TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT VĨNH LONG

KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO

MÔN: TRÍ TUỆ NHÂN TAO

ĐỀ TÀI: A.I CHƠI GAME BẰNG TRÍ TUỆ VÀ TRẢI NGHIỆM

Sinh viên thực hiện: Lê Nguyễn Quang Bình 21022010

Lớp: ĐH. KHMT 2021

Khóa: 2021 - 2025

Giáo viên hướng dẫn: Lê Hoàng An

Nguyễn Khắc Tường

Vĩnh Long - Năm 2023

Mục lục

Danh mục hình ảnh, đồ thị

Danh mục bảng

NHẬN XÉT & ĐÁNH GIÁ ĐIỂM CỦA NGƯỜI HƯỚNG DẪN

Ý thức thực hiện:.....

.....

.....

.....

Nội dung thực hiên:.....

.....

.....

.....

Hình thức trình bày: .....

.....

.....

.....

Tổng hợp kết quả: .....

.....

.....

.....

☐ Tổ chức báo cáo trước hội đồng

☐ Tổ chức chấm thuyết minh

Vĩnh Long, ... ngày .... tháng ... năm ....

Người hướng dẫn

(Ký và ghi rõ họ tên)

LỜI CAM ĐOAN

Em, nhóm nghiên cứu, xin cam kết rằng nội dung trong bài báo cáo này là kết quả của quá trình nghiên cứu, tổng hợp và đánh giá cẩn thận và tỉ mỉ của em. Em đã tham khảo từ các nguồn tài liệu có nguồn gốc rõ ràng và đáng tin cậy, và em đã chỉ sử dụng thông tin từ những nguồn đã được cho phép. Mọi nguồn được trích dẫn đều đã được ghi rõ ràng và đầy đủ trong bài báo cáo này.

Em hiểu rằng việc cung cấp thông tin sai lệch có thể ảnh hưởng tới mục tiêu và tính xác thực của bài báo cáo này. Do đó, em xin chịu trách nhiệm hoàn toàn nếu có bất kỳ thông tin sai lệch nào được phát hiện trong báo cáo của mình. Em xin trân trọng cảm ơn và hy vọng bài báo cáo này sẽ đáp ứng được mọi yêu cầu và kỳ vọng.

Vĩnh Long, ngày 29 tháng 11 năm 2023

LỜI CẢM ƠN

Đầu tiên và quan trọng nhất, em xin gửi lời cảm ơn chân thành đến Trường Đại học Sư Phạm Kỹ Thuật Vĩnh Long vì đã tích cực đưa môn Trí Tuệ Nhân Tao vào chương trình giảng dạy. Trên hết, em muốn bày tỏ lòng biết ơn sâu sắc đến Giảng viên Lê Hoàng An và giảng viên Nguyễn Khắc Tường, người đã tận tâm hướng dẫn em trong suốt quá trình học tập và hoàn thành báo cáo học phần Trí Tuệ Nhân Tao.

Khi tham gia lớp Trí Tuệ Nhân Tao do thầy/cô phụ trách, em đã học hỏi được nhiều kiến thức bổ ích, thúc đẩy tinh thần học tập hiệu quả và nghiêm túc. Những kiến thức này không những quý báu mà còn là hành trang vững chắc cho em tiến bước trong tương lai. Trong quá trình thực hiện báo cáo, các thầy đã rất nhiệt tình góp ý, giúp em nâng cao chất lượng bài viết, đồng thời truyền đạt thêm nhiều kiến thức có ích trong suốt quá trình học và hoàn thành học phần Trí Tuệ Nhân Tao.

Bộ môn Trí Tuệ Nhân Tao là một môn học cực kỳ thú vị, bổ ích và thực tế. Nó không chỉ đảm bảo cung cấp đầy đủ kiến thức mà còn liên kết chặt chẽ với nhu cầu thực tế của sinh viên. Tuy nhiên, do hạn chế về kiến thức và khả năng tiếp thu thực tế, em còn gặp nhiều khó khăn.

Dù em đã cố gắng hết sức, nhưng không thể tránh khỏi những thiếu sót và những phần chưa chính xác trong báo cáo. Vì vậy, em rất mong nhận được sự đóng góp ý kiến nhiệt tình từ thầy/cô và các bạn, giúp em hoàn thiện hơn và rút kinh nghiệm cho những học phần sau.

Em xin chân thành cảm ơn thầy/cô!

I. Giới Thiệu

1 Đặt vấn đề

Với sự phát triển mạnh mẽ trong lĩnh vực Trí Tuệ Nhân Tạo (A.I), học máy ngày càng chứng tỏ sức mạnh và khả năng áp dụng rộng rãi. Trong lĩnh vực học tăng cường(Reinforcement Learning - RL), chúng ta chứng kiến sự xuất hiện của những mô hình máy học có khả năng thích ứng với môi trường khác nhau và từ đó, tự động tối ưu hoá hiệu suất theo thời gian, để dễ dàng nhận thấy tôi đã cho mô phỏng trong môi trường game. Lĩnh vực này mở ra những triển vọng mới cho việc phát triển hệ thống máy tính có khả năng đưa ra quyết định tự động và linh hoạt.

Ứng dụng của học tăng cường không chỉ giới hạn trong thế giới ảo, mà còn mở ra những triển vọng mới cho thực tế. Ví dụ, trong lĩnh vực y tế, các mô hình RL có thể được sử dụng để tối ưu hóa lịch trình điều trị cho bệnh nhân dựa trên phản hồi và thay đổi trong tình trạng sức khỏe. Ngoài ra, trong tự động hóa và quản lý tài nguyên, học tăng cường đóng vai trò quan trọng trong việc đưa ra quyết định linh hoạt và tự động.

2 Mục tiêu và ý nghĩa của nghiên cứu

Mục tiêu chủ yếu của đề tài này là nghiên cứu chuyên sâu về cơ chế hoạt động của học tăng cường trong lĩnh vực trí tuệ nhân tạo, đặc biệt là khi ứng dụng vào môi trường game. Tôi đặt ra mục tiêu nghiên cứu để làm sáng tỏ cách thức mà các mô hình học tăng cường có thể thích ứng với môi trường game, học thông qua tương tác và linh hoạt ứng phó với những thử thách đặc biệt của môi trường game.

Ý nghĩa của nghiên cứu này không chỉ là ở mức lý thuyết mà còn ở mức thực tiễn, cụ thể là đối với việc thiết kế hệ thống thông minh có khả năng tương tác tự động và linh hoạt trong môi trường game. Điều này có thể mở ra những triển vọng mới về việc xây dựng trải nghiệm người dùng linh hoạt và tối ưu hoá.

3 Phạm vi và hạn chế của đề tài

Đề tài tập trung vào nghiên cứu các mô hình học tăng cường và áp dụng chúng trên môi trường game Pong. Phạm vi của đề tài bao gồm việc ứng dụng lý thuyết học tăng cường vào môi trường game nhằm đạt được mức độ tự động hoá và hiệu quả cao. Tuy

nhiên, hạn chế về quy mô mô hình và mức độ phức tạp của môi trường game đã làm giảm tính khả thi của đề tài nghiên cứu. Hơn nữa, sự khác biệt về tài nguyên có thể ảnh hưởng đến khả năng thực hiện và đánh giá các mô hình trong môi trường khác trên thực tế.

II. Lý Thuyết Cơ Bản

1 Trí tuệ nhân tạo và máy học

Trí tuệ nhân tạo AI(artificial intelligence) là khả năng máy tính hoặc robot điều khiển bằng máy tính thực hiện các nhiệm vụ thường liên quan đến quá trình trí tuệ đặc trưng của con người, chẳng hạn như khả năng suy luận. Thông thường, thuật ngữ "trí tuệ nhân tạo" thường được sử dụng để mô tả các máy chủ móc (hoặc máy tính) có khả năng bắt chước các chức năng "nhận thức" mà con người thường phải liên kết với tâm trí, như "học tập" và "giải quyết vấn đề".!

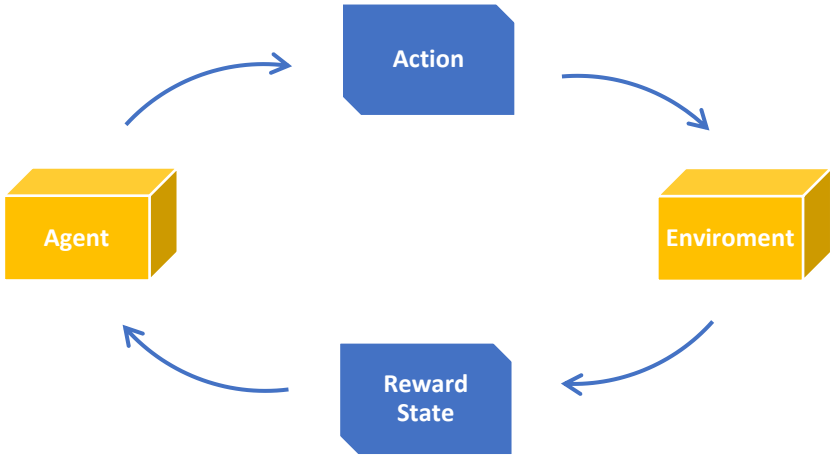
Máy học(machine learning) là một lĩnh vực của trí tuệ nhân tạo liên quan đến việc nghiên cứu và xây dựng các kĩ thuật cho phép các hệ thống "học" tự động từ dữ liệu để giải quyết những vấn đề cụ thể. Các thuật toán máy học xây dựng một mô hình dựa trên dữ liệu mẫu, được gọi là dữ liệu huấn luyện, để đưa ra dự đoán hoặc quyết định mà không cần được lập trình chi tiết về việc đưa ra dự đoán hoặc quyết định này. Ví dụ như các máy có thể "học" cách phân loại thư điện tử xem có phải thư rác (spam) hay không và tự động xếp thư vào thư mục tương ứng. Máy học rất gần với suy diễn thống kê (statistical inference) tuy có khác nhau về thuật ngữ.!

2 Học tăng cường

Các thuật toán học máy thường được phân thành 3 loại lớn: học có giám sát, học không giám sát và học tăng cường. Nếu như học giám sát là học tập từ một tệp các dữ liệu được gắn nhãn để suy luận ra quan hệ giữa đầu vào và đầu ra, thì học không giám sát không được cung cấp các dữ liệu được gắn nhãn ấy, thay vào đó chỉ được cung cấp dữ liệu mà thuật toán tìm cách mô tả dữ liệu và cấu trúc của chúng. Học tăng cường là phương pháp tập trung vào việc làm thế nào để cho một tác tử trong môi trường có thể hành động sao cho lấy được phần thưởng nhiều nhất có thể. Khác với học có giám sát,

học tăng cường không có cặp dữ liệu gán nhãn trước làm đầu vào và cũng không có đánh giá các hành động là đúng hay sai.

Trong lĩnh vực trí tuệ nhân tạo nói chung và lĩnh vực học máy nói riêng, học tăng cường (RL - Reinforcement Learning) là một cách tiếp cận tập trung vào việc học để hoàn thành được mục tiêu bằng việc tương tác trực tiếp với môi trường. Các thuật toán học tăng cường sẽ cố gắng tìm một chiến lược ánh xạ các trạng thái của thế giới tới các



Hình 2.1 Cách hoạt động của thuật toán học tăng cường

hành động mà agent nên chọn trong các trạng thái đó.

- Agent: được định nghĩa là "*anything that can be viewed as perceiving its environment through sensors and acting upon that environment through actuators"(Là máy quan sát được môi trường và trả về một hành động tương ứng).*
- Enviroment: môi trường là không gian xung quanh của Agent, nơi Agent tồn tại và có thể tương tác.
- State: là trạng thái của môi trường mà Agent nhận được.
- Action: hành động của Agent dựa vào State cho phép cho phép Agent tương tác và thay đổi môi trường.

- Policy: Chính sách là yếu tố xác định cách thức hoạt động của agent tại một thời điểm nhất định. Nói cách khác, chính sách là một ánh xạ từ các trạng thái (state) của môi trường đến các hành động sẽ được thực hiện khi ở trong các trạng thái đó
- Reward: là phần thưởng được môi trường gửi đến Agent sau mỗi hành động, mục tiêu của Agent là tối đa lượng điểm nhận được trông khoản thời gian, reward sẽ xác định cho Agent sự kiện nào tốt và sự kiện nào sấu để Agent làm cơ sở thay đổi chính sách.

3 Advantage Actor-Critic (A2C)

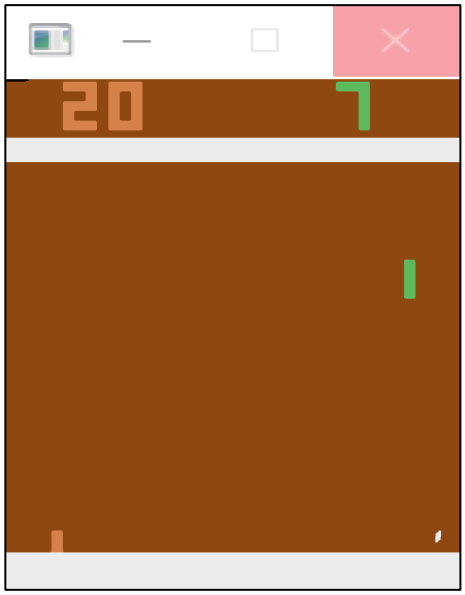
A2C là một thuật toán học máy tăng cường được thiết kế để đào tạo mô hình chính sách (Actor) và giá trị (Critic) đồng thời trong môi trường học máy tăng cường, là một phiên bản synchronous của A3C(Asynchronous Advantage Actor Critic) các actor của A3C sẽ trải nghiệm độc lập và không đồng bộ trong khi A2C sẽ đợi cho tất cả các actor hoàn thành trải nghiệm của mình rồi mới đưa ra các quyết định kế tiếp nhằm cải thiện sự ổn định của quá trình đào tạo. Thuật ngữ "Advantage" trong tên gọi đề cập đến việc sử dụng lợi nhuận (advantage) để đánh giá lợi ích của hành động so với giá trị dự đoán của mô hình.

4 Proximal Policy Optimization (PPO)

PPO, hay Proximal Policy Optimization, là một thuật toán học máy thuộc loại Policy Optimization, được thiết kế để đào tạo mô hình chính sách (policy) trong môi trường học máy tăng cường lấy ý tưởng từ thuật toán A2C. Mục tiêu của PPO là cải thiện chính sách hiện tại một cách an toàn và ổn định. Nó ra đời để giải quyết một số vấn đề mà các thuật toán trước đây, như TRPO (Trust Region Policy Optimization), gặp phải, đặc biệt là vấn đề về sự không ổn định trong quá trình đào tạo.

III. Phương Pháp Nghiên Cứu

1 Mô tả môi trường và bài toán



Hình 3.1 Giao diện môi trường trò chơi Pong

1.1 Môi trường của trò chơi Pong trong Gym:

- Trạng thái (State): Trạng thái của môi trường được biểu diễn dưới dạng hình ảnh, thường là một ma trận pixel mô tả cảnh trong trò chơi có dạng:

  Box(0, 255, (210, 160, 3), uint8)

- Hành động (Action): Người chơi có thể thực hiện một trong sáu hành động sau:

Giá trị	Hành động	Giá trị	Hành động	Giá trị	Hành động
0	Đứng yên	1	Phát thẳng	2	Di xuống
3	Di lên	4	Phát dưới	5	Phát trên

Bảng 3.1 Mô tả hành động trong môi trường Pong

- Phần thưởng (Reward): Người chơi nhận được phần thưởng dựa trên hiệu suất của mình trong trò chơi. Thưởng có thể được cung cấp khi ghi được điểm hoặc phạm lỗi trong phạm vi (-1 : 1).

- Kết thúc trò chơi (Done): Trò chơi kết thúc khi một trong hai bên đạt được số điểm cố định hoặc khi trò chơi đạt đến một số bước tối đa.
- Thông tin bổ sung (Info): Cung cấp thông tin bổ sung về trò chơi, như thông tin về trạng thái của trò chơi, số khung hình.

1.2 Bài toán của trò chơi Pong

Bài toán chính của trò chơi Pong là thiết lập một chính sách (policy) cho thanh điều khiển sao cho tỷ lệ ghi điểm là cao nhất. Một chính sách là một chiến lược quyết định hành động dựa trên trạng thái hiện tại của môi trường.

2 Kiến trúc A2C(Advantage Actor-Critic)

$$\boxed{J(\theta) = \mathbb{E}_\tau \left[\sum_{t=0}^{T} A_t \nabla_\theta \log \pi_\theta(a_t | s_t) \right]}$$

E_t: giá trị kỳ vọng(là một biến ngẫu nhiên là trung bình có trọng số là xác suất của tất cả các giá trị cụ thể của biến đó).

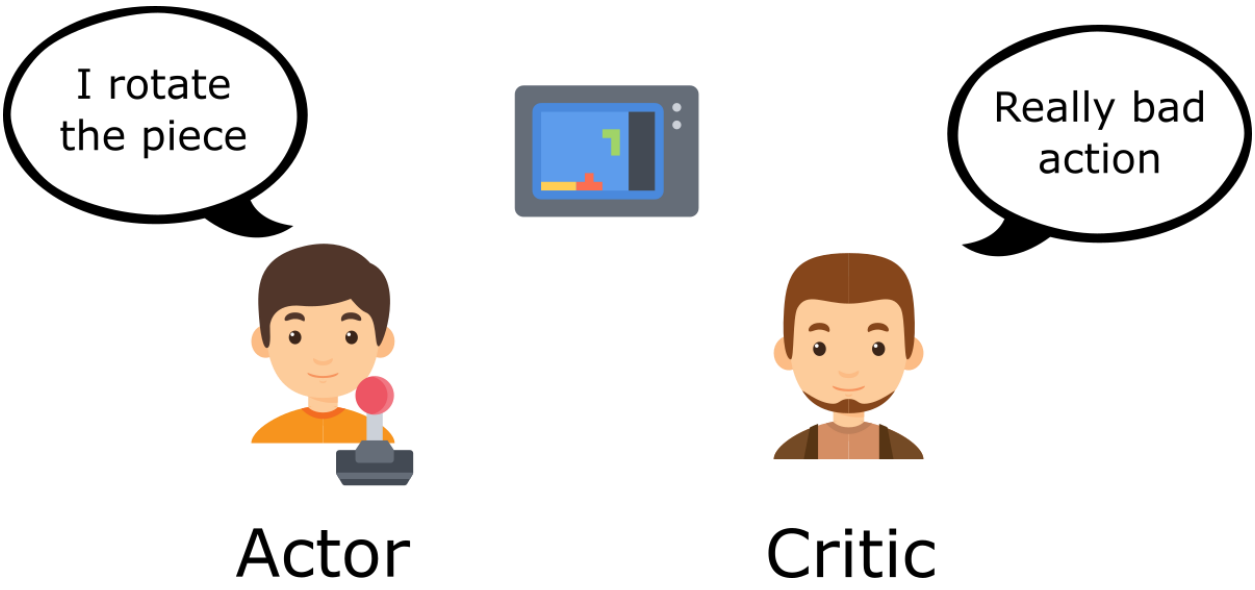
$\pi_\theta(a_t|s_t)$: là xác suất của hành động a_t tại trạng thái s_t dưới chính sách được tham số hóa bởi θ.

a_t: hành động agent tại thời điểm t.

s_t: là trạng thai tại thời điểm t.

Kiến trúc của A2C gồm 2 phần:

- **Phương pháp Actor-Critic:** khi agent thử một số hành động ngẫu nhiên, Critic sẽ đưa ra phản hồi dựa trên hành động đó, sau đó agent sẽ rút kinh nghiệm từ phản hồi và cập nhật chính sách của mình, Critic cũng sẽ cập nhật để có thể đưa ra phản hồi tốt hơn.



Hình 3.2 Critic đưa ra phản hồi dựa trên hành động của Actor

- **Hàm ưu tiên Advantage**: Để ổn định việc học sẽ sử dụng hàm ưu tiên thay gì Critic. Dùng để tính phần thưởng chênh lệch với phần thưởng trung bình tại trạng thái đó.

$$A_t=Q_t-V(s)$$

A_t: là ưu tiên tại thời điểm *t*.

Q_t: là giá trị Q của hành động tại trạng thái *t*.

V(s): giá trị điểm trung bình tại trạng thái *t*.

Nếu A_t lớn hơn 0 thì định hướng chính sách theo hướng đó.
Nếu A_t bén hơn 0(phần thưởng từ hành động mang lại bé hơn phần thưởng trung bình) thì định hướng khác hướng đó.

3 Kiến trúc PPO(Proximal Policy Optimization)

$$J(\theta) = \mathbb{E}_t \left[\min \left(\rho_t(\theta) A_t, \text{clip}(\rho_t(\theta), 1-\epsilon, 1+\epsilon) A_t \right) \right]$$

E_t: giá trị kỳ vọng(là một biến ngẫu nhiên là trung bình có trọng số là xác suất của tất cả các giá trị cụ thể của biến đó).

$\pi_\theta(a_t|s_t)$: là xác suất của hành động a_t tại trạng thái s_t dưới chính sách được tham số hóa bởi θ.

$p_t(\theta)$: là tỷ lệ xác suất mới so với xác suất cũ của hành động được chọn.

$$p_t(\theta)=\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta old}(a_t|s_t)}$$

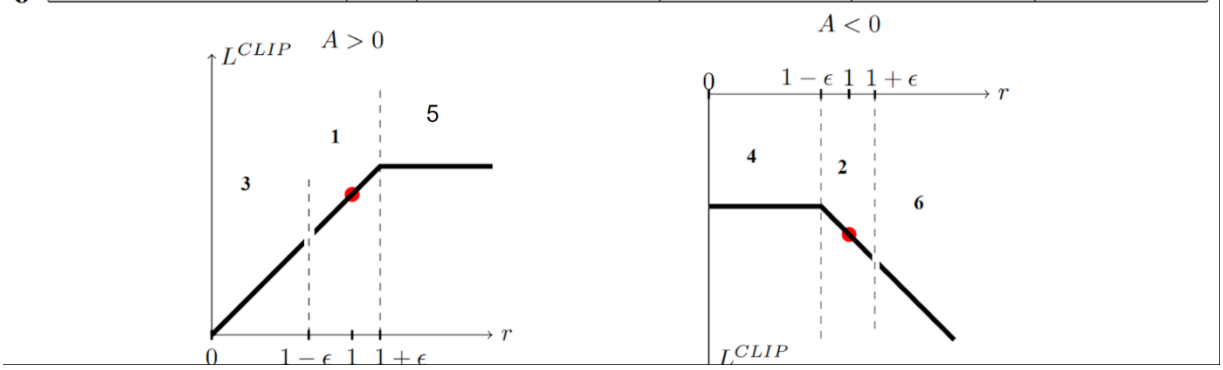
A_t: là ưu tiên tại thời điểm t.

ε: là siêu tham số thường là 0,1 hoặc 0,2.

PPO kế thừa từ A2C và bổ sung thêm 1 hàm

- **Hàm clip**: Để hạn chế việc chính sách thay đổi quá nhiều dẫn đến bất ổn định thì hàm clip sẽ giới hạn phạm vi cập nhật.

	$p_t(\theta) > 0$	A_t	**Return Value of** min	**Objective is Clipped**	**Sign of Objective**	**Gradient**
1	$p_t(\theta) \in [1-\epsilon, 1+\epsilon]$	$+$	$p_t(\theta)A_t$	no	$+$	✓
2	$p_t(\theta) \in [1-\epsilon, 1+\epsilon]$	$-$	$p_t(\theta)A_t$	no	$-$	✓
3	$p_t(\theta) < 1-\epsilon$	$+$	$p_t(\theta)A_t$	no	$+$	✓
4	$p_t(\theta) < 1-\epsilon$	$-$	$(1-\epsilon)A_t$	yes	$-$	**0**
5	$p_t(\theta) > 1+\epsilon$	$+$	$(1+\epsilon)A_t$	yes	$+$	**0**
6	$p_t(\theta) > 1+\epsilon$	$-$	$p_t(\theta)A_t$	no	$-$	✓



Hinh 3. 3 Hình biểu thị cắt bớt của hàm clip

Trường hợp 1 và 2: tỷ lệ nằm trong khoảng

Trong tình huống 1 và 2, việc cắt bớt không áp dụng vì tỷ lệ nằm trong$[1-\epsilon ,1+\epsilon]$

Tình huống 1, chúng ta có lợi thế tích cực: hành động tốt hơn mức trung bình của tất cả các hành động ở trạng thái đó. Vì vậy, chúng ta nên khuyến khích chính sách hiện tại của mình tăng khả năng thực hiện hành động đó trong trạng thái đó.

Tình huống 2, chúng ta có lợi thế tiêu cực: hành động tệ hơn mức trung bình của tất cả các hành động ở trạng thái đó. Do đó, chúng ta nên ngăn cản chính sách hiện tại của mình thực hiện hành động đó trong trạng thái đó.

Trường hợp 3 và 4: tỷ lệ nằm dưới khoảng

Tình huống 3, ước tính lợi thế là dương (A>0), thì bạn muốn tăng xác suất thực hiện hành động đó ở trạng thái đó.

Tình huống 4, ước tính lợi thế là âm thì chúng ta không muốn giảm thêm xác suất thực hiện hành động đó ở trạng thái đó. Nên chúng ta không cập nhật trọng số của mình.

Trường hợp 5 và 6: tỷ lệ nằm trên phạm vi

Tình huống 5, lợi thế là tích cực thì chúng ta không muốn quá tham lam . Chúng tôi đã có xác suất thực hiện hành động đó ở trạng thái đó cao hơn chính sách trước đây. Nên chúng ta không cập nhật trọng số của mình.

Tình huống 6, lợi thế là âm, chúng ta muốn giảm xác suất thực hiện hành động đó ở trạng thái đó.

IV. Thực Nghiệm và Kết Quả

1 Môi trường

1.1 Môi trường game

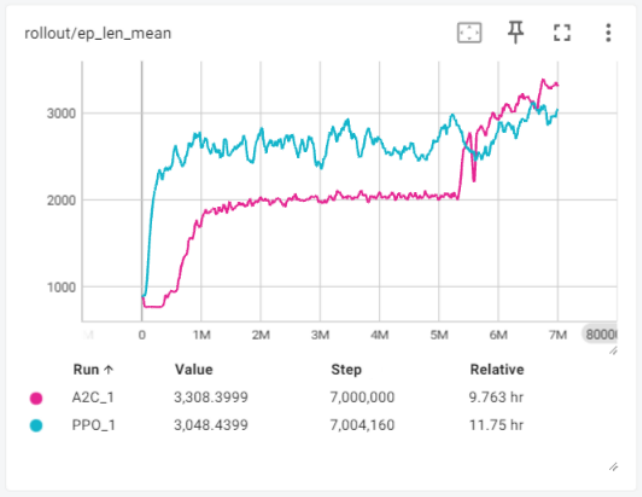
Môi trường game Pong phiên bản ALE/Pong-v5 với sự khác biệt về trạng thái, số khung hình bỏ qua hoặc tỉ lệ lập lại hành động. ALE/Pong-v5 loại trạng thái quan sát là "rbg", số khung hình bỏ qua là 4 và tỉ lệ lập lại hành động của máy là 25%.

1.2 Môi trường huấn luyện

Kaggle là một nền tảng dành cho việc học và thực hành khoa học dữ liệu và học máy. Nó được thành lập bởi Anthony Goldbloom và Ben Hamner vào năm 2010, và sau đó được Google mua lại vào năm 2017. Kaggle cung cấp một cộng đồng nơi các nhà khoa học dữ liệu và máy học có thể tương tác, chia sẻ ý tưởng, thực hiện các dự án, và tham gia vào các cuộc thi khoa học dữ liệu.

2 So sánh kết quả thực hiện huấn luyện

2.1 Biểu đồ episode length mean(độ dài trung bình của 1 màng)



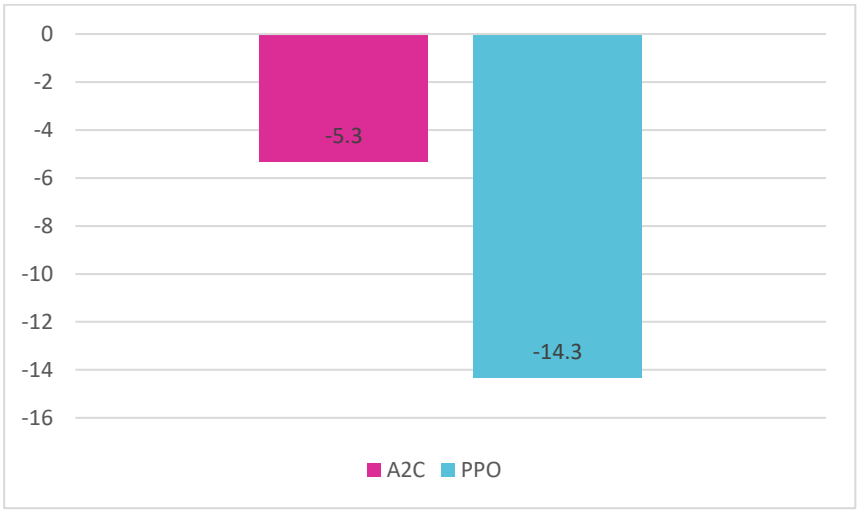
Hình 4.1 Biểu đồ episode length mean

2.2 Biểu đồ episode reward mean(phần thưởng trung bình 1 màng)

Hình 4.2 Biểu đồ episode reward mean

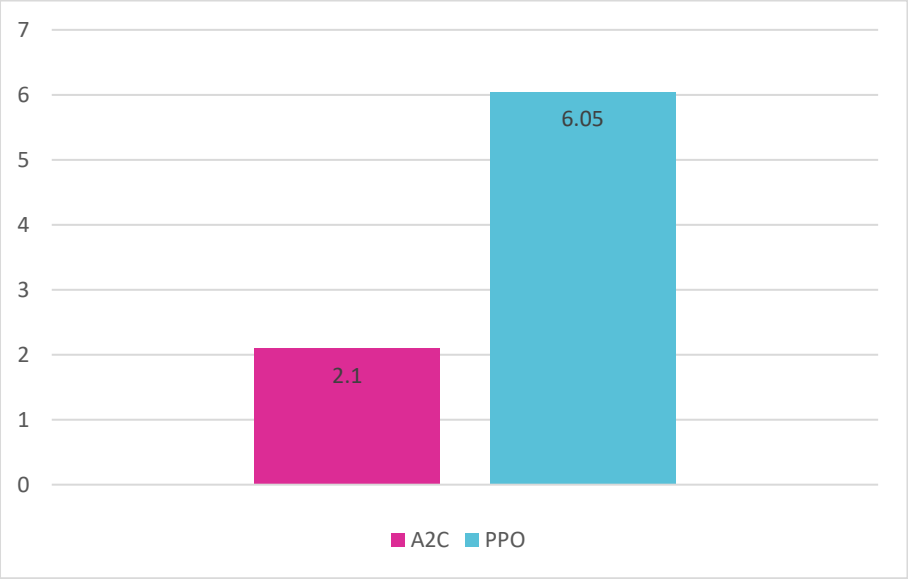
3 Kết quả thực nghiệm

3.1 Điểm tổng trung bình



Hình 4.3 Biểu đồ so sánh điểm tổng trung bình

3.2 Điểm trung bình



Hình 4.4 Biểu đồ so sánh điểm trung bình

4 Đánh giá

	A2C	PPO
Ưu điểm	▪ Tốc độ huấn luyện nhanh chống. ▪ Dễ dàng triển khai.	▪ Tính ổn định cao hơn. ▪ Dễ dàng hơn trong việc tinh chỉnh tham số phù hợp.
Nhược điểm	▪ Dễ dàng xuất hiện việc bất ổn định. ▪ Khó tinh chỉnh tham số phù hợp.	▪ Cần nhiều thời gian hơn để huấn luyện. ▪ Khó triển khai hơn.

Bảng 4.1 so sánh ưu nhượt điểm A2C và PPO

TÀI LIỆU THAM KHẢO

[1] Schulman, John, et al. "Proximal policy optimization algorithms." arXiv preprint arXiv:1707.06347 (2017).

[2] Mnih, Volodymyr, et al. "Asynchronous methods for deep reinforcement learning." *International conference on machine learning*. PMLR, 2016.

[3] Huang, Shengyi, et al. "A2C is a special case of PPO." *arXiv preprint arXiv:2205.09123* (2022).

[4] Holubar, Mario S., and Marco A. Wiering. "Continuous-action reinforcement learning for playing racing games: Comparing SPG to PPO." *arXiv preprint arXiv:2001.05270* (2020).

[5] Bick, Daniel, and M. A. Wiering. Towards Delivering a Coherent Self-Contained Explanation of Proximal Policy Optimization. Diss. Master's thesis, 2021.[Online]. Available: https://fse. studenttheses. ub. rug. nl/25709, 2021.