ADVERSARIAL ATTACKS INTO DEEP LEARNING MODELS USING PIXEL TRANFORMATION

Truong Phi Ho[1], Hoang Thanh Nam[1], Tran Quang Tuan[2], Pham Minh Thuan[3], Pham Duy Trung[1*]
[1]Vietnam Academy of Cryptography Techniques, [2]Telecommunications University
[3]Vietnam Government Committee for Information Security

ARTICLE INFO

Received: 22/11/2022
Revised: 26/12/2022
Published: 26/12/2022

KEYWORDS

Deep learning
Adversarial attack
Black-box attack
DNN
Trained model

ABSTRACT

Deep learning is currently developing and being studied by many groups of authors, but deep learning models have potential security risks that can become serious vulnerabilities for applications. At present, the designed antagonist pattern to deceive the neural network in the deep neural network is misbehaving compared to the original design, and the success probability of the antagonist pattern is very worrying, which raises security concerns for machine learning models. Studying and understanding adversarial attacks enhances the security of machine learning models. Most research on antagonistic attacks can fool black box models. The article uses pixel transformation to perform a counterattack, from which it is possible to attack and fool the deep learning system. By the way, the pixel transform method using the Cats and Dogs dataset was tested on the InceptionV3 model. The results demonstrate that the proposed method has a high success rate that causes the deep learning model to misrecognize in the specified target direction.

TẤN CÔNG ĐỐI KHÁNG VÀO MÔ HÌNH HỌC SÂU SỬ DỤNG PHƯƠNG PHÁP BIẾN ĐỔI ĐIỂM ẢNH

Trương Phi Hồ[1], Hoàng Thanh Nam[1], Trần Quang Tuấn[2], Phạm Minh Thuấn[3], Phạm Duy Trung[1*]
[1]Học viện Kỹ thuật Mật mã, [2]Trường Đại học Thông tin Liên lạc, [3]Ban Cơ yếu Chính phủ

THÔNG TIN BÀI BÁO

Ngày nhận bài: 22/11/2022
Ngày hoàn thiện: 26/12/2022
Ngày đăng: 26/12/2022

TỪ KHÓA

Học sâu
Tấn công đối kháng
Tấn công hộp đen
Mạng thần kinh sâu
Huấn luyện mô hình

TÓM TẮT

Học sâu hiện nay đang phát triển và được nhiều nhóm tác giả quan tâm nghiên cứu, tuy nhiên các mô hình học sâu có những rủi ro tiềm tàng về an toàn có thể trở thành những lỗ hổng nghiêm trọng cho các ứng dụng. Hiện nay mẫu đối kháng được thiết kế đánh lừa mạng nơ ron trong mạng thần kinh sâu hoạt động sai so với thiết kế ban đầu, và xác suất thành công của các mẫu đối kháng là rất đáng lo ngại, điều này đặt ra những lo ngại về bảo mật cho các mô hình học máy. Việc nghiên cứu và hiểu rõ các tấn công đối kháng giúp tăng cường độ an toàn cho các mô hình học máy. Thực tế, hầu hết các nghiên cứu về các cuộc tấn công đối kháng có thể đánh lừa các mô hình hộp đen. Bài báo sử dụng phương pháp thay đổi điểm ảnh để thực hiện một cuộc tấn công đối kháng, từ đó có thể tấn công và đánh lừa hệ thống học sâu. Bằng cách này, phương pháp biến đổi điểm ảnh sử dụng tập dữ liệu Mèo và Chó thử nghiệm trên mô hình InceptionV3. Kết quả chứng minh phương pháp đề xuất có tỉ lệ thành công cao khiến mô hình học sâu nhận dạng sai theo hướng mục tiêu đã được chỉ định.

DOI: https://doi.org/10.34238/tnu-jst.6964

* Corresponding author. *Email: trungpd@actvn.edu.vn*

1. Giới thiệu

Mạng nơ ron (NN) đã đạt được những tiến bộ tích cực và thể hiện hiệu suất cao trong phân loại ảnh và trong nhiều ứng dụng khác. Tuy nhiên, mặc dù được sử dụng rộng rãi, nó vẫn rất dễ bị đánh lừa khi đối phương sử dụng hình ảnh đối kháng hay còn được gọi là mẫu đối kháng, đó là các mẫu hoặc hình ảnh chứa nhiễu nhỏ, được sử dụng để đánh lừa NN, làm cho nó phân loại sai. Những mô hình mạng thế này khi hoạt động trong các ứng dụng thực tiễn cũng có thể bị tấn công, bằng cách sửa đổi cấu trúc các đối tượng phải được phân loại [1], [2]. Vấn đề này không chỉ trong phạm vi đối với các NN hoạt động trong phân loại hình ảnh, mà còn đối với các NN trong xử lý ngôn ngữ tự nhiên như giọng nói, văn bản...

Phương pháp tấn công đối kháng có thể được nhóm thành hai loại theo sự hiểu biết về mô hình: tấn công hộp trắng và tấn công hộp đen [3], [4]. Đối với tấn công hộp trắng, đây là phương pháp tấn công khi đối phương có đầy đủ kiến thức về mô hình và quy trình đào tạo của mô hình học máy. Trong trường hợp tấn công hộp trắng, đối phương có thể dựa vào bất kì thông tin nào trong chính mô hình, chẳng hạn như gradient cho một mẫu nhất định hoặc dựa vào chính trọng số của mô hình. Tuy nhiên phương pháp tấn công hộp trắng thường ít được sử dụng trong thực tế vì những kẻ muốn tấn công thường khó có thể biết được các thông số hoặc thông tin chính xác của mô hình theo giả thuyết. Trong phần này, tác giả giới thiệu sơ lược những phương pháp tấn công mà một mô hình học máy có thể gặp phải. Một hướng tấn công khác theo sự hiểu biết về mô hình, trong trường hợp tấn công hộp đen, đối phương chỉ có thể truy cập và tác động vào đầu vào và đầu ra của mô hình, và do đó đây là phương pháp thường được sử dụng trong thực tế nhiều hơn.

Những tiến bộ gần đây trong học máy và mạng thần kinh sâu cho phép các nhà nghiên cứu giải quyết nhiều vấn đề thực tế quan trọng trong phân loại hình ảnh, video, văn bản... Tuy nhiên, như đã nêu ở trên thì hầu hết các mô hình học máy nói chung và học sâu nói riêng hiện có rất dễ bị tấn công theo nhiều cách khác nhau [5]-[8]. Mẫu đối kháng là một ví dụ khi dữ liệu đầu vào đã được sửa đổi theo cách đơn giản nhằm mục đích đánh lừa mô hình học máy, làm cho nó phân loại sai. Trong nhiều trường hợp, những thay đổi này có thể tinh vi đến mức khó có thể nhận biết bằng mắt của con người, tuy nhiên lại đánh lừa mô hình học máy một cách hiệu quả.

Các mẫu đối kháng đặt ra những lo ngại về bảo mật vì các mẫu này có thể được sử dụng để thực hiện một cuộc tấn công vào hệ thống học máy, ngay cả khi kẻ tấn công không có quyền truy cập vào mô hình ở dạng cơ bản. Hơn nữa trong các nghiên cứu [9], [10] đã chỉ ra rằng nó có thể được những kẻ tấn công sử dụng thực hiện tấn công ngay cả vào các hệ thống học máy đang hoạt động trong thực tiễn và nhận dạng hay xác thực thông qua các thiết bị còn đơn sơ, thay vì đọc dữ liệu kỹ thuật số chính xác. Trong tương lai, các mô hình học máy và hệ thống trí tuệ nhân tạo sẽ được cải thiện trở nên mạnh mẽ và an toàn hơn. Các lỗ hổng bảo mật của học máy nói chung, cũng như các mẫu đối kháng nói riêng có thể được sử dụng dùng trong huấn luyện và giúp cho hệ thống trí tuệ nhân tạo tự cải thiện, đảm bảo an toàn hơn. Do đó, mẫu đối kháng là một trong những nghiên cứu quan trọng trong vấn đề đảm bảo tính an toàn cho các hệ thống trí tuệ nhân tạo nói chung, học máy nói riêng.

Nghiên cứu về các cuộc tấn công và cách chống tấn công đối với phương pháp đối kháng đang là hướng nghiên cứu mới và khó tiếp cận với nhiều người bởi nhiều lý do. Trong đó, một lý do là việc đánh giá các cuộc tấn công được đề xuất hoặc các biện pháp phòng thủ được đề xuất không đơn giản. Học máy truyền thống, với giả định về tập huấn luyện và tập thử nghiệm đã được chọn lọc, do đó dễ dàng đánh giá bằng cách đo lường những nguy cơ trên bộ thử nghiệm. Để học máy sử dụng mẫu đối kháng, các mô hình này phải đối mặt với một vấn đề tương đối mới, trong đó kẻ tấn công sẽ gửi đầu vào từ một nguồn không xác định. Và số lượng mẫu đối kháng không đủ để đánh giá được khả năng phòng thủ và chống lại một cuộc tấn công đơn lẻ hoặc thậm chí một loạt các cuộc tấn công được kẻ tấn công chuẩn bị trước với thời gian dài vượt qua cả giới hạn của người nghiên cứu để đề xuất giải pháp chống lại.

Ngay cả khi phòng thủ hoạt động tốt trong một số thử nghiệm, thì nó vẫn có thể bị đánh bại bởi một cuộc tấn công mới hoạt động theo cách mà người phòng thủ đã không biết và dự đoán trước. Trong điều kiện lý tưởng, một biện pháp phòng thủ hiệu quả có thể được chứng minh, nhưng học máy nói chung và mạng thần kinh sâu nói riêng rất khó phân tích hay chứng minh về mặt lý thuyết. Do đó, một số cuộc thi được đề xuất nhằm mang lại một hình thức đánh giá trung gian: phòng thủ được đọ sức trước các cuộc tấn công được xây dựng bởi các nhóm tấn công độc lập, với cả nhóm phòng thủ và đội tấn công đều được khuyến khích để giành chiến thắng. Những đánh giá kiểu như vậy không phải là kết luận theo hình thức chứng minh về mặt lý thuyết, mà nó là chứng minh dưới dạng thực nghiệm; một kịch bản an toàn trong thực tế được đánh giá cao hơn so với một minh chứng về lý thuyết trong trường hợp tấn công sử dụng mẫu đối kháng.

Trong bài báo này, tác giả đề xuất một cuộc tấn công hộp đen đơn giản nhưng đạt hiệu quả gọi là chuyển đổi điểm ảnh, phương pháp này có khả năng tấn công mô hình một cách hiệu quả bằng cách đánh giá chỉ với một mẫu đầu vào, độ tin cậy của dự đoán, cũng có thể coi như là một xác suất. Phương pháp này dựa trên tìm kiếm ngẫu nhiên và nó tạo ra một hình ảnh đối kháng hiệu quả bằng cách sắp xếp lại một tập hợp nhỏ các điểm ảnh tại 1 vị trí trong hình ảnh, cũng cho thấy rằng một hình ảnh thường chứa tất cả các điểm ảnh cần thiết để làm cho mô hình phân loại sai nó, sử dụng một số lần lặp lại, điều này được thể hiện trong kết quả thử nghiệm của chúng tôi.

Phần còn lại của bài báo được tổ chức như sau. Đầu tiên chúng tôi trình bày một số kiến thức tổng quan trong mục 2 và phương pháp nghiên cứu được đề xuất trong 2.2. Mục 3 trình bày về thử nghiệm và kết quả. Kết luận bài báo và các công việc trong tương lai sẽ được nhóm tác giả trình bày trong mục 4.

2. Phương pháp nghiên cứu

2.1. Lý thuyết tổng quan về vấn đề nghiên cứu

Vấn đề bảo mật trong học máy luôn là vấn đề quan trọng nhằm phát triển và cải tiến các mô hình trở nên mạnh mạnh mẽ hơn [11], [4]. Mô hình học máy đầu tiên bị tấn công là Máy học vectơ hỗ trợ [12], sau đó, trong các tác giả phát hiện ra rằng NN có xu hướng tấn công một cách khá đơn giản bằng cách sử dụng một số thuật toán dựa trên gradient [13], bằng cách thu thập thông tin gradient của hình ảnh bị tấn công và do đó đánh lừa mô hình. Sau những khám phá này, vấn đề an toàn đối với NN đã trở thành một chủ đề quan trọng trong thực tiễn triển khai. Trong phần này, tác giả muốn giới thiệu về lý thuyết cơ bản nhất của tấn công sử dụng mẫu đối kháng, sau đó ở phần 2.2 tác giả giới thiệu cụ thể hơn về kỹ thuật tấn công hộp đen, phương pháp được đề xuất trong bài báo này cũng được phân loại ở dạng tấn công này.

2.1.1. Mẫu đối kháng và phân loại tấn công

Các mô hình học máy hoàn thiện thường bị tấn công bằng cách nhận dữ liệu đầu vào là các mẫu đối kháng. Một mẫu đầu vào là "sạch" nếu nó là mẫu thô, chẳng hạn như một bức ảnh được lấy từ tập dữ liệu ImageNet. Nếu đối phương cố ý sửa đổi một mẫu để mô hình học máy phân loại sai thì mẫu đó được gọi là "mẫu đối kháng" (AE). Dựa trên các khía cạnh khác nhau, ta chia tấn công đối kháng thành 3 nhóm dưới đây:

Trước hết, các cuộc tấn công có thể được phân loại theo loại kết quả mà đối thủ mong muốn:

Tấn công không mục tiêu: Trong trường hợp này, mục tiêu của AE là khiến bộ phân loại dự đoán không chính xác đối với bất kì đối tượng nào mà không quan trọng đối tượng cụ thể.

Tấn công có mục tiêu: Trong trường hợp này, kẻ thù muốn thay đổi dự đoán của bộ phân loại thành một số loại mục tiêu cụ thể.

Thứ hai, các kịch bản tấn công có thể được phân loại theo lượng kiến thức mà đối thủ có về mô hình:

Tấn công hộp trắng: Trong tấn công hộp trắng, đối thủ có đầy đủ kiến thức về mô hình bao gồm loại mô hình, kiến trúc mô hình và các giá trị của tất cả các tham số và trọng số huấn luyện.

Hộp đen có đầu dò: Trong trường hợp này, đối phương không biết nhiều về mô hình, nhưng có thể thăm dò hoặc truy vấn mô hình, tức là cung cấp một số đầu vào và quan sát đầu ra. Có nhiều biến thể của kịch bản này - đối thủ có thể biết kiến trúc nhưng không biết các thông số hoặc đối thủ thậm chí có thể không biết kiến trúc mà thực hiện quan sát xác suất đầu ra cho từng lớp, hay đối thủ chỉ lựa chọn quan sát các lớp có khả năng nhất.

Hộp đen không thăm dò: Trong kịch bản hộp đen không thăm dò, đối thủ có kiến thức hạn chế hoặc không có kiến thức về mô hình sẽ tấn công và không được phép thăm dò hoặc truy vấn mô hình trong khi xây dựng các AE. Trong trường hợp này, kẻ tấn công phải xây dựng các AE để đánh lừa hầu hết các mô hình máy học.

Thứ ba, các cuộc tấn công có thể được phân loại bằng cách cung cấp dữ liệu đầu vào mô hình máy học:

Tấn công kỹ thuật số: Trong trường hợp này, đối thủ có quyền truy cập trực tiếp vào dữ liệu thực tế được đưa vào mô hình. Nói cách khác, đối thủ có thể chọn các giá trị như float32 cụ thể làm đầu vào cho mô hình. Trong cài đặt thực tế, điều này có thể xảy ra khi kẻ tấn công tải tệp PNG lên dịch vụ web và cố tình thiết kế tệp để khiến cho tệp đó được đọc không chính xác.

Tấn công vật lý: kẻ thù không có quyền truy cập trực tiếp vào biểu diễn kỹ thuật số được cung cấp cho mô hình. Thay vào đó, mô hình được cung cấp đầu vào thu được bởi các cảm biến như máy ảnh hoặc micrô. Kẻ thù có thể đặt các đối tượng trong môi trường vật lý mà máy ảnh nhìn thấy hoặc tạo ra âm thanh mà micrô nghe thấy. Biểu diễn kỹ thuật số chính xác mà cảm biến thu được sẽ thay đổi dựa trên các yếu tố như góc máy ảnh, khoảng cách đến micrô, ánh sáng xung quanh hoặc âm thanh trong môi trường… Điều này có nghĩa là kẻ tấn công có quyền kiểm soát kém chính xác hơn đối với đầu vào được cung cấp cho máy học.

2.1.2. Tấn công hộp đen

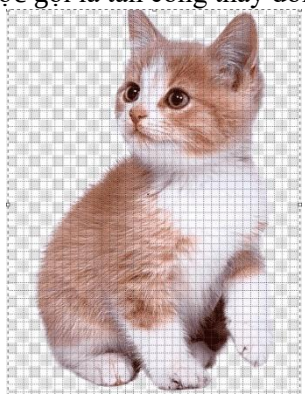
Các AE được sinh ra giữa các mô hình khác nhau [14]. Nói cách khác, một phần đáng kể các AE khi đã đánh lừa được một mô hình thì cũng lừa được một mô hình khác. Thuộc tính này được gọi là "khả năng xâm nhập" và được sử dụng để tạo ra các AE trong tấn công hộp đen. Số lượng các AE có thể xâm nhập thực tế có thể thay đổi từ vài phần trăm và đạt đến gần 100% tùy thuộc vào mô hình nguồn, mô hình mục tiêu, tập dữ liệu và các yếu tố khác. Những kẻ tấn công bằng kịch bản hộp đen có thể huấn luyện mô hình của riêng họ trên cùng một tập dữ liệu với mô hình mục tiêu, hoặc thậm chí huấn luyện mô hình của họ trên một tập dữ liệu khác được rút ra từ cùng một phân lớp. Khi đó, các AE cho mô hình của đối thủ có cơ hội tốt để đánh lừa một mô hình mục tiêu không xác định.

Nếu kẻ tấn công không có tình huống hộp đen hoàn toàn mà được phép sử dụng các thiết bị thăm dò, để huấn luyện bản sao của chính kẻ tấn công của mô hình mục tiêu [15], [8] được gọi là "kẻ thay thế". Cách tiếp cận này rất hiệu quả vì các mẫu đầu vào được gửi dưới dạng thăm dò không cần phải là các mẫu để huấn luyện thực tế; thay vào đó, chúng có thể là các giá trị đầu vào do kẻ tấn công chọn để tìm ra chính xác giới hạn quyết định của mô hình mục tiêu nằm ở đâu. Do đó, mô hình của kẻ tấn công được huấn luyện không chỉ trở thành một bộ phân loại tốt mà còn thực sự thiết kế ngược các chi tiết của mô hình mục tiêu, do đó, hai mô hình được định hướng một cách có hệ thống để có số lần xâm nhập cao. Trong kịch bản hộp đen hoàn chỉnh mà kẻ tấn công không thể gửi các đầu dò, một cách để tăng tốc độ truyền là sử dụng một tập hợp nhiều mô hình làm mô hình nguồn cho các AE [16]. Ý tưởng cơ bản là nếu một AE đánh lừa mọi mô hình trong tập hợp, thì nhiều khả năng nó sẽ tổng quát hóa và đánh lừa các mô hình bổ sung.

Cuối cùng, trong kịch bản hộp đen với các đầu dò, có thể chỉ chạy các thuật toán tối ưu hóa mà không sử dụng gradient để tấn công trực tiếp mô hình mục tiêu [17], [9]. Nhìn chung, thời gian cần thiết để tạo ra một AE cao hơn nhiều so với khi sử dụng một kẻ thay thế, nhưng nếu chỉ cần một số lượng nhỏ các AE được yêu cầu, thì các phương pháp này có thể có lợi thế hơn vì chúng có chi phí cố định ban đầu không cao khi huấn luyện thay thế.

2.2. Phương pháp được đề xuất

Các tác giả đã đề xuất một phương pháp đơn giản tạo ra AE bằng cách xáo trộn các phân đoạn ngẫu nhiên của hình ảnh để tạo ra AE [18]. Trong công bố này [18], các tác giả đề xuất ý tưởng để xác định các nhiễu tần suất thấp nhằm cải thiện hiệu quả truy vấn khi tấn công một mô hình. Một nhóm phương pháp khác được bao gồm bởi phương pháp tiếp cận ước tính gradient của mô hình mà không cần truy cập trực tiếp vào bên trong mô hình [19] – [22], mô phỏng các cuộc tấn công hộp trắng. Trong bài báo này, tác giả miêu tả các cuộc tấn công có mục tiêu và hộp đen, chúng tôi quan tâm và phát triển được gọi là tấn công thay đổi điểm ảnh. Được mô tả như sau:



Hình 1. *Hình ảnh mèo khi được chia nhỏ thành nhiều bản vá*

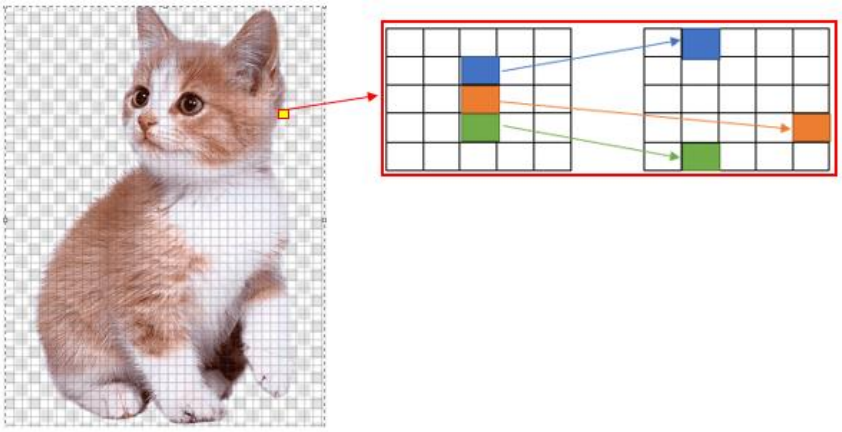
Thay đổi điểm ảnh là một cuộc tấn công hộp đen dựa trên tìm kiếm ngẫu nhiên. Đầu tiên, sử dụng một ảnh $x \in R^{(3,w,h)}$ như Hình 1, ý tưởng chính của thuật toán là lấy mẫu một phần ngẫu nhiên p các điểm ảnh liền kề từ nó, để dễ hiểu tác giả gọi chung là bản vá, sau đó sắp xếp lại một số điểm ảnh bên trong nó vào các vị trí khác được tính toán, đối với mỗi điểm, bằng cách sử dụng một hàm được xác định trước m, tác giả tạm gọi là hàm map. Cho một hình ảnh x, một bản vá hay bản xáo trộn được định nghĩa bởi 4-tuple $p = (w_p, o_y, w_p, h_p)$, trong đó $0 < o_x \leq w$ và $0 < o_y \leq h$ là các tọa độ trên hình ảnh được sử dụng làm điểm gốc của bản vá trên ảnh, khi w_p và h_p lần lượt là chiều ngang và chiều cao của các mảnh vụn bên trong bản vá. Dưới dạng P danh sách các chỉ mục (i, j), được xác định bởi các tọa độ của bản vá, được xác định nằm trong một hình chữ nhật có các đỉnh là $[(o_x, o_y), (o_x + w_p, o_y), (o_x, o_y + h_p), (o_x + w_p, o_y + h_p)]$. Chúng tôi gán danh sách này là $P = [(o_x + i, o_y + j)] \forall i \in \{0, \ldots, w_p\}, j \in \{0, \ldots, h_p\}$, có giá trị $|P| = w_p \times h_p$ giả sử rằng các bản vá không vượt quá biên của hình ảnh, bằng không, nó được di chuyển theo cách sao cho các chỉ mục đều bên trong hình ảnh. Hàm ánh xạ m được định nghĩa theo (1):

$$m: (i, j) \in P \mapsto \mathbb{N}_+^2 \in [1, h] \times [1, w], \tag{1}$$

đối với mỗi điểm trong bản vá, vị trí nơi điểm ảnh phải được di chuyển trong hình ảnh gốc. Với chức năng này, chúng tôi xác định mỗi điểm của hình ảnh đối kháng theo (2):

$$\bar{x}_{i,j} = \begin{cases} m(i, j) & (i, j) \in P \\ x_{i,j} & \text{otherwise} \end{cases} \tag{2}$$

Quy trình tiếp xúc chỉ thay đổi các điểm ảnh đích, được ghi đè lên điểm nguồn. Phương pháp này hiệu quả nếu chúng ta muốn tìm kiếm hình ảnh đối kháng làm giảm khoảng cách từ bản gốc. Nếu không quan tâm đến khía cạnh này, tác giả mong muốn tăng tốc độ hội tụ, một cách tiếp cận khả thi khác là để hoán đổi vị trí của điểm ảnh nguồn và điểm ảnh đích tại đồng thời sử dụng chức năng ánh xạ như trước đây. Theo cách này, không có thông tin dư thừa được đưa vào hình ảnh đối kháng. Các quy trình tổng thể, lấy mẫu một bản vá và tính toán ánh xạ chức năng tương tự. Thuật toán thay đổi điểm ảnh được hiển thị trực quan như Hình 2. Các kết quả và thử nghiệm cụ thể được trình bày trong phần tiếp theo.

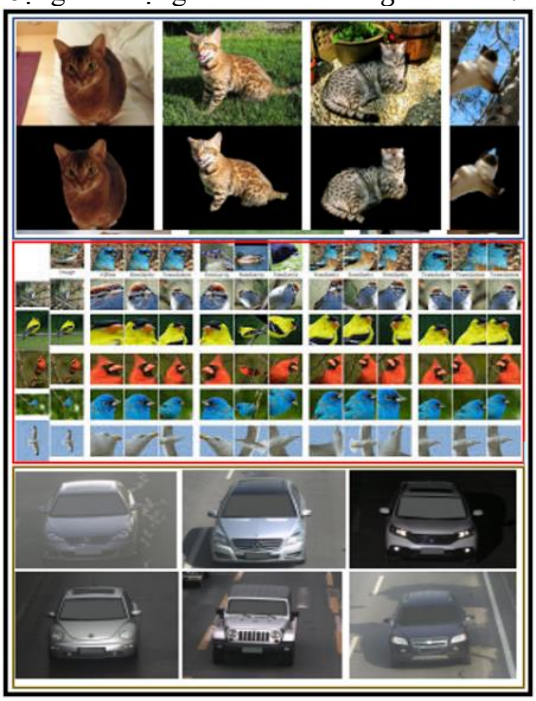


Hình 2. *Chuyển đổi điểm ảnh tại điểm biên của hình gốc*

3. Thực nghiệm và kết quả

3.1. Thực nghiệm

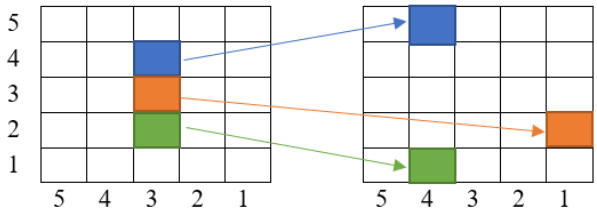
Có rất nhiều bộ dữ liệu được tạo sẵn từ các tác giả. Trong phần thực nghiệm của chúng tôi sử dụng tập dữ liệu gồm nhiều hình ảnh khác nhau: chim, ô tô, chó và mèo,... Ví dụ bộ dữ liệu CompCars [23] có 163 đặc trưng ô tô với 1.716 kiểu ô tô, với mỗi ô tô chú thích và được gán nhãn. Với nhiều hơn 1000 hình ảnh các loại. Trong đó chúng tôi tập trung vào việc kiểm tra Tính khả thi của thuật toán từ ba tập dữ liệu được nêu [6], [23]. Hình 3 cho thấy một số ảnh đại diện của hình ảnh từ tập dữ liệu được sử dụng để trong thực nghiệm của chúng tôi, có mục tiêu đánh lừa hệ thống học sâu nhận dạng đối tượng thành "Lò nướng bánh mì".



Hình 3. *Một số hình ảnh từ 3 tập dữ liệu được chọn*

Chúng tôi mô tả thuật toán chi tiết bằng hình ảnh của mèo sử dụng từ tập dữ liệu [24]. Sử dụng ma trận biến đổi tại điểm trên hình ảnh con mèo và tải mô hình nhận dạng từ Keras được miêu tả trực quan trong Hình 2. Hình 4 mô tả cụ thể hơn ma trận Keras và thuật toán biến đổi

điểm ảnh có thể hiểu đơn giản như sau: lấy các điểm ảnh ở trung tâm và di chuyển chúng đến rìa của bản vá như đã nêu.



Hình 4. *Thuật toán biến đổi điểm ảnh tại các bản vá*

Trong thử nghiệm, tác giả chạymột vòng lặp. Đầu tiên, chọn ngưỡng thích hợp T, sau đó chúng tôi sử dụng vòng lặp để chạy chương trình sử dụng thuật toán đề xuất đến khi chi phí (cost) đạt được kết quả như mong muốn là ngưỡng T, được thể hiện theo đoạn mã giả dưới đây:

Input: Cat image, Keras in InceptionV3 model, Threshold T

Clip border at pixels, cost = 0

For cost < **T** do

Pixel transformation using (1) or (2)

End

Output: Adversarial examples

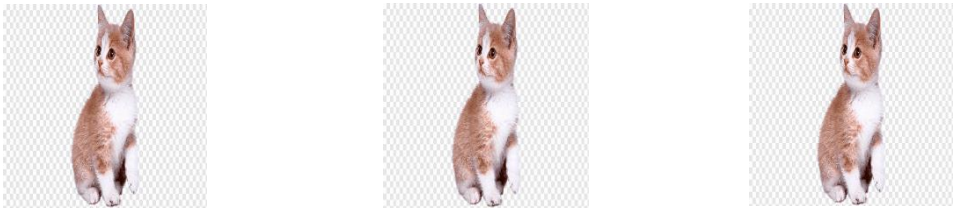
3.2. Kết quả

Khi chuyển đổi kết quả sau khi chuyển đổi điểm ảnh với nhằm mục đích đánh lừa hệ thống nhận dạng với thành lò nướng bánh mì. Kết quả đối với tập dữ liệu chim [6] đạt trung bình 90,23% khi thử nghiệm với 10% dữ liệu ngẫu nhiên. Tỉ lệ trung bình là 99,94% đối với tập dữ liệu chó và mèo [24] khi thử nghiệm với 10% dữ liệu ngẫu nhiên, và kết quả khi thử nghiệm trên tập dữ liệu ô tô [23] đạt trung bình 87,56% khi thử nghiệm trên 5% ảnh ngẫu nhiên. Được thể hiện tổng quát trong Bảng 1.

Bảng 1. *Thực nghiệm thay đổi và tỉ lệ thành công*

Tập dữ liệu	Tỉ lệ thành công trung bình
Birds	90,23%
Cats and Dogs	99,94%
CompCars	87,56%
Tỉ lệ trung bình thực nghiệm	**92,58%**

Không có sự khác biệt đáng kể của hình ảnh sau khi biến đổi với hình ảnh ban đầu theo cảm nhận của mắt người như trong Hình 5. Khi chuyển đổi thành công theo tỉ lệ đối với từng tập dữ liệu theo Bảng 1 đánh lừa mô hình máy học khi sử dụng biến đổi điểm ảnh trên hình ảnh gốc. Kết quả phân loại là một máy nướng bánh mì. Không có sự khác biệt đáng kể về mẫu đối kháng được sinh ra với chi phí đã đặt ở 2 ngưỡng khác nhau nếu quan sát bằng mắt thường. Thời gian chạy chương trình để tạo một mẫu đối kháng sẽ dài hơn khi thiết đặt chi phí cao hơn cho tỉ lệ thành công cao hơn.



Hình 5. *Hình ảnh so sánh giữa ảnh gốc và ảnh được biến đổi với các ngưỡng khác nhau*

4. Kết luận và hướng phát triển

Hầu hết các nghiên cứu về các cuộc tấn công được trên thế giới đều mong muốn và hướng tới tấn công hộp trắng. Trong bài báo này, tác giả đã đề xuất một cuộc tấn công hộp đen mới được gọi là thay đổi điểm ảnh. Trong thử nghiệm tấn công của chúng tôi, mặc dù chỉ sử dụng xác suất dự đoán được trả về bởi mô hình bị tấn công, được chứng minh là có hiệu quả trên nhiều loại tập dữ liệu và kiểu kiến trúc mô hình khác nhau, và trên nhiều chỉ số, chẳng hạn như số lần lặp lại bắt buộc và số lượng điểm ảnh được sửa đổi. Kết quả thông qua quá trình biến đổi từng nhóm điểm ảnh ở biên cho kết quả là 92,58%, đây là giá trị trung bình khi thử nghiệm trên phần trăm dữ liệu của ba bộ dữ liệu đã được chọn. Tỉ lệ thành công cao hơn so với các cuộc tấn công hộp đen khác có tỉ lệ cao nhất đạt 88,9% (xem bảng 2 trong [9]), hay đạt 90% trong nghiên cứu của Zhu và cộng sự [25]. Tác giả mong muốn sẽ thử nghiệm được trên nhiều dữ liệu và các mô hình khác nhau trong nghiên cứu tiếp theo.

So với một số cuộc tấn công hộp trắng như dấu gradient (FGSM) trong [7]. Để kiểm tra rằng các AE có thể được tìm thấy chỉ bằng cách sử dụng ước lượng tuyến tính của mục tiêu mong muốn. L-BFGS, một trong những phương pháp đầu tiên để tìm ra các AE cho mạng nơ ron đã được đề xuất trong [14]; hay đối với phương pháp lặp lại (BIM) được giới thiệu [9], người ta quan sát thấy rằng với đủ số lần lặp lại cuộc tấn công này hầu như luôn thành công trong việc đánh vào lớp mục tiêu [9]. Tuy nhiên, tỉ lệ thành công khi tạo thành mẫu đối kháng không phải là tuyệt đối. Do đó, phương pháp được đề xuất trong bài báo này đạt được hiệu quả cao hơn một số phương pháp vừa được liệt kê. Thông qua thử nghiệm của tác giả, cũng có thể tạm thời kết luận rằng: chưa có phương pháp phòng thủ nào để chống lại các AE hoàn toàn tuyệt đối. Đây vẫn là một lĩnh vực nghiên cứu thời sự và cần được tiếp cận.

Trong tương lai, chúng tôi mong muốn tìm hiểu mối tương quan giữa lớp bị tấn công và các điểm ảnh trong hình ảnh, đặc biệt là khi thực hiện một cuộc tấn công có mục tiêu. Nhóm tác giả đề xuất nên phát triển thêm cho hệ thống chức năng phát hiện sự bất thường trước khi nhận dạng hoặc cung cấp đầu vào bổ sung mẫu đối kháng cho mô hình huấn luyện. Hướng nghiên cứu tiếp theo, chúng tôi sẽ đề xuất một phương pháp xác định các mẫu bất thường để cảnh báo trước khi đưa chúng vào phân loại.

TÀI LIỆU THAM KHẢO/ REFERENCES

[1] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok, "Synthesizing robust adversarial examples," in *International conference on machine learning*, PMLR, 2018, pp. 284-293.

[2] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song, "Robust physicalworld attacks on deep learning visual classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1625-1634 .

[3] S. Bhambri, S. Muku, A. Tulasi, and A. B. Buduru, "A survey of black-box adversarial attacks on computer vision models," *arXiv preprint arXiv:1912.01667*, 2019.

[4] M. Barreno, B. Nelson, R. Sears, A. D. Joseph, and J. D. Tygar, "Can machine learning be secure?," in *Proceedings of the 2006 ACM Symposium on Information, computer and communications security*, 2006, pp. 16-25.

[5] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, and F. Roli, "Evasion attacks against machine learning at test time," in *Joint European conference on machine learning and knowledge discovery in databases*, Springer, 2013, pp. 387-402.

[6] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," *Technical Report CNS-TR-2010-001,* California Institute of Technology, 2011.
[7] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.
[8] N. Papernot, P. McDaniel, and I. Goodfellow, "Transferability in machine learning: from phenomena to black-box attacks using adversarial Samples," *arXiv preprint arXiv:1605.07277*, 2016.
[9] P. Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C. J. Hsieh, "Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models," in *Proceedings of the 10th ACM workshop on artificial intelligence and security*, 2017, pp. 15-26.
[10] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition," in *Proceedings of the 2016 acm sigsac conference on computer and communications Security*, 2016, pp. 1528-1540.
[11] M. Barreno, B. Nelson, A. D. Joseph, and J. D. Tygar, "The security of machine learning," *Machine Learning,* vol. 81, pp. 121-148, 2010.
[12] B. Biggio, B. Nelson, and P. Laskov, "Poisoning attacks against support vector machines*," arXiv preprint arXiv:1206.6389*, 2012.
[13] T. Chen, C. Luo, and L. Li, "Intriguing properties of contrastive losses," *Advances in Neural Information Processing Systems,* vol. 34, pp. 11834-11845, 2021.
[14] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.
[15] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, 2017, pp. 509-519.
[16] Y. Liu, X. Chen, C. Liu, and D. Song, "Delving into transferable adversarial examples and black-box attacks," *arXiv preprint arXiv:1611.02770*, 2016.
[17] W. Brendel, J. Rauber, and M. Bethge, "Decision-based adversarial attacks: Reliable attacks against black-box machine learning models," *arXiv preprint arXiv:1712.04248,* 2017.
[18] N. Narodytska and S. P. Kasiviswanathan, "Simple Black-Box Adversarial Attacks on Deep Neural Networks," in *CVPR Workshops*, 2017, doi: 10.1109/CVPRW.2017.172.
[19] A. Ilyas, L. Engstrom, A. Athalye, and J. Lin, "Blackbox adversarial attacks with limited queries and information," in *International Conference on Machine Learning*, PMLR, 2018, pp. 2137-2146 .
[20] M. Alzantot, Y. Sharma, S. Chakraborty, H. Zhang, C.J. Hsieh, and M. B. Srivastava, "Genattack: Practical black-box attacks with gradient-free optimization," in *Proceedings of the Genetic and Evolutionary Computation Conference*, 2019, pp. 1111-1119.
[21] J. Su, D. V. Vargas, and K. Sakurai, "One pixel attack for fooling deep neural networks," *IEEE Transactions on Evolutionary Computation*, vol. 23, no. 5, pp. 828-841, 2019.
[22] R. Storn and K. Price, "Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces," *Journal of Global Optimization,* vol. 11, no. 4, pp. 341-359, 1997.
[23] L. Yang, P. Luo, C. C. Loy, and X. Tang, "A large-scale car dataset for fine-grained categorization and verification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3973-3981.
[24] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. V. Jawahar, "Cats and Dogs," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, doi: 10.1109/CVPR.2012.6248092.
[25] J. Y. Zhu, C. Xiao, B. Li, W. He, M. Liu, and D. Song, "Spatially transformed adversarial examples," *arXiv:1801.02612v2*, 2018.