基於深度強化學習之 CivRealm 戰鬥策略決策模型設計與實作

A Deep Reinforcement Learning–Based Combat Strategy Decision Model for CivRealm

指導教師：温育瑋 博士

學生：王柏勝 賴柏丞 劉暉曜

國立聯合大學資訊工程學系

苗栗市南勢里聯大2號

ywwen@nuu.edu.tw

摘要

本研究針對4X 類策略遊戲 CivRealm 之 minigame 戰鬥任務環境，設計並實作一套結合卷積神經網路（CNN）與深度 Q 網路（DQN）的戰鬥決策模型。通過將地圖轉換為四通道的類影像二維資料後，利用 CNN 轉換出對應動作的 Q 值向量，並以 DQN 在 ε-greedy 策略下進行訓練。整體而言，本研究完成一套可重複使用的訓練框架，為未來戰略 AI 的改良提供了一個方向。

Abstract

This study targets the minigame combat mission environment of the 4X strategy game CivRealm, designs and implements a combat decision-making model that integrates a Convolutional Neural Network (CNN) with a Deep Q-Network (DQN). By converting the map into a four-channel, image-like 2D representation, the CNN generates a Q-value vector corresponding to possible actions, and the DQN is trained under an ε-greedy strategy. Overall, this study presents a reusable training framework that provides a direction for improving strategic AI in the future.

一、緒論

隨著複雜決策問題的解決能力在近十年獲得重大突破，從棋類遊戲到即時戰略遊戲，遊戲環境逐漸成為研究高維度決策、長期回饋以及複雜策略生成愛用的標準測試平台。

相對於圍棋與即時戰略遊戲，4X 類文明策略遊戲（如 Civilization, Freeciv）具有更高的複雜度，包括極長回合跨度、多層次決策結構、多代理互動、外交與隱藏資訊等特性，使其成為更貼近真實世界決策的研究場域。

本研究希望透過 CNN 與 DQN 結合成的卷積式深度 Q 網路（Convolutional Deep Q-Network, CNN-DQN）的架構，引入以地圖特徵為中心的資料表示方式，讓代理能從視覺化空間中自主學習戰鬥判斷能力，在僅使用地圖型態觀測的前提下，是否能透過適當 state 和 reward 設計來學得具備實用價值的戰鬥策略。

二、文獻探討

早期的電腦圍棋依賴大量人工設計特徵與蒙地卡羅模擬。Maddison 等人的深度卷積網路工作成功證明，從純粹的棋盤圖像中直接學習有效的策略分布是可行的，並且能以較低的搜尋成本達到接近專家水準的走法選擇能力[1]。而 Silver 等人後續的 AlphaGo 系統在此基礎上加入策略網路與價值網路，將深度神經網路提供的先驗知識嵌入蒙地卡羅樹搜尋，使得搜尋過程更具方向性並能降低探索成本[3]–[5]。AlphaGo Zero 的研究更進一步展示深度強化學習可在無任何人類知識的情況下快速自我提升，並最終超越所有人類與先前系統的水準[6]。在 4X 類遊戲研究中，

Freeciv 作為開源且高度模組化的 Civilization 類遊戲，長期被用於研究策略 AI 的學習與推理能力。Bonde 等人開發的 Freeciv-Python 介面[11]使 Freeciv 能以張量形式與深度 RL 系統連結，並指出遊戲巨大的狀態空間與高度策略組合。Voss 提出的 knowledge-based reinforcement learning (KB-RL)方法，結合顯式知識庫與強化學習，能在完整 Freeciv 對局中穩定擊敗內建 AI，並在多種遊戲設定中取得顯著改善[8]–[10]。然而，KB-RL 嚴重依賴手工知識庫，難以擴展至未見過的地圖、策略或規則。為克服上述問題，CivRealm 將 Freeciv-web 平台重新設計成支援深度學習代理與語言模型代理的通用環境。該研究指出文明類遊戲兼具長期策略、動態外交、多代理和決策與語言互動需求，因此單純依賴 DRL 並不足以處理完整遊戲情況。此外，初步實驗證明，現有 DRL 架構雖能處理小型戰術型 minigame，但在完整 CivRealm 中難以建立有效策略，顯示文明類遊戲的學習需求遠超過既有遊戲 AI 基準[13]。現有 Civilization/Freeciv 研究多半在小型戰術任務中取得進展，但在完整文明遊戲中，深度 RL 系統仍無法有效學得穩健策略。基於上述觀察，本研究定位於探索一種能整合深度學習、搜尋、卷積化特徵擷取的架構，使代理能在 CivRealm 這類高複雜度、長時序、多代理環境中展現更有效的決策能力。

三、系統架構

(一) 擷取地圖特徵轉為多通道圖

透過將所有地塊與單位資訊，編碼成一張類似影像的二維資料，並設計四個 channel 來盡可能地還原地圖的有用資訊。地形 channel 將地塊的 terrain 編碼正規化。血量 channel 根據單位的 health 屬性，以最大血量為基準進行歸一化，並使用正負號區分我方與敵方。綜合戰鬥能力通道則將攻擊力、防禦力、火力與射程整合成單一數值，並同樣使用正負號表達敵我差異，提供模型對「戰力優勢區」與「戰力弱勢區」的直覺。單位價值 channel 則透過兵種名稱的關鍵字推估不同單位的重要性，讓模型能對「值得保護或優先擊殺的高價值目標」有更清楚的感知。

表1 多通道地圖特徵表示表

channel	名稱	說明
0	地形層	移動利損耗需求
1	血量	我方(正)/敵方(負)血量
2	戰鬥力(攻擊力)	我方(正)/敵方(負)戰鬥力
3	單位價值	生產這個單位所需耗損的資源

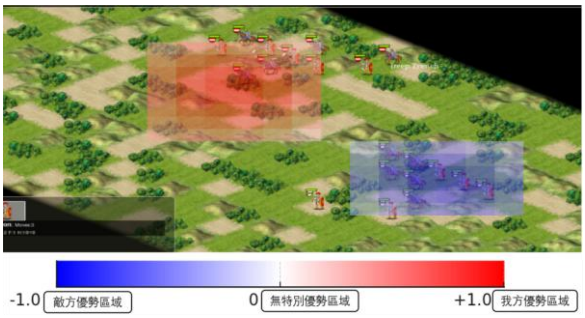

圖1 戰鬥力 channel 模擬圖

(二) CNN 將地圖特徵轉為 state

CNN 在讀取上述多通道地圖後，會輸出對應全域動作索引(global action indices)的 Q 值向量。網路結構採用兩層3×3卷積與 ReLU 非線性，接著將所有 feature map 展開為一個長度128×20×20的向量，並透過兩層全連接層逐步壓縮到中間的512維與256維。為了處理 CivRealm 中動作名稱是字串、且不同局面可用動作集合不固定的情況，系統額外維護一個動作詞彙表，將可用動作名稱映射到固定長度的整數索引空間。在每次互動中，系統會先建立當前可用動作名稱與參數的列表，再將這些名稱轉換為全域動作索引 形成一個在當前狀態下「合法的動作索引集合」。

（三）DQN 訓練

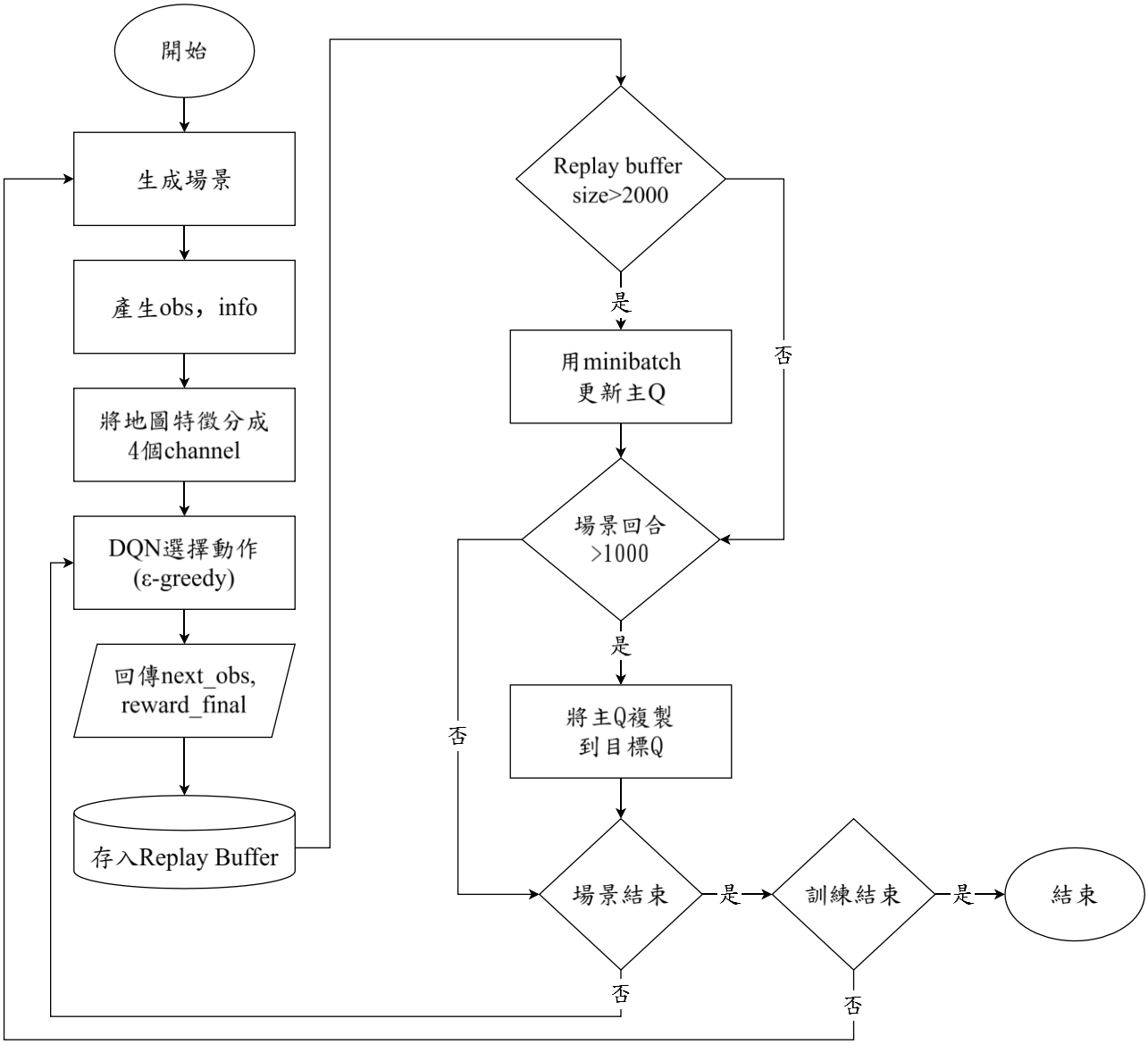

圖2 訓練流程圖

本研究實作了一個容量固定的經驗重播緩衝區(Replay Buffer)，將每一步的狀態、動作索引、塑形後回饋、下一個狀態以及終止標記存入其中。當緩衝區累積到足夠多樣的經驗後，才開始進行 DQN 的參數更新。每次更新會隨機抽樣一批 transition，將其中的 map state 與 next map 組成兩個批次張量，分別送入線上網路與目標網路計算 Q 值與目標值。線上網路的輸出透過 gather[12]選取出實際執行動作對

，

透過 Adam 最佳化器進行參數更新[2]。目標網路的權重則間隔固定步數同步一次應的 Q 值，目標網路則透過對下一狀態的所有動作求最大值形成 bootstrap 項，兩者之間的差以均方誤差作為損失函數，確保訓練過程中的目標值不會變化過於頻繁。訓練流程中，用全局步數計算 epsilon 的指數衰減，使代理在訓練初期大量探索不同策略，在後期逐漸轉向利用已學得的價值估計；並在每個回合開始進行時間懲罰的計算，讓 reward 塑形函數能對拖延戰鬥作出負面回饋。每一回合結束時，系統會判斷是因為單位全數陣亡、分數達到預設

上限，或是環境回報的終止或截斷條件達成，並在 log 檔中紀錄該局的總回饋與結束原因。系統也會依照預先設定的步數間隔定期儲存模型檔案，包括線上網路與目標網路的權重以及最佳化器狀態。

四、實驗

(一) 獎勵設計

在本研究的戰鬥情境中，我們採用「延遲回饋（delayed reward）」結合「血量差分塑形（HP-difference shaping）」的方式設計強化學習的獎勵函數。首先在每一步行動後，根據前後兩個觀測中的單位血量變化計算一個基礎獎勵 。我們透過單位的 owner 欄位區分我方與敵方。

接著，我們以敵方總血量的下降作為「造成傷害」的指標，以我方總血量的下降作為「承受傷害」的指標，並線性組合成塑形回饋：擊殺敵方給予+1.0的獎勵，對敵方血量的減少給予+0.1×敵方損失血量的獎勵，而我方血量的損失則以−0.12×我方損失血量的懲罰；同時加入與時間步數成正比的微小負回饋每回合−0.01，鼓勵代理人以較少步數結束戰鬥，以上的總和為基礎獎勵。這個結果不會立即存入 replay buffer，而是暫存每個單位在出手時的狀態、動作與當下血量，待敵方行動結束或回合終止時，再重新計算該單位血量的實際損失ΔHP actor ，並與基礎獎勵加總後，才將最終獎勵與後續狀態一起寫入 replay buffer。

(二)實驗設計與結果

我們通過與 CivRealm 中內建的不同難度原始 AI 進行對戰來檢驗訓練的成果。在進入評估模式時，模型會先載入指定的權重檔案，凍結網路參數，不再進行反向傳播與梯度更新，並將決策策略切換為純 greedy（即選擇 Q 值最大的動作，不再使用 ε-greedy 探索）來進行遊戲。

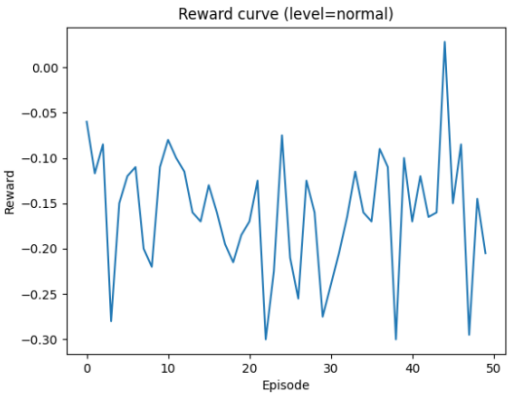

圖3 與 AI 對戰的 reward 變化圖

在 normal 難度下，評估50局對戰的 reward 多數為負值且波動較大，僅少數局數接近零或略為正值，顯示目前模型在中等難度的原版AI面前尚未能穩定取得優勢，戰鬥效率與損失控制仍有明顯改善空間

五、結論與建議

本研究提出的多通道地圖編碼方法與 CNN-DQN 結合的流程設計，能有效解決在對棋盤類地圖進行特徵抓取時，出現維度爆炸的情況，提升代理的訓練穩定性。然而，結果也指出系統仍具進一步改善的空間。我們為了簡化實作的過程，只針對戰鬥的部分進行相關的設計，當把這套訓練流程班上完整的遊戲上時，能否正確的運作仍有更多可以進行優化和研究的地方。

目前的多通道 CNN 編碼模型雖能捕捉地形、血量與戰鬥能力等主要戰術特徵，但在複雜局面下仍可能不足以完整描述所有有效的地圖資訊。reward 塑形策略方面，目前的方法對於細節性戰術行為的引導稍顯不足。在判斷高價值單位保護、火力集中效率、地形占領等方面，現有 reward 仍較為粗略。未來可設計更精細的戰術分解式 reward 或採用基於學習的 reward 模型（如 IRL），以提升代理在高階戰略判斷的準確性。

整體而言，本研究成功實現了 CivRealm 原本未提供的 CNN 特徵提取與 DQN 訓練結合的決策流程，使 DQN 能在動作空間非固定、狀態高度離散的情況下

仍保持訓練穩定性，為策略遊戲中的深度強化學習研究提供了新的技術方向。此外，我們也建構了一套可重複使用的訓練流程與模組化架構，使研究者能在不需大量修改底層程式的情況下，快速加入新的特徵通道、替換決策模型或更動 reward 設計。此系統減輕了策略遊戲強化學習的研究門檻，對後續相關工作具有實用價值。

六、專題作品影音平台網址

https://www.youtube.com/watch?v=9-alDazzKQ0

七、參考文獻

[1] C. Maddison et al., “Move Evaluation in Go Using Deep Convolutional Neural Networks,” ICLR, 2015.

[2] D. P. Kingma and Ba, “Adam: A Method for Stochastic Optimization,” in Proc. 3rd Int. Conf. Learn. Representations (ICLR), 2015.

[3] D. Silver et al., “Mastering the game of Go with deep neural networks and tree search,” Nature, vol. 529, pp. 484–489, 2016.

[4] D. Silver et al., “Mastering the game of Go without human knowledge,” Nature, vol. 550, pp. 354–359, 2017.

[5] D. Silver et al., “Mastering the game of Go with deep neural networks,” Communications of the ACM, vol. 62, no. 3, pp. 114–120, 2019.

[6] D. Silver et al., “Mastering chess and shogi by self-play with a general reinforcement learning algorithm,” arXiv:1712.01815, 2017.

[7] J. Kaiser, “HUBERT: A Freeciv-based Intelligent Agent Architecture,” arXiv:2001.05018, 2020.

[8] M. Voss and A. Khosravi, “Knowledge-Based Reinforcement Learning for Freeciv,” IEEE Trans. Games, vol. 14, no. 3, pp. 314–326, 2022.

[9] M. Voss, “Knowledge-Based Reinforcement Learning for Strategy Games,” Ph.D. dissertation, University of Central Florida, 2020.

[10] M. Voss, “Optimizing Strategy Game Behavior Using KB-RL,” arXiv:1909.11122, 2019.

[11] R. Bonde et al., “Freeciv-Python: A Python Interface for Freeciv to Enable AI Research,” GitHub repository, 2020.

[12]V. Mnih et al, “Human-level control through deep reinforcement learning,” Nature, vol. 518, pp. 529–533, 2015.

[13] Z. Qi et al., “CivRealm: A Learning and Reasoning Odyssey in Civilization for Decision-Making Agents,” ICLR, 2024.