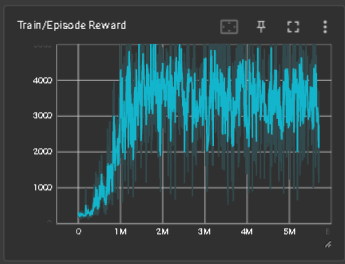
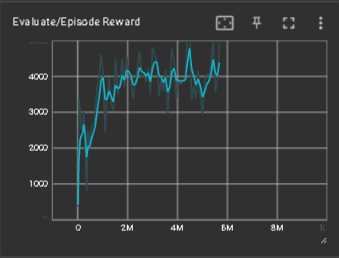
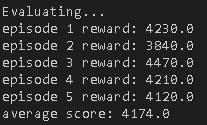
**RL Lab2-DQN  
學號:313554044 姓名:黃梓誠**

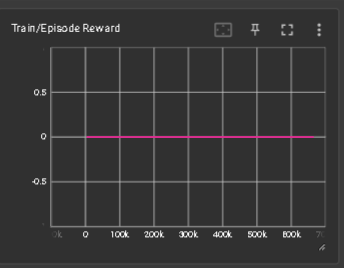
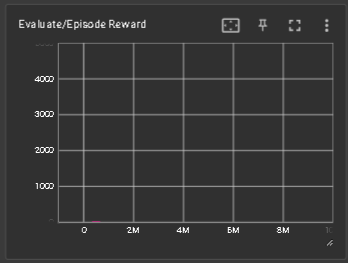
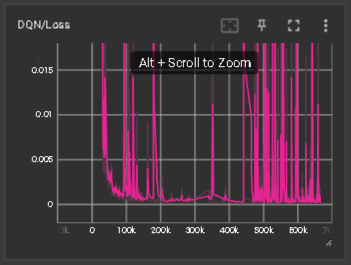
**Screenshot of Tensorboard training curve and testing results on DQN**

Training curve:

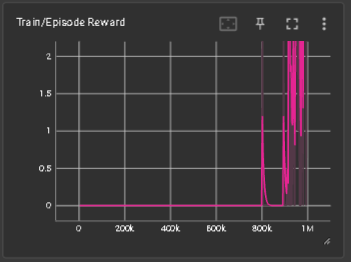
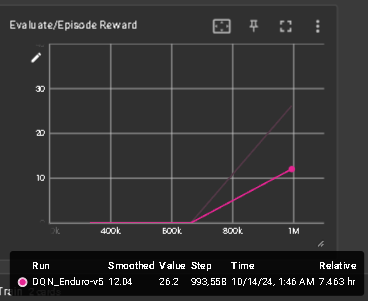
Testing results (5 games):



**Screenshot of Tensorboard training curve and testing results on Enduro-v5 using DQN (10%).**



6\* 10E5 epoch 後 : 開始有獲得零星reward



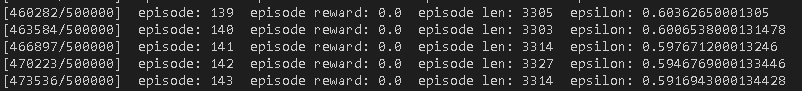
我在最後一兩天測試Enduro-v5 DQN，但發現他無法取得reward。以下是我分析的幾點原因與改進方向

1. Enduro-v5 中reward相當sparse：

Enduro是一款獎勵稀少的賽車遊戲。與Ms. Pacman頻繁獲得獎勵（吃點點數）不同，在 Enduro 中，agent僅收到過去車輛的獎勵。這意味著agent可能會經歷很長的sequence而不會收到任何獎勵，特別是在訓練的早期，它還不擅長超越其他車輛時。

可能的解決方案：確保在訓練的早期有足夠大的replay buffer和足夠的exploration（使用更高的 epsilon 值），以使agent有更多機會學習到罕見的獎勵。

2. **Long Episode Lengths**:



在log中，**Episode Length**相當大（大約 3300），但reward為 0。這可能代表agent在遊戲中沒有取得進展，可能只是在沒有超車的情況下駕駛。或許agent沒有學習到如何改進。

可能的解決方案：可以考慮調整獎勵結構

（例如，對停留時間過長而沒有超車的情況給予輕微的負獎勵），以鼓勵agent採取更快地超車的行動。

3.動作空間複雜度：

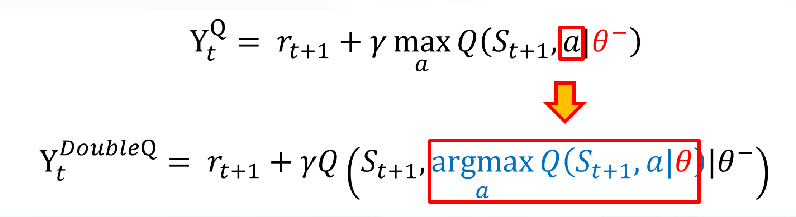
與Ms. Pacman相比，Enduro 的動作空間相對較小。這使得agent更容易探索所有可能的操作。然而，DQN 可能會在環境具有時間依賴性時陷入劣勢，意思是agent需要學習做出一系列動作才能超越汽車，而使用標準 DQN 可能需要更長的時間來學習。

可能的解決方案：使用Double DQN (DDQN)或Dueling DQN，等幫助agent做出更明智決策的能力。

4.Enduro 中的獎勵訊號太稀疏，DQN 可能很難學習。

可能的解決方案：嘗試獎勵標準化或獎勵裁剪，以確保獎勵在 DQN 可以處理的範圍內。

**◆ Screenshot of Tensorboard training curve and testing results on DDQN, and discuss the difference between DQN and DDQN (3%).**

觀念:

1. DQN suffers from over-estimation:

- DQN 使用同一個 Q network來選擇和評估最佳動作，這會導致高估 Q 值，並進而影響決策品質。

2. Behavior and Target Network (行為網絡與目標網絡)\*\*:

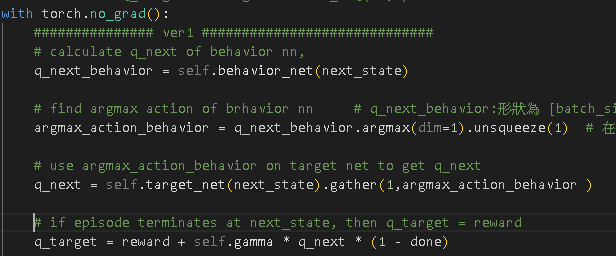
- DDQN 使用兩種networks：

1. behavior network : 用於根據當前策略選擇動作

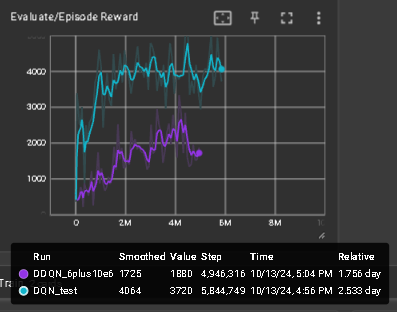
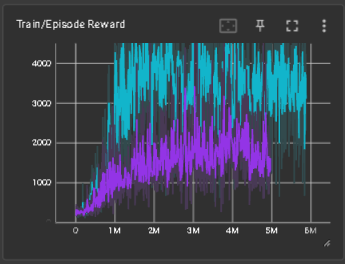
2. target network : 用於提供更穩定的 Q 值估計。目標網絡更新的頻率較低，以保持穩定性。

3. Reduce the over-estimation problem:

在 DDQN 中，行為網絡選擇最佳動作，而目標網絡則評估該動作的 Q 值，這樣可以降低 Q 值被高估的可能性，並改善學習效果。

實作:

實作中的結果意外發現DQN在一開始的訓練反而比較好，因為設備緣故我無法訓練完1e8 epoch，我認為DDQN較差可能是訊練時長不夠所以沒有發揮效果。

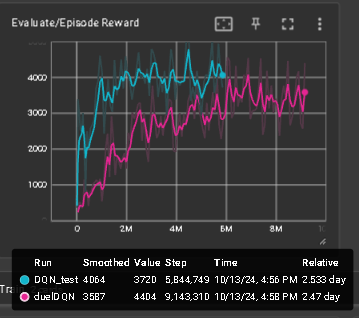


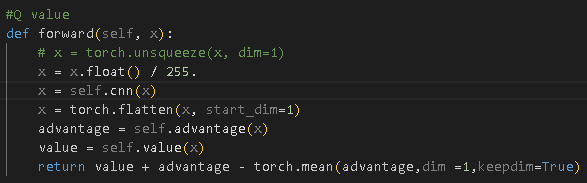
**◆ Screenshot of Tensorboard training curve and testing results on Dueling DQN,and discuss the difference between DQN and Dueling DQN**

觀念:

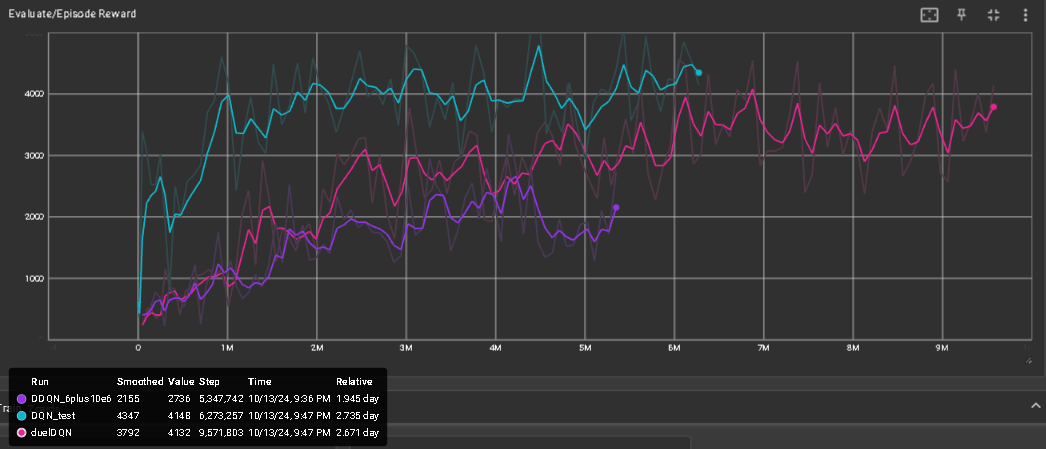
1. 在 Dueling DQN中，Q-value 被分成兩個獨立的分支：一個是Advantage function，另一個是 Value function。Advantage function代表在給定狀態下每個動作的相對重要性，而 Value function則表示該狀態的整體價值。
2. 由於A值的數值範圍相對較小，對模型更新更加敏感，使得模型更容易考慮與其他動作相關的相對變化。
3. 另外，Dueling DQN包含一個機制來限制 Advantage function的值，從而避免 Q-value 中的誤判。這是用減去所有動作的 Advantage值的平均來實現的。這確保了 Q-value 能夠更好地被normal，並更有效地反映動作之間的**相對差異**，避免 Advantage值可以通過加上一個常數，而不改變 Q-value 的問題。

實作:





可以看到duel是穩健上升的而DQN雖然比較快上升但接著就一直卡了，如果時間尺度拉長或許可以有更好的效果

**◆ Training curve comparison (DQN vs. DDQN vs. Dueling DQN)**