ĐÀO SÂU VÀO CÁC CUỘC TẤN CÔNG ADVERSARIAL TRÊN CÁC CHÍNH SÁCH SÂU

Jernej Kos

Đại học Quốc gia Singapore

Dawn Song

Đại học California, Berkeley

TÓM TẮT

Các ví dụ Adversarial đã được chứng minh tồn tại đối với nhiều kiến trúc học sâu khác nhau. Học tăng cường sâu đã cho thấy kết quả đáng kỳ vọng trong việc huấn luyện chính sách tác nhân trực tiếp trên đầu vào thô như điểm ảnh hình ảnh. Trong bài báo này, chúng tôi trình bày một nghiên cứu mới vào các cuộc tấn công Adversarial trên các chính sách học tăng cường sâu. Chúng tôi so sánh tính hiệu quả của các cuộc tấn công sử dụng các ví dụ Adversarial so với nhiễu ngẫu nhiên. Chúng tôi trình bày một phương pháp mới để giảm số lần cần tiêm các ví dụ Adversarial để tấn công thành công, dựa trên hàm giá trị. Chúng tôi tiếp tục khám phá cách làm mới trên nhiễu ngẫu nhiên và sự biến động FGSM ảnh hưởng đến khả năng chống lại các ví dụ Adversarial.

1 GIỚI THIỆU

Các ví dụ Adversarial đã được chứng minh tồn tại đối với nhiều kiến trúc học sâu khác nhau. Chúng là các biến động nhỏ trên các đầu vào ban đầu, thường khó nhận thấy với mắt thường, nhưng được chế tạo cẩn thận để đánh lừa mạng nơ-ron để tạo ra đầu ra không chính xác. Công trình ban đầu của Szegedy et al. (2013) và Goodfellow et al. (2014), cũng như nhiều công trình gần đây khác, đã chứng minh rằng các ví dụ Adversarial rất phổ biến và dễ tìm thấy. Mạng nơ-ron sâu đã được sử dụng trong học tăng cường sâu (DRL) với kết quả đáng kỳ vọng trong việc huấn luyện các chính sách trực tiếp trên đầu vào nguyên thủy như điểm ảnh hình ảnh. Một trong những thuật toán thành công nhất để huấn luyện chính sách sâu là A3C (Mnih et al., 2016), cho phép cập nhật không đồng bộ các trọng số chính sách, dẫn đến một thực thi song song hiệu quả. Vì các chính sách có thể điều khiển các tác nhân tự động khác nhau như ô tô tự lái trong thế giới thực, các cuộc tấn công Adversarial có thể còn quan trọng hơn nữa.

Bài báo của chúng tôi là một trong những bài viết đầu tiên điều tra các ví dụ Adversarial trên các chính sách DRL, chỉ ra rằng những chính sách sâu này dễ bị đánh lừa bởi các cuộc tấn công Adversarial với biến động Adversarial rất nhỏ. Ngoài ra, trong bài báo này, chúng tôi khám phá ba chiều mới về các cuộc tấn công Adversarial trên các chính sách DRL mà bất kỳ công trình nào cũng chưa giải quyết trước đó. Đầu tiên, chúng tôi so sánh các ví dụ Adversarial với nhiễu ngẫu nhiên và chỉ ra rằng những ví dụ Adversarial là hiệu quả tấn công chính sách DRL nhiều lần.

Một chiều quan trọng khác trong các hệ thống DRL là thời gian. Nếu kẻ tấn công cần phải tiêm các biến động Adversarial ít thường xuyên hơn, thì cuộc tấn công sẽ dễ thực hiện hơn. Để đạt được mục đích này, chúng tôi khám phá việc sử dụng hàm giá trị của chính sách như một hướng dẫn để tiêm các biến động. Các thí nghiệm của chúng tôi cho thấy rằng, với sự hướng dẫn này, kẻ tấn công có thể tiêm các biến động chỉ trong một phần nhỏ các khung hình, và hiệu quả hơn khi so sánh với việc tiêm các biến động cùng tần suất nhưng không có hướng dẫn. Kết quả của chúng tôi cho thấy rằng các cuộc tấn công

Adversarial có thể phức tạp hơn rất nhiều trong môi trường học tăng cường so với các môi trường khác đã được nghiên cứu trước đó như phân loại hình ảnh.

Chiều thứ ba là sự bền vững của chính sách thông qua huấn luyện lại. Chúng tôi trình bày các kết quả sơ bộ cho thấy các tác nhân có thể trở nên bền vững hơn đối với cuộc tấn công FGSM (phương pháp dấu hiệu gradient nhanh) thông qua huấn luyện lại với cả nhiễu ngẫu nhiên và các biến động FGSM, trong khi huấn luyện lại với các biến động FGSM có thể hiệu quả hơn so với huấn luyện lại với nhiễu ngẫu nhiên. Tuy nhiên, các tác nhân đã được huấn luyện lại có thể vẫn dễ bị tấn công bởi các phương pháp tấn công khác như tấn công dựa trên tối ưu, tuy nhiên, các phương pháp tấn công khác này thường rất chậm để thực hiện, thường khiến cho các cuộc tấn công rất chậm đặc biệt là trong trường hợp tác nhân.

Cùng một lúc và độc lập với công việc của chúng tôi (được đệ trình cho cùng hội thảo ICLR), Huang et al. (2017) cũng trình bày một nghiên cứu về các cuộc tấn công Adversarial trên các chính sách DRL, chỉ ra rằng khi một kẻ tấn công tiêm các biến động Adversarial nhỏ vào mỗi khung hình, tác nhân đã học sẽ thất bại.

Do giới hạn không gian, chúng tôi tập trung vào các tác nhân được huấn luyện trên nhiệm vụ Atari Pong bằng thuật toán A3C và các biến động Adversarial FGSM. Công việc của chúng tôi là một bước đầu tiên để hiểu rõ hơn về những thách thức và giới hạn của DRL dưới đầu vào Adversarial.

2 MỤC TIÊU NGHIÊN CỨU

Độ hiệu quả của Cuộc tấn công Adversarial so với Nhiễu Ngẫu Nhiên Chúng tôi nghiên cứu cách tiêm nhiễu ngẫu nhiên vào môi trường so với việc tiêm biến động Adversarial FGSM.

Sử dụng Hàm Giá Trị để Hướng Dẫn Tiêm Biến Động Adversarial Chúng tôi muốn xem liệu giảm tần suất tiêm biến động Adversarial có thể tạo ra một cuộc tấn công hiệu quả hay không. Chúng tôi nghiên cứu ba phương pháp khác nhau: a) chúng tôi chỉ tiêm biến động Adversarial mỗi N khung hình và các khung hình trung gian không có bất kỳ biến động nào, b) chúng tôi chỉ tính toán lại biến động Adversarial mỗi N khung hình và tiêm biến động đã tính toán cuối cùng vào các khung hình trung gian; và c) chúng tôi sử dụng hàm giá trị, được tính trên đầu vào ban đầu, để ước tính thời điểm tiêm biến động Adversarial sao cho hiệu quả nhất, và chỉ tiêm biến động Adversarial khi ước tính này vượt qua một ngưỡng nhất định.

Độ Hiệu Quả của Việc Huấn Luyện Lại với Các Ví Dụ Adversarial và Nhiễu Ngẫu Nhiên Chúng tôi nghiên cứu xem liệu các tác nhân có thể được huấn luyện lại trên một môi trường với nhiễu ngẫu nhiên hoặc các biến động Adversarial đã tiêm vào để làm cho chúng trở nên bền vững hơn đối với các biến động Adversarial khác. Ngoài ra, chúng tôi nghiên cứu xem sự bền vững này có thể chuyển sang các môi trường với các độ lớn và loại biến động khác nhau (ví dụ: một tác nhân được huấn luyện trên nhiễu ngẫu nhiên có bền vững hơn đối với các biến động Adversarial FGSM không).

3 ĐÁNH GIÁ THÍ NGHIỆM

Để thực hiện các thí nghiệm của chúng tôi, chúng tôi sử dụng một phiên bản TensorFlow của thuật toán A3C (Mnih et al., 2016). Chúng tôi đánh giá phương pháp trên nhiệm vụ Atari Pong, trong đó các pixel ảnh đầu vào ban đầu được cắt bớt và tỉ lệ về kích thước 42x42. Cuối cùng, độ sáng được tính từ các giá trị RGB, cho chúng ta các kích thước khung hình là 42x42x1.

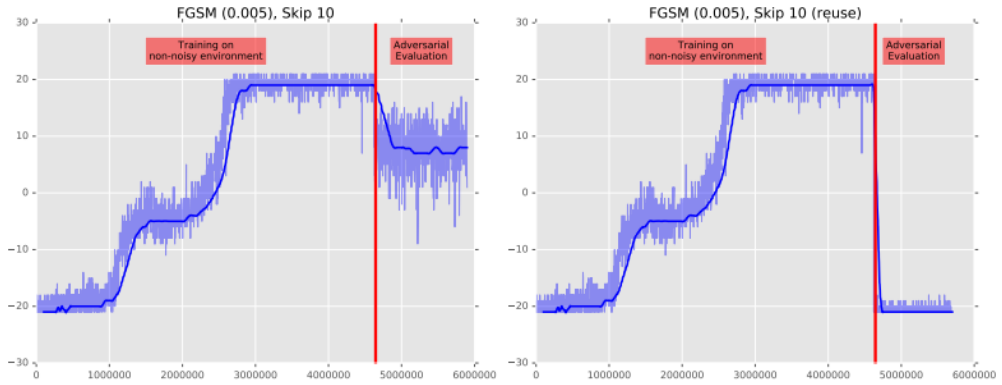
Để tạo ra các biến động Adversarial, chúng tôi sử dụng phương pháp đạo hàm theo dấu nhanh (FGSM) được phát triển ban đầu bởi Goodfellow et al. (2014). FGSM yêu cầu một hàm mất mát $J(\theta, x, y)$ để tính

đạo hàm của nó ∇x. Chúng tôi sử dụng hàm mất mát entropy chéo giữa y (một vector của các logits, đại diện cho trọng số cho mỗi hành động, được tạo ra bởi chính sách) và mã hóa one-hot của arg max y. Điều này có nghĩa là cuộc tấn công cố gắng tạo ra một đầu vào để đẩy đầu ra của chính sách ra khỏi hành động tối ưu.

Trong tất cả các thí nghiệm của chúng tôi, tác nhân được huấn luyện trên một môi trường cơ sở (không có nhiễu) cho đến khi đạt được một phần thưởng tối ưu trong một số tập (tác nhân cơ sở). Sau đó, để tạo ra các biến động FGSM, chúng tôi đặt một giá trị epsilon phù hợp và tính ưu đạo hàm sgn ∇xJ(θ, x, y). Đối với việc tạo ra nhiễu ngẫu nhiên, chúng tôi lấy mẫu từ phân phối đều Unif(0, β), trong đó chúng tôi thiết lập giá trị β dựa trên độ mạnh cần thiết.

Độ hiệu quả của cuộc tấn công của ví dụ Adversarial so với Nhiễu Ngẫu Nhiên được đánh giá trên tác nhân cơ sở trên một phiên bản thay đổi của môi trường, trong đó nhiễu ngẫu nhiên hoặc biến động FGSM được chèn vào mỗi khung hình. Hình 3 trong phần phụ lục cho thấy sự khác biệt trong hiệu quả của cuộc tấn công giữa nhiễu ngẫu nhiên và biến động FGSM. Trong khi mức độ thấp của nhiễu ngẫu nhiên (β ≤ 0,02) không ảnh hưởng nhiều đến hiệu suất của tác nhân, việc sử dụng nhiễu ngẫu nhiên với độ lớn lớn hơn (β ≥ 0,05) gây suy giảm hiệu suất nghiêm trọng. Biến động Adversarial FGSM là hiệu quả nhiều lần so với nhiễu ngẫu nhiên cho các cuộc tấn công thành công, thành công trong tấn công tác nhân cơ sở ở mức độ biến động thấp hơn rất nhiều.

Việc Sử dụng Hàm Giá Trị để Hướng Dẫn Chèn Biến động Adversarial: Trước hết, chúng tôi khám phá cách tần suất chèn biến động adversarial ảnh hưởng đến sự thành công của cuộc tấn công. Trong thí nghiệm này, chúng tôi chèn biến động FGSM chỉ sau mỗi 10 khung hình và sử dụng các khung hình gốc ở giữa, hoặc tính lại biến động sau mỗi 10 khung hình và sử dụng biến động tính toán cuối cùng ở giữa.

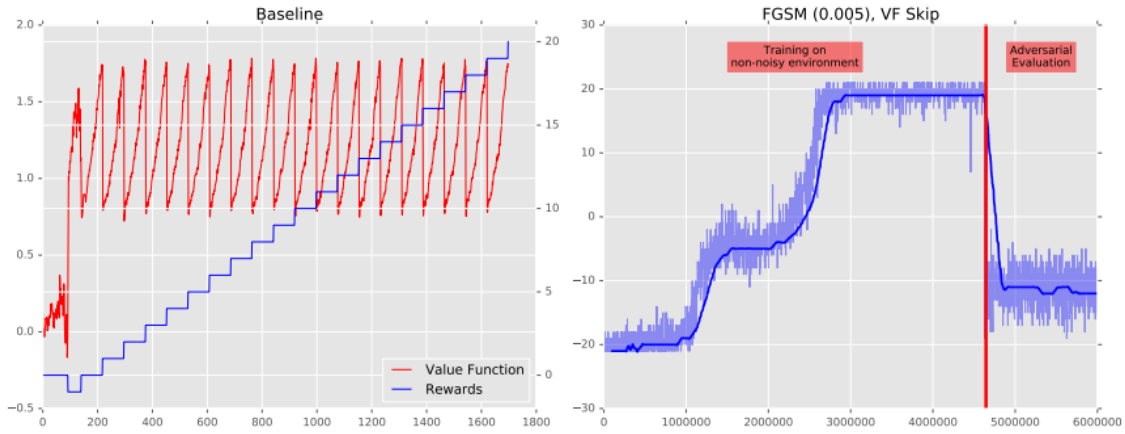


Hình 1: Độ hiệu quả của cuộc tấn công khi biến động FGSM chỉ được chèn vào mỗi khung hình thứ 10 (bên trái) và khi các biến động được tính lại sau mỗi khung hình thứ 10 và được sử dụng lại trong các khung hình trung gian (bên phải).

Các thí nghiệm được thực hiện với được thiết lập là 0,001. Kết quả của chúng tôi cho thấy chỉ chèn biến động FGSM vào mỗi khung hình thứ 10 không có vẻ là một cuộc tấn công đặc biệt hiệu quả (Hình 1,

bên trái). Trong khi đó, tính lại biến động sau mỗi khung hình thứ 10 và sử dụng lại biến động trước đó trong các khung hình trung gian cũng có hiệu quả như cuộc tấn công gốc (Hình 1, bên phải).

Chúng tôi cũng phát triển một phương pháp tấn công (VF) trong đó chúng tôi chỉ tiêm các ảnh chênh lệch gây ảnh hưởng vào chính sách khi giá trị hàm, được tính trên khung gốc, vượt qua một ngưỡng nhất định (trong thí nghiệm này, chúng tôi đặt ngưỡng này là 1,4). Lý do đằng sau là chúng tôi chỉ muốn gây rối cho tác nhân trong những khoảnh khắc quan trọng, khi nó gần như đạt được phần thưởng. Hình 2 cho thấy tính hiệu quả của phương pháp này, chứng tỏ rằng phương pháp tấn công VF rất hiệu quả trong khi chỉ tiêm các ảnh chênh lệch gây ảnh hưởng vào một phần nhỏ trong số các khung. Chúng ta có thể so sánh phương pháp VF với việc tiêm chênh lệch mù quáng vào mỗi khung thứ mười (Hình 1, bên trái). Mặc dù cả hai phương pháp đều tiêm chênh lệch số lần tương đương trung bình trong một tập (120 cho phương pháp VF và 125 cho phương pháp mù quáng), phương pháp VF đã được chứng minh là hiệu quả hơn nhiều. Điều này chứng tỏ rằng một kẻ tấn công có thể sử dụng hàm giá trị để thực hiện một cuộc tấn công hiệu quả hơn phương pháp tấn công truyền thống, trong đó các ảnh chênh lệch gây ảnh hưởng được tiêm vào mỗi khung (như trong (Huang et al., 2017)). Điều này cũng cho thấy rằng các cuộc tấn công gây ảnh hưởng có thể phức tạp hơn nhiều trong môi trường học tăng cường so với các môi trường đã được nghiên cứu trước đó như phân loại hình ảnh.



Hình 2: Ước lượng hàm giá trị của chính sách trong một tập phim gốc (trái). Hiệu quả của việc tiêm sự rối loạn FGSM chỉ trong các khung hình có hàm giá trị vượt ngưỡng (phải).

Chúng tôi cũng nghiên cứu xem liệu các agent có thể được huấn luyện lại để cải thiện độ bền với cả nhiễu ngẫu nhiên và các độ méo phá hoại bất hợp pháp FGSM. Chúng tôi cũng khám phá xem sự đàn hồi này có chuyển sang các độ méo phá hoại khác với độ lớn và kiểu khác nhau không. Trong các thử nghiệm này, sau khi huấn luyện ban đầu trong môi trường không nhiễu, agent được phép huấn luyện lại trong khi chúng tôi tiêm nhiễu ngẫu nhiên hoặc độ méo phá hoại FGSM trên mỗi khung hình. Sau khi agent đạt được hiệu suất tốt, nó sẽ được đóng băng và đánh giá trong một môi trường ồn ào mới, hoặc với nhiễu ngẫu nhiên hoặc độ méo phá hoại FGSM.

Hình 4 trong phụ lục cho thấy trong cài đặt này, đối tượng được đào tạo căn bản có thể chống lại một số cấp độ nhất định của sự nhiễu FGSM sau khi được đào tạo lại trên môi trường nhiễu trong một số tập tin, với độ nhiễu ngẫu nhiên đủ hoặc các lệnh nhiễu FGSM được thêm vào trong quá trình đào tạo lại. Thú vị hơn, thí nghiệm của chúng tôi cho thấy rằng đối tượng được đào tạo lại cũng có khả năng chống

lại sự nhiễu FGSM có độ lớn (hoặc nhỏ) hơn rất nhiều so với độ lớn của sự nhiễu FGSM được sử dụng trong quá trình đào tạo lại. Chúng tôi cũng trực quan hóa các hành động được dự đoán bởi chính sách trong không gian hình ảnh (xem Hình 5).

REFERENCES

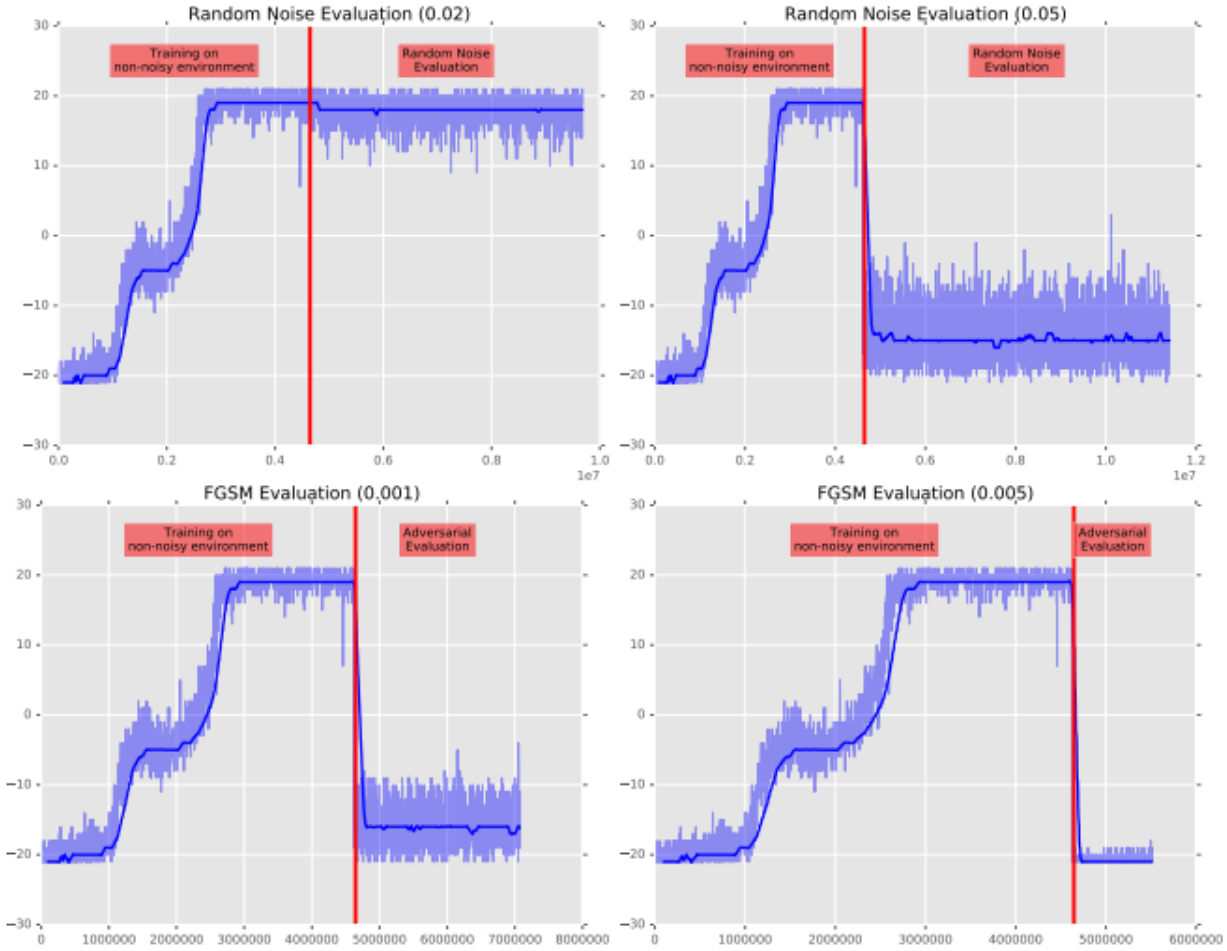
Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Sandy Huang, Nicolas Papernot, Ian Goodfellow, Yan Duan, and Pieter Abbeel. Adversarial attacks on neural network policies. *arXiv preprint arXiv:1702.02284*, 2017.

Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy P Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International Conference on Machine Learning*, 2016.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

A PHỤ LỤC

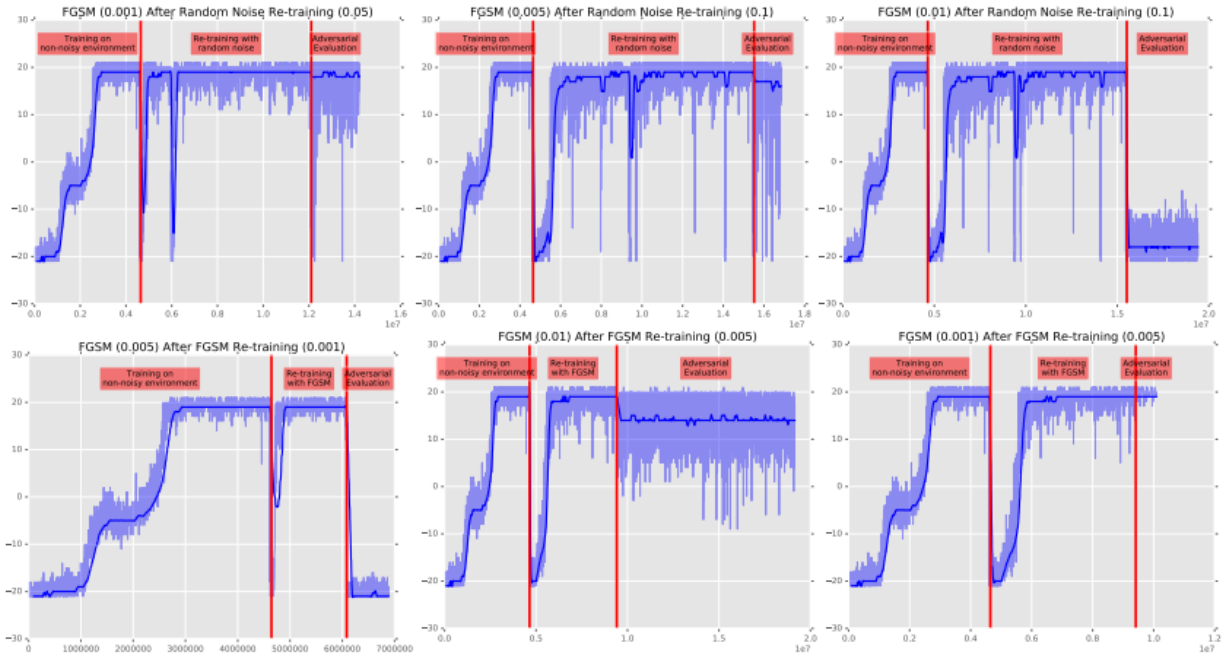


Hình 3: Độ hiệu quả của tấn công nhiễu ngẫu nhiên với giá trị β là 0.02 và 0.05 (phía trên) so với độ hiệu quả của các đột biến phá hoại FGSM với giá trị là 0.001 và 0.005 (phía dưới).

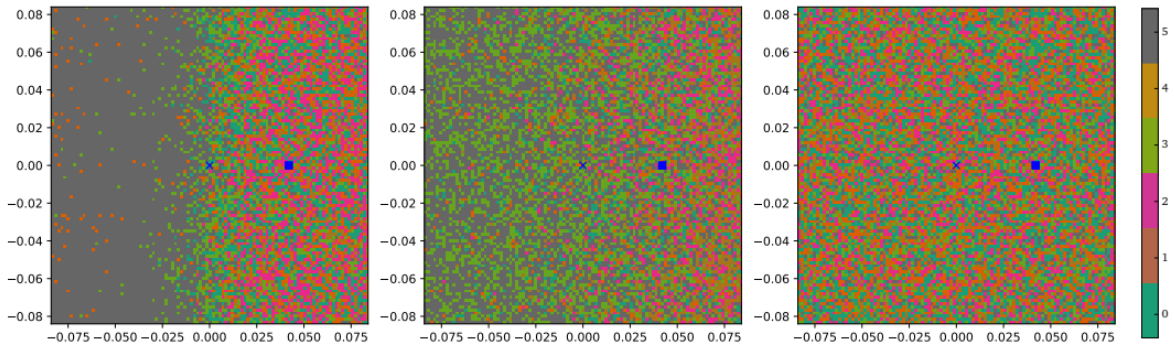
A.1 TRỰC QUAN HÓA RỘNG CỦA HÀNH ĐỘNG MẠNG CHÍNH SÁCH

Chúng tôi tiếp tục nghiên cứu cách thức biên giới hành động trông như thế nào đối với mạng chính sách và cách re-training ảnh hưởng đến nó. Để làm được điều này, chúng tôi chuẩn bị một hình ảnh của hành động dự đoán trong không gian hình ảnh (Hình 5). Chúng tôi tạo ra biểu đồ bằng cách định nghĩa hai vector chuẩn hóa, d1 và d2, chia tách không gian hình ảnh đầu vào. Vector trên trục x trỏ sang hướng của sự tác động nghịch đảo (d1), trong khi vector trên trục y trỏ vào hướng ngẫu nhiên được chọn ngẫu nhiên (d2). Các điểm trong mặt phẳng biểu thị cho các hành động được dự đoán bởi mạng chính sách cho đầu vào x+ud1 +vd2, trong đó x là hình ảnh gốc (một khung hình duy nhất). Vì A3C là ngẫu nhiên, chúng tôi lấy mẫu các dự đoán từ mạng chính sách 7 lần cho mỗi đầu vào và hiển thị hành động phổ biến nhất. Mỗi hành động rời rạc (hành động 0 đến 5) được đại diện bởi màu riêng của nó, được hiển thị trong hình bên phải. Các giá trị trên trục là các giá trị của biến u và v.

Biểu đồ trực quan cho thấy không gian quyết định bị phân mảnh. Những biến đổi nhỏ trong đầu vào có thể khiến hành động tối ưu được chọn bởi mạng chính sách thay đổi đáng kể. Sự đào tạo lại trong môi trường có nhiễu (cả nhiễu ngẫu nhiên và nhiễu đối kháng FGSM) dường như không làm cho đường biên hành động trở nên mượt mà hơn.



Hình 4: Thí nghiệm huấn luyện lại cho agent. Ban đầu, agent được huấn luyện trên một môi trường không gây nhiễu. Đầu: Sau khi huấn luyện lại lần đầu với nhiễu ngẫu nhiên (với giá trị β là 0.05 và 0.1). Dưới: Sau khi huấn luyện lại lần đầu với các nhiễu FGSM (với giá trị epsilon là 0.001 và 0.005). Sau khi huấn luyện lại, agent được đánh giá trên các nhiễu FGSM (với giá trị epsilon là 0.001, 0.005 và 0.01).



Hình 5: Hình dung các hành động trong không gian ảnh cho một khung hình đơn. Trục x ở hướng của ví dụ phá hoại được tạo ra (FGSM, epsilon = 0.001) cho mạng mục tiêu. Trục y nằm trong một hướng vuông góc ngẫu nhiên. Mỗi điểm là kết quả của việc lấy mẫu mạng chính sách 7 lần với ảnh tại điểm đó là đầu vào và hiển thị hành động được sản xuất phổ biến nhất bởi mạng (các hành động khác có màu sắc khác nhau). Đánh dấu "x" màu xanh lam chỉ vị trí khung hình ban đầu, trong khi hình vuông màu xanh lam chỉ vị trí của ví dụ phá hoại. Thanh màu bên phải hiển thị ánh xạ các màu sắc với các hành động rời rạc. Bên trái: Mạng cơ sở mà không có bất kỳ huấn luyện lại nào (hành động cho đầu vào gốc là hành động 5). Giữa: Mạng với huấn luyện lại trên nhiễu ngẫu nhiên (β = 0,1, hành động cho đầu vào gốc là hành động 5). Phải: Mạng với huấn luyện lại trên sự nhiễu FGSM (epsilon = 0,005, hành động cho đầu vào gốc là hành động 0).

Quyết định của mạng chính sách được biểu diễn bằng ranh giới quyết định. Điểm ảnh đầu vào nhỏ trong ranh giới quyết định có thể làm cho hành động tối ưu được chọn bởi mạng chính sách thay đổi một cách đáng kể. Tuy nhiên, việc đào tạo lại mô hình trong môi trường nhiễu (cả nhiễu ngẫu nhiên và nhiễu đối kháng FGSM) không có vẻ làm cho ranh giới quyết định trở nên mịn màng hơn và không gian dường như trở nên ngày càng tách rời hơn (Hình 5, giữa và bên phải).

Chúng tôi cũng điều chỉnh trực quan hóa cho ngữ nghĩa hành động. Lý do đằng sau điều này là mặc dù không gian hành động chứa 6 hành động hợp lệ, các hành động thực sự bị trùng lặp (ví dụ: nhiều hành động thực sự có cùng tác động lên môi trường). Chúng tôi đã kiểm tra tác động của mỗi hành động lên môi trường và ánh xạ các hành động tương ứng. Ba hành động là: không làm gì cả, di chuyển vợt lên và di chuyển vợt xuống. Hình 6 cho thấy rằng ngay cả khi chúng tôi điều chỉnh cho sự trùng lặp, không gian vẫn bị tách rời.

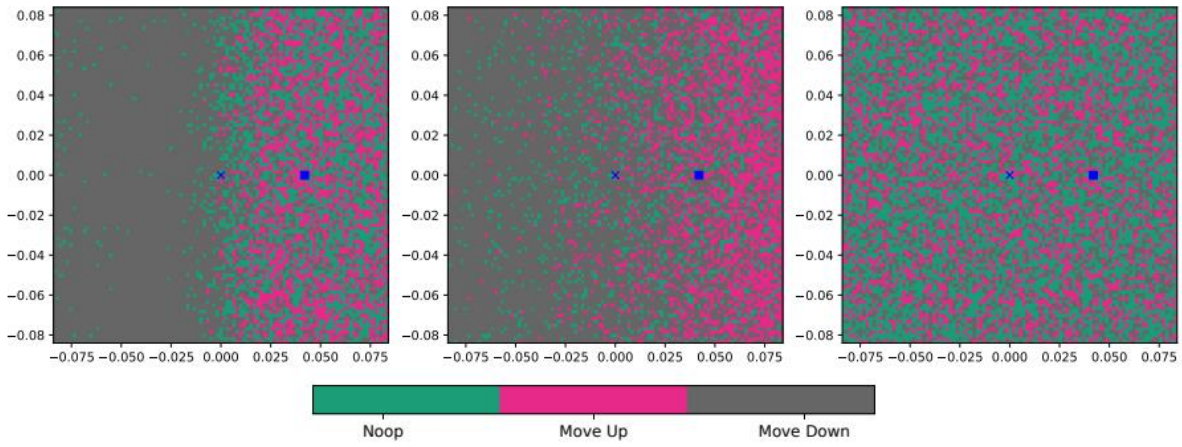


Figure 6: Hình minh họa cho các hành động, đã điều chỉnh theo ngữ nghĩa hành động, trong không gian hình ảnh cho một khung hình duy nhất. Trục x hướng vào hướng của ví dụ đối kháng được tạo ra (FGSM, = 0.001) cho mạng mục tiêu. Trục y hướng theo một hướng vuông góc ngẫu nhiên. Mỗi điểm là kết quả của việc lấy mẫu mạng chính sách 7 lần cho hình ảnh tại điểm đó là đầu vào, và hiển thị hành động phổ biến nhất được đầu ra bởi mạng (các hành động khác có màu sắc khác nhau, được ánh xạ dựa trên ngữ nghĩa hành động). "X" màu xanh đánh dấu vị trí khung hình gốc, trong khi hình vuông màu xanh đánh dấu vị trí ví dụ đối kháng. Thanh màu dưới cùng hiển thị ánh xạ màu cho các hành động rời rạc. Trái: Mạng cơ sở mà không có bất kỳ huấn luyện lại nào (hành động cho đầu vào ban đầu là "di chuyển xuống"). Giữa: Mạng với huấn luyện lại trên nhiễu ngẫu nhiên (β = 0.1, hành động cho đầu vào ban đầu là "di chuyển xuống"). Phải: Mạng với huấn luyện lại trên đối kháng FGSM perturbations (= 0.005, hành động cho đầu vào ban đầu là "không làm gì").