 512 個 unit 的 FC Layer 後接 2 個 unit 的 Output Layer，因此 action size 為 2(up and down)，另外在決定 action 時，我們有 ε 的機率會採取 random action，而 ε 會從 1 隨著 step decline 到 0.025，其他參數設定如下：Replay Memory Size = 10000、Perform Update Current Network Step = 4、Perform Update Target Network Step = 1000、Learning Rate = 1e-4、Batch Size = 32。

- **Plot the learning curve to show the performance of your Deep Q Learning on Breakout (1%)**

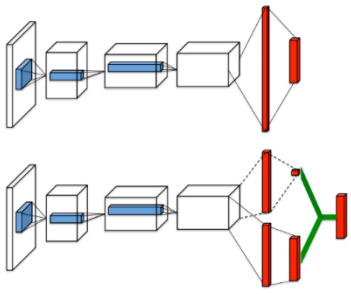


- **Implement 1 improvement method on page 6**
  - **Describe your tips for improvement (1%)**
    - **A. Double DQN**

      使用 current network 來決定 action，使用 target network 來計算 Q value，這樣不會高估 accumulated expected reward。 $Y_t^{\text{DoubleDQN}} \equiv R_{t+1}+\gamma Q(S_{t+1}, \operatorname{argmax}_a Q(S_{t+1}, a; \boldsymbol{\theta}_t), \boldsymbol{\theta}_t^-)$

    - **B. Dueling DQN**

      將 network 的 Output 層進行修改，將原先 Output 層只有兩個 unit 改為 3 個 unit，第一個 unit 成為 Value，後兩 units 為 Advantage，計算 Advantage-mean(Advantage)+Value 後成為最終的 Output。

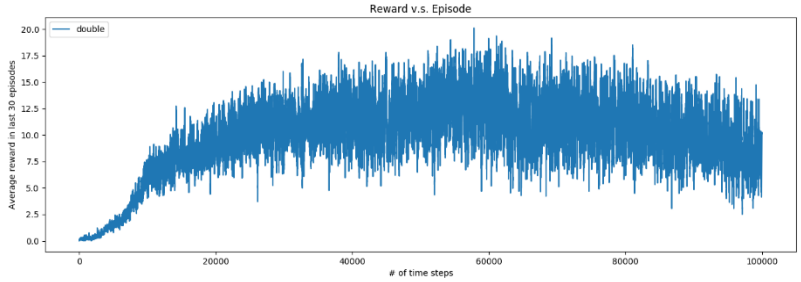


    - **C. Double Dueling DQN**

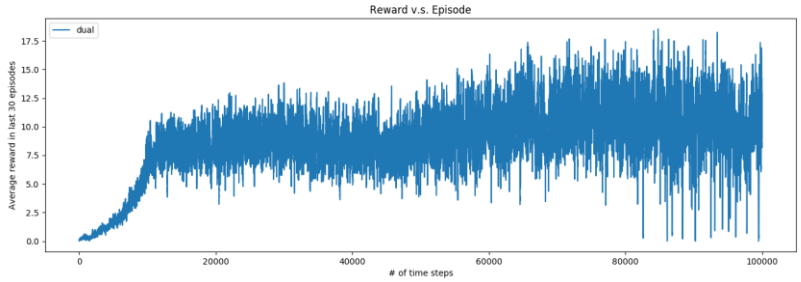
      綜合 Double 和 Dueling 的方法。

■ **Learning curve (1%)**

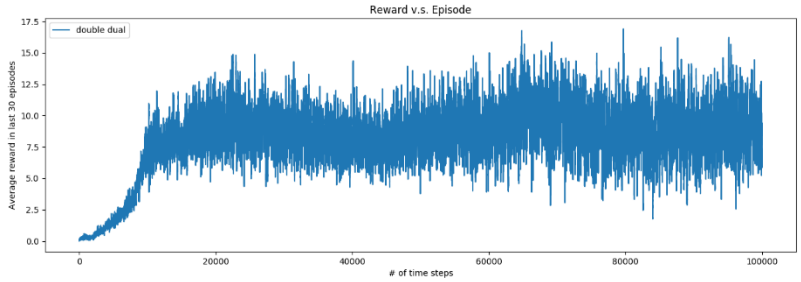
A. Double DQN



B. Dueling DQN



C. Double Dueling DQN



■ **Compare to origin Deep Q Learning (1%)**

A. DQN V.S. Double

Double 在剛開始和最終的 reward 並沒有與 DQN 差很多，顯示 Double 與 DQN 最終會收斂到差不多的結果，主要在中間段時，Double 會表現比 DQN 好一些，在不會高估 Q value 的情形下確實獲得的 reward 會高一些。

B. DQN V.S. Dueling

Dueling 和 DQN 並沒有顯著的差距，不過可以看到最後確實會收斂到一個較好的值，這與 Output Layer 拆解成 Value 與 Advantage 有關。

C. DQN V.S. Double Dueling

看來綜合兩種方法並未得到較好的結果。

3. HW4-3 Actor-Critic (6%)

- **Describe your actor-critic model on Pong and Breakout (2%)**

  Loss = loss_clip - c_1 * loss_vf + c_2 * entropy

  c_1 = 1，c_2 = 0.01

  Optimizer = tf.train.AdamOptimizer (learning_rate=1e-4, epsilon=1e-5)

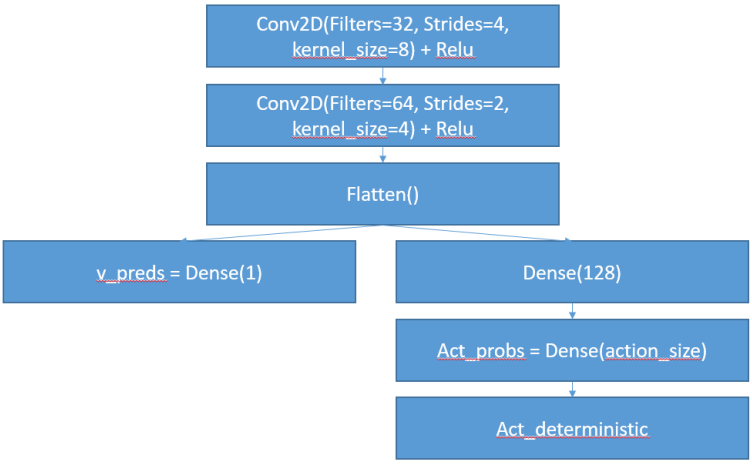


圖 3.1 Actor-Critic 架構圖

- **Plot the learning curve and compare with 4-1 and 4-2 to show the performance of your actor-critic model on Pong & Breakout (2%)**
  - **Compare with 4-1**

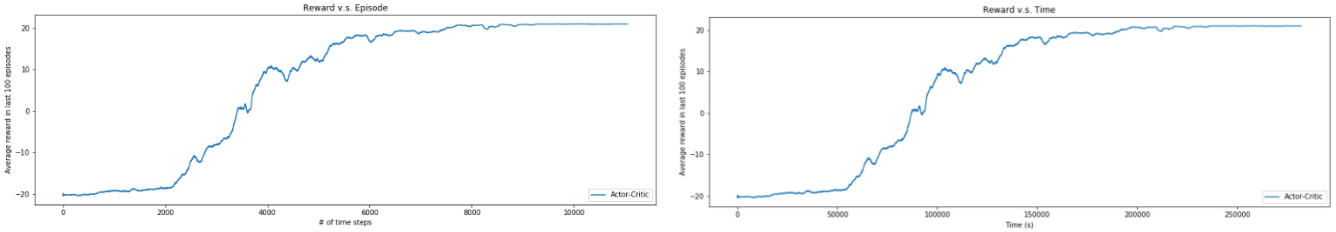
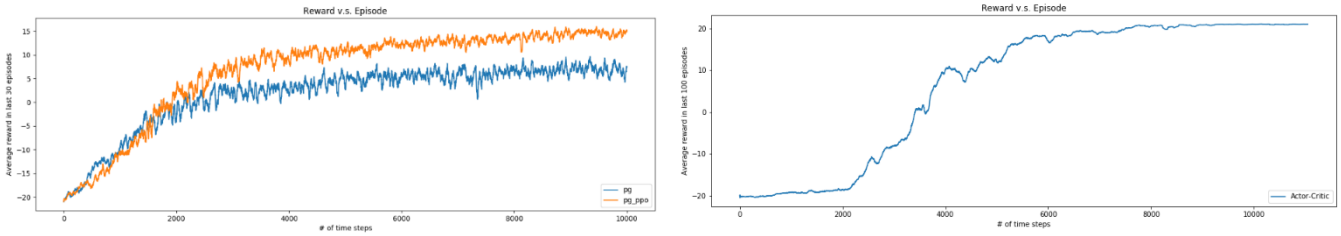


圖 3.2 Pong 之 Learning Curve AC (左 Reward vs. Episode；右 Reward vs. Time)



	PG 及 PG+PPO	AC
穩定性	較容易震盪	**較穩定**
收斂速度	較快	較慢
一萬步表現	將近 15 分	**滿分 21 分**

■ **Compare with 4-2**

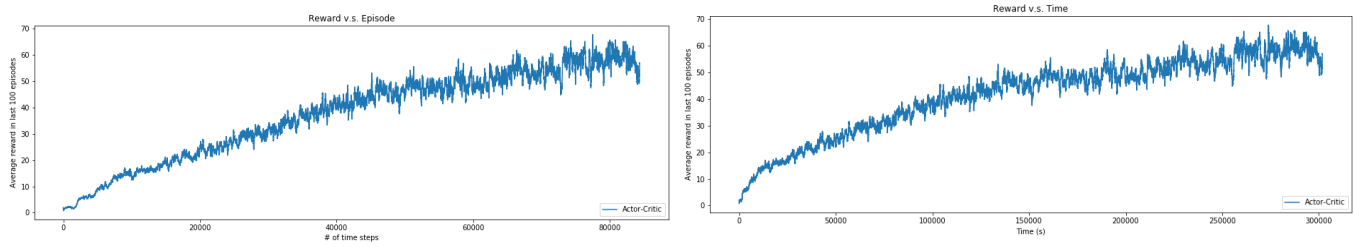
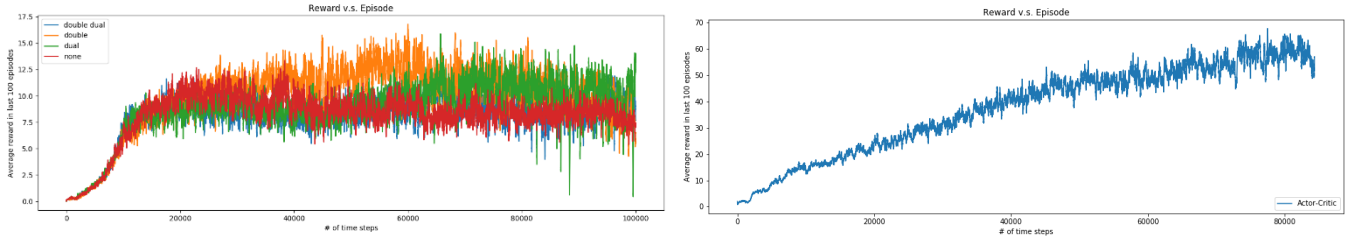


圖 3.3 Breakout 之 Learning Curve AC (左 Reward vs. Episode；右 Reward vs. Time)



	DQN 及其變形	AC
穩定性	較容易震盪	**較穩定**
收斂速度	慢	中
一萬步表現	5~10 分震盪(clip)	40~50 分震盪(unclip)

● **Reproduce 1 improvement method of actor-critic (Allow any resource)**

■ **Describe the method (1%)**

我們使用 A3C (Asynchronous Advantage Actor-Critic)這個 Tip。

■ **Plot the learning curve and compare with 4-1 and 4-2, 4-3 to show the performance of your improvement (1%)**

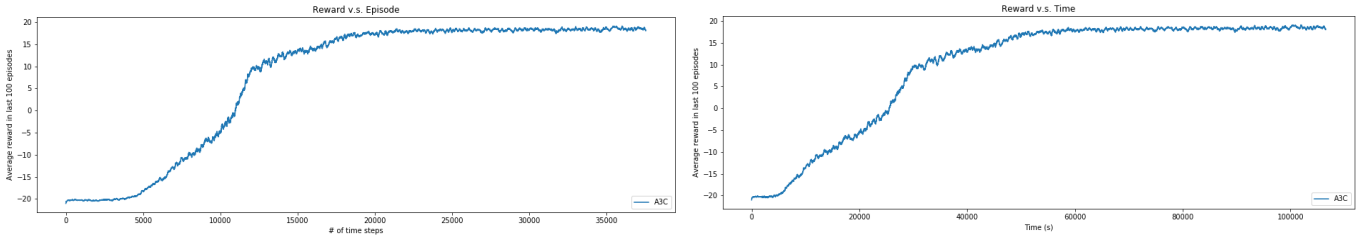


圖 3.4 Pong 之 Learning Curve A3C (左 Reward vs. Episode；右 Reward vs. Time)

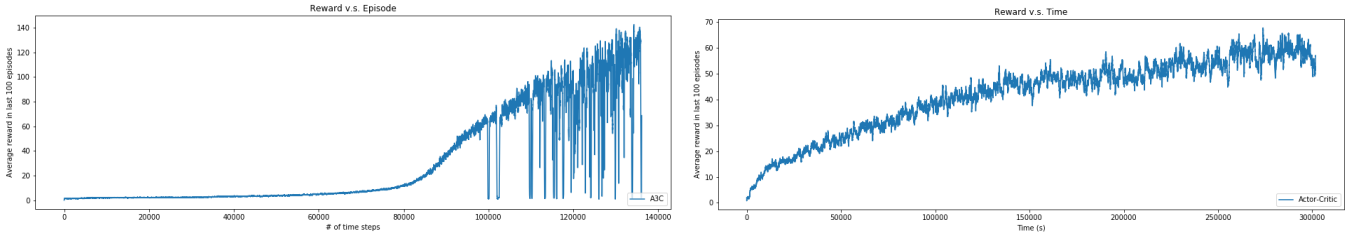
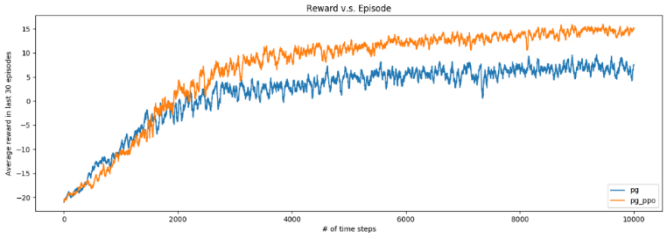
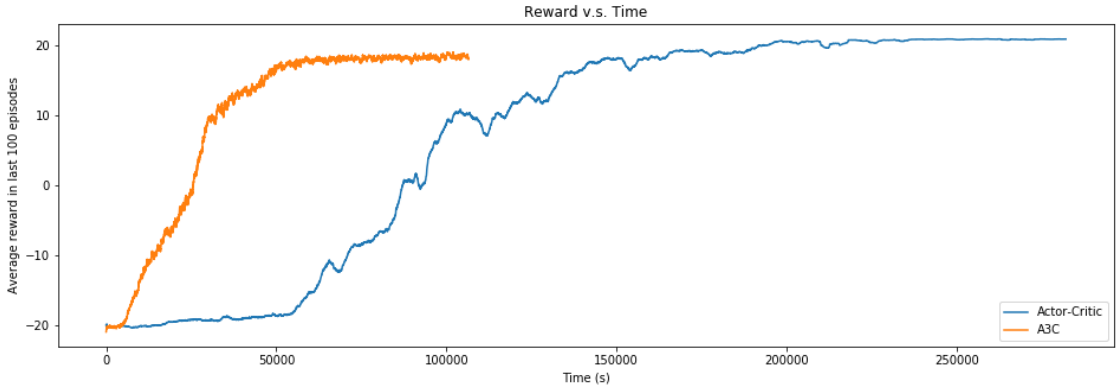


圖 3.5 Breakout 之 Learning Curve A3C (左 Reward vs. Episode；右 Reward vs. Time)

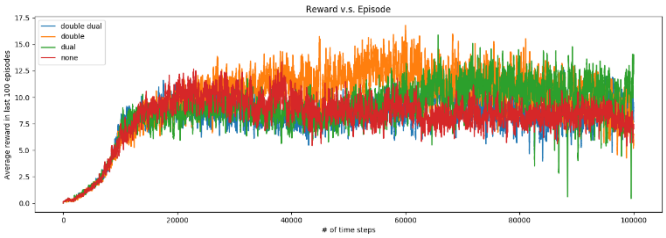
◆ Compare with 4-1 and 4-3

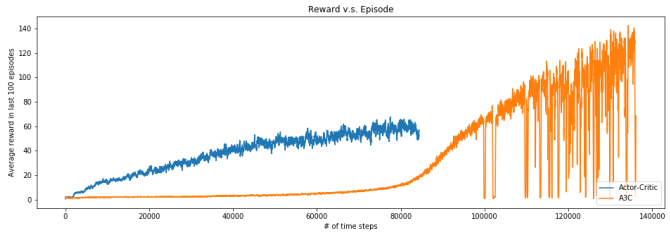


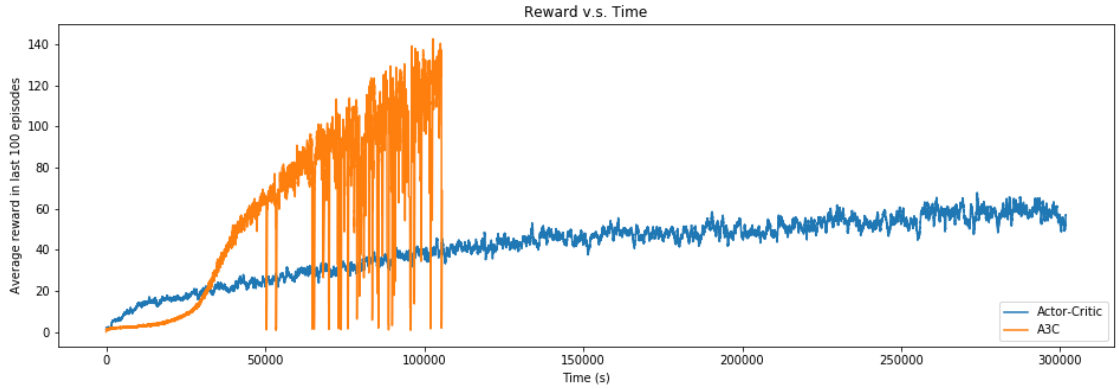


	PG 及 PG+PPO	AC	A3C
穩定性	較容易震盪	**較穩定**	較容易震盪
收斂速度	慢	中	**快**
最後表現	15 分震盪	**滿分 21 分**	17 分震盪

◆ Compare with 4-2 and 4-3







	DQN 及其變形	AC	A3C
穩定性	較容易震盪	**較穩定**	較容易震盪
收斂速度	慢	中	**快**
最後表現	5~10 分震盪(clip)	40~50 分震盪(unclip)	**100~140** 分震盪(unclip)