

# 深度強化式學習技術之應用研究

計畫主持人：[吳毅成教授](#)

## 概要

近年來，深度強化式學習（Deep Reinforcement Learning, DRL）已被應用於許多人工智慧應用問題中。其中一個重大的成果，就是 AlphaGo Zero，在本計畫中稱為「Zero 技術」，從「零知識」開始學圍棋，超越所有人類棋士以及圍棋 AI 程式，是突破性的成果。本計畫將聚焦在 DRL 和 Zero 技術的五大研究主題：(1) 持續研發高段圍棋程式 CGI。(2) 運用 Zero 技術研發其他類遊戲 AI。(3) 結合 Zero 技術與確切解之研究。(4) 研發電玩遊戲的 AI bot。(5) 研發機器手臂工件夾取技術。

## 關鍵字

深度強化式學習、強化式學習、深度學習、蒙地卡羅樹搜尋、AlphaGo Zero、電腦對局、圍棋、電玩遊戲、賽車遊戲、機器人、機器手臂工件夾取技術

## 創新

- 提出一種新的價值網路—多標籤價值網路，能為圍棋遊戲輸出不同貼目下的盤面價值，同時也能降低訓練均方誤差。
- 開發基於蒙地卡羅樹搜尋（MCTS）的棋力調整方法。並進行理論分析，透過使用閾值比率，能夠保證調整後 MCTS 所選擇的棋步具有一定品質。
- 驗證 Zero 技術能夠套用於非確定性遊戲，開發 2×4 中國暗棋 Zero 程式。
- 提出雙曲正切衰減學習率調整機制，可應用於隨機梯度下降（SGD）訓練。
- 提出狀態離散化方法，可以對環境變化的感知進行離散化，並生成狀態轉移圖。
- 發展新的權重交叉熵方法，在機器手臂夾取任務中可以達到近 100% 的成功率（DDPG 僅 70%）。
- 開發新的 end-to-end 混合動作空間 DRL 方法—參數化近端策略優化，應用於機械手臂夾取與推送任務，可大幅提升成功率至 99%。

## 效益

- 透過結合多標籤價值網路與 CGI 圍棋程式，此計畫開發了世界上第一個在不同貼目下皆能對弈的圍棋程式，此結果已發表於 IEEE Transactions on Games。

- 開發圍棋終身學習系統，是世界上第一個能夠提供不同等級（從初學者到超越職業棋士）的電腦圍棋系統。此結果參加 2018 未來科技展，並且發表於頂級會議 AAAI-19（錄取率僅為 1,150 / 7,095 = 16.2%）。
- 本研究開發的 2x4 中國暗棋 Zero 程式為世界上第一個隨機遊戲 Zero 程式，該篇論文也獲得 TAAI 2018 國際會議的最佳論文獎。
- 雙曲正切衰減（Hyperbolic-Tangent Decay）的論文已發表於 IEEE WACV 2019 會議。
- 研發分散式 end-to-end DRL 演算法並成功應用於產學合作計畫，發展賽車遊戲，並取得超過最頂尖測試玩家之表現。
- 發展狀態離散化方法，預計將可應用於許多 DRL 研究。
- 參數化近端策略優化發表於 NeurIPS 2018 會議的 Infer2Control Workshop。

## Studies of Applications with Deep Reinforcement Learning Technologies

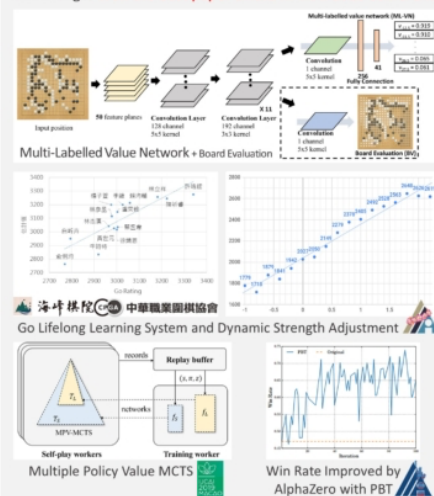
Principal Investigator: I-Chen Wu

Recently, **Deep Reinforcement Learning (DRL)** has been applied to many AI applications. One of the successful achievements is the **AlphaZero**, called the **Zero method in this project**, was presented to learn without human knowledge and surprisingly surpass all the human players and all the AI programs.

This project focus on three classes of DRL applications: **(1) Lightweight model**, e.g., Go program, Zero method, and exact methods. **(2) Heavyweight model**, e.g., AI bot of video games. **(3) Real-world model**, e.g., DRL applications for robotics.

### Class 1: Lightweight-Model Applications

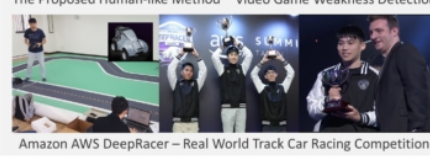
- Developed the **first-ever Go program that can play under different komis** by using the multi-labelled value network (ML-VN). **Accepted by IEEE Transactions on Games.**
- Proposed the multiple policy value MCTS (MPV-MCTS), which combines networks of various sizes to retain their advantages. **Accepted by IJCAI-19. (acceptance rate: 850/4,752 = 17.8%)**
- Improved AlphaZero with population based training, **achieved stat-of-the-art of 20 blocks ResNet. Accepted by AAAI-20 with an oral presentation. (acceptance rate: 1,591/7,737 = 20.6%; oral presentation rate: 453/7,737 = 5.85%)**
- With above techniques, the strength of our Go program **CGI surpassed Facebook's ELF Open Go v2 by a win rate of 74%.**
- Developed a computer Go lifelong learning system, which is the first Go system that **provides different strengths from beginners to super-human**, and dynamically analyzes the player's strength. The system is **selected for the 2018 Future Tech**, and has been **accepted by AAAI-19. (acceptance rate: 1,150/7,095 = 16.2%)**
- Developed the 2x4 Chinese Dark Chess Zero program, which is the first Zero program for stochastic games in the world to our knowledge. **Won the best paper award in TAAI 2018 conference.**



### Class 2: Heavyweight-Model Applications

- Developed a high-performance distributed end-to-end DRL algorithm, which is the first algorithm that **can surpass human experts with just screen image input in a racing game.**
- Proposed a new DRL method to **make AI bots behave like humans without affecting the performance**, and successfully applied the method to industrial-university joint projects.
- Detect weakness for video games based on trial and error.
- We won the **3rd place of Amazon AWS DeepRacer 2019 World Championship Cup**, the **1st place of Taipei summit circuit**, and also the **1st place of October virtual circuit (with a lap time 7.172s, which is the fastest record on all virtual circuits).**

Case	Score	Shake	Spin	Games
Human	29.14	8.16	0.03	80
Baseline (PPO)	35.61	39.60	0.56	240
Biological Constrain	33.08	0.22	3.47	240
Human-like method	37.24	0.45	1.00	240



### Class 3: Real-World-Model Applications

- Proposed a new weighted cross entropy method (WCCEM) and combine it with DDPG. **DDPG+WCCEM greatly improves the success rate of robotics grasping tasks to 96%**, while DDPG can only reach 70%.
- Proposed a new end-to-end hybrid action space DRL method, Parameterized Proximal Policy Optimization (P3O), that greatly **improves the accuracy of grasping and pushing tasks to nearly 99%. Accepted by Infer2Control @ NeurIPS 2018.**
- Proposed a new hyperbolic-tangent learning rate decay (HTD) which can be applied with SGD. **Accepted by IEEE WACV 2019.**

