112學年度工程科學系大學部專題研究競賽

專題名稱 **運用DQN與PPO演算法於離散動作空間問題-以象棋為例**

專題生 王瑞辰　　指導教授 陳牧言

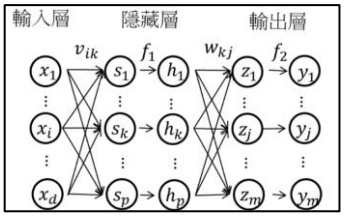
研究動機

2017年之後，AI的發展速度呈指數性上升。強化式學習在解決複雜問題上取得了令人矚目的成就。我希望由這個專題，探索相關技術與理論。

選擇PPO和DQN兩種不同演算法以及個別使用變體神經網絡處理象棋這項複雜的離散動作空間問題，個別對其產生的model進行比較，讓一般人也能從中學習AI的思考模式。

神經網絡相關模型

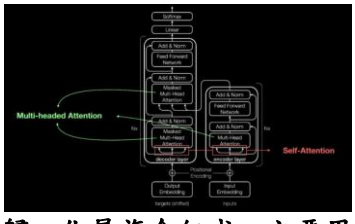
- **MLP（Multilayer Perceptron）**



由多個神經元層組成，包括輸入層、隱藏層以及輸出層。每個神經元都與前後層的神經元相連接，利用反向傳遞學習調整權重，信息不會循環連接。

MLP通常用多個全連接層作為神經元之間的連接方式。

- **Transformer**

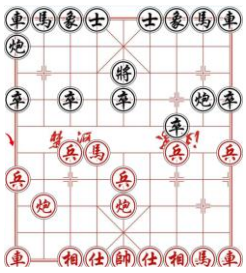


為編碼器-解碼器架構，由多頭注意力機制(堆疊多個self-attention組成)、前饋神經網絡、殘差連接和歸一化層複合組成。主要用於處理或生成順序輸入資料資料。

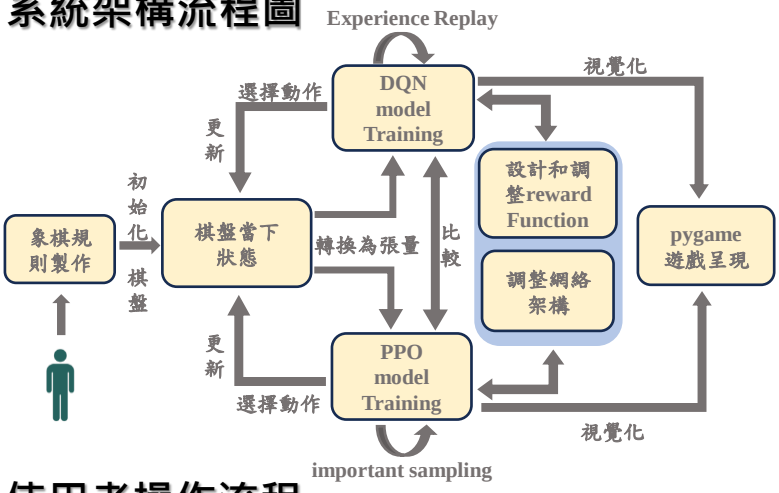
self-attention:透過增強減弱權重，關注資料中最重要的部分。根據query、key和value自適應的更改權重大小。

成果展現

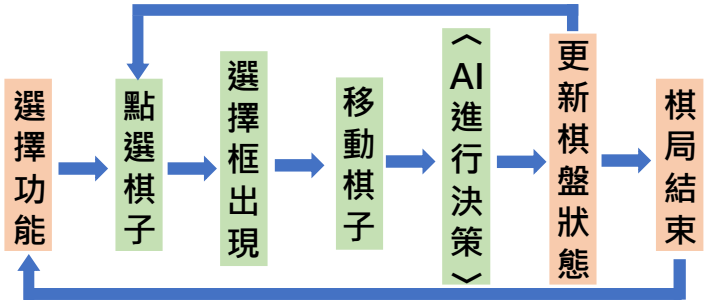
相關數值圖表　demo影片



系統架構流程圖



使用者操作流程



演算法比較

- DQN關注值函數的學習，選擇最高Q值的動作。
  PPO關注策略的學習，由動作的概率選擇策略。
- DQN由Q network和Target network組成。
  PPO由Policy Network和Value Network組成。
- DQN會使用非當次迭代的經驗樣本。
  PPO只會使用當次迭代的經驗樣本。
- DQN通過最小化Q值的均方誤差更新參數。
  PPO 通過最大化優勢函數和KL散度更新參數。

未來展望

象棋的動作維度很多，導致開局選擇動作的正確率不佳，到了殘局才有較好的表現。希望未來當我對象棋與神經網絡有更深了解後，能訓練出媲美AlphaZero的model，在變幻莫測的棋局給予精準的策略選擇。

更進一步發展擁有語音教學模式的AI，協助所有對象棋有興趣的玩家能學習如何針對不同的棋局做出決策，並推廣這項古老的遊戲。