應用強化學習與生成網路於自動駕駛訓練

實驗介紹

前言

1. 自動駕駛在2025年美國消費性電子展(CES)中，成為全場最受注目的焦點。
2. 強化學習（Reinforcement Learning）是讓智能體(Agent)從環境中不斷試錯，學習策略的 AI 演算法。
3. 因應高維度問題，採用 DQN（Deep Q Network）取代 Q-Table；為了使DQN訓練更穩定，加入經驗回放（Replay Buffer）與固定Q目標（FixedQ）。
4. 現實中自動駕駛中運用感知融合與 VAE 技術，故本研究利用該演算法展示深度強化學習在遊戲與自動駕駛領域的應用潛力，並為未來研究提供基礎。

訓練環境介紹

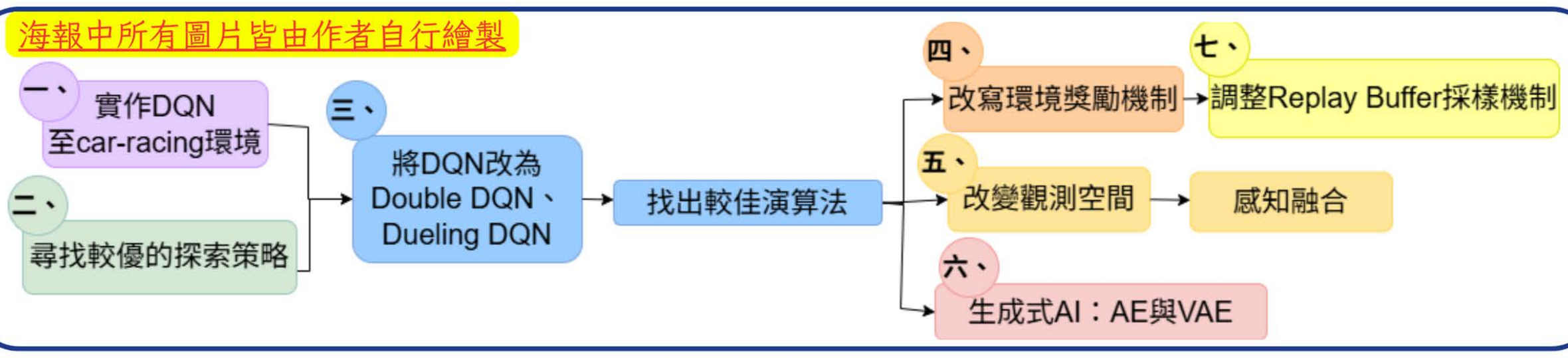
1. 使用環境：OpenAI Gymnasium 的 Car Racing
2. 狀態空間：為寬高各96像素、含有RGB的三維影像
3. 動作空間：可採取5種離散動作
4. 獎勵機制：每一幀 -0.1，每踩一塊賽道格子 +1000/總賽道格子數
5. 終止條件：當賽車開完賽道或駛離地圖

研究設備及器材

1. Nvidia Jetson Orin 16GB：訓練模型使用的遠端平台及GPU
2. Python3.10：程式的主要撰寫語言
3. Pytorch：搭建及訓練網路的框架

研究過程或方法

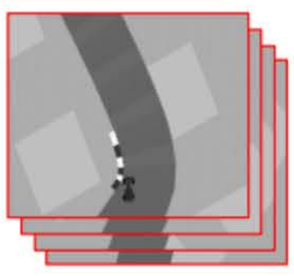
海報中所有圖片皆由作者自行繪製



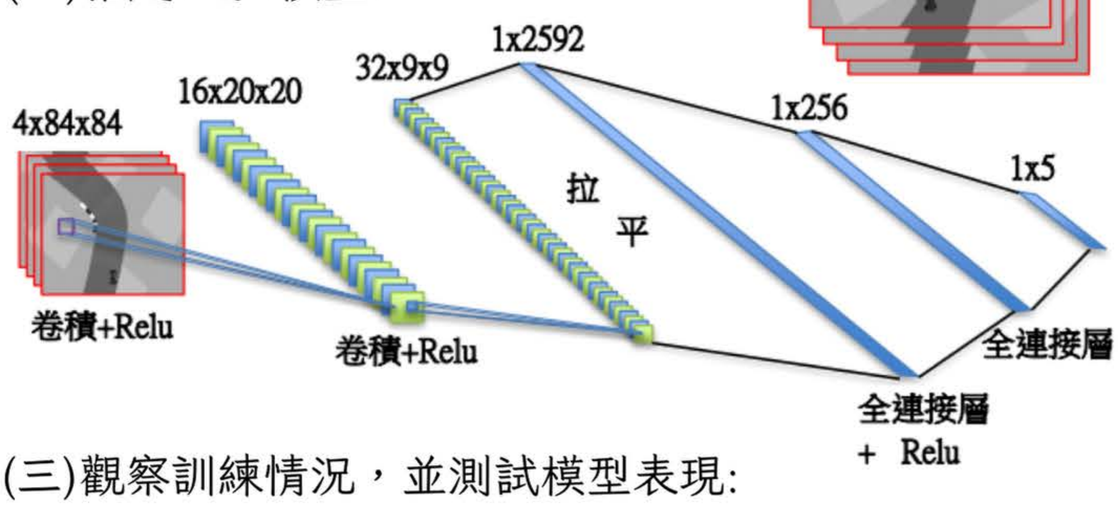
一、實作DQN網路，解決Car_racing模擬環境

(一)圖片預處理:

裁掉底部訊息條，轉為灰階4幀圖片



(二)搭建DQN模型:



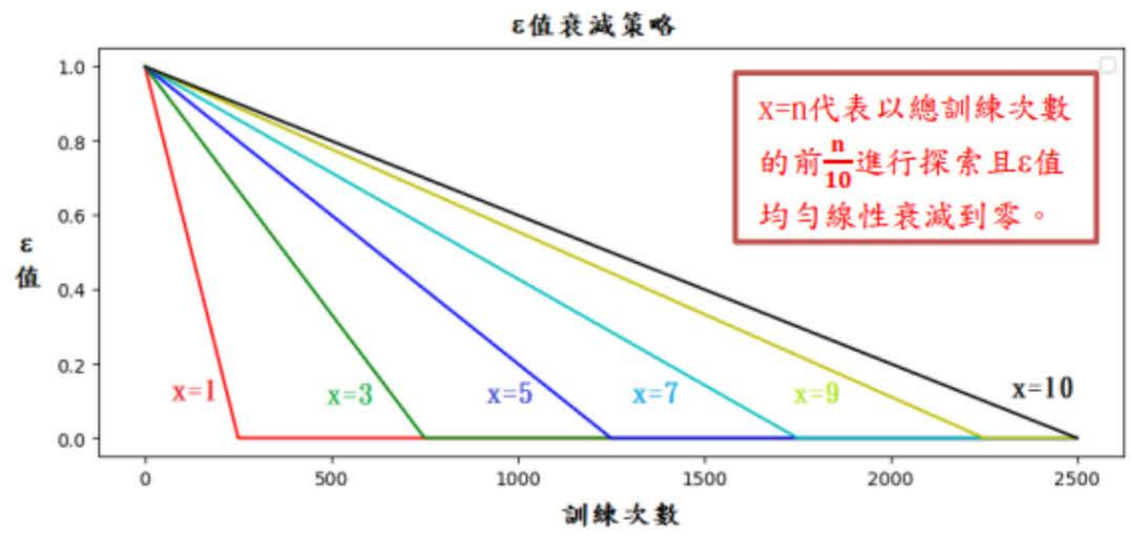
(三)觀察訓練情況，並測試模型表現:

訓練2500回合，探索策略為前1250次，ε 由1衰減到0

二、探索(exploration)與利用(exploitation)的難題

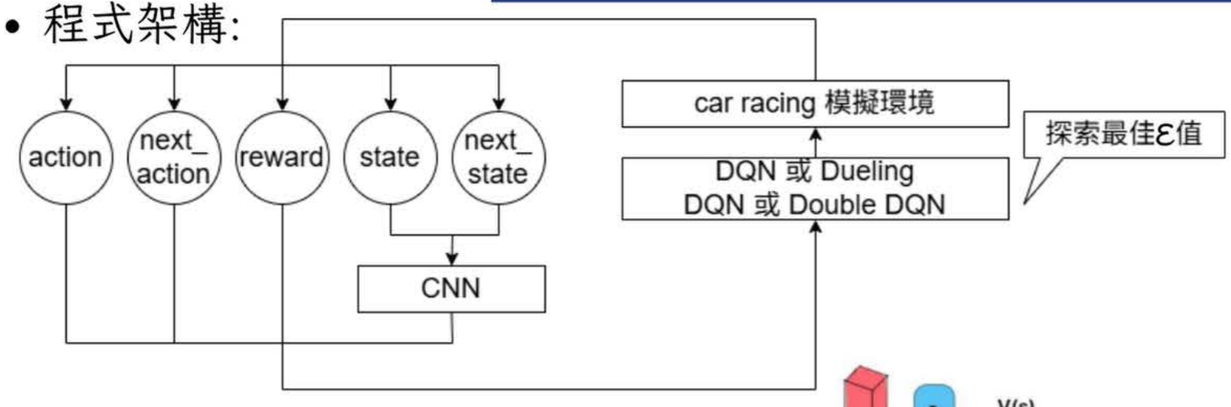
(一)尋找最適合此環境的衰減策略:

以衰減係數x為1、3、5、7、9、10的線性衰減策略訓練模型，並討論分析其訓練結果。

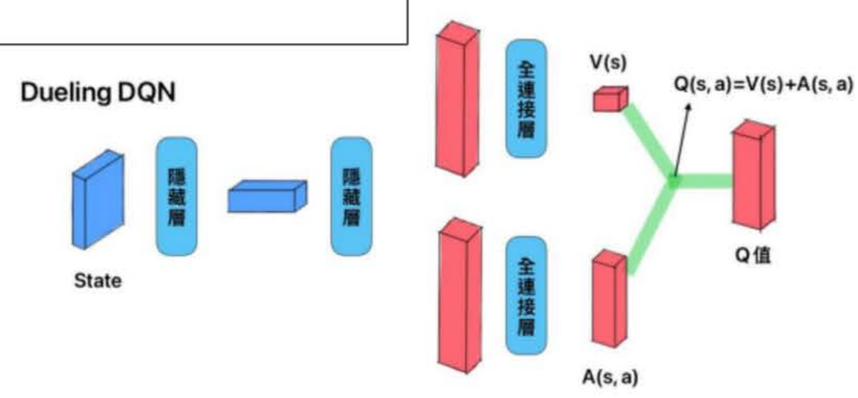


三、DQN、Double DQN和DuelingDQN比較

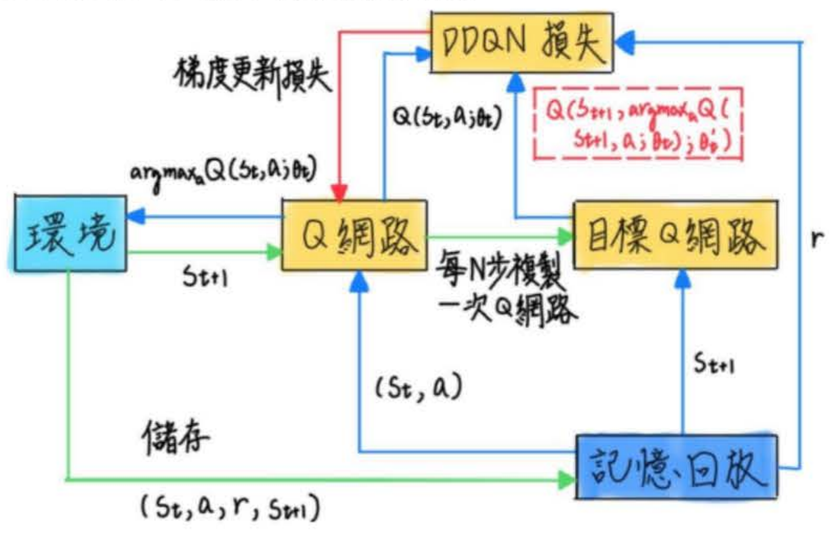
- 程式架構:



- Dueling DQN 網路架構:



- Double DQN 網路架構:



四、改寫環境的獎勵機制

(一)問題:

車子只要在賽道內，環境都會給予正向獎勵。

(二)解決:

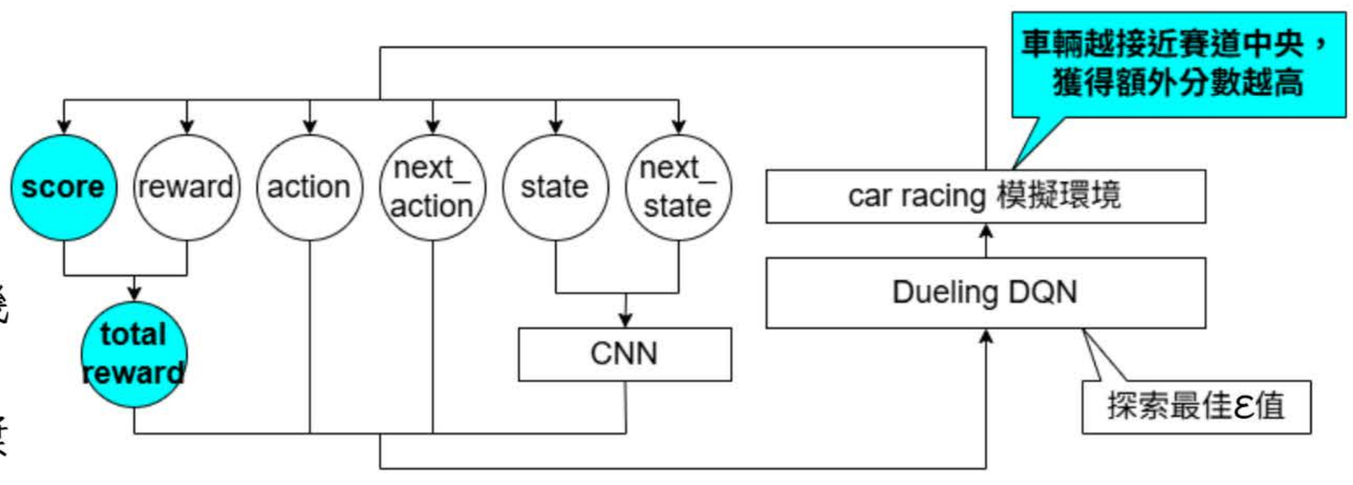
修改獎勵機制，使車子盡量在賽道中央

(三)實作:

(1)設定變數，作為車輛加扣分的分界線

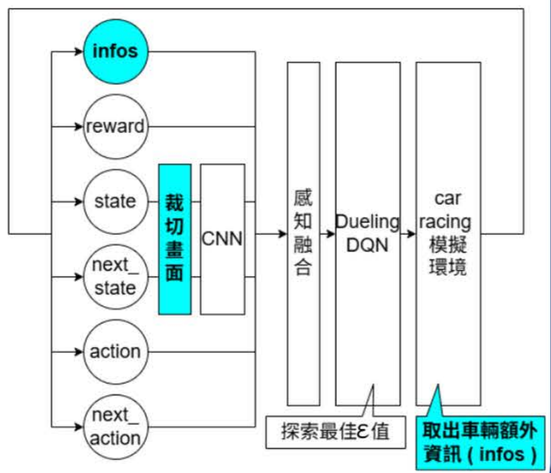
(2)計算賽道頂點平均值作為賽道中點，並用歐幾里得公式計算車子與賽道中點距離，並歸一化

(3)越靠近中點得愈多分，反之扣愈多分，並將獎勵同乘一倍數，使差距拉大



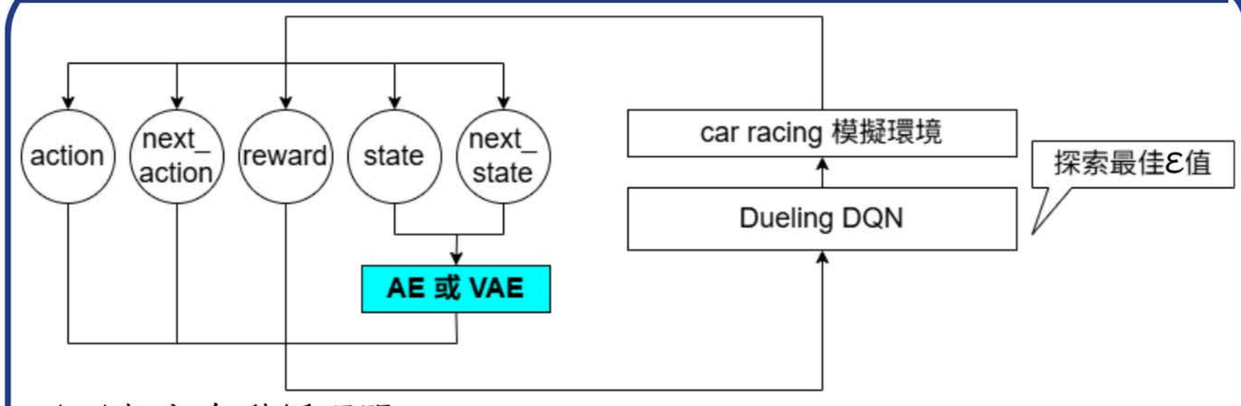
五、不同觀測空間的訓練與感知融合

(一)裁切圖片:



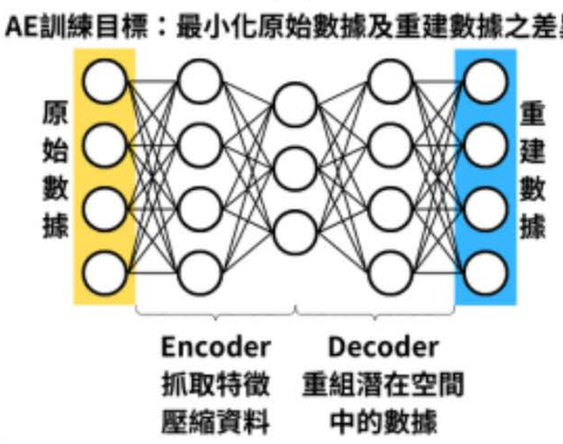
(二)加入感知融合(Sensor Fusion):
(1)修改原環境程式，並將資料回傳
(2)在模型中新增儲存車子資料區
(3)將車前圖像和車子資料合併

六、使用生成式AI協助強化學習環境感知的訓練

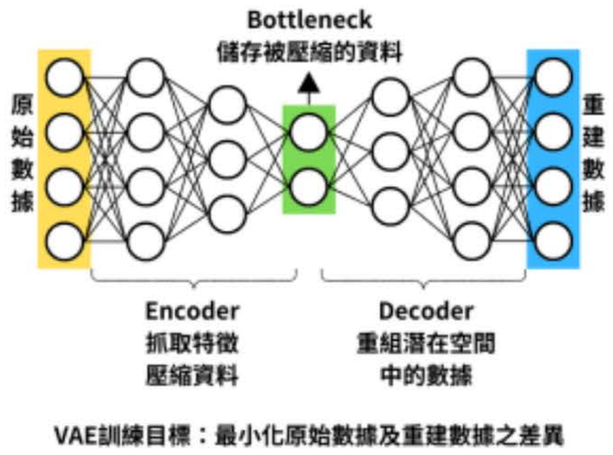


(一)加入自動編碼器:
智能體能用更短的時間及更高的效率做出判斷，減少車子因判斷時間誤差而造成的意外，提升自動駕駛應用的潛力

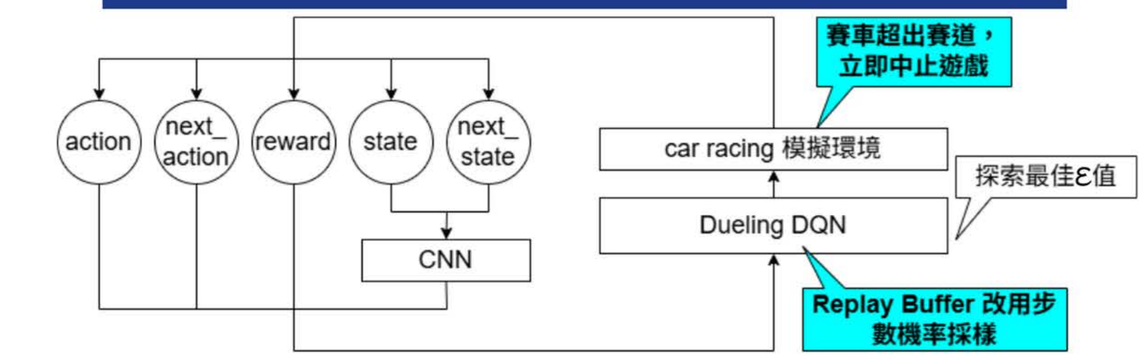
AE架構圖



VAE架構圖



七、調整Replay Buffer採樣機制

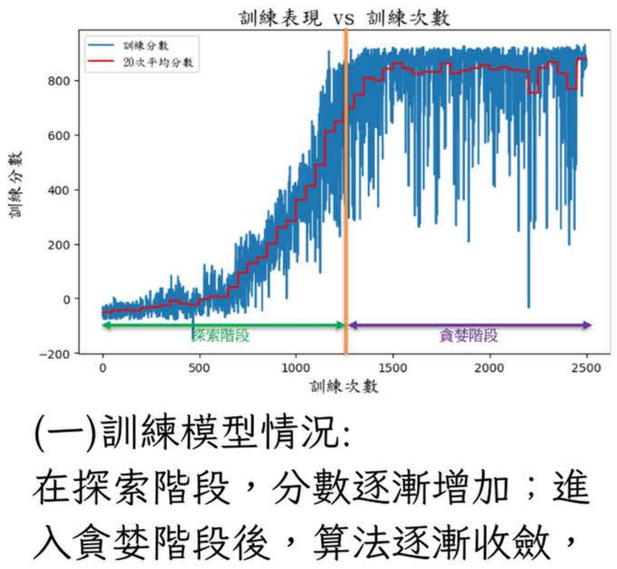


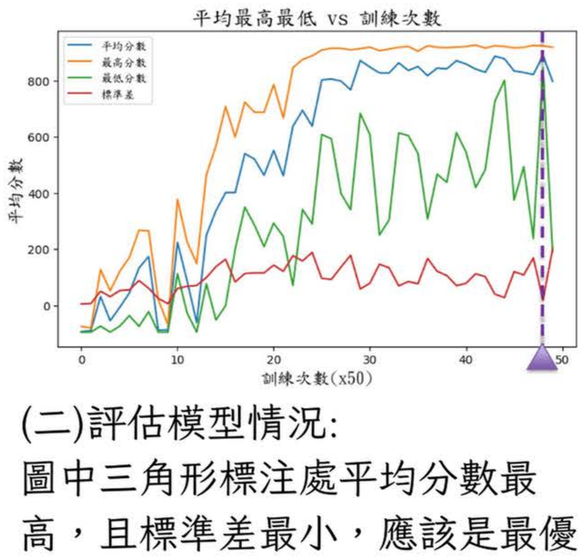
(一)改寫環境:
遊戲中的車輛跑出賽道時，立即終止遊戲
(二)實作:
利用射線法與函式判斷車子是否在賽道內及是否超出賽道，若超出賽道則終止遊戲

(三)問題:
(1)模型分數難以突破上限
(2)推測原因: 模型較難訓練到遊戲後期環境，導致模型在後段賽道表現不佳
(3)解決:經驗回放改使高步數經驗有較大機率被採樣

研究結果

一、實作DQN網路，解決Car_racing模擬環境



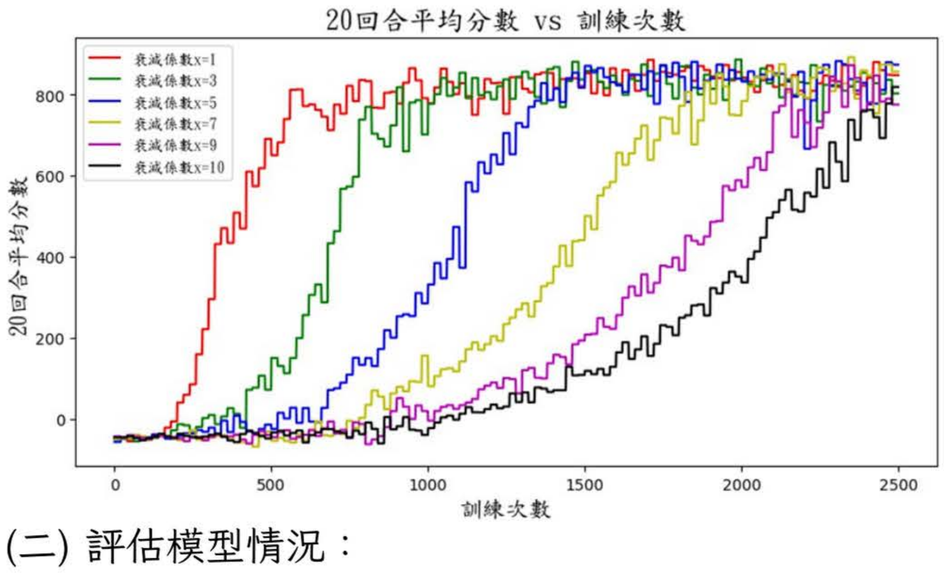


(一)訓練模型情況:
在探索階段，分數逐漸增加；進入貪婪階段後，算法逐漸收斂，分數漸趨於上限約900分，平均800分。

(二)評估模型情況:
圖中三角形標注處平均分數最高，且標準差最小，應該是最優模型

二、探索(exploration)與利用(exploitation)的難題

(一)訓練模型情況：
(1)所有平均分數值皆有上升，且其最大值皆能高於800分
(2)衰減係數越小，其學習速度越快(即平均分數較快上升)；衰減係數較大的模型，算法收斂比較慢，最後分數表現較不佳

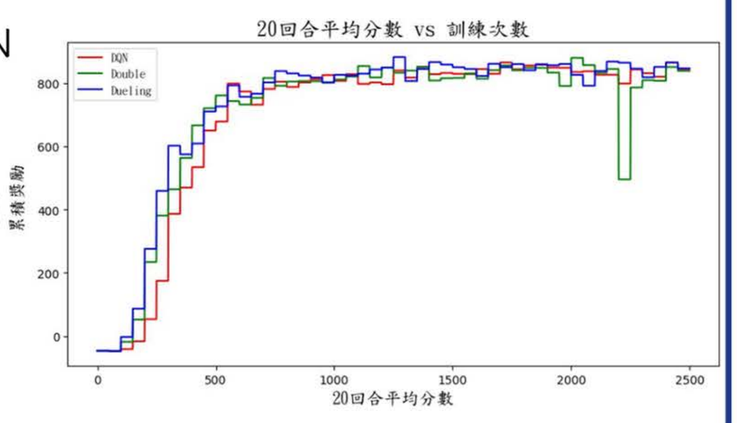


(二) 評估模型情況：

衰減係數	x=1	x=3	x=5	x=7	x=9	x=10
平均分數	832	829	846	816	754	695
最高分	913	915	920	916	908	894
最低分	464	458	511	394	461	259
標準差	110	113	97	138	135	180

三、DQN、DoubleDQN和DuelingDQN比較

(一) 訓練模型情況：
探索階段的DuelingDQN的平均分數上升最快，DQN則最慢；貪婪階段，的DoubleDQN平均分數突然下降，顯示其不穩定，其他趨勢大約相同



(二) 評估模型情況：

算法	DQN	DoubleDQN	DuelingDQN
平均分數	832	789	858
最高分	913	913	920
最低分	464	438	602
標準差	110	146	77

四、改寫環境的獎勵機制

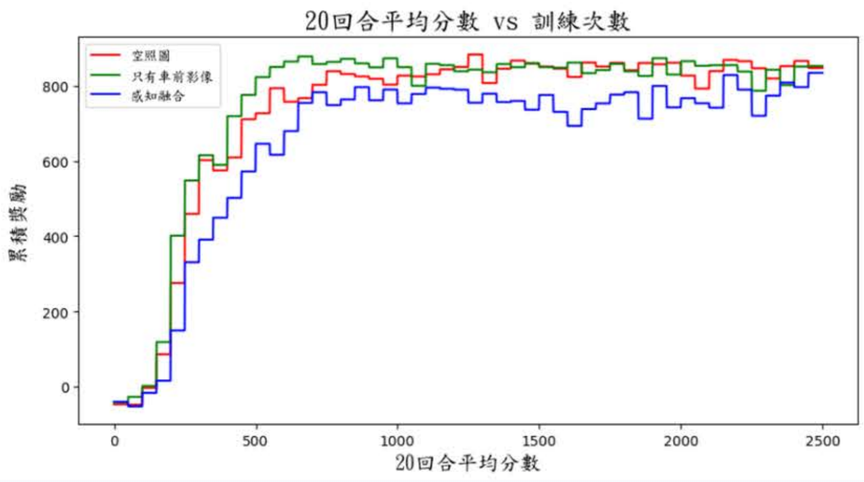
(一)訓練模型情況：
在探索階段，有修改獎勵機制的平均分數比原始獎勵機制上升較快；在貪婪階段，兩種獎勵機制相差不大

(二) 評估模型情況：
在探索期間，修改獎勵機制的測試平均分數和最高分數都比較高，可見此機制能加速學習；在貪婪階段修改獎勵機制的表現平均分數和最高分相差不大，但表現波動性較大

五、不同觀測空間的訓練與感知融合

(一)訓練模型情況:

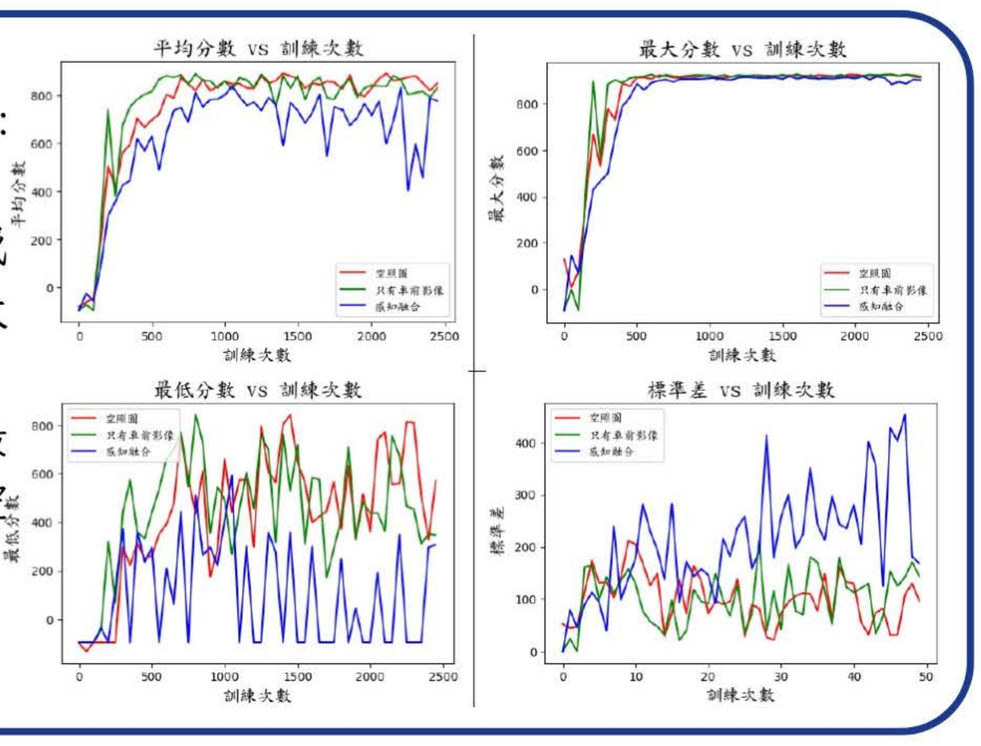
在探索階段，只有車前影像的平均分數上升最快，空照圖次之，感知融合則是最慢；在貪婪階段，只有車前影像和空照圖表現差不多，感知融合則表現略差



(二)評估模型情況:

(1)三者最高分表現相差不大，但感知融合標準差略大

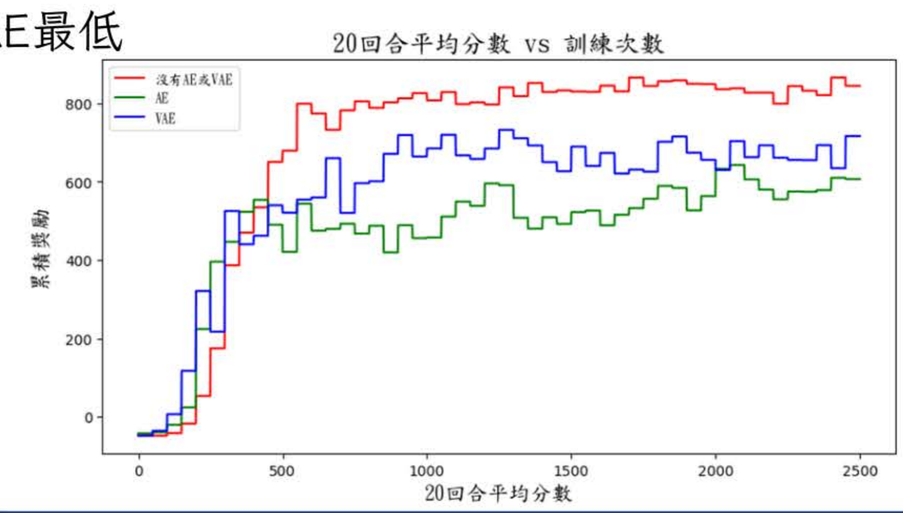
(2)車前影像在收斂前平均分數和最高分數明顯優於空照圖



六、使用生成式AI協助強化學習環境感知的訓練

(一)訓練模型情況:

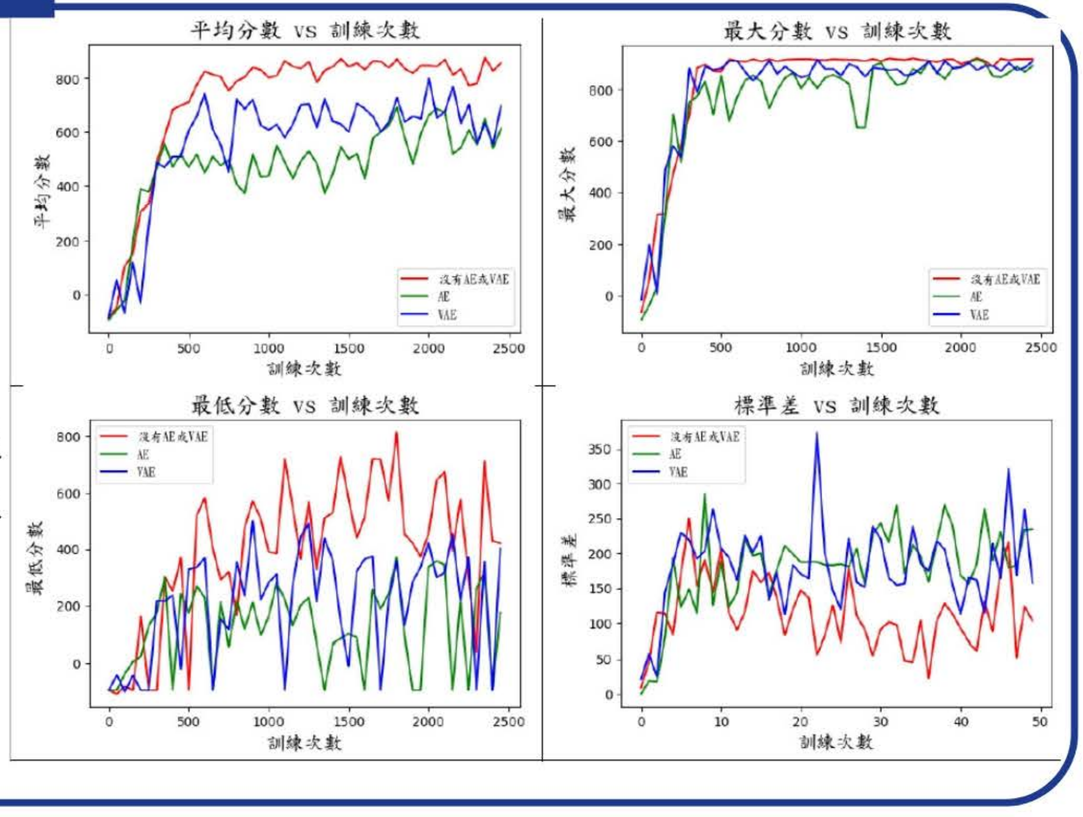
在探索階段，VAE平均分數上升最快，AE次之，未使用AE或VAE則最慢；在貪婪階段，未使用AE或VAE平均分數最高，VAE次之，AE最低



(二)評估模型情況:

(1)未使用AE或VAE的表現優於有使用VAE和AE，而VAE又優於AE

(2)VAE和AE在訓練過程中，表現逐步提升，最高分也能和未使用AE或VAE相當，唯穩定性較差，仍代表其可行性

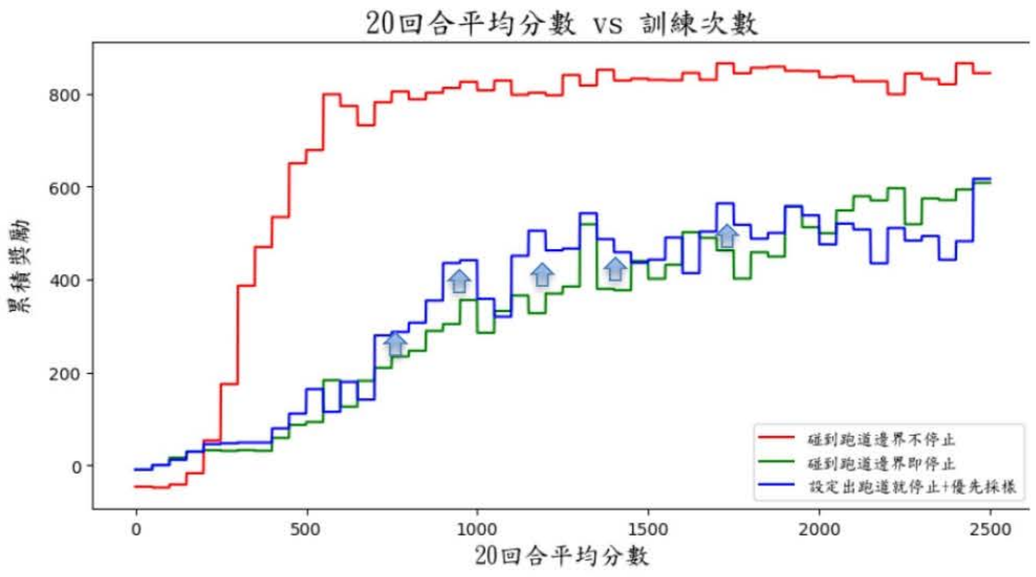


七、調整ReplayBuffer採樣機制

(一)訓練模型情況:

(1)新終止條件的獎勵低於未設置此條件的模型，但仍有緩慢的上升趨勢

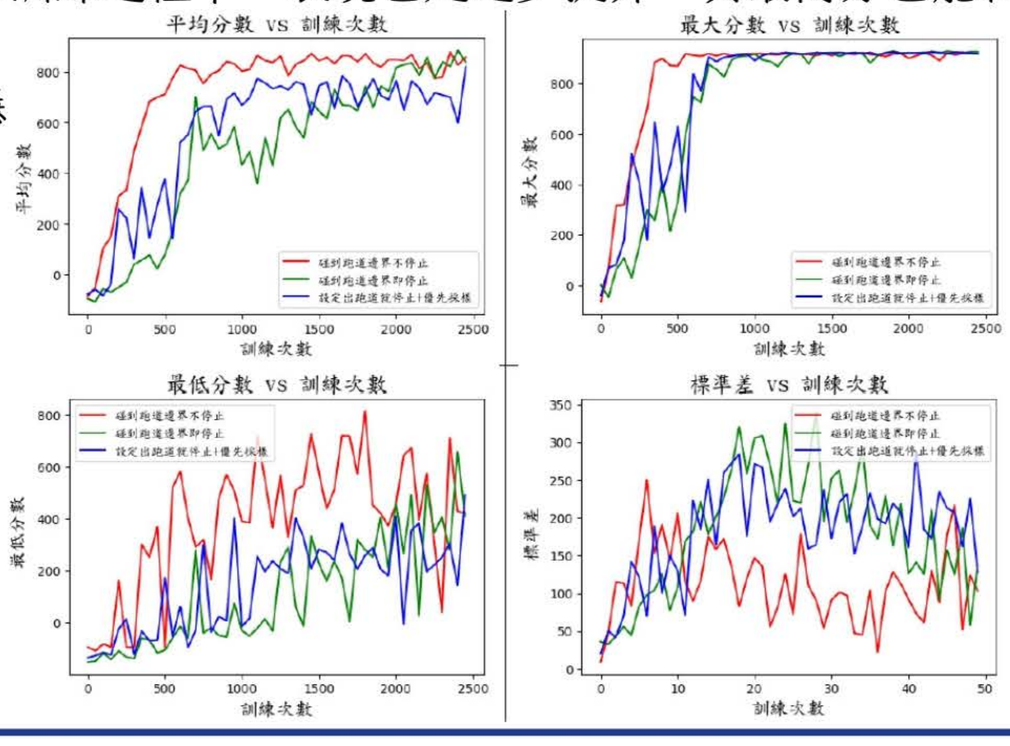
(2)加入優先採樣後，累計分數較高的車道環境被採樣訓練的機率較大



(二)評估模型情況:

(1) 由上面的各項指標綜合觀察，沒有使用AE或VAE的表現優於有使用VAE和AE；而VAE又優於AE

(2) VAE和AE在訓練過程中，表現也是逐步提升，其最高分也能和沒有使用AE或VAE相當，唯穩定性較差，仍代表其可行性



討論與結論

討論

1. 發現環境原始碼有bug，debug過程加深對程式理解
2. 換用 NumPy array 取代 deque 提升訓練效率
3. 減少經驗回放的 MaxSize 以減少主記憶體使用空間
4. DQN 雖適合離散動作，但不適用連續控制，可能造成精細操作困難
5. CarRacing 環境過於單純，導致使用進階演算法效果不顯著
6. 需要更大算力的GPU支持Agent訓練

結論

1. 規劃適合的探索衰減策略非常重要
2. 簡單環境建議使用較小的 ε 衰減係數
3. Dueling DQN 在三種模型中表現最優
4. 鼓勵車輛靠近賽道中央的獎勵設計可有效加快學習速度
5. 感知融合能提升模型擬真度但穩定性略差
6. 裁切僅保留車前影像可提升初期學習速度
7. 新終止條件能提升補救能力，但不夠真實
8. 任務採樣可提升訓練效率，但需防止過度擬合
9. VAE 較 AE 更具泛化能力與訓練穩定性
10. 複雜模型在簡單環境易過度擬合，需調整訓練設計

參考文獻

1. 李茹楊, 彭慧民, 李仁剛, 趙坤. 強化學習演算法與應用綜述. 電腦系統應用, 2020, 29 (12): 13-25.
2. Mnih, V., Kavukcuoglu, K., Silver, D., et al. (2015). Human-level control through deep reinf orcement learning. Nature, 518(7540), 529-533.
3. Sutton, R. S., Barto, A. G. (1992, May 31). Reinforcement Learning: An Introduction.
4. Pykes, K. (2024, November 6). Understanding the Bellman Equation in Reinforcement Learn ing. Datacamp.
5. Wang, Z.,Schaul, T., Hessel, M., Hasselt, H. v., Lanctot, M., & Freitas, N. d.(2016).Dueling Network Architectures for Deep Reinforcement Learning.
6. Watson, D. (2024, January 1). What Is a Double Deep Q Network? THE ENGINEERING PR OJECTS.
7. 劉智皓. (2023, April 25). 深度學習Paper系列(04)：Variational Autoencoder (VAE). Medi um.
8. Nick. (n.d.). [Day5] VAE，好久不見. IT邦幫忙.
9. Bergmann, D., & Stryker, C. (2024, June 12). What Is a Variational Autoencoder? IBM.
10. Lee, H.-Y. (2016, December 7). ML Lecture 18: Unsupervised Learning - Deep Generative Model (Part II). YouTube.
11. Udacity team. (2020, August 25). Sensor Fusion Algorithms Explained. UDACITY.
12. 算法集市. (2019, January 7). 判斷一點是否在多邊形內部：射線法. 每日頭條.
13. Tsai, Y.-R. (2019, March 9). What Are Autoencoders? Medium.
14. Liu, Y., & Diao, S., An Automatic Driving Trajectory Planning Approach in Complex Traffic Scenarios Based on Integrated Driver Style Inference and Deep Reinforcement Learning. (2024, January 25). https://doi.org/10.1371/journal.pone.0297192
15. Agrawal, S. (n.d.). Sensor Fusion Software in Self Driving Cars: A Binmile Study. Binmile. https://binmile.com/blog/sensor-fusion-software-in-self-driving-cars/