適應性跨環境訓練強化學習框架(ARC)

專題編號：113-CSIE-S013
執行期限：112 年第 1 學期至 113 年第 1 學期
指導教授：林惠勇
專題參與人員： 110590001　郭丞軒
110590004　林奕廷

一、摘要

在強化學習（Reinforcement Learning, RL）研究中，代理人泛化能力的提升是亟待解決的問題，尤其在多環境場景中，單一環境訓練的模型難以應對變異性強的情境。為此，我們提出了一個創新的自適應跨環境訓練框架 Adaptive Cross-Environment Reinforcement Training (ARC)，該框架允許代理人通過動態切換策略在多個相似但具差異的環境中交替或混和學習。這一設計提升了代理人在不同環境中的泛化能力，並在多環境場景中提供了靈活且高效的訓練方案。該框架支援多種強化學習演算法和不同類型的環境，具有高度的可擴展性。

關鍵詞：強化學習、跨環境訓練、泛化能力、框架設計

二、緣由與目的

在當前的強化學習領域中，大多數研究專注於提升代理人在單一環境中的表現。然而，許多實際應用場景要求代理人在不同的環境或情境中具備一致且穩定的決策能力，這需要模型具備強大的泛化能力。傳統的強化學習演算法在這一點上存在顯著限制，代理人在面對新的或變化的環境時經常表現出過擬合、學習效率低下等問題[1]。

為解決這一挑戰，我們提出了一個專門用於多環境訓練的強化學習框架。該框架的核心在於允許代理人在多個環境中進行交替學習，並通過動態策略切換機制來促進策略的一致性與通用性。我們的目標是提供一個靈活且強大的訓練框架，能夠提升代理人在不同任務中的適應性與性能，並為多環境強化學習的應用提供理論支持和技術實施方案。

三、框架設計與技術細節

(一) 多環境訓練機制

我們提出的框架基於兩種主要的多環境訓練概念，包含多環境交替訓練與多環境混和方法。這兩種方法能夠幫助代理人在多個相似但略有差異的環境中學習到更具泛化能力的策略，避免過度依賴單一環境的特徵並有效應對不同環境中的變異。以下為兩種方法的詳細說明。

1. 多環境交替訓練

在多環境交替訓練中，代理人依次在不同的環境中進行學習。框架通過動態環境切換機制根據代理人在當前環境中的學習狀態進行自動切換。這樣的設計能夠防止代理人在單一環境中陷入過擬合，促進策略的通用性和穩定性。

2. 多環境混和

多環境混和是另一種提升代理人泛化能力的方法。該方法參考 Mixreg[2]之概念，將來自多個環境的觀察信息（如畫面、行為、獎勵等）進行整合，創建一個包含多環境特徵的混和環境。在這個泛化的環境中，代理人將接受來自不同環境的綜合學習信號，幫助其訓練出能夠在多樣化情境下適應的策略。這種方式進一步提升代理人應對未知情境的靈活性與適應能力。

(二) 強化學習演算法支持

我們的框架具有高度的靈活性，並接入了主流的強化學習框架，如 Stable Baselines 3[3]和 DI-engine[4]，這些框架內置了眾多先進的強化學習演算法，並提供豐富的配置選項，適應各種不同的訓練需求。這兩大框架內置了各類先進的強化學習技術和工具，並具有完善的日誌記錄、模型儲存、性能監控等功能，為我們的研究提供了強大的基礎設施支持。通過集成這些強大的框架，我們的訓練系統能夠在

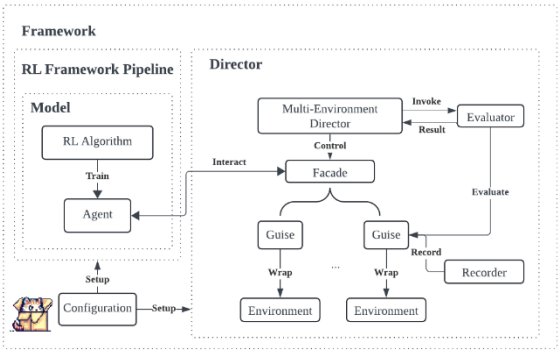
各種環境下靈活選擇和配置合適的演算法，確保代理人能夠在多環境學習中達到最佳效果。

(三) 配置與擴展性

框架的另一個特點是其高度靈活的配置系統，允許用戶根據具體應用場景和任務需求調整環境參數、演算法設置以及切換策略。在環境選擇方面，我們的框架支持來自多個平台（如 Gym[5] 及 Gymnasium[6]等）的多樣化環境，包括簡單到複雜的遊戲場景，如 Atari 遊戲環境等。這些環境在學習動作與策略方面具有一定的相似性，便於在不同環境間進行有效的交替訓練與混和學習。

此外框架支持在 Windows 和 Linux 平台上運行，並通過 TensorBoard 進行實時監控與性能分析，這使得其具備良好的擴展性與可操作性。

(四) 框架架構



四、 實驗結果

在實驗中，我們選擇 Gymnasium 中的 Atari 街機遊戲環境進行實驗，分別為 Demon Attack、Phenix 以及 Galaxian。我們將選擇兩組環境進行多環境訓練，分別為 Demon Attack 和 Phoenix (DP)以及 Demon Attack 和 Galaxian (DG)。在這兩組環境中，模型將以交替式訓練與混合式訓練進行學習，且所有訓練的總步數皆設為一千萬步（10M）。訓練完成後，我們將對訓練好的模型在每個單一環境中進行測試與評估。

實驗結果如 Table 1 顯示，無論是在哪個遊戲環境下，混合式學習和交替式學習的表現均明顯優於隨機策略，證明了這兩種學習方式在提升代理人泛化能力方面的有效性。

Table 1 混合式學習、交替式學習及隨機策略分數比較

Env		Blend 10M	Switch 10M	Random
DP	D	1005	416.9	152.1
	P	1894	2822.5	761.4
DG	D	1002	432	152.1
	G	1783	1654	764.3

註：D = Demon Attack、P = Phoenix、G = Galaxian

五、 結論

我們提出的自適應跨環境訓練框架為強化學習模型提供了一個強大的訓練平台，通過動態切換策略與多環境交替學習，代理人能夠有效學習到具備泛化性的策略，並在不同任務中展現出較高的適應能力。該框架具有高度的靈活性和擴展性，能夠適應多種應用需求，為未來的多任務強化學習研究提供了堅實的技術基礎。

未來的研究工作將進一步探索該框架在更多複雜環境中的應用，並計畫引入更多先進的演算法（如 MuZero），以進一步提高訓練效率與模型性能。我們也將嘗試結合元學習技術，來提升代理人在新環境中的快速適應能力。

六、 參考文獻

[1] Cobbe, K., Hesse, C., Hilton, J., & Schulman, J. (2020). Leveraging Procedural Generation to Benchmark Reinforcement Learning. In *Proceedings of the 37th International Conference on Machine Learning* (pp. 2048–2056). PMLR.

[2] Wang, K., Kang, B., Shao, J., & Feng, J. (2020). Improving Generalization in Reinforcement Learning with Mixture Regularization. arXiv:2010.10814

[3] Raffin, A., Hill, A., Gleave, A., Kanervisto, A., Ernestus, M., & Dormann, N. (2021). Stable Baselines3: Reliable

Reinforcement Learning Implementations. Journal of Machine Learning Research, 22, 1–8. https://github.com/DLR-RM/stable-baselines3

[4] DI-engine Contributors. (2021). DI-engine: Reinforcement Learning platform [Software]. OpenDILab. https://github.com/opendilab/DI-engine

[5] Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., & Zaremba, W. (2016). OpenAI Gym. OpenAI. https://gym.openai.com/

[6] Farama Foundation. (2022). Gymnasium (Version 0.26.3) [Software]. https://gymnasium.farama.org/