THÔNG TIN CHUNG CỦA NHÓM

- Link YouTube video của báo cáo: https://youtu.be/Z55JDlaQYCY
- Link slides: https://github.com/Skizdukion/CS2205.CH183/blob/main/slide.pdf

- Họ và Tên: Phạm Thăng Long
- MSSV: 240101016

- Lớp: CS2205.CH183
- Tự đánh giá (điểm tổng kết môn): 8.5/10
- Số buổi vắng: 1
- Số câu hỏi QT cá nhân: 3
- Link Github: https://github.com/Skizdukion

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

Mở rộng Alpha Zero vào các trò chơi có thông tin không hoàn hảo.

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

Adapt Alpha Zero idea in to imperfect information games

TÓM TẮT *(Tối đa 400 từ)*

Sự thành công của Alpha Zero trong việc tạo ra một thuật toán đã đánh bại con người trong các môn cờ cổ điển như chess, Go, và shogi đã đánh dấu một bước ngoặt mới trong lĩnh vực trí tuệ nhân tạo và đặc biệt trong lĩnh vực học tăng cường nói chung. Tuy nhiên, khi mở rộng sang các trò chơi có thông tin không hoàn hảo như Poker hay Liar Dice, kiến trúc Alpha Zero gặp phải thách thức lớn. Thuật toán cây tìm kiếm (MCTS) của Alpha Zero vốn hiệu quả trong môi trường thông tin hoàn hảo, trở nên không hiệu quả khi tương tác với các thông tin ẩn trong trò chơi. Đề cương này trình bày một phương pháp tiên tiến có khả năng giải quyết vấn đề này, mở rộng khả năng của Alpha Zero sang các trò chơi thông tin không hoàn hảo. Chúng tôi đề xuất một kiến trúc mạng neuron mới, tích hợp khả năng suy luận về hành động của đối thủ dựa trên việc lý giải thông tin ẩn.

Điểm cốt lõi của phương pháp này là cải tiến thuật toán tìm kiếm cây với kiến trúc mạng mới, cho phép hệ thống lý luận hiệu quả trong môi trường thông tin không chắc chắn. Trong quá trình duyệt cây, thuật toán duy trì phân phối xác suất của các trạng thái ẩn có thể có của đối thủ. Sau đó mở rộng cây tìm kiếm bằng cách dự đoán xác suất các hành động đối thủ, từ đó giúp người chơi đưa ra chiến thuật đối đầu hợp lý và linh hoạt. Phương pháp này hứa hẹn mở ra hướng đi mới trong việc ứng dụng học tăng cường vào các bài toán phức tạp trong thực tế, nơi thông tin thường không đầy đủ và hành vi của đối tượng khó dự đoán.

GIỚI THIỆU *(Tối đa 1 trang A4)*

Alpha Zero đã đạt được thành công ấn tượng trong các trò chơi cờ vua và cờ vây bằng cách kết hợp Deep neural network và thuật toán tìm kiếm tìm kiếm cây (MCTS). Mạng neural, đóng vai trò bộ não, dự đoán khả năng các nước đi hợp lý và ước lượng giá trị trạng thái (VD: vị thế ván cờ) từ các thông tin trong trò chơi và MCTS dùng với mạng neural để tìm kiếm, chọn nước đi tốt nhất

Tuy nhiên, phương pháp này gặp phải thách thức lớn khi áp dụng cho các trò chơi có thông tin không hoàn hảo như Poker hay Liar Dice. Trong các trò chơi thông tin hoàn hảo, trạng thái trò chơi luôn được hiển thị đầy đủ cho tất cả người chơi, cho phép thuật toán có được chiến lược tối ưu thông qua việc duyệt cây tìm kiếm. Ngược lại,

trong các trò chơi thông tin không hoàn hảo, người chơi chỉ có quyền truy cập vào một phần thông tin, đòi hỏi khả năng suy luận thông tin ẩn, dự đoán hành động của đối thủ, và đánh giá mức độ rủi ro dựa trên thông tin không đầy đủ. Sự khác biệt cốt lõi này khiến cho việc trực tiếp áp dụng Alpha Zero vào các trò chơi thông tin không hoàn hảo trở nên cực kỳ kém hiệu quả

Đề cương này tập trung vào việc khắc phục hạn chế này bằng cách đề xuất một phương pháp mới, mở rộng kiến trúc của Alpha Zero để hoạt động hiệu quả trong môi trường thông tin không hoàn hảo. Ý tưởng chính là bổ sung cho Alpha Zero khả năng suy luận về thông tin ẩn và hành vi của đối thủ. Để đạt được mục tiêu này, chúng tôi đề xuất một vài thay đổi ở kiến trúc mạng neural trong phương pháp gốc và cùng với đó là một thuật toán cây tìm kiếm phù hợp với kiến trúc mạng neural này để có thể lý luận được thông tin ẩn của đối thủ trong quá trình duyệt cây, do đó tìm kiếm hiến lược tối ưu.

Và chúng tôi tích hợp các phương pháp mới này dựa trên cơ chế tự chơi của Alpha Zero, từ đó tối ưu hóa chiến lược đối đầu trong môi trường thông tin không hoàn hảo. Kết quả của nghiên cứu này hứa hẹn sẽ mở ra một hướng đi mới cho việc áp dụng học tăng cường vào các trò chơi có thông tin không hoàn hảo và các bài toán ra quyết định trong thực tế.

MỤC TIÊU *(Viết trong vòng 3 mục tiêu)*

- Nghiên cứu, đề xuất phương pháp để cho kiến trúc Alpha Zero có khả năng suy luận thông tin ẩn
- Hiện thực kiến trúc mới và huấn luyện mô hình trên Leduc Poker (một biến thể đơn giản của Poker)
- So sánh đánh giá được kiến trúc mới so với các mô hình hiện tại

NỘI DUNG VÀ PHƯƠNG PHÁP

Để đạt được các mục tiêu trên, nghiên cứu này sẽ tập trung vào các nội dung và phương pháp sau:

1. Thiết kế và tích hợp kiến trúc mạng nơ-ron có khả năng lý luận thông tin ẩn

Phương pháp:

- Thay vì kiến trúc một mạng nơ ron ở Alpha Zero, chúng tôi đề xuất bằng một kiến trúc mới gồm 2 mạng nơ ron:
    - Mạng học chiến lược và định giá hành động: Mục tiêu của mạng này là đưa ra chiến lược tối ưu cho người chơi và ước tính giá trị dựa trên trạng thái hiện tại. Ở kiến trúc gốc thì chỉ cần một mạng này là đủ để cho hệ

thống có thể lý luận được chiến lược của đối thủ bằng cách duyệt cây tìm kiếm
    - Mạng suy luận thông tin ẩn: Mục tiêu của mạng này là suy luận về các thông tin ẩn của đối thủ, chẳng hạn như bài riêng trong Poker. Phương pháp là sử dụng mạng neuron sâu để phân tích thông tin công khai (ví dụ: hành động đã thực hiện, các lá bài chung) và trạng thái ẩn của người chơi để ước tính phân phối xác suất của các trạng thái ẩn có thể có của đối thủ.

2. Thiết kế, chỉnh sửa thuật toán tìm kiếm cây cho phù hợp với lý luận thông tin ẩn

Phương pháp:

- Thuật toán cây tìm kiếm mới sẽ được thiết kế để trực tiếp sử dụng thông tin suy luận từ mạng suy luận thông tin ẩn:
    - Cụ thể, ở khâu expandsion, sau khi thuật toán lựa chọn một node để mở rộng, thuật toán sẽ đưa thông tin từ node này vào cả 2 mạng học chiến lược và định giá hành động, và mạng suy luận thông tin ẩn
    - Sau đó từ xác suất trạng thái ẩn được lấy từ output của mạng suy luận thông tin ẩn thuật toán thực hiện sample để được một trạng thái ẩn
    - Nếu trạng thái được sample chưa tồn tại trong cây thì thuật toán sẽ tạo một nốt mới dựa trên trạng thái đó và chuyển sang độ sâu tiếp theo
    - Nếu trạng thái đã tồn tại nốt đã tồn tại thì sẽ di chuyển sang nốt đó và thực hiện expandsion như trong thuật toán gốc của Alpha Zero

3. Huấn luyện mô hình

Phương pháp:

- Tích hợp cơ chế self play vào kiến trúc
- Thiết kế hàm loss cho quá trình huấn luyện bao gồm:
    - Mạng học chiến lược và định giá hành động $L = (z-v)^2 - \pi^T \log p + c \,//\, \theta \,//\, ^2$, trong đó:
        - **Loss cho giá trị $(z-v)^2$:** z là kết quả thực tế của trò chơi (1 nếu thắng, -1 nếu thua, 0 nếu hòa), v là giá trị ước lượng từ mạng (dự đoán khả năng chiến thắng).
        - **Loss cho chính sách $\pi^T \log p$:** π là phân phối xác suất hành động từ MCTS, p là phân phối hành động được dự đoán bởi mạng chính sách.
        - **Regularization $c \,//\, \theta \,//\, ^2$:** θ là các tham số của mạng nơ-ron. c là một hệ số điều chỉnh.
    - Mạng suy luận thông tin ẩn $L = \sum y_i \log(y\hat{}_i)$, trong đó:

        - y_i là giá trị thực thông tin ẩn của đối thủ (one-hot, chỉ có 1 trạng thái ẩn bằng 1, còn lại là 0).
        - $\hat{y}_i$ là xác suất mô hình dự đoán (softmax output).
- Thực hiện hyperparameters searching để tìm kiếm tham số hiệu quả cho Leduc Poker

4. Đánh giá kiến trúc mới

Phương pháp:

- Training một số thuật toán trong học tăng cường như PPO, RPG, NeuRD hay thử triển khai một số thuật toán SOTA trong Poker như Deepstack, Pluribus trong Leduc Poker
- Thiết kế môi trường cho phép kiến trúc mới tương tác với các thuật toán đã triển khai
- Lưu trữ, so sánh các chỉ số như Expected Value (EV) hoặc Chips Won Per Hand (BB/100)

KẾT QUẢ MONG ĐỢI

- Xây dựng được môi trường trò chơi Leduc Poker cho phép người thật tương tác với mô hình
- Kiến trúc đề xuất có thể huấn luyện được mô hình đạt tới Nash Equilibrium trong Leduc Poker
- Mở ra một hướng nghiên cứu mới trong việc huấn luyện mô hình trong môi trường thông tin không hoàn hảo

TÀI LIỆU THAM KHẢO *(Định dạng DBLP)*

[1] Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., et al. (2017). Mastering chess and shogi by self-play with a general reinforcement learning algorithm. arXiv preprint arXiv:1712.01815.
[2] Zinkevich, M., Johanson, M., Bowling, M., Piccione, C.: Regret minimization in games with incomplete information. In: Advances in Neural Information Processing Systems, pp. 1729– 1736 (2008)
[3] Brown, N., Lerer, A., Gross, S., Sandholm, T.: Deep counterfactual regret minimization. In: International Conference on Machine Learning, pp. 793–802 (2019)
[4] Noam Brown, Anton Bakhtin, Adam Lerer, and Qucheng Gong. Combining deep reinforcement learning and search for imperfect-information games. In Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.
[5] H. Fu, W. Liu, S. Wu, Y. Wang, T. Yang, K. Li, J. Xing, B. Li, B. Ma, Q. Fu et al., "Actor-critic policy optimization in a large-scale imperfect-information game," in International Conference on Learning Representations, 2021, pp. 1–12.
[6] Brown N, Sandholm T. Superhuman ai for multiplayer poker. Science. 2019;365:eaay2400.
[7] Lanctot, M., Waugh, K., Zinkevich, M., Bowling, M.: Monte Carlo sampling for regret minimization in extensive games. In: Advances in Neural Information Processing Systems, pp. 1078–1086 (2009)
[8] Moravcˇ´ık, M. et al. Deepstack: Expert-level artificial intelligence in heads-up nolimit poker. Science 356, 508–513 (2017).
[9] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347, 2017.