台中，中華民國 111 年 11 月 05 日
Taichung, November 5th, 2022
論文編號：YY-XX

強化學習應用於無人機姿態控制

Apply reinforcement learning for UAV attitude control

吳柏勳 [a]、蕭富元 [b]
[a, b] 淡江大學航空太空工程學系

Wu, Po-Hsun[a], Hsiao, Fu-Yuen[b]
[a, b]Department of Aerospace Engineering, Tamkung University

摘要

本研究採用強化學習實現無人機的姿態控制，利用 OpenAI GYM package 建立強化學習的環境後，再使用強化學習演算法與 Python/Tensorflow 對環境進行學習。

關鍵字：無人飛行載具、強化學習、OpenAI GYM、Python/Tensorflow

一、緒論

1.1 研究動機

近年來機器學習技術日漸成熟和電腦運算速度的提升，機器學習開始大量的應用於影像辨識、自然語言處理、文本分析…等領域，透過大量的訓練資料和機器學習演算法來訓練模型使其達成我們所希望達到的目標。

而在控制領域通常都需要將非線性的模型線性化後，再運用 PID 或 LQR…等方法來設計出控制器，而設計出的控制器也會因為線性化模型的緣故，在遠離平衡點時，容易與真實狀況不符合導致系統發散。

若我們利用機器學習演算法針對非線性的數學模型於電腦上進行大量模擬和訓練，即可得到一個以神經網路為基礎的控制器，也因為在訓練的過程中使用的模型是非線性的，所以在狀態遠離平衡點後就使控制器無法有效的達到目標。

1.2 文獻回顧

[1] [2] [3]

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

1.3 研究方法

本研究是利用強化學習演算法來訓練決策使其可以有效的控制無人機的姿態，首先利用 Python/Tensorflow 建立一個神經網路來作為產生動作的決策，再使用 OpenAI GYM 架構建立強化學習的環境，而無人機的動態模擬是使用文獻 [3] 所計算出的動態方程式進行建模，最後利用強化學習演算法蒐集決策與環境間互動的資料進行計算，最終神經網路藉由改變神經網路的參數來最佳化獎勵函數來達到我們控制無人機姿態的目標。

二、強化學習

2.1 介紹

強化學習 (Reinforcement learning, RL) 屬於機器學習的一種，與其他機器學習方法不同的是，強化學習是基於與環境 (Environment) 進行互動來獲得獎勵 (Reward)，藉由強化學習演算法改善決策 (Policy) 最終得到一個能夠最大化獎勵函數的決策。

如圖 1，強化學習主要由決策、環境和演算法組成，決策會從環境中獲得狀態 (State) 後，根據不同的狀態反饋給環境不同的動作 (Action)，而環境得到決策所給予的動作後，也會計算出下一個狀態和獎勵值給予決策，而演算法會去蒐集每個時間下的狀態、動作和獎勵值，藉由獎勵值的大小來改變決策，獎勵值大的動作會使決策增加該動作的出現機率，反之，獎勵值小的動作則會減小決策執行該動作的機率，藉由大量的學習的過程使每次的動作都能產生最大的獎勵值。

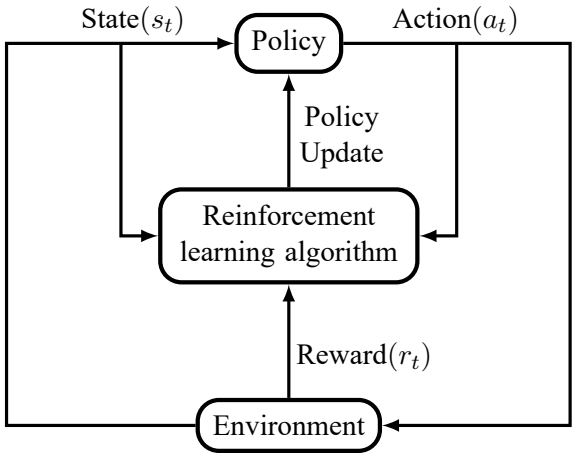

圖 1: 強化學習架構

2.2 PPO 演算法

PPO 演算法 (Proximal Policy Optimization alogrithm) 是由 OpenAI 於 2017 年提出的演算法，PPO 是由 TRPO(Trust region policy optimization) 將最佳化問題簡化後所得出，經文獻 [4] 證實 PPO 演算法具有 TRPO，而在文獻 [4] 的結果裡，PPO 演算法比起其他種強化學習的演算法具有更好的整體性能。

三、建模與訓練

3.1 OpenAI GYM

3.2 訓練結果

四、結論

參考文獻

[1] W. Koch, “Flight controller synthesis via deep reinforcement learning,” *CoRR*, vol. abs/1909.06493, 2019.

[2] W. Koch, R. Mancuso, R. West, and A. Bestavros, “Reinforcement learning for UAV attitude control,” *CoRR*, vol. abs/1804.04154, 2018.

[3] 廖晉揚, “定翼無人機的最佳順滑模態控制,” 淡江大學航空太空工程學系碩士班, 2022.

[4] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, “Proximal policy optimization algorithms,” *CoRR*, vol. abs/1707.06347, 2017.