AlphaGo đến AlphaZero: Kẻ hạ bệ trí tuệ con người?

Phần 2: Bắt đầu từ con số 0!

Nối tiếp phần 1 – *Tổng quan về sự tiến hóa từ AlphaGo đến AlphaZero*, trong bài viết lần này, chúng ta sẽ cùng đi sâu vào cách thức mà AlphaGo Zero và AlphaZero đã vượt khỏi giới hạn kiến thức của loài người.

Vào tháng 10 năm 2017, DeepMind xuất bản một bài báo với tựa đề *"Mastering the game of Go without human knowledge"* và phiên bản nâng cấp **AlphaGo Zero** đã chính thức được trình làng. Nó mang sức mạnh vượt mặt tất cả các phiên bản tốt nhất của AlphaGo chỉ trong vỏn vẹn 40 ngày với tỉ số toàn thắng 100-0. Điều đặc biệt chính là, kiến thức của AlphaGo Zero bắt đầu từ **"con số 0", không cần bất cứ nước đi chuyên gia của loài người.**

Vì **AlphaZero** là phiên bản tổng quát và mở rộng của AlphaGo Zero để có thể tương tác với các thể loại trò chơi khác ngoài cờ vây như cờ vua, cờ shogi, các trò chơi Atari. Vậy nên hôm nay, chúng ta sẽ cùng đi thẳng vào bóc tách nguyên lí hoạt động của **AlphaZero** luôn nhé!

AlphaZero

Bắt đầu từ con số 0!

Như đã đề cập ở phần 1, toàn bộ nền tảng của AlphaGo dựa trên 4 Deep Convolutional Neural Network (Mạng nơ-ron tích chập sâu), và kết quả từ Mạng trạng thái và Mạng giá trị được đưa vào một cây tìm kiếm Monte Carlo.

AlphaZero có một cách tiếp cận hoàn toàn khác so với AlphaGo. Trong đó, AlphaZero đã:

- Loại bỏ mạng học có giám sát trên nước đi của chuyên gia, **dữ liệu hoàn toàn tự sinh dựa trên cơ chế tự học, tự chơi (self-play).**
- **Cây tìm kiếm Monte Carlo** vẫn được sử dụng để đánh giá chính sách và giá trị của tình trạng bàn cờ, đồng thời nó cũng được cải tiến mạnh mẽ.
- Kết hợp Mạng chính sách và Mạng giá trị thành **một hệ thống mạng nơ-ron học sâu đơn lẻ duy nhất**.
- Chỉ sử dụng các thuật toán tổng quát và những quy tắc/luật chơi cơ bản của trò chơi.
- Cơ chế Đấu trường được áp dụng. Mỗi agent sẽ được đấu với những agent khác, để lựa chọn ra mô hình tốt nhất, mạnh mẽ nhất và tổng quát nhất.

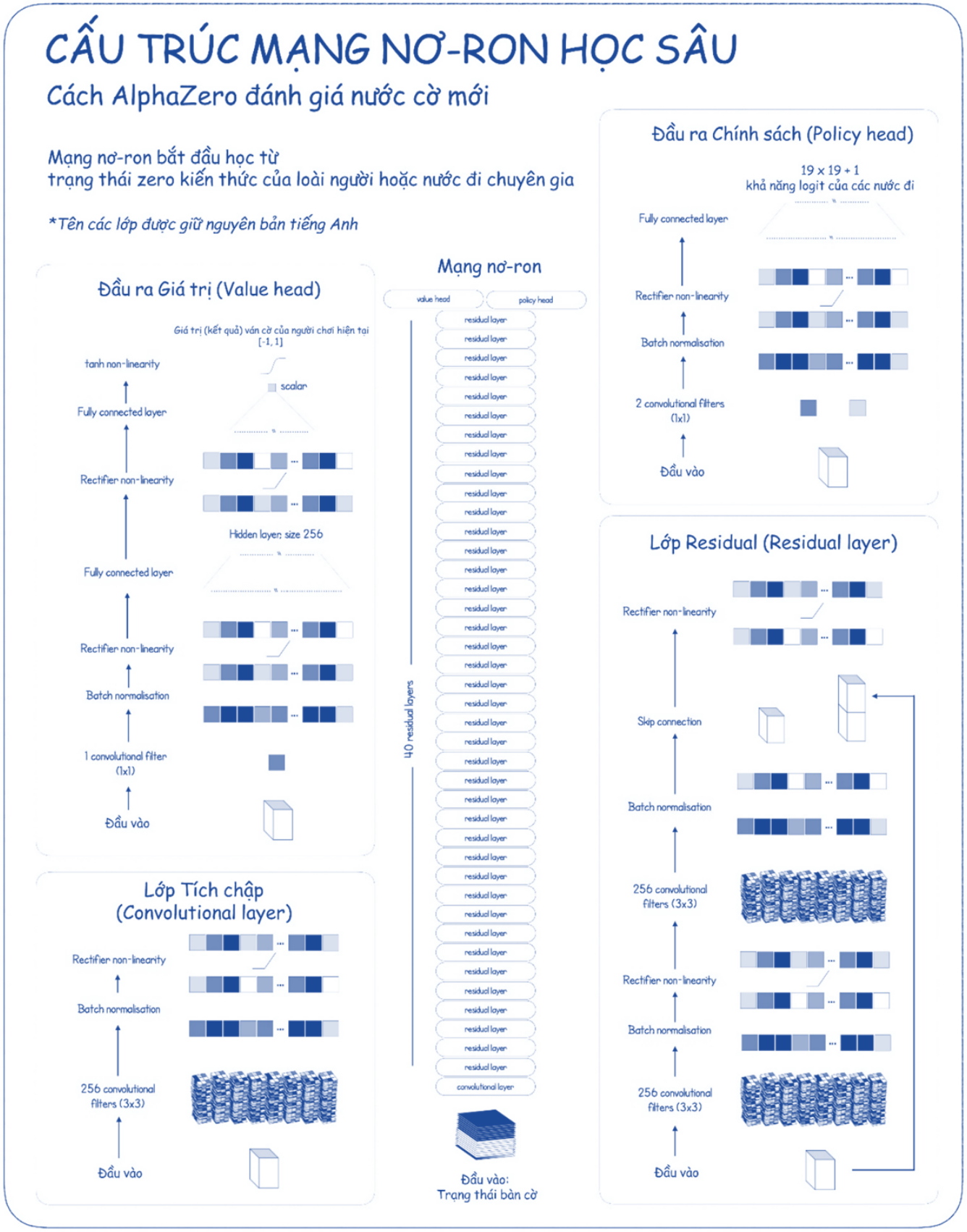
1. Mạng nơ-ron

Cũng như mọi mô hình trí tuệ nhân tạo khác, mạng nơ-ron đóng vai trò cốt lõi trong AlphaZero.

Cấu trúc mạng nơ-ron bên trong AlphaZero được đánh giá là rất sâu, với rất nhiều lớp, và được lấy ý tưởng từ mạng ResNet. Về tổng quan, đầu vào của mạng nơ-ron là tình trạng bàn cờ (vị trí các quân cờ trên bàn). Dữ liệu thông tin của bàn cờ tiếp tục được đưa vào 1 lớp Convolutional (tích chập), theo sau bởi 19 hoặc 39 lớp Residual. Cuối cùng, mạng nơ-ron xuất ra hai đầu là Value Head (Đầu ra Giá trị) và Policy Head (Đầu ra Chính sách). Đây chính là điểm nâng cấp rất quan trọng ở AlphaZero so với AlphaGo. AlphaZero đã **kết hợp Mạng chính sách và Mạng giá trị** để cùng chia sẻ các tham số.

Ý tưởng của mạng nơ-ron là để học hỏi xem những trạng thái nào sẽ dẫn đến kết quả thắng hoặc thua. Ngoài ra, quá trình học Chính sách cung cấp ước tính khả thi hơn về những nước cờ tốt nhất tại trạng thái bàn cờ hiện tại. Kiến trúc của mạng nơ-ron nói chung sẽ phụ thuộc hoàn toàn vào trò chơi. Hầu hết các loại trò chơi bàn cờ như cờ vây, cờ vua, cờ shogi... đều có thể sử dụng kiến trúc mạng tích chập nhiều lớp như trong hình 1.

Với một cấu trúc dày đặc như vậy, nhiều nhà nghiên cứu trên thế giới tranh luận rằng, sức mạnh thực sự ở AlphaZero nằm ở khả năng tính toán khổng lồ mà DeepMind (dưới sự hậu thuẫn của Google) đã áp dụng. Có một điều hài hước là, trong bài báo của DeepMind, họ ghi rằng, AlphaZero chỉ cần được đào tạo trong **4 tiếng đồng hồ**. Nhưng thực chất, đó là 4 tiếng huấn luyện trên bộ máy với cấu hình tính toán khổng lồ: **5000 TPU** thế hệ thứ nhất cho cơ chế tự chơi và **64 TPU** thế hệ thứ hai để đào tạo mạng nơ-ron. Không chỉ vậy, 1 TPU (Tensor Processing Unit) của Google còn mang sức mạnh **gấp ít nhất 100 lần so với GPU cao cấp nhất**. Chỉ với những con số như thế thôi, hẳn bạn cũng có thể tưởng tượng được AlphaZero đã được huấn luyện bằng bộ máy có sức mạnh siêu việt đến thế nào.



Hình 1. *Cấu trúc mạng nơ-ron của AlphaZero*

2. Cây tìm kiếm Monte Carlo & Cơ chế Self-play (tự chơi)

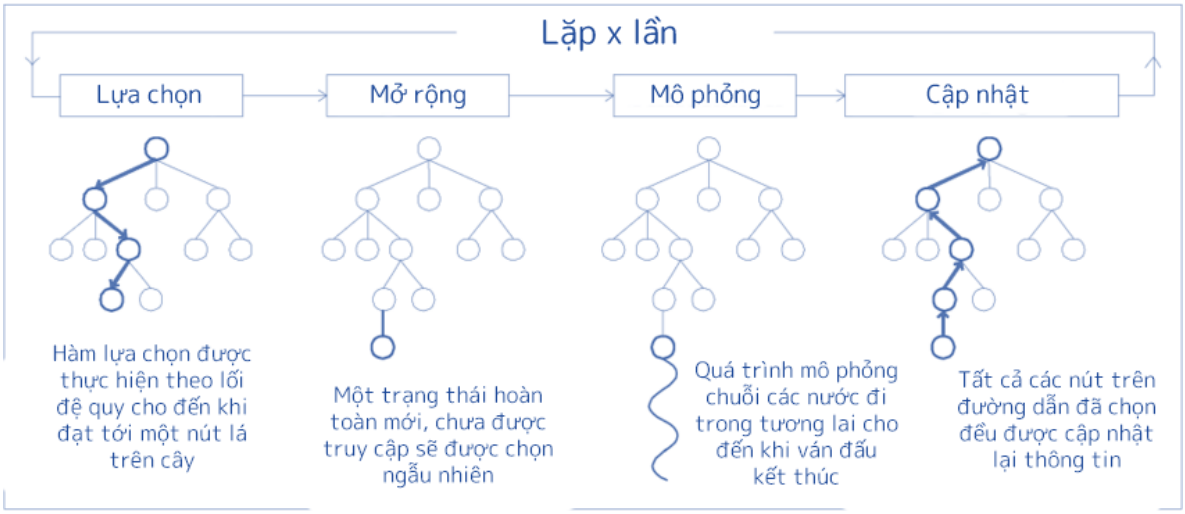
Một vấn đề với agent trí tuệ nhân tạo nói chung và AlphaGo nói riêng là việc luôn sử dụng kiến thức của con người làm dữ liệu. Điều này **quá tốn kém, không đáng tin cậy** hoặc đơn giản là không tồn tại trong một số tình huống nhất định.

Chính bởi vậy, **AlphaZero** đã được đào tạo trên một mạng nơ-rôn duy nhất với "zero" kiến thức của trò chơi. Đây cũng chính là lí do DeepMind đặt tên cho đứa con của mình là AlphaZero.

AlphaZero học mà không hề có sự giám sát từ dữ liệu loài người nào cả — nó chỉ đơn giản là **chơi với chính nó** (cơ chế self-play) và mau chóng có thể **dự đoán các nước đi của chính mình** cộng với sự ảnh hưởng của các nước cờ đến kết quả của ván đấu.

Cây tìm kiếm Monte Carlo vẫn được DeepMind tin tưởng để sử dụng làm nòng cốt của quá trình tự chơi của AlphaZero. Trên cây tìm kiếm, mỗi nút trên cây được định nghĩa là một trạng thái của bàn cờ. Cơ chế tự chơi được phối hợp với cây tìm kiếm Monte Carlo bao gồm 4 bước như sau:

1) **Lựa chọn:** Những nước đi được lựa chọn dựa trên những thông tin đã có trên cây Monte Carlo. Hàm lựa chọn được thực hiện theo lối đệ quy cho đến khi đạt tới một nút lá trên cây, đồng nghĩa hàm lựa chọn đã tìm đến trạng thái chưa được truy cập.
2) **Mở rộng:** một nút, hay một trạng thái hoàn toàn mới, chưa được truy cập sẽ được chọn ngẫu nhiên và được gắn thêm vào cây tìm kiếm.
3) **Mô phỏng:** Từ trạng thái vừa được mở rộng, một quá trình mô phỏng chuỗi các nước đi hoàn toàn ngẫu nhiên trong tương lai, cho đến khi ván đấu khép lại và quyết định người chiến thắng. Với cờ vua, giá trị kết thúc của bàn cờ có thể được quy là +1 (cờ trắng thắng), 0 (hòa cờ) và -1 (cờ đen thắng).
4) **Cập nhật:** Tất cả các nút trên đường dẫn đã chọn đều được cập nhật thông tin, tương ứng với kết quả của ván đấu thu được từ quá trình mô phỏng.



Hình 2. *4 bước trong chu trình hình thành cây tìm kiếm Monte Carlo*

Tại trạng thái bàn cờ đầu tiên, cây tìm kiếm bắt đầu từ nút gốc của cây và thực hiện quá trình mô phỏng với độ sâu tối đa của cây là 1600 nước đi trong tương lai – dựa theo thông số trong bài báo của DeepMind.

Trong quá trình mô phỏng, hành động tối ưu nhất sẽ được chọn, tùy thuộc vào chỉ số khám phá (đi thử những nước đi mới) hoặc khai thác (chọn nước đi có phần thưởng cao nhất).

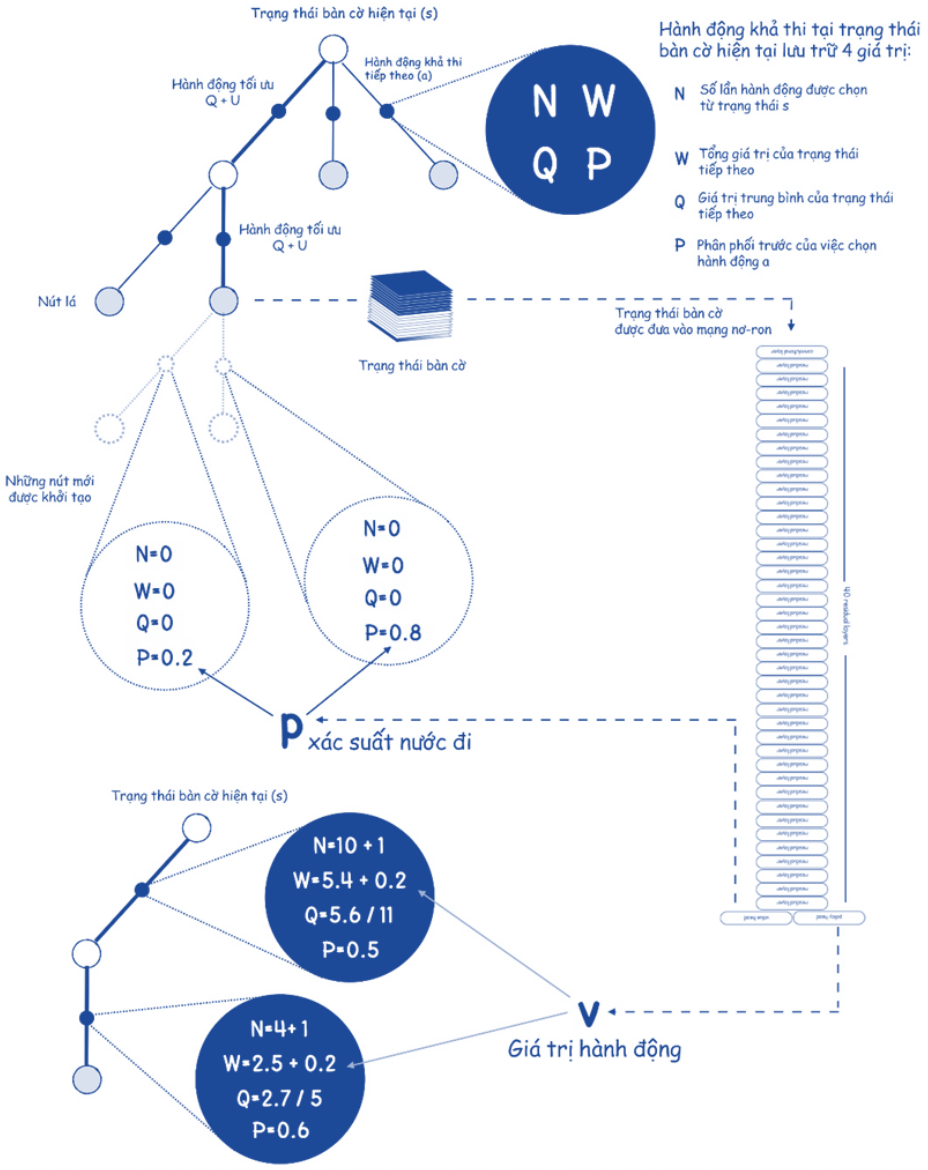
Sau đó, trạng thái bàn cờ hiện tại được đưa vào mạng nơ-ron và xuất ra dự đoán hai giá trị sau: Xác suất của các nước đi (Policy) và Giá trị của trạng thái (Value). Các nước cờ khả thi tại trạng thái bàn cờ mới sẽ được gắn thêm giá trị xác suất.

Đồng thời, tất cả các nước cờ dẫn tới trạng thái bàn cờ hiện tại sẽ được cập nhật lại ba giá trị khác nhau: Số lần thực hiện nước cờ, Tổng giá trị của nước cờ và Giá trị trung bình của nước cờ.

Cuối cùng, nước cờ tiếp theo sẽ được quyết định dựa vào yếu tố khai thác hay khám phá. Chi tiết quá trình này được mô tả trong hình dưới đây.

CÂY TÌM KIẾM MONTE CARLO

Cách AlphaZero chọn nước đi tiếp theo



Hành động khả thi tại trạng thái bàn cờ hiện tại lưu trữ 4 giá trị:

- **N** Số lần hành động được chọn từ trạng thái s
- **W** Tổng giá trị của trạng thái tiếp theo
- **Q** Giá trị trung bình của trạng thái tiếp theo
- **P** Phân phối trước của việc chọn hành động a

Đầu tiên, thực hiện quá trình mô phỏng với độ sâu là 1600

Bắt đầu từ nút gốc (trạng thái bàn cờ đầu tiên)

1. Chọn hành động tối ưu giá trị sau:

$$Q + U$$

Giá trị trung bình của trạng thái bàn cờ tiếp theo

Hàm của P và N tăng khi hành động tiếp theo chưa được khám phá nhiều so với các hành động còn lại, hoặc nếu xác suất trước của hành động cao

Trong giai đoạn đầu của mô phỏng, U sẽ chi phối (khám phá nhiều), nhưng về sau, Q sẽ quan trọng hơn (khai thác nhiều)

2. Tiếp tục cho đến nút lá (trạng thái bàn cờ mới)

Trạng thái bàn cờ tại lá nút sẽ được đưa vào mạng nơ-ron, với đầu ra dự đoán hai giá trị sau:

- **p** Xác suất của nước đi
- **v** Giá trị của trạng thái (cho người hiện tại)

Xác suất của nước đi p được gắn vào những hành động khả thi tại trạng thái bàn cờ mới

3. Cập nhật các cạnh trước đó

Mỗi cạnh trước đó dẫn tới nút lá được cập nhật lại như sau:

$$N \rightarrow N + 1$$
$$W \rightarrow W + v$$
$$Q = W / N$$

...và cuối cùng sẽ quyết định nước cờ

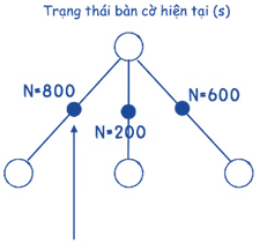
Sau quá trình mô phỏng, nước cờ tiếp theo sẽ được quyết định dựa vào:

Khai thác
Chọn hành động với giá trị N cao nhất

Khám phá
Chọn hành động dựa vào phân phối dưới đây

$$\pi \sim N^{1/\tau}$$

τ là thông số nhiệt độ, điều chỉnh chỉ số khám phá



Chọn nước đi này nếu là Khai thác
Nếu là khám phá, chọn hành động dựa vào phân phối phân loại
π với xác suất (0.5, 0.125, 0.375)

Điểm lưu ý khác

- Nhánh cây từ nước đi được chọn sẽ được giữ lại để tính toán các nước đi tiếp theo
- Phần còn lại của cây sẽ bị loại bỏ

Hình 3. Quá trình chọn nước cờ mới từ trạng thái bàn cờ hiện tại

Cơ chế Đấu trường (Arena)

Sau khi toàn bộ cây tìm kiếm Monte Carlo đã được hoàn thiện, mô hình mạng nơ-ron cũng đã được tối ưu, một phiên bản AlphaZero ra đời. Tuy nhiên, làm thế nào để đánh giá cây tìm kiếm Monte Carlo này đã có thể bao quát tất cả các nước đi, thế cờ? DeepMind đã áp dụng cơ chế Đấu trường (Arena) để cho các phiên bản AlphaZero khác nhau thi đấu với nhau. Phiên bản nào giành được số ván thắng vượt trội hơn sẽ được lựa chọn làm nhà vô địch. Các phiên bản mới khác sẽ đóng vai kẻ thách đấu với hi vọng lật đổ phiên bản vô địch trước đó. Đến cuối cùng, phiên bản vô đối nhất, đã đánh bại hàng trăm phiên bản khác sẽ được chọn là phiên bản cuối cùng.

3. Sự tương đồng giữa AlphaZero và con người

Hẳn là phần lí thuyết bên trên vẫn còn quá hàn lâm và mơ hồ phải không? Đến đây, tôi xin tóm gọn quy trình tổng thể của AlphaZero như sau:

> Đầu tiên, bằng *cơ chế tự chơi*, AlphaZero sẽ chơi những nước cờ nằm trong dự kiến. Đặc biệt, với *cơ chế cân bằng khám phá và khai thác*, nó ưu tiên những nước đi hứa hẹn, đồng thời cân nhắc xem liệu đối phương sẽ phản ứng với nước đi của mình như thế nào. Song song, nó vẫn tiếp tục khám phá và thử nghiệm những nước đi mới mẻ.
>
> Khi gặp một thế cờ lạ, AlphaZero sẽ đánh giá mức độ thuận lợi của các nước đi khả thi và xếp hạng điểm số chuỗi nước đi dẫn tới thế cờ hiện tại.
>
> Sau khi đã suy nghĩ xong về các khả năng trong tương lai, AlphaZero sẽ ra tay đi nước cờ tiếp theo. Cho đến khi ván đấu khép lại, ta sẽ quay lại và đánh giá xem mình đã đánh giá sai ở đâu, giá trị của các vị trí trong tương lai và cập nhật kiến thức của bản thân cho phù hợp.

Nghe có vẻ giống như cách loài người chúng ta học chơi cờ phải không? Khi chúng ta chơi một nước đi "tù", có thể là do chúng ta đã đánh giá sai khả năng của các nước đi trong tương lai. Hoặc là chúng ta đã đánh giá thấp đối phương, nên đã không lường trước những nước đi bá đạo của họ, khiến bản thân phải thốt lên "Nước đi hay đấy, sư huynh!".

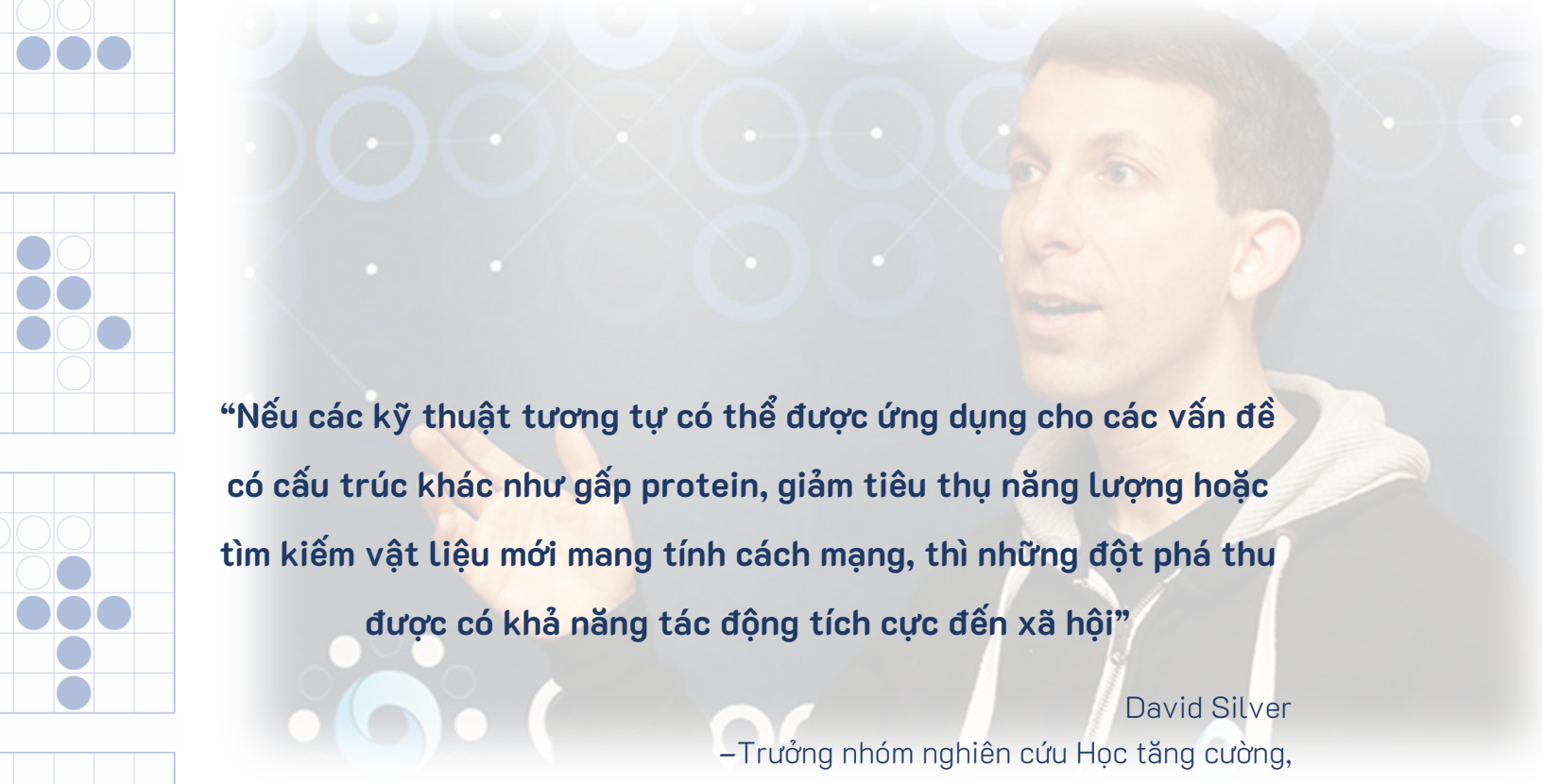
Lời kết

Khi AlphaZero ra mắt, nó đã nhanh chóng dễ dàng đánh bại các phần mềm chơi cờ tốt nhất thời bấy giờ: Stockfish (phần mềm chơi cờ vua có chỉ số ELO cao nhất), Elmo (cờ shogi), và chiến thắng chính phiên bản tiền nhiệm AlphaGo Zero.

Các đại kiện tướng cờ vua đã bày tỏ sự phấn khích tột độ về AlphaZero. Đại kiện tướng Đan Mạch Peter Heine Nielsen (người đoạt danh hiệu Grandmaster trao bởi FIDE vào năm 1994 và là huấn luyện viên của Magnus Carlsen – kì thủ cờ vua số 1 thế giới) đã ví lối chơi của AlphaZero tựa như một kì thủ ngoài hành tinh siêu hạng. Kiện tướng người Na Uy Jon Ludvig Hammer cũng đã mô tả lối chơi của AlphaZero là "tấn công điên cuồng" với trí tuệ và hiểu biết sâu sắc về từng vị trí, từng thế cờ. Nhà cựu vô địch cờ vua người Nga Garry Kasparov cũng cho hay "Đây quả là một thành tích đáng kể, quả đúng như mong đợi sau AlphaGo."

Song cũng có những phát biểu đối lập. Kiện tướng Hikaru Nakamura không tỏ ra bất ngờ lắm, anh cho rằng "AlphaZero về cơ bản đang sử dụng siêu máy tính của Google và Stockfish không được chạy trên phần cứng đó; Stockfish đang chạy ngay trên chiếc máy tính xách tay của tôi. Nên nếu để so sánh, ta phải có Stockfish chạy trên một siêu máy tính. "

Dẫu có nhiều ý kiến đa phương đa luồng như vậy, tiềm năng của AlphaZero cũng rất hứa hẹn với tương lai được áp dụng vào từng ngóc ngách của đời sống con người.



“Nếu các kỹ thuật tương tự có thể được ứng dụng cho các vấn đề có cấu trúc khác như gấp protein, giảm tiêu thụ năng lượng hoặc tìm kiếm vật liệu mới mang tính cách mạng, thì những đột phá thu được có khả năng tác động tích cực đến xã hội”

David Silver
–Trưởng nhóm nghiên cứu Học tăng cường,
Đội trưởng đội nghiên cứu AlphaGo, AlphaZero và AlphaStar tại DeepMind

Tuy nhiên, một số bạn hiểu biết về trí tuệ nhân tạo, đặc biệt là lĩnh vực Học tăng cường, vẫn sẽ cảm giác rằng, cách học hỏi của AlphaZero vẫn chưa thực sự “con người”. Bởi vì, dù DeepMind khẳng định chắc nịch rằng, kiến thức của AlphaZero xuất phát điểm từ con số 0, song việc đưa môi trường của trò chơi vào để AlphaZero có thể đánh giá trạng thái của bàn cờ cũng đồng nghĩa **AlphaZero đã hiểu biết sẵn luật chơi**. Việc này tựa như một đứa trẻ chưa tập nói nhưng đã hiểu biết tất tần tật về ngữ pháp vậy.

Chính điều này đã thúc đẩy DeepMind tiếp tục đào sâu nghiên cứu và cho ra đời thế hệ tiếp theo của hệ máy chơi cờ, với tên gọi là **MuZero**, chính thức vượt ra khỏi gia đình “Alpha”. Đến nay, cơ cấu hoạt động của MuZero vẫn còn là một bí ẩn, chưa được bóc tách rõ ràng. Rất nhiều trường Đại học, Viện nghiên cứu trên thế giới đang chạy đua để có thể tái tạo lại mô hình MuZero, với hi vọng tiệm cận với trình độ của MuZero.

Hồi cuối của loạt bài viết, “*AlphaGo đến AlphaZero, kẻ hạ bệ trí tuệ loài người?*” sẽ kết thúc với **Phần 3: MuZero – Kẻ lật đổ vương triều Alpha**.

Hi vọng sẽ được các bạn đón đọc!

Tham khảo

[1] Van Gelder, Tim; (1998), “In to the Deep Blue yonder”, Quadrant, 41 (2-1), pp. 33-39.

[2] Ciancarini, Paolo; (2005), “Il computer gioca a scacchi”, Mondo Digitale, V. 3, p. 9.

[3] Poe, Edgar A.; (1836), The Maelzel’s chess player, Createspace Independent Pub (2014).

[4] Rudolf, Anna; (2018), “AlphaZero’s Attacking Chess”, https://www.youtube.com/watch?v=nPexHaFL1uo&t=1312s Accessed 27.05.2019, 12.45.

[5] Kasparov, Garry; (2018), “Class of 2006 War Studies Conference”, West Point – Military Accademy, https://www.youtube.com/watch?v=QSyKlzh9Zl8&t=1734s Accessed 27.05.2019, 12.49.

[6] For a very detailed analysis: Hassabis, Silver, et. All (2018) “A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play”, Science 362 (6419), 1140-1144. DOI: 10.1126/science.aar6404

[7] I strongly suggest to watch entirely the insightful video: Nando de Freitas, (2017), “DeepMind’s Nando de Freitas – Learning to Learn”, https://www.youtube.com/watch?v=5yNirTp92Uk, Accessed 27.05.2019, 12.54.

[8] Kasparov, Garry, (2013), “Garry Kasparov on “Achieving Your Potential””, https://www.youtube.com/watch?v=NPTOvg_Jl8Q, Accessed 27.05.2019, 13.00.

[9] Hassabis, Silver, et. All (2018) “A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play”, Science 362 (6419), 1140-1144. DOI: 10.1126/science.aar6404

[10] Rudolf, Anna; (2018), “AlphaZero’s Attacking Chess”, https://www.youtube.com/watch?v=nPexHaFL1uo&t=1312s Accessed 27.05.2019, 12.45.

[11] Pili, Giangiuseppe; (2012), Un mistero in bianco e nero – La filosofia degli scacchi, Le Due Torri, Bologna, Chap. 10.