# 基於強化學習之膠囊內視鏡自動化牽引研究

嚴聲揚 R06921096、謝忱 R06921088、黃浩恩 F05921120、賴棹沅 D05921011

***摘要*—為達到膠囊內視鏡之自動牽引，透過學習模擬環境之建置與強化學習神經網路的應用，使電腦知曉控制膠囊內視鏡之邏輯，於模擬環境中完成任務，並評估使用於實務上之可行性。**

***Keywords*—**

## I. 簡介

隨國人飲食變化、作息之改變，台灣大腸癌之發生機率已經成為世界第一。據衛生福利部國民健康署於 105 年公布之癌症登記報告中指出，平均每 4 分鐘 58 秒就有 1 人罹癌，並於 10 大癌症排行榜中，大腸癌已連續 11 年居冠[1]。

由於現今之醫學與醫材之進步，透過腸鏡之檢查，發現並治癒，生還機率皆可到達 90%[2]。現今台灣大型醫院、診所多數皆使用傳統之大腸內視鏡來進行檢測與篩檢，但有礙於傳統內視鏡之侵入性、所造成的疼痛性，常使病人感受到恐懼、延緩篩檢時辰；傳統大腸內視鏡之操作亦有人員操作之熟練度差異，除了可能使病人不適外，亦可能因技術不純熟而造成傷害。

基於此，膠囊內視鏡成為現今熱門的研究課題。本研究團隊透過可磁控之膠囊內視鏡來進行人體腸道內之牽引，除降低病人之不適感、疼痛感與恐懼感外，亦可透過多項電腦輔助系統與技術來輔助醫生之操作與檢測，提升檢測品質與可靠性。

本研究團隊透過自主開發之磁力牽引平台(Magnetic Field Navigator, MFN)，將磁控膠囊進行牽引與控制。於控制策略中亦提出自動化牽引之目標，並結合強化學習(Reinforcement learning)之技術，使機器透過特定之模擬環境的建設、獎勵等設計，達到自走之目的。

## II. 研究方法

### A. *模擬環境建置*

為要有效的模擬、訓練機器，本研究團隊使用 pyglet 建置模擬環境，如圖 1，其中紅色之區塊路徑為模擬之大腸路徑，綠色方框為模擬機械手臂之座標位置，淡藍色方框則為模擬膠囊內視鏡之座標。於環境的設計中，將本研究團隊開發之磁力差動探測技術之特性整併於模擬環境，設計以下三種之環境狀態、獎勵制度與其可行動作，期許透過這三種模型，達到膠囊自動牽引之目的。

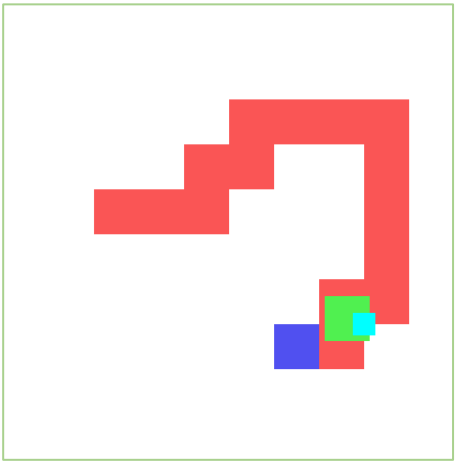

圖 1：模擬環境之 pyglet 建置介面

*1) 模擬環境之狀態 (State)*

<u>模型 1</u>：

模擬環境之狀態維度設計為 5 個狀態，分別為：
機械手臂之 X 座標、
機械手臂之 Y 座標、
膠囊內視鏡之 X 座標、
膠囊內視鏡之 Y 座標、
膠囊是否於手臂之正下方訊號。

其中膠囊是否於手臂正下方之訊號之來源，來自模擬差動探測技術之訊號。

<u>模型 2</u>：

模擬環境之狀態維度設計為 6 個狀態，分別為：
機械手臂之 X 座標與膠囊內視鏡之 X 座標之差、
機械手臂之 Y 座標與膠囊內視鏡之 Y 座標之差、
機械手臂與膠囊內視鏡之座標差的斜率、
前一個動作與狀態的獎勵、
膠囊是否於手臂之正下方訊號、
機械手臂之強力磁場與膠囊內視鏡小磁鐵之吸引力。

其中兩者磁鐵的吸引力之來源，來自模擬差動探測技術之訊號。

<u>模型 3</u>：

模型 3 之環境狀態設計為圖像輸出，格式為 3*80*80 之圖像矩陣。圖像資訊主要描述手臂與膠囊內視鏡行走過的路徑與現行已知的狀態，如圖 2(右)。該模型之狀態設計為：膠囊可行走之路徑將會標記為綠色，並隨著經過的次

數增加；倘若磁力差動探測技術判斷目前之位置為腸壁，則該位置將會於圖像矩陣中標示為紅色；若膠囊沒有拜訪過的座標位置，圖像矩陣將會標示為黑色。其中淡藍色之方框為手臂之位置，白色之方框為膠囊內視鏡於座標之位置。

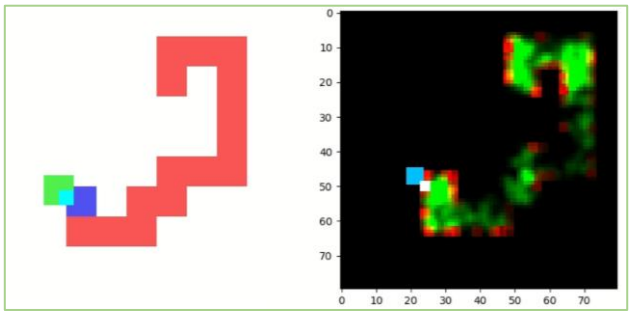


圖 2：模型 3 之環境建置。(左) pyglet 建置介面，(右)模型 3 之狀態圖像輸出

*2) 模擬環境之獎勵制度 (Reward)*

模型 1：

模型 1 之獎勵制度分成兩種狀態：

第一種為膠囊內視鏡脫離手臂之控制，則獎勵制度 R 設計如下：

$$R = -\sqrt{(arm['x'] - obj['x'])^2 + (arm['y'] - obj['y'])^2}$$

第二種狀況為膠囊內視鏡被控制於手臂下：

$$R = \text{Distance of path from the start point}$$

模型 2：

模型 2 之獎勵制度分成五種狀態：

第一種為膠囊內視鏡脫離手臂之控制，則獎勵制度 R 設計如下：

$$R = -\sqrt{(arm['x'] - obj['x'])^2 + (arm['y'] - obj['y'])^2}$$

第二種狀況為膠囊內視鏡被控制於手臂下，並走在已走過的路徑上：

$$R -= 0.5$$

第三種狀況為膠囊內視鏡被控制於手臂下，並走在未走過的路徑上：

$$R = 2$$

第四種狀況為膠囊內視鏡被控制於手臂下，並走在已走過的附近(半徑 5 個 pixel)：

$$R -= 0.5$$

第五種狀況為膠囊內視鏡被控制於手臂下，並走在已走過的附近(半徑 5 個 pixel)：

$$R += 10$$

模型 3：

模型 3 之獎勵制度分成兩種狀態：

第一種為膠囊內視鏡脫離手臂之控制，則獎勵制度 R 設計如下：

$$R = -\sqrt{(arm['x'] - obj['x'])^2 + (arm['y'] - obj['y'])^2}$$

第二種狀況為膠囊內視鏡被控制於手臂下，並透過圖像矩陣之走過路徑之綠色、紅色含量作為評分對象：

$$R += 125 - (imageTable[obj['x','y'][:,:,1]) - (imageTable[obj['x','y'][:,:,0])$$

*3) 模擬環境之控制動作 (Action)*

模型 1：

模型 1 之控制動作設計為 4 個動作，分別為：機械手臂向上 5 個移動單位、機械手臂向下 5 個移動單位、機械手臂向左 5 個移動單位、機械手臂向右 5 個移動單位。

模型 2：

模型 3：

模型 2、3 之控制動作設計為 8 個動作，分別為：機械手臂向上、右上、右、右下、下、左下、左、左上肢方向移動 5 個移動單位。

*B. 強化學習神經網路*

本研究採用以 DQN 為架構之兩種神經模型，為 DQN 與 Dueling DQN。

*1) DQN 模型*

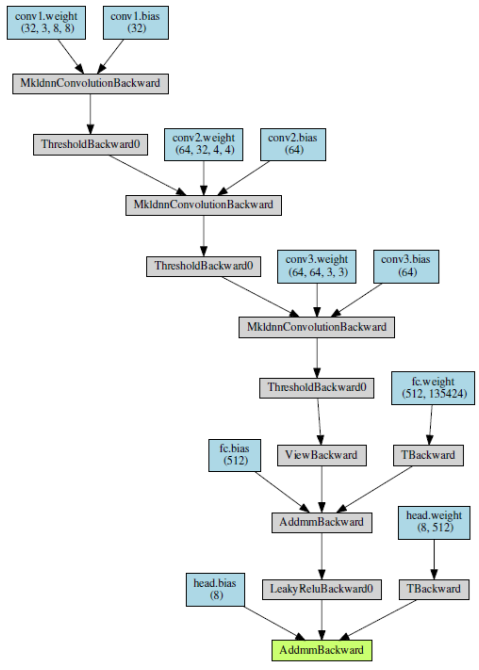


圖 3：DQN 模型架構

*2) Dueling DQN 模型*

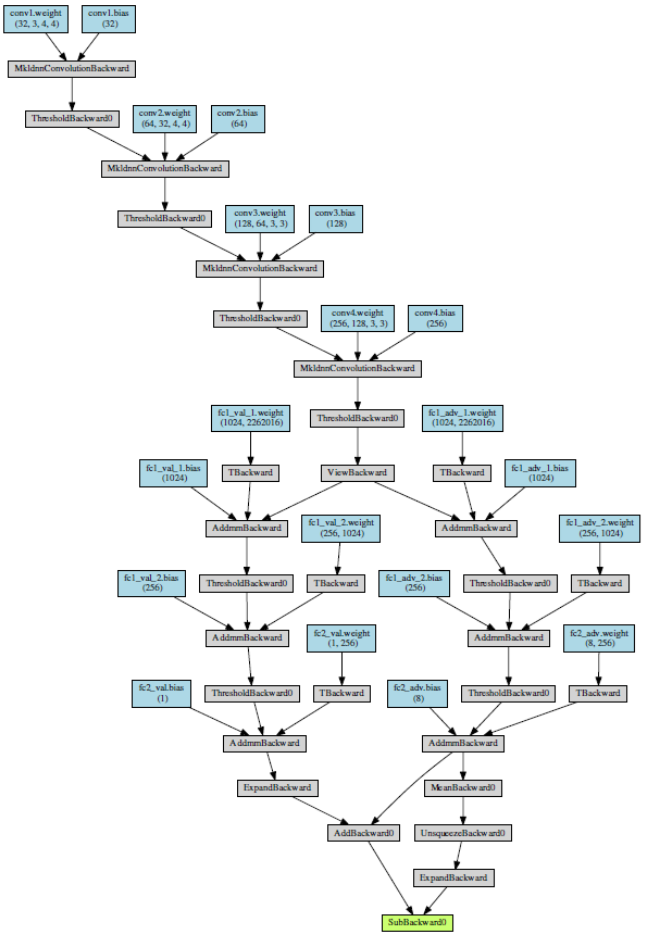


圖 4：Dueling DQN 模型架構

## III. 實驗模擬與結果

模型 1：

模型 1 因環境狀態(State)之設計不完整，雖然電腦已學會持續將機械手臂去抓取膠囊內視鏡，但電腦亦將隨機產生之模擬地圖之實際座標資訊記下，而使模擬之結果不盡理想。故繼續將模型 1 之 State 加以更改、優化。

模型 2：

因獎勵機制(Reward)經過特殊的設計與評估，使模型 2 學會靠著道的邊界來進行移動，其訓練過程之 Reward 曲線如圖 5。大約於 20 個 epoch 後，進步已趨近平緩，平均單次走完地圖之動作約為 2000 步，如圖 6。

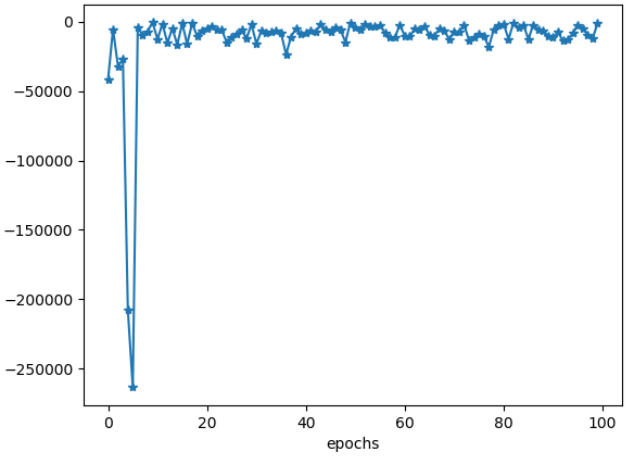


圖 5：模型 2 於 DQN 下之訓練結果

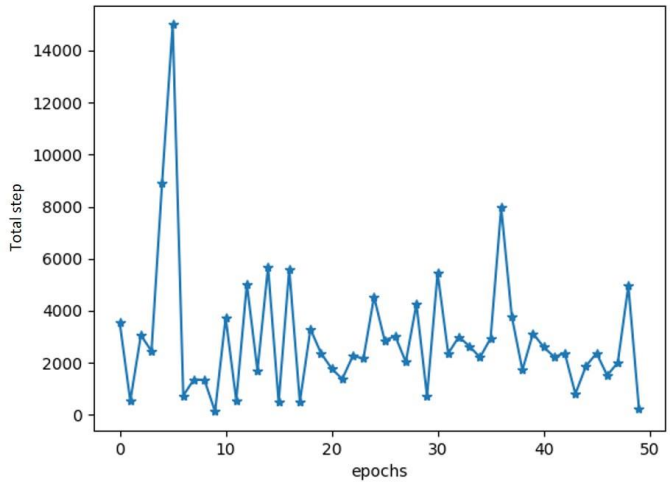


圖 6：模型 2 單次回合之 step 次數

模型 3：

模型 3 之 State 設計較特別，因此訓練也較緩慢。大約至 150 個 epoch 後進步才趨近平緩，如圖 7。平均單次走完地圖之動作約為 2000 步，如圖 8。

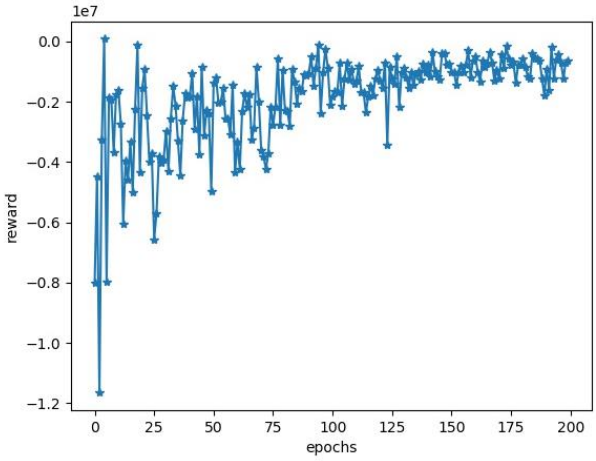


圖 7：模型 3 於 Dueling DQN 下之訓練結果

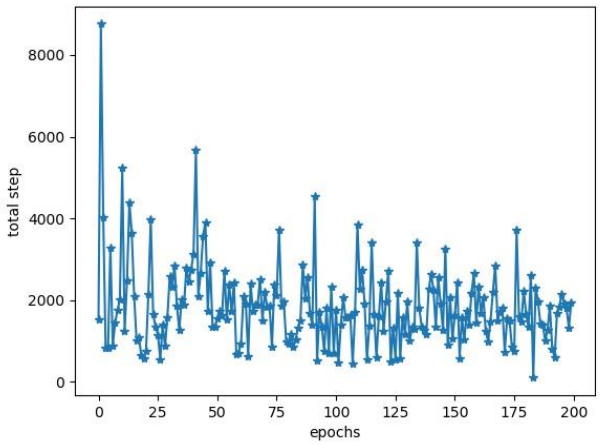


圖 8：模型 3 次回合之 step 次數

## IV. 工作分配

於此次之 Final Project 工作分配如下表：

| 姓名 | 工作內容 |
|---|---|
| 嚴聲揚 | 模型 2 之 Environment 參數更改、模型 2 之 Model 建置與訓練 |
| 謝忱 | 模型 1 之參數更改、Reward 設計與更改 |
| 黃浩恩 | 模型 3 之 Environment 建置、Reward 設計與更改、Dueling DQN 建置與訓練 |
| 賴棹沅 | 模型 2 之 Model 參數更改、文書報告 |

## V. 結論

透過模擬環境之建置與相關參數之設計，使強化學習可以透過該模型中完成期望之任務。由模型 2、模型 3 之模擬結果中可以知曉，雖然神經網路有學習到相關之狀態與關係間的關係，但平均單一地圖之動作次數皆接近 2000 步，倘若採用該方法至實際場合，可能將導致機械手臂頻繁振擺，並整體運作時間拉長，運用於實務上可能需再精進。

## REFERENCES

[1] CANCER REGISTRY ANNUAL REPORT, 2016. HEALTH PROMOTION ADMINISTRATION MINISTRY OF HEALTH AND WELFARE, TAIWAN. December 2018.

[2] Lucarini G, Ciuti G, Mura M, et al. A New Concept for Magnetic Capsule Colonoscopy base