人工智慧期末專題計劃書

A.I. Final Project Proposal

教授: 林傑森

基於強化式學習(Reinforcement Learning)的五子棋 - Beta

Gobang

第 8 組

洪郁修 4108056036

江尚軒 4108056005

王思正 4108056004

要做什麼?

五子棋遊戲 A.I.我們預計經由 RL 方式訓練來訓練五子棋 A.I.模型 - Beta GoBang，成品的命名參考著名的圍棋 A.I.阿爾法 Go (AlphaGO)，透過合適的訓練演算法、優化器與強化學習模型訓練 A.I.，令其能夠下五子棋並與玩家對弈，在最後將遊戲透過 Pygame 或其他套件完成渲染並製作成 App 發布或是以網頁的形式展現我們的作品。

背景

現在人工智慧在實際應用上已有不少的案例，比如影像辨識領域、自然語言處理等，這些背後是許多機器學習研究的成果，其中之一就是強化學習(Reinforcement learning)，強化學習是基於動態規劃與馬可夫決策過程(Markov decision process)，其獎勵函數會考慮過去的樣本，因此能有效解決動態的最佳化問題。強化學習訓練出來的模型本質上是對一個動態問題的策略(Policy)，可以用來解決相似的問題而不需要重新再訓練或計算，具有通用性，我們預計會參考 AlphaGO 的設計理念與相關的演算法(蒙地卡羅搜尋等)來實作我們的 Beta Gobang。

在這次期末專題中，我們想要嘗試使用強化學習 (Reinforcement Learning) 技術，其中較為人知的應用就是在遊戲上，比如說 AlphaGo，在實戰中一步一步將柯潔等數位人類棋王逼入絕境，更可怕的是 AlphaGo 還會藉由學習變得越來越強，這也促進了強化學習的發展，也因此誕生出更新更強的圍棋 A.I.，比如 AlphaGo Zero 等。

動機

我們查到許多關於 Reinforcement Learning 的遊戲 A.I.與相關實作分享，但大多都過於複雜，而且訓練資料難以取得，加上又必須在短時間內拿出成品，於是我們選擇相對較少人做過，複雜度又相對簡易的五子棋遊戲作為環境來訓練我們的 A.I.，在強化學習的框架與遊戲環境互動的訓練下，打造出屬於自己的五子棋遊戲 A.I. - Beta Gobang。

如此一來，以後自己一個人感到無聊時，也可以和 Beta Gobang 來場戰況膠著的五子棋了!

難度

難度方面應該為**中上**:

1. 已有其他人實作過類似的遊戲 A.I.
2. 遊戲的狀態變化函數是自定義的，訓練複雜度上可以控制
3. 訓練資料要由強化學習的 Agent 自行與環境互動並調整策略
4. Agent 需與自己對弈並學習，訓練初期須檢視和評估訓練成效
5. 人工智慧模型訓練需耗費時間
6. 已有成熟的開源函式庫供強化學習的模型開發

實作方法

我們先講解整個遊戲的基本架構，在 Environment 與 Agent 建置下，讓兩台電腦就扮演 Agent 部分，行動就是依據現在棋盤的狀態，在棋盤上選擇要下的位置，並輸入到遊戲環境中，Environment 方面就是依據電腦選擇的位置，輸出改變後的棋盤樣貌，再將其推回給下個 Actor。

因此，我們會先完成整個遊戲的環境設定，並將棋盤樣貌加上選擇位置組成做作為智慧體得到的資訊(fully observed)，並在收到 game over 訊號(白子或黑子獲勝)後自動收集成一系列形式為(狀態 $_1$, 動作 $_1$)、(狀態 $_2$, 動作 $_2$)、(狀態 $_3$, 動作 $_3$)...的時間序列(Time Sequence)，因此我們需要針對這些時間序列來做採樣，常見的方法有蒙地卡羅搜尋、TD 等，其中蒙地卡羅搜尋比較適合我們的五子棋遊戲，主要是因為五子棋與圍棋類似，是要在五顆棋子連成一線後才知道結果，遊戲環境給的獎勵具有高度的延遲性，因此 TD 法不好估算每一步的獎勵價值為何。

資料收集完畢後作為訓練資料回傳給 DQN 中的決策網路來更新參數，同時儲存到一個有限的緩衝區中作為經驗回放(Experience Replay)，並額外儲存一個相同結構的網路來暫時固定訓練目標供原本的決策網路做擬合(Fitting)。

依據初期訓練效果來決定兩個互相對奕的智慧體是否需要共用同一個 DQN，或是分別訓練兩個相異的網路(黑子和白子)，類似對抗式生成(GAN)，兩者皆可部署到我們發布的 App 或網頁中。

關於訓練模型部分:

1. **定義遊戲數值及規則**

五子棋的規則與勝利條件:

* 棋盤大小為 9*9
* 黑子先下，再換白子
* 若五顆相同顏色的棋子連成一線(對角亦算)則執棋方勝利
* 預計會使用 OpenAI/gym 的套件來進行遊戲環境的開發
* 提供強化學習模型學習的虛擬環境

2. **定義 RL agent 的 reward function 與其他參數 (Learning Rate 等)**

A. Reward Function:

* 贏家為正，輸家為負

B. Hyperparameter(包含但不限於):

* 強化學習模型策略的框架 (Ex: DQN/A3C 等)
* 合適的採樣演算法 (Ex: MCTS、TD)
* 網路最佳化函數 (Ex: PPO、PPO-clip、Bellman Update 等)

C. 實作方式:

* 預計會使用 OpenAI/baselines 的套件來開發我們的智慧體
* 預計會利用蒙地卡羅搜尋來讓智慧體對我們的環境進行採樣
* 預計 DQN 網路來做為智慧體的決策函數

3. **檢視訓練結果並繪製圖表 (Loss 函數變化圖等)**

* 不定回合數利用人工的方式與訓練中的模型對弈
* 人工方式評估模型目前的棋力表現
* 隨著訓練筆數動態調整學習率
* 將訓練過程繪製成圖表以利檢視訓練狀況
* 若遇到瓶頸則思考如何調整超參數或訓練方式
* 與其他網路上之五子棋 A.I. 對弈，檢視模型強度
* 預計使用 Matplotlib 繪製圖表來呈現訓練狀況、數據

4. **將遊戲渲染出來並做成網頁形式或 APP**

* 將已經訓練好的模型與遊戲製作成 APP 或網頁形式
* 可選擇與電腦對弈或與真人實戰(以網頁形式才支援)
* 發布 App 或網址供其他人遊玩和 Demo

預計使用到的工具與程式語言

* Python : 程式語言
* PyGame：視覺化界面
* Pytorch：模型訓練
* TensorFlow：模型訓練
* OpenAI: 強化學習框架
* VScode: 輕量化編譯器
* Anaconda: 建立人工智慧模型及遊戲所需之套件的虛擬環境
* Colab: Google 提供的雲端平台，有 TPU/GPU 等硬體裝置加速訓練過程
* Django: 用以提供建立網頁前後端、資料庫等 Python 套件與開發框架

參考資料

如何使用自對弈強化學習訓練一個五子棋機器人 Alpha Gobang Zero https://www.cnblogs.com/zhiyiYo/p/14683450.html 2022/10/15

GitHub - Alpha Gobang Zero https://github.com/zhiyiYo/Alpha-Gobang-Zero 2022/10/15

【機器學習 2021】概述增強式學習 (Reinforcement Learning, RL) https://www.youtube.com/watch?v=XWukX-ayIrs&list=PLJV_el3uVTsMhtt7_Y6sgTHGHp1Vb2P2J&index=30&ab_channel=Hung-yiLee 2022/10/15

強化學習 五子棋演算法 https://www.gushiciku.cn/pl/p3jQ/zh-tw 2022/10/15