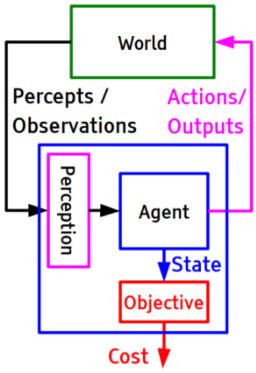
工業物聯網 應用查找功課

廖沁旋 N96104080

一、 Reinforcement Learning & Q-learning



A. 概念: 增強式學習以白話來說就是讓 AI 從通過每一次的錯誤來進行學習。在 AI 系統中可以分成三個主要組成: 1. Agent, 負責接受環境的相關資訊，並選擇最大化獎勵的行動，2. Perception: 評估環境的狀況 3. Objective: agent 想要找到的目標。(更詳盡的內容可以在 datacamp 上看到，但因為這不是本次功課的重點所以就附上連結就好: https://pse.is/3s3n4x)

B. 背景技術:

I. MDP (Markov Decision Process) 馬可夫決策過程:

因為 Agent 一開始並不知道環境的狀態，因此只能從曾經歷的 observation，action，reward 跟現在所得到的 observation, reward 來去當作現在的狀態，以數學公式呈現即 $s_t = f(o_1, r_1, a_1, ..., a_{t-1}, o_t, r_t)$。但如果需要預測評估下一個狀態(St+1)就要把 S1~St 的所有狀態給考慮進去，這樣模型便會非常的大，所以在理想情況下可以假設下一個狀態只跟現在這個狀態有關，就能把模型給縮小，也可以說是 MDP 的一種模型，數學呈現為此 $P(s_{t+1} \mid s_t) = P(s_{t+1} \mid s_1, ..., s_t)$ 。

II. Q-learning

Q-Learning 是一種基於著名的貝爾曼方程的離策略、無模型 RL 算法：$v(s) = \mathbb{E}[R_{t+1} + \lambda v(S_{t+1})|S_t = s]$上式中的 E 指的是期望，而 λ 指的是貼現因子。我們可以把它改寫成 Q 值的形式：

$$Q^{\pi}(s,a) = \mathbb{E}[r_{t+1} + \lambda r_{t+2} + \lambda^2 r_{t+3} + \ldots | s,a]$$

$$= \mathbb{E}_{s'}[r + \lambda Q^{\pi}(s',a')|s,a]$$

用 Q* 表示的最優 Q 值可以表示為：

$$Q^*(s,a) = \mathbb{E}_{s'}[r + \lambda \max_{a'} Q^*(s',a')|s,a]$$

目標是最大化 Q 值。

(參考資料: https://pse.is/3tppnc

https://pse.is/3syml2
https://pse.is/3tj3bf)

二、 Application: Aerobatic Helicopter Flight

1. 專案概要: 此專案讓直升機能透過增強式學習學會自主前空翻、低軸速滾動、tail-in funnel, and nose-in funnel，直升機相關的研究都有高維度、不對稱、嘈雜、非線性、非最小相位動態等的問題需要解決，因此需要有完善的演算法及模型才能控制得以。

2. 該實際應用之程式架構或流程圖

a. 資料蒐集

- 蒐集飛行員在操作直升機時的動作數據資料

- 根據現有模型找到一個仿真的控制器，並在直升機上測試是否有效

b. 模型訓練

- 直升機狀態由(x, y, z)方向組成、速度 (x•,y•,z•) 和角速度 (ωx, ωy, ωz)，並且由 4 維控制動作空間（u1、u2、u3、u4），通過使用循環俯仰 (u1, u2) 和尾槳 (u3) 控制，這些參數可以使飛行員可以圍繞其每個主軸旋轉直升機並將直升機帶到任何方向。

- 由於需要考慮到來自不同方向帶給直升機的力、加速度及飛行時的角度造成的影響，在訓練模型時需要將資料進行降維度(如做積分以獲得直升機隨時間變化的狀態)來降低模型學習問題的維度，運用方式如下:

$$
\begin{aligned}
\ddot{x}^b &= A_x \dot{x}^b + g_x^b + w_x, \\
\ddot{y}^b &= A_y \dot{y}^b + g_y^b + D_0 + w_y, \\
\ddot{z}^b &= A_z \dot{z}^b + g_z^b + C_4 u_4 + E_0 \|(\dot{x}^b, \dot{y}^b, \dot{z}^b)\|_2 + D_4 + w_z,
\end{aligned}
$$

$$
\begin{aligned}
\dot{\omega}_x^b &= B_x \omega_x^b + C_1 u_1 + D_1 + w_{\omega_x}, \\
\dot{\omega}_y^b &= B_y \omega_y^b + C_2 u_2 + C_{24} u_4 + D_2 + w_{\omega_y}, \\
\dot{\omega}_z^b &= B_z \omega_z^b + C_3 u_3 + C_{34} u_4 + D_3 + w_{\omega_z}.
\end{aligned}
$$

c. 設計控制器

- Reinforcement Learning Formalism and Differential Dynamic Programming (DDP): 用馬爾可夫決策過程 (MDP) 來描述，它包括六元組 (S, A, T, H, s(0), R)。這裡 S 是狀態集；A 是一組動作或輸入； T 是動力學模型，它是一組概率分佈;H 是范圍內的時間步數； $s(0) \in S$ 是初始狀態；$R : S \times A \rightarrow R$ 是獎勵函數。

- 線性反饋控制器可以使用動態規劃有效地計算

d. DDP 設計選擇

e. 獎勵函數中的權衡

獎勵函數包含 24 個特徵，包括平方誤差狀態變量、平方輸入，連續時間步之間輸入的平方變化，以及平方積分錯誤狀態變量。利用逆強化式學習去調整參數，並且人工調整迭代的次數。

4. 此應用能再做何種延伸或使用在哪些領域

由於我的實驗室是做自動控制相關的，因此我認為增強式學習也可以用在自走車(自駕車)上，並且透過與此篇論文相同的方式進行參數調整，即可以讓車隊進行自動編隊。若能與影像辨識進行結合，那麼就能將其應用在現在一般道路的紅綠燈上，用影像去辨識車流、人流量，在依照過去的資料訓練去進行紅綠燈的轉換、秒數的設定。此外，增強式學習也有運用在我們熟知的 DeepMind(擊

敗世界上排名最高的圍棋棋手)、Google X 的學習機器人、車輛導航系統及協助工業物流自動化的流程中，因此我認為只要設計到”自動控制”的系統都可以使用增強式學習來進行自動化處理。