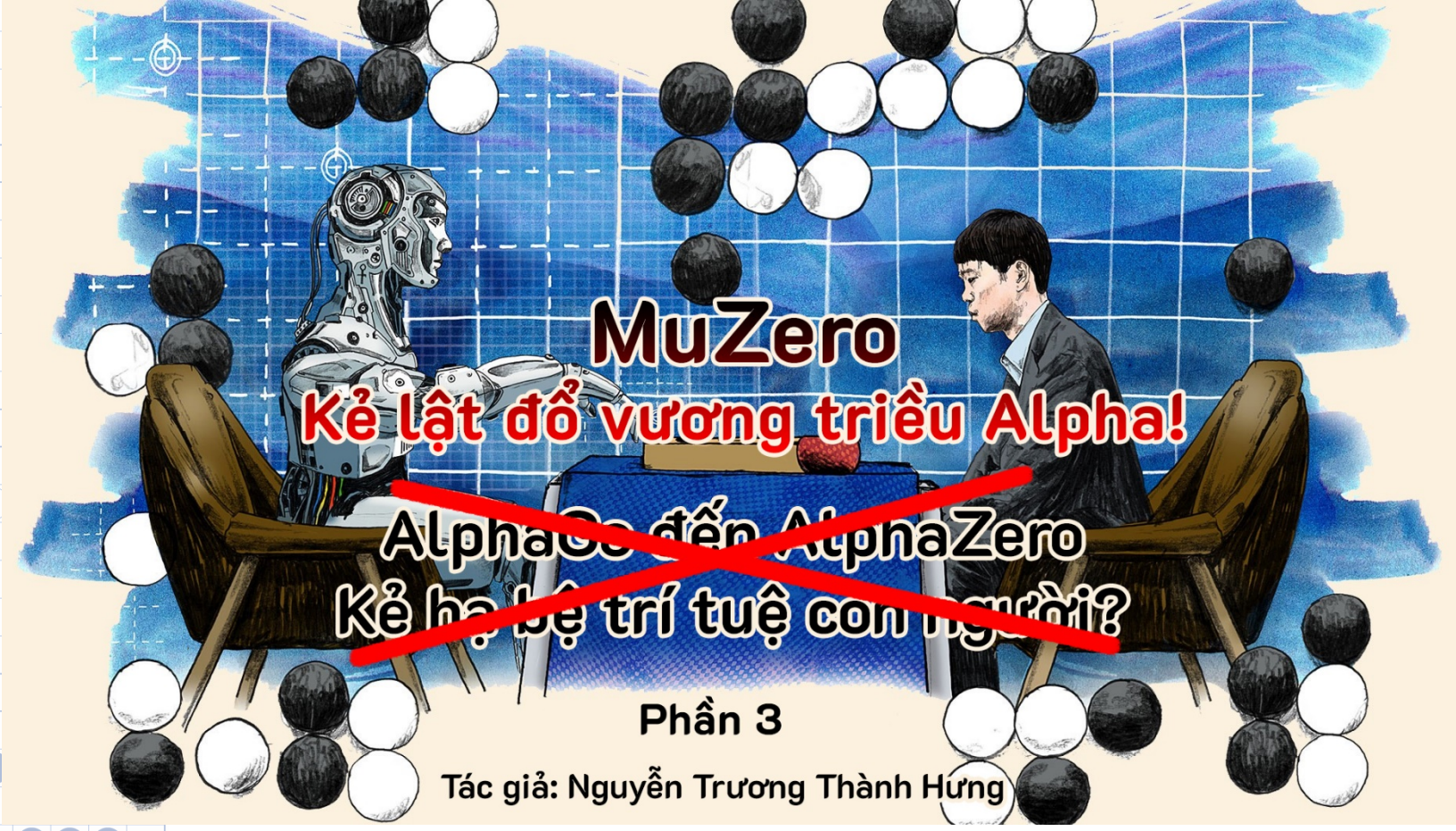


Phần 3: MuZero - Kẻ lật độ vương triều Alpha!

Đồng hành qua hai bài viết của series, hẳn các bạn đã phần nào hiểu được cấu trúc, cơ chế hoạt động và "bí thuật" được sử dụng bên trong những kì thủ vô tri, từ AlphaGo đến AlphaZero. Khi đã nắm bắt được muôn hình vạn trạng của Trí tuệ nhân tạo, hi vọng rằng nhiều bạn sẽ không còn cảm thấy lo sợ về tương lai của loài người trước Trí tuệ nhân tạo nữa. Trí tuệ nhân tạo sẽ dần thay con người thực hiện những công việc lặp đi lặp lại nhàm chán, và trở thành bàn đạp cho con người chúng ta vươn tới những vùng trời kiến thức rộng lớn hơn.

Chính nguồn động lực, khát khao khám phá chân trời mới đã kích thích đội ngũ xây dựng gia đình Alpha tại DeepMind tiếp tục đào sâu nghiên cứu và cải tiến AlphaZero trở nên "con người" nhất có thể.

Quả nhiên, đội ngũ DeepMind đã làm được điều đó! Cuối năm 2019, thế hệ kì thủ vô tri mới nhất – **MuZero** đã nhanh chóng ra đời để truất ngôi vương của triều đại Alpha và đáp trả trước câu hỏi hóc búa *"Liệu trí tuệ nhân tạo có thể tự đào tạo bản thân y như cách một đứa trẻ học đi, học nói?"*.

Nào! Chúng ta cùng tìm hiểu xem **MuZero** "bá đạo" đến mức nào để có thể phế truất vương triều Alpha nhé!

MuZero

Tháng 11 năm 2019, DeepMind xuất bản bài báo với tựa đề *"Mastering Atari, Go, Chess and Shogi by Planning with a Learned Model"* để giới thiệu về MuZero – thế hệ kì thủ kế vị AlphaZero. Bài báo này đã gây được tiếng vang lớn trong cộng đồng Trí tuệ nhân tạo vào thời điểm bấy giờ. Điều gì đã khiến MuZero trở nên khác biệt với AlphaGo hay AlphaZero đến vậy?

Như đã đề cập ở phần trước, tuy AlphaZero không cần học hỏi nước cờ của con người, nhưng AlphaZero vẫn sử dụng môi trường trò chơi được định nghĩa rõ ràng như luật chơi, nước đi khả thi... Điều này làm khả năng áp dụng của AlphaZero bị giới hạn ở những trò chơi 2 người chơi có luật rõ ràng, có kết quả chung cuộc, ví dụ như cờ vua, khi ván cờ khép lại, ta luôn có kết quả thắng/thua/hòa để đánh giá lại.

Nhưng trong nhiều lĩnh vực thực tế như robot, điều khiển công nghiệp, trợ lí thông minh hay các trò chơi điện tử (như bộ trò chơi *Atari* – bộ những trò chơi điện tử dành cho một người chơi), chúng ta sẽ không có kết quả cuối cùng mà sẽ nhận được intermediate rewards (phần thưởng tức thì) sau mỗi hành động. Vậy làm sao để có thể khái quát hóa khả năng học hỏi của AlphaZero lên những bài toán đời thực?

"Trong nhiều vấn đề của thế giới thực, chúng ta không biết "thế giới" hoạt động như thế nào, đó là một nơi không có bất kỳ quy tắc nào cả. Vì vậy, chúng tôi đã phát triển một biến thể mới có tên MuZero, học bằng cách tự chơi và học trong môi trường của riêng nó. Đó là một mô hình học theo cách thực sự chỉ tập trung vào những gì hữu ích để tối đa hóa phần thưởng. "

David Silver

–Trưởng nhóm nghiên cứu Học tăng cường tại DeepMind

Câu trả lời là: **MuZero không nhận bất cứ kiến thức nào về luật chơi cả!** Bí mật trong cách tiếp cận mới của MuZero là kiến thức về quy luật trò chơi không còn được "mã hóa" nữa. Ngược lại, MuZero sẽ tự học luật chơi thông qua cơ chế self-play (tự chơi). Tuy cơ chế tự chơi đã xuất hiện ở AlphaZero, nhưng phải đến MuZero, cơ chế này mới thực sự được làm chủ và độc lập.

Cơ chế lựa chọn hành động dựa trên cây tìm kiếm Monte Carlo được tái sử dụng. Tuy nhiên với MuZero, các quy tắc của trò chơi không cần phải lập trình trực tiếp vào cây tìm kiếm Monte Carlo nữa, mà **MuZero sẽ tự mình quản lý**

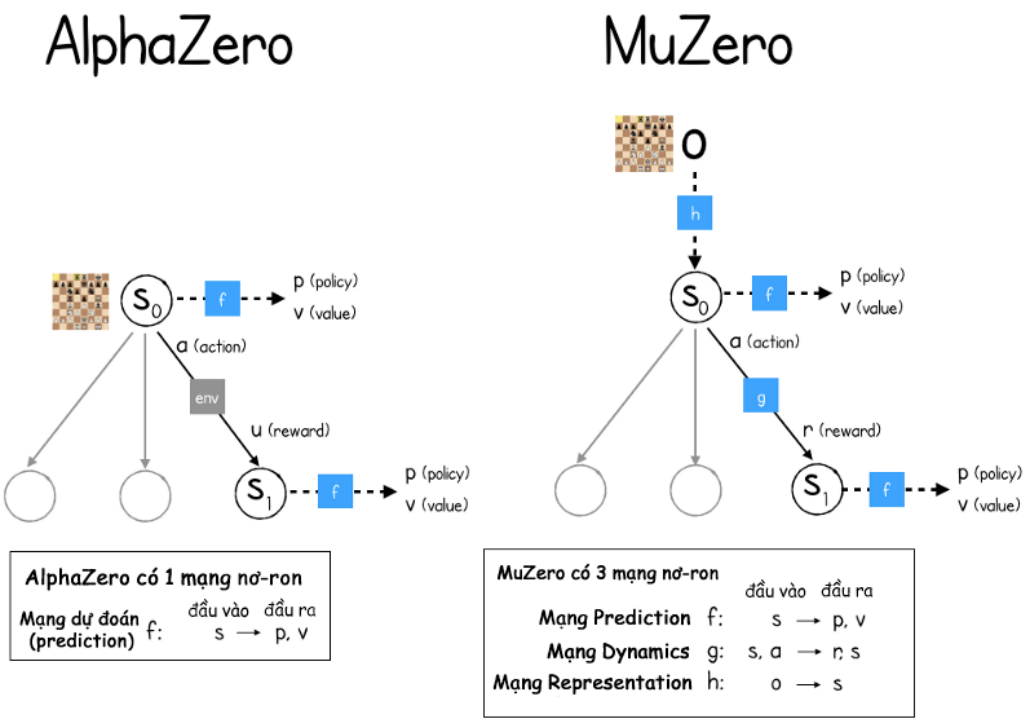
những representation (biểu diễn) về luật chơi. Điều này cũng tương tự một đứa trẻ đang tự học ngữ pháp theo cách hiểu "con nít" của riêng chúng nó vậy. Có thể nói, đây là một bước đột phá đáng ghi nhận trong lĩnh vực Học tăng cường nói riêng và Trí tuệ nhân tạo nói chung.

Vậy để hiểu hơn về cách MuZero tự học luật chơi, chúng ta sẽ cùng lột trần điểm cải tiến mạnh mẽ nhất ở MuZero là Hệ thống mạng nơ-ron và cơ chế Representation Learning (Học biểu diễn) nhé!

Hệ thống mạng nơ-ron & Học biểu diễn

Nếu MuZero không được biết trước quy luật của trò chơi, vậy làm sao để nó có thể quyết định được đâu là nước cờ tốt nhất từ bàn cờ hiện tại?

MuZero đã có một bước phát triển đáng kinh ngạc, đó là: MuZero học luật chơi bằng cách áp dụng Representation Learning và Dynamic Model. Nói dễ hiểu, MuZero đang tự **tạo ra trí tưởng tượng "tiềm ẩn" của riêng mình về môi trường và quy luật trò chơi**. 3 mạng nơ-ron bao gồm: Representation, Dynamic và Prediction sẽ là công cụ để **MuZero tối ưu hóa, "hiện thực hóa" trí tưởng tượng của mình trở nên giống với môi trường thực tế nhất.**



Hình 1. *Sơ đồ so sánh mạng nơ-ron cùng cây tìm kiếm Monte Carlo trong AlphaZero và MuZero.*

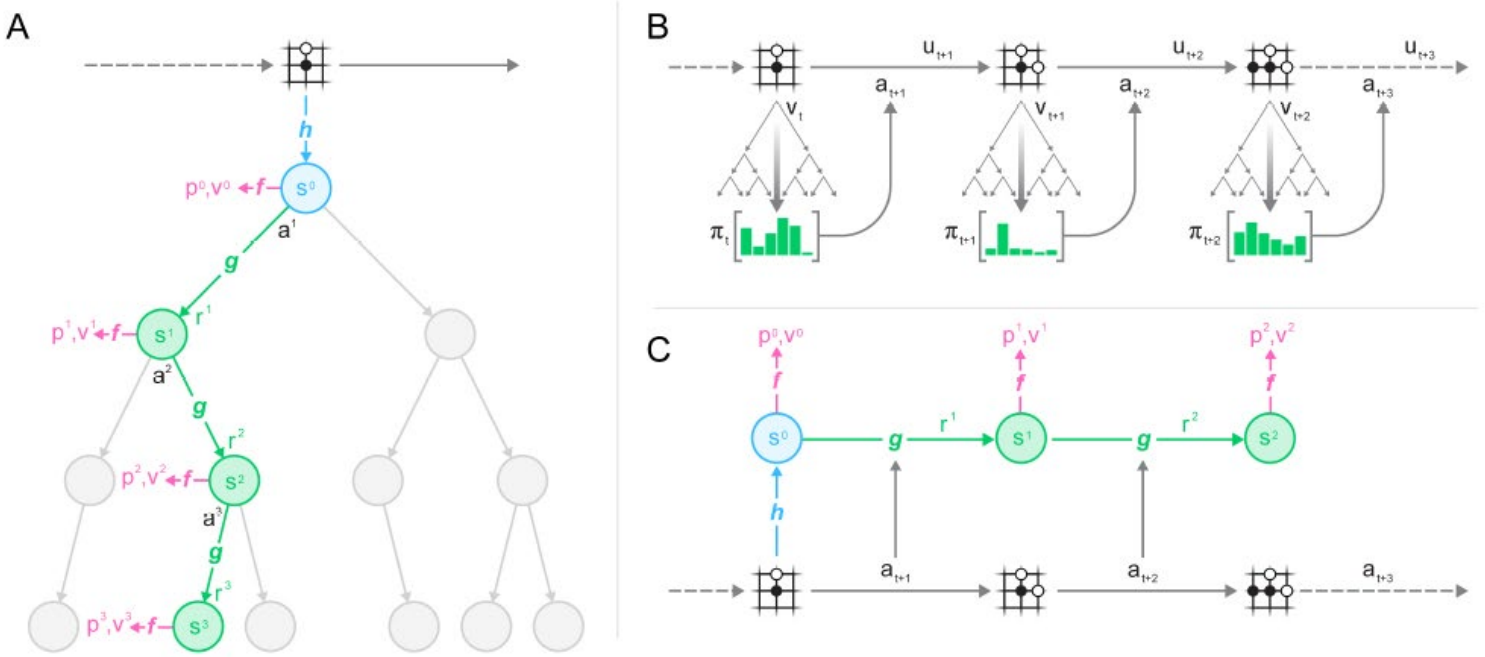
Ban đầu, **MuZero sẽ được phép quan sát** một số lượng bàn cờ nhất định (đối với board game) hoặc hình ảnh chụp lại quá trình chơi game (đối với các bộ trò chơi Atari) và đưa vào một mạng Biểu diễn (Representation) để **hình thành trạng thái trò chơi đầu tiên.** Đặc biệt hơn, hình ảnh từ các trò chơi trong bộ Atari còn được mã hóa vào không gian tiềm ẩn (Latent space) với mục đích là giảm số chiều và kích thước của đầu vào, giúp quá trình đào tạo mạng "nhanh gọn lẹ" hơn.

Như đã đề cập ở phần 2 về AlphaZero, công việc của mạng Prediction là dự đoán giá trị Policy và Value tại một trạng thái của bàn cờ. Tất cả giá trị Policy sẽ được phân phối trên những nước đi khả thi, được lấy từ luật của cờ.

Tương tự, MuZero cũng có một mạng Prediction, nhưng MuZero không có được môi trường trò chơi – một khối kiến thức xa xỉ nữa. Mà nay, trạng thái của bàn cờ sẽ là một biểu diễn tiềm ẩn. MuZero sẽ học cách **phát triển biểu diễn tiềm ẩn này thông qua một mạng nơ-ron Dynamic**. Mạng Dynamic sẽ nhận trạng thái tiềm ẩn (hidden state) của bàn cờ hiện tại và hành động (action) đã chọn để tính toán phần thưởng (reward) nhận được và một trạng thái bàn cờ tiếp theo (next state).

Quá trình này giống việc đứa trẻ kết nối "trí tưởng tượng" của mình về cách ăn, cách nói với những gì nó quan sát được từ thế giới thực.

Cuối cùng, các giá trị **Policy, Value và Reward sẽ được dùng để đào tạo lại 3 mạng nơ-ron**. Riêng giá trị Reward của các loại board game (cờ vua, cờ vây, cờ shogi...) là kết quả chung cuộc – thắng/thua/hòa, còn với các bộ trò chơi Atari hay trong các bài toán thực tế là kết quả đạt được sau một số hành động nhất định.



Hình 2.

Hình A) **Cách MuZero sử dụng model để lên kế hoạch.** MuZero đưa trạng thái bàn cờ mới vào 3 mạng nơ-ron Represetation (h), Dynamics (g) và Prediction (f) để đưa ra những giá trị Policy, Value và Reward cần thiết.

Hình B) **Cách MuZero tương tác với môi trường.** MuZero sử dụng cây tìm kiếm Monte Carlo (tương tự AlphaZero) để chọn ra hành động tiếp theo. Quá trình cứ lặp lại cho đến khi trò chơi kết thúc.

Hình C) Cách MuZero đào tạo model. Tại bước đầu tiên, MuZero sử dụng hình ảnh quan sát trong quá khứ để hình thành mạng Representation. Sau một số bước nhất định, trạng thái tiềm ẩn từ các những nước cờ trước đó cùng hành động tương ứng để đào tạo lại 3 mạng nơ-ron.

Hẳn phần nội dung về mô hình mạng nơ-ron của MuZero quá "bá đạo" và có phần khó hiểu phải không? Quả thật vậy, mô hình mạng của MuZero thực sự rất phức tạp và được áp dụng hầu hết những công nghệ mới nhất trong lĩnh vực Học tăng cường, tiêu biểu là Học biểu diễn (Representation Learning), sử dụng Không gian tiềm ẩn (Latent space), cùng nhiều "chiêu trò" khác. Đó là lí do, đến hiện tại, rất nhiều đội ngũ, nhóm chuyên gia về Học tăng cường trên thế giới cố gắng tái tạo lại MuZero dựa trên thông tin ít ỏi trong bài báo xuất bản bởi DeepMind. Song thực sự là chưa có ai dám công bố kết quả của mình tiệm cận với MuZero của DeepMind.

Kết quả và Đón nhận

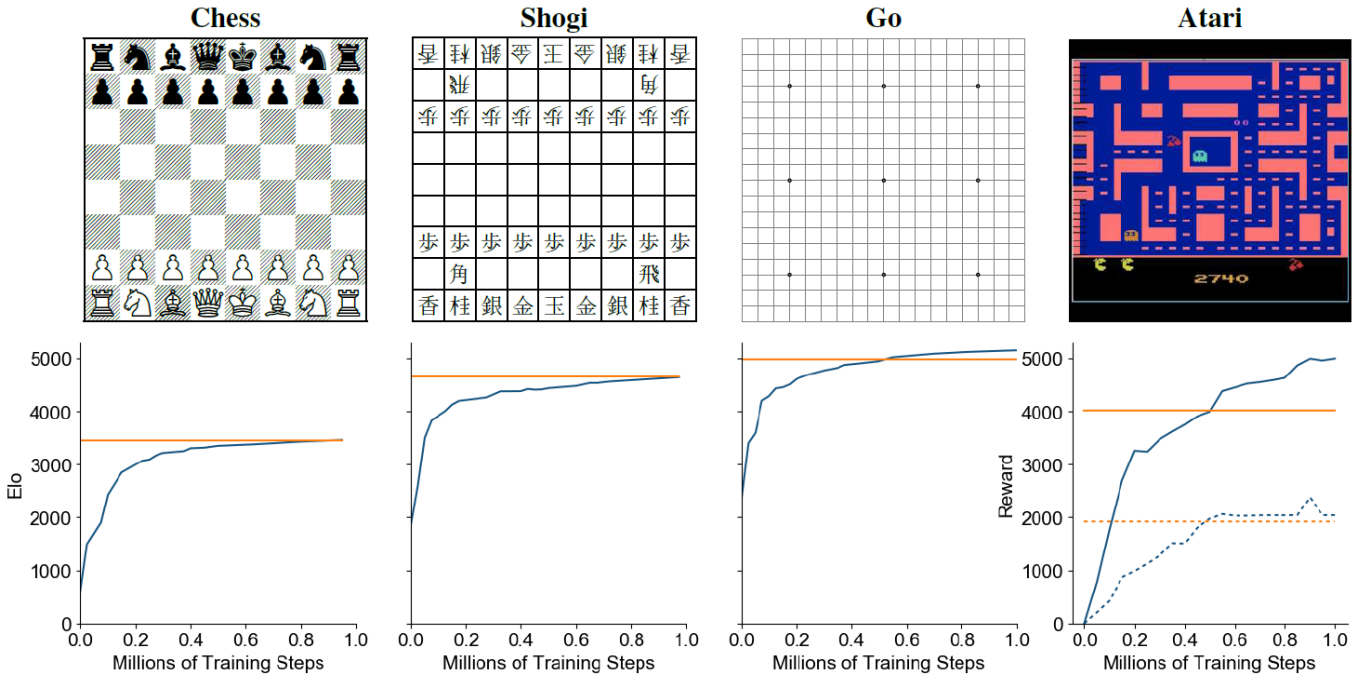
MuZero đã sử dụng **16 TPU thế hệ thứ ba để đào tạo** và **hơn 1000 TPU để tự chơi** đối với board game với 800 lần mô phỏng (simulation) tại mỗi bước, và 8 TPU để đào tạo và 32 TPU để tự chơi đối với bộ trò chơi Atari với 50 lần mô phỏng tại mỗi bước. Trong khi đó, AlphaZero đã sử dụng 64 TPU thế hệ đầu tiên để đào tạo và 5000 TPU thế hệ thứ hai để tự chơi. Khi thiết kế của TPU đã được cải thiện, TPU thế hệ thứ ba mạnh gấp 2 lần so với TPU thế hệ thứ hai, với những tiến bộ hơn nữa về băng thông và mạng, có thể nói **"tương quan lực lượng" của MuZero và AlphaZero là ngang nhau.**

Bởi vì MuZero là một **bước thay đổi lớn về mặt khái quát hóa khả năng học luật chơi và áp dụng được ở nhiều lĩnh vực**, nên kết quả đạt được của MuZero được hi vọng sẽ mở rộng hơn AlphaZero ở nhiều trò chơi khác nhau.

MuZero sánh ngang với thành tích của AlphaZero trong cờ vua, cờ shogi sau khoảng 1 triệu bước đào tạo. Thậm chí MuZero còn vượt trội hơn AlphaZero ở

bộ môn sở trường là cờ vây với 1 triệu bước đào tạo. Cũng chỉ với 1 triệu bước, MuZero cũng bỏ xa kĩ thuật tốt nhất lúc bấy giờ là Recurrent Replay Distributed Deep Q-Network (R2D2) ở hầu hết các trò chơi điện tử trong bộ trò chơi Atari.

Có thể nói, MuZero đã được cộng đồng đón nhận là một bước tiến mạnh mẽ so với AlphaZero, đặc biệt là khả năng khái quát hóa trong các kỹ thuật học tập không giám sát.

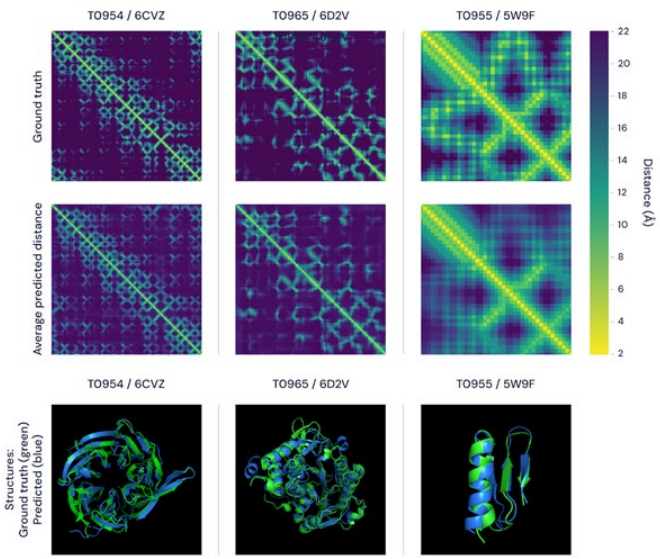


Hình 3. *Biểu đồ kết quả của MuZero so với AlphaZero ở cờ vua, cờ shogi và cờ vây (đánh giá dựa trên chỉ số Elo) và so với R2D2 ở bộ trò chơi Atari (đánh giá dựa trên giá trị Reward)*

Lời kết

MuZero đã trở thành con át chủ bài của DeepMind để chứng tỏ vị thế top đầu của họ trong lĩnh vực Trí tuệ nhân tạo trên bản đồ thế giới. Không cần quy tắc rõ ràng, tự chơi với góc quan sát đầy thông minh, có thể tóm gọn MuZero với 3 tính từ: **thanh lịch, giản đơn và thậm chí rất kì quặc!** Tiềm năng của MuZero bây giờ đã vượt xa khỏi các thể loại board game, nhờ cách tiếp cận tổng quát và linh hoạt hơn rất nhiều.

Cá nhân tôi là một người tập trung nghiên cứu về lĩnh vực Học tăng cường, tôi cũng phải thừa nhận rằng khả năng áp dụng Học tăng cường vào những bài toán thực tế vẫn còn gặp rất nhiều trắc trở, từ khả năng tính toán của máy móc đến tính thực tiễn. Sự ra mắt của MuZero đã khiến tôi đứng ngồi không yên, thức cả đêm cố gắng thấu hiểu được nội dung bài báo (thực ra đến bây giờ tôi vẫn chưa hiểu trọn được "cô gái MuZero" này nữa), và cảm nhận được tương lai mở ra cho lĩnh vực Học tăng cường. Chỉ trong vòng 4 năm, kể từ lúc AlphaGo ra đời và gây sóng gió trên khắp các mặt trận truyền thông, thì khi đến MuZero ra đời, con người chúng ta cũng chỉ tặc lưỡi cho qua, bảo nhau rằng vì nó là máy tính, nó thừa sức đánh bại con người là điều dĩ nhiên. Thậm chí đến thời điểm tôi viết bài viết này, đã có nhiều bộ máy khác mạnh mẽ hơn MuZero ở cả độ chính xác lẫn độ khái quát hóa như Atari57, OpenAI Five...



AlphaFold (phiên bản ứng dụng của AlphaZero) đóng góp to lớn trong lĩnh vực gấp protein. Protein là những phân tử lớn, phức tạp chỉ đạo gần như mọi chức năng trong cơ thể chúng ta. Chức năng của chúng phụ thuộc vào cấu trúc 3D độc đáo và mô hình hóa các nếp gấp của protein là một trong những lĩnh vực nghiên cứu y sinh hiện đại thú vị nhất.

Song, **bây giờ mới là giai đoạn Trí tuệ nhân tạo học hỏi cách con người học tập, lao động để thay thế những công việc nhàm chán của chúng ta.** Tôi hi vọng, qua series bài viết lần này, các bạn độc giả sẽ có cái nhìn gần gũi và thiện cảm hơn với Trí tuệ nhân tạo, và các bạn đam mê về Trí tuệ nhân tạo sẽ không ngần ngại và tự tin bước đi trên con đường rộng mở nhưng đầy chông gai này.

Cuối cùng, tôi xin gửi lời cảm ơn tới tất cả các bạn đã ủng hộ series của tôi, tới những đội ngũ đã và đang nghiên cứu về Trí tuệ nhân tạo.

Xin chào và hẹn gặp lại các bạn ở những bài viết tiếp theo!

Tham khảo

[1] "DeepMind's MuZero teaches itself how to win at Atari, chess, shogi, and Go | VentureBeat." https://venturebeat.com/2019/11/20/deepminds-muzero-teaches-itself-how-to-win-at-atari-chess-shogi-and-go/ (accessed Nov. 22, 2020).

[2] "How To Build Your Own MuZero AI Using Python (Part 1/3) | by David Foster | Applied Data Science | Medium." https://medium.com/applied-data-science/how-to-build-your-own-muzero-in-python-f77d5718061a (accessed Nov. 22, 2020).

[3] J. Schrittwieser *et al.*, "Mastering Atari, Go, Chess and Shogi by Planning with a Learned Model," Nov. 2019, Accessed: Nov. 22, 2020. [Online]. Available: http://arxiv.org/abs/1911.08265.

[4] "MuZero Explained | Papers With Code." https://paperswithcode.com/method/muzero (accessed Nov. 22, 2020).

[5] "MuZero: A new revolution for Chess? – Mathieu Acher – Associate Professor in Computer Science." http://blog.mathieuacher.com/MuZeroChess/ (accessed Nov. 22, 2020).

[6] "How an algorithm became superhuman at Go — but not StarCraft — and then moved on to modeling proteins - Heidelberg Laureate Foundation." https://www.newsroom.hlf-foundation.org/blog/article.html?tx_news_pi1%5Baction%5D=detail&tx_news_pi1%5Bcontroller%5D=News&tx_news_pi1%5Bnews%5D=205&cHash=2bcc541af83676edca1f8677f146086e (accessed Nov. 22, 2020).