PROPOSED METHOD TO REMOVE ADVERSARIAL PERTURBATION USING GENERATIVE MODEL BASED ON DEEP LEARNING

Tran Duc Su[1], Nguyen Tien Dung[2], Dinh Duy Khanh[3*]
[1]*Posts and Telecommunications Institute of Technology*
[2]*School of Information and Communication Technology - Hanoi University of Science and Technology*
[3]*College of Cryptographic Techniques*

ARTICLE INFO

Received: 06/11/2024
Revised: 18/12/2024
Published: 18/12/2024

KEYWORDS

Generative adversarial network
Deep learning
Adversarial perturbation
Adversarial attack
Adversarial defense

ABSTRACT

With the rapid advancement of information technology, artificial intelligence has found extensive applications in various fields, including object recognition, facial recognition, autonomous vehicle operation, and healthcare. However, deep neural networks, which serve as the foundation of many artificial intelligence systems, are highly vulnerable to adversarial examples. These adversarial examples are crafted by introducing subtle and imperceptible perturbations into clean images, effectively deceiving artificial intelligence models and exposing critical weaknesses. Addressing this challenge, the authors propose a new method to remove adversarial perturbation present in the images. This method employs a data generator that learns features directly from the input images, enabling the reconstruction of clean (adversarial perturbations has been removed). The research results demonstrate that this method not only effectively mitigates noise in individual adversarial examples but also counters attacks utilizing adversarial images. This approach opens a new pathway to enhance the accuracy and security of artificial intelligence applications in practice.

ĐỀ XUẤT GIẢI PHÁP LOẠI BỎ NHIỄU ĐỐI KHÁNG SỬ DỤNG MÔ HÌNH TẠO SINH DỰA TRÊN HỌC SÂU

Trần Đức Sự[1], Nguyễn Tiến Dũng[2], Đinh Duy Khanh[3*]
[1]*Học viện Công nghệ Bưu chính Viễn thông*
[2]*Trường Công nghệ thông tin và Truyền thông - Đại học Bách khoa Hà Nội*
[3]*Trường Cao đẳng Kỹ thuật mật mã*

THÔNG TIN BÀI BÁO

Ngày nhận bài: 06/11/2024
Ngày hoàn thiện: 18/12/2024
Ngày đăng: 18/12/2024

TỪ KHÓA

Mạng tạo sinh
Học sâu
Nhiễu đối kháng
Tấn công đối kháng
Phòng thủ đối kháng

TÓM TẮT

Với sự tiến bộ nhanh chóng của công nghệ thông tin, trí tuệ nhân tạo ứng dụng rộng rãi trong nhiều lĩnh vực như nhận dạng đối tượng, nhận dạng khuôn mặt, vận hành xe tự hành và chăm sóc sức khỏe. Tuy nhiên, mạng nơ-ron sâu, vốn là nền tảng của nhiều hệ thống trí tuệ nhân tạo, lại dễ bị tổn thương trước các mẫu đối kháng. Các mẫu đối kháng được tạo ra bằng cách thêm các nhiễu loạn khó nhận thấy vào hình ảnh sạch, đánh lừa hiệu quả các mô hình trí tuệ nhân tạo và thể hiện các điểm yếu của mô hình. Để giải quyết thách thức này, các tác giả đề xuất một phương pháp mới để loại bỏ nhiễu đối kháng có chứa trong hình ảnh. Phương pháp này sử dụng mô hình tạo dữ liệu học các đặc trưng trực tiếp từ hình ảnh đầu vào, cho phép tái tạo hình ảnh sạch. Kết quả nghiên cứu cho thấy phương pháp này không chỉ khắc phục nhiễu hiệu quả trên các mẫu đối kháng riêng lẻ mà còn chống lại các cuộc tấn công sử dụng ảnh đối kháng. Điều này mở ra một hướng tiếp cận mới nhằm nâng cao độ chính xác và tính an toàn của các ứng dụng trí tuệ nhân tạo trong thực tế.

DOI: https://doi.org/10.34238/tnu-jst.11486

* Corresponding author. *Email: duykhanh09099085@gmail.com*

1. Giới thiệu

Mạng nơ-ron sâu (DNN) đã được sử dụng rộng rãi trong nhiều ứng dụng thị giác máy tính, chẳng hạn như nhận dạng hình ảnh [1], xử lý hình ảnh [2], phân đoạn [3] và hợp nhất hình ảnh [4]. Tuy nhiên DNN có thể dễ dàng bị đánh lừa bởi các hình ảnh gọi là mẫu đối kháng (AE); đó là những mẫu, hình ảnh chứa nhiễu nhỏ, khó có thể nhận biết bằng mắt thường.

Shi và cộng sự [5] đã đề ra một phương pháp tấn công hiệu quả để đánh lừa bộ phân loại hình ảnh bằng cách phân bổ kích thước bước nhiễu dựa trên thông tin độ dốc. Hay Xiao và nhóm nghiên cứu [6] phát triển một phương pháp tấn công đối kháng để đánh lừa DNN. Ứng dụng của DNN rất nhạy cảm và cần độ chính xác cao như nhận diện đối tượng, nhận diện gương mặt, lái xe tự động hay phân tích y tế,v.v. Do vậy, nghiên cứu phát triển AE và các phương pháp chống lại AE luôn mang tính thời sự.

Các phương pháp chống lại đối kháng điển hình như: huấn luyện đối kháng [7], chuẩn hóa gradient [8] và phương pháp dựa trên đầu độc đầu vào [9]. Phương pháp huấn luyện đối kháng và chính quy hóa gradient cần phải huấn luyện lại hoặc chỉnh sửa bộ phân loại. So với 2 phương pháp trên, các phương pháp dựa trên đầu độc đầu vào tập trung điều chỉnh, sửa đổi đầu vào trước khi đưa vào bộ phân loại, do đó, phương pháp này có tính ứng dụng hơn. Jia và nhóm nghiên cứu [10] tập trung vào việc khắc phục các AE bằng các mô hình đã được huấn luyện dựa trên tập dữ liệu huấn luyện lớn. Tuy nhiên, phương pháp phòng thủ này chủ yếu dựa trên các ưu tiên bên ngoài đã học được từ tập dữ liệu huấn luyện lớn mà bỏ qua chính những ưu tiên phong phú bên trong đầu vào. Dữ liệu huấn luyện đã được thống kê không thể tổng quát cho mọi loại tấn công, vậy nên việc ứng dụng các phương pháp phòng thủ này cũng bị hạn chế.

Ngoài phương pháp phòng thủ đối kháng đã được phát triển như huấn luyện đối kháng và đầu độc đầu vào, một số phương pháp khác nhằm cải thiện tính mạnh mẽ của mô hình học sâu đã thêm các mẫu đối kháng vào dữ liệu huấn luyện nhưng khả năng khái quát hóa kém đối với các cuộc tấn công chưa biết trước [11]. Để khắc phục vấn đề này, Xie và nhóm tác giả [12] đề xuất thêm các khối khử nhiễu đặc trưng cho bộ phân loại.

Một phương pháp phòng thủ khác là dựa trên đầu độc đầu vào [13], [14], không yêu cầu huấn luyện lại hoặc sửa đổi bộ phân loại. Phương pháp này nhằm mục đích loại bỏ nhiễu đối kháng từ đầu vào trước khi đưa vào bộ phân loại. Trong [15], [16], các tác giả đã sử dụng các biến đổi đầu vào khác nhau gồm độ sâu bit màu, làm mờ hình ảnh và nén JPEG, để có được hiệu suất bảo vệ tốt. Tuy nhiên, các phương pháp này bị mất thông tin hình ảnh và không hoạt động tốt với nhiễu đối kháng mạnh.

Trong nghiên cứu này, nhóm tác giả đề xuất phương pháp phòng thủ bằng cách loại bỏ nhiễu đối kháng có trong hình ảnh đầu vào. Dựa vào hình ảnh đầu vào thông qua việc khai thác các đặc trưng bên trong ảnh đối kháng đầu vào riêng lẻ; có thể tái cấu trúc hình ảnh ban đầu và loại bỏ nhiễu đối kháng do kẻ tấn công thêm vào. Việc tái tạo hình ảnh tuân theo chiến lược học hai giai đoạn: giai đoạn chuẩn hóa hình ảnh nhằm trích xuất những đặc trưng của ảnh, giai đoạn sau học đặc trưng của ảnh sau khi chuẩn hóa nhằm tái tạo ảnh loại bỏ nhiễu đối kháng.

Ngoài ra, phương pháp này chủ yếu học đặc trưng của ảnh sau khi đã được chuẩn hóa, sau đó làm mịn ảnh. Từ đó có thể khắc phục nhiễu đối với AE riêng lẻ và chống lại một số cuộc tấn công dùng nhiễu đối kháng. Nhóm tác giả tiến hành thực nghiệm trên tập dữ liệu Hybrid CIFAR-10 [17], sử dụng 5 mô hình học sâu hiện đại để đánh giá độ chính xác. Từ đó, tác giả đã chứng minh rằng phương pháp đề xuất có thể giúp tái tạo lại hình ảnh thông qua loại bỏ nhiễu đối kháng. Điều này làm cho hình ảnh sau khi khôi phục được nhận dạng đúng với nhãn nguyên bản.

Phần còn lại của bài báo được cấu trúc như sau. Trong phần 2, tác giả trình bày khái quát về một số thuật toán tấn công đối kháng được tác giả chia thành cổ điển và hiện đại; đồng thời tác giả trình bày về cách tiếp cận, chi tiết phương pháp được đề xuất trong nghiên cứu này. Phần 3 trình bày thử nghiệm phương pháp đề xuất và kết quả chứng minh tính khả thi của phương pháp. Kết luận bài viết và các công việc trong tương lai được trình bày trong phần 4.

2. Tấn công đối kháng và đề xuất phương pháp chống lại tấn công đối kháng

2.1. Tấn công đối kháng và các phương pháp phòng chống

2.1.1. Phương pháp tấn công đối kháng cổ điển và hiện đại

Có nhiều phương pháp phân loại các cuộc tấn công đối kháng khác nhau. Trong nghiên cứu này, nhóm tác giả trình bày 2 hình thức tấn công đối kháng phân theo thời gian nghiên cứu và công bố được tác giả định nghĩa: cổ điển và hiện đại.

- *Tấn công đối kháng cổ điển:*

Phương pháp tấn công điển hình có thể kể tới là phương pháp ký hiệu Gradient nhanh (Fast Gradient Sign Method - FGSM) và các biến thể của phương pháp này. Phương pháp FGSM là một trong những phương pháp tấn công đầu tiên được giới thiệu bởi Goodfellow và cộng sự [18]. Phương pháp FGSM tạo ra các mẫu đối kháng bằng cách sử dụng thuật toán như sau: cho một hàm mục tiêu $L(X+\rho;\theta)$, trong đó, θ biểu thị các tham số của mạng, mục tiêu là tối đa hóa hàm mất mát theo công thức (1).

$$arg\max_{\rho\in\mathcal{R}^m} L(X+\rho;\theta) \tag{1}$$

FGSM là phương pháp tấn công một bước và nhằm mục đích tìm ra nhiễu đối kháng bằng cách di chuyển theo hướng ngược lại với độ dốc của hàm mất mát $L(X+\rho;\theta)$, với ϵ là tỉ lệ nhiễu loạn; mẫu đối kháng X_{adv} được sinh ra theo công thức (2).

$$X_{adv} = X + \epsilon \times sign\big(\nabla L(X_i+\rho;\theta)\big) \tag{2}$$

Hàng loạt các biến thể của FGSM được các nhà nghiên cứu phát triển như PGD [19], tăng cường tấn công bằng động lượng [20]. Các phương pháp này có thể được coi là cổ điển vì nghiên cứu và công bố vào thời kỳ đầu tiên của tấn công đối kháng.

- *Tấn công đối kháng hiện đại sử dụng mạng tạo sinh và thuật toán khuếch tán:*

Một số phương pháp tấn công đối kháng được nghiên cứu gần đây (được tác giả định nghĩa là hiện đại) như sử dụng mạng tạo sinh đối kháng (Generative Adversarial Network-GAN) hay sử dụng thuật toán khuếch tán (Diffusion) trong việc tạo ra hình ảnh đối kháng. Zhang và cộng sự [21] đã tạo ra các hình ảnh đối kháng từ MNIST và CIFAR-10 bằng LSGAN, tạo ra một tập dữ liệu có thể đánh lừa các mô hình. Hay như Jordan và các cộng sự [22] đã giới thiệu tập dữ liệu CIFAKE, sử dụng các thuật toán khuếch tán để tạo ra các hình ảnh giống CIFAR-10. Cả 2 phương pháp này đều thể hiện sự tiến bộ, mang tính thời sự và được đông đảo cộng đồng nghiên cứu quan tâm và phát triển. Hiện nay, chúng được coi là các phương pháp tấn công đối kháng hiện đại điển hình.

2.1.2. Một số phương pháp phòng thủ đối kháng cơ bản và cách tiếp cận của tác giả

Một số phương pháp phòng thủ dựa trên chuyển đổi đầu vào trước đây [14], [23] cố gắng tinh lọc hoặc sửa đổi các mẫu đối kháng thành hình ảnh sạch bằng DNN. Trong [24], Liao và nhóm nghiên cứu đề xuất một bộ khử nhiễu cấp cao (HGD) để loại bỏ nhiễu đối kháng. Trong [14], [25], các tác giả đã tận dụng các mô hình tổng quát để làm sạch các mẫu đối kháng, từ đó biến những hình ảnh bất lợi thành hình ảnh rõ ràng.

Samangouei và cộng sự đã sử dụng mạng sinh đối kháng (Defend-GAN) [26] được đề xuất chiếu các mẫu đối kháng vào không gian của một máy mô hình hóa việc phân phối các hình ảnh sạch. Tuy nhiên, Defense-GAN tồn tại một số nhược điểm như yêu cầu một lượng lớn dữ liệu không được đánh nhãn để huấn luyện và khả năng tính toán cao. Từ đó, dẫn đến mô hình có thể học tập những đặc trưng bên ngoài hình ảnh cần khử nhiễu bị thống kê sai lệch. Đây cũng là hạn chế của phương pháp vừa nêu khi ứng dụng vào thực tế.

Cách tiếp cận của nghiên cứu này không giống với những phương pháp phòng thủ trước đây mà tập trung vào việc sử dụng các đặc trưng bên trong hình ảnh để giảm tác động nhiễu đối kháng. Tuân theo chiến lược học đặc trưng sau khi hình ảnh đã được chuẩn hóa, đảm bảo hiệu suất của nó đối với các cường độ nhiễu và các loại tấn công khác nhau. Hơn nữa, phương pháp được đề xuất

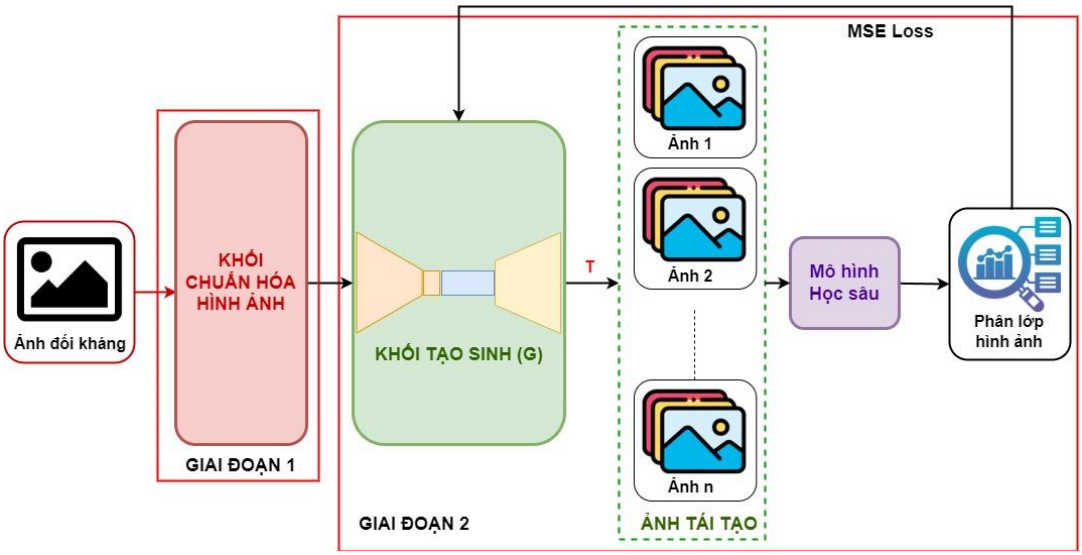
chỉ yêu cầu và áp dụng trên một mẫu đối kháng riêng lẻ. Do đó, phương pháp này có thể linh hoạt hơn để phòng thủ trước các loại tấn công với đa dạng mô hình tấn công khác nhau.

2.2. Đề xuất phương pháp loại bỏ nhiễu đối kháng sử dụng mô hình tạo sinh

Tổng quát về phương pháp được đề xuất, quá trình loại bỏ nhiễu đối kháng gồm 2 giai đoạn:

- **Giai đoạn 1.** Chuẩn hóa hình ảnh trước khi đưa vào tái tạo sử dụng mạng tạo sinh.
- **Giai đoạn 2.** Sử dụng mô hình tạo sinh để loại bỏ nhiễu đối kháng, tái tạo hình ảnh.

Hình 1 trực quan hóa quy trình loại bỏ nhiễu đối kháng thông qua 2 giai đoạn vừa nêu. Có thể tóm tắt thuật toán loại bỏ nhiễu đối kháng như sau: cứ mỗi T bước nhảy trong tổng số F vòng lặp, ghi lại hình ảnh được sinh bởi khối tạo sinh G thu được n hình ảnh; phân lớp n hình ảnh thu được và chọn xác suất nhãn được dự đoán là cao nhất, từ đó chọn ra được phân lớp ảnh chính xác của ảnh nguyên bản khi chưa có nhiễu đối kháng.



Hình 1. *Mô hình tổng quát quá trình loại bỏ nhiễu đối kháng*

Kiến trúc mô hình tạo sinh (được tác giả viết tắt là G) thể hiện như trong Hình 2. Quan sát Hình 2, có thể thấy G gồm 3 thành phần chính như sau:

1) **Encoder:** có chức năng trích chọn những đặc trưng bậc cao của ảnh trước khi đưa vào Bottleneck để lưu giữ những thông tin có giá trị được sử dụng cho quá trình tái tạo ảnh.

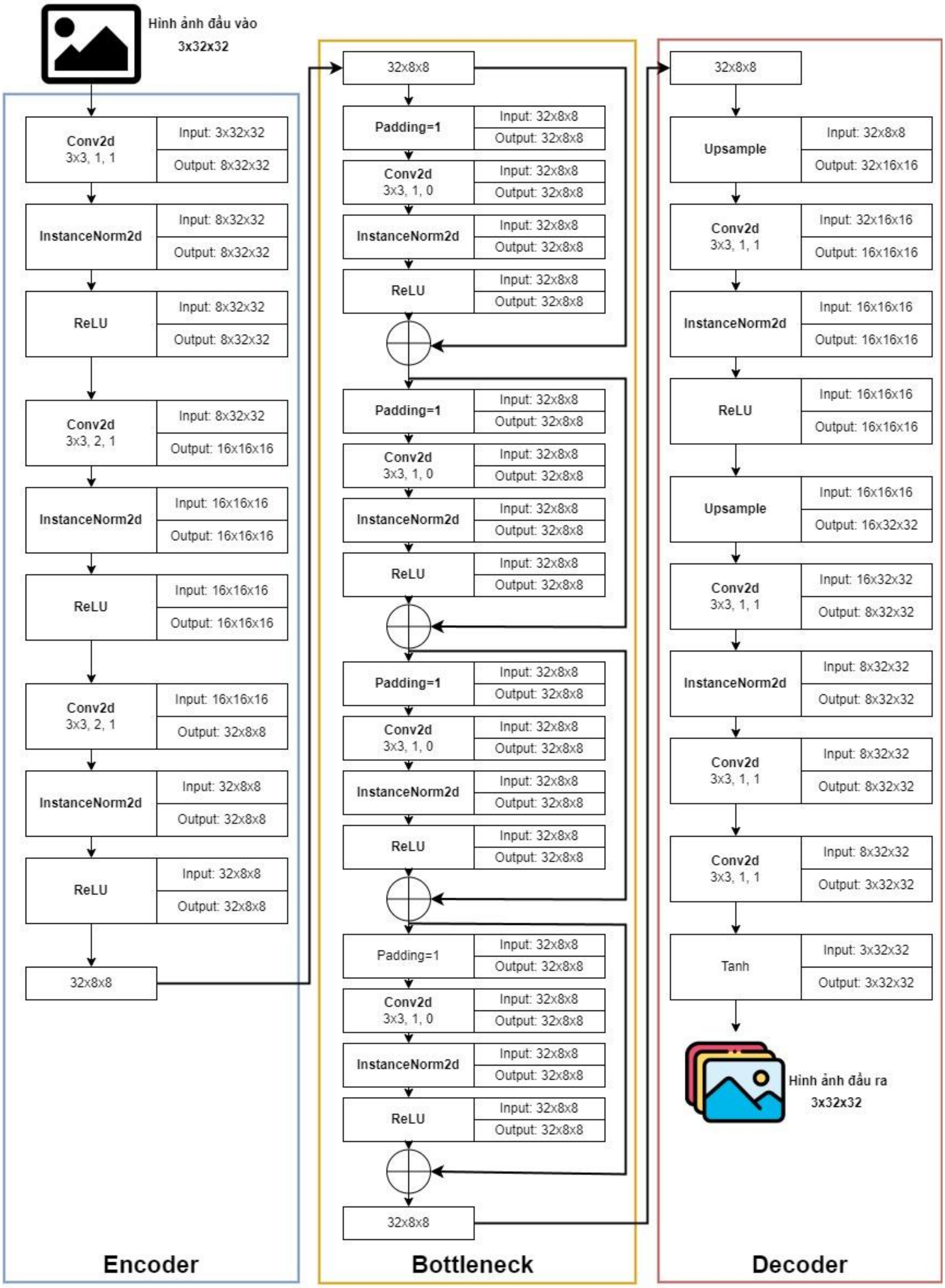
2) **Bottleneck:** được tác giả thiết kế dựa trên kiến trúc ResNet với các kết nối tắt giúp hạn chế hiện tượng mất dấu "gradient", đây là không gian ẩn lưu trữ những đặc trưng bậc cao do Encoder trích chọn trước khi đưa vào tái tạo ảnh ở Decoder.

3) **Decoder:** chức năng tái tạo ảnh khi nhận những thông tin từ Bottleneck, qua đó ảnh được sinh ra đã được lọc bỏ nhiễu đối kháng, nhiễu này không có giá trị về mặt thông tin, đặc trưng vốn có của hình ảnh nguyên bản.

Nghiên cứu này sử dụng trình tối ưu hóa Adam, với hệ số học "learning rate = 0,001". Tác giả cũng sử dụng hàm mục tiêu MSE Loss (Mean Square Error) (3) cho quá trình huấn luyện mô hình.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \quad (3)$$

trong đó, y_i là giá trị thực tế của mẫu thứ i, $\hat{y}_i$ là giá trị dự đoán của mô hình cho mẫu thứ i, và N là tổng số mẫu trong tập dữ liệu.



Hình 2. *Kiến trúc mô hình tạo sinh (G)*

3. Thực nghiệm và kết quả

3.1. Phương pháp thực nghiệm

Nhóm tác giả thực nghiệm trên Python3, dùng PyTorch 2.3.1, CUDA 12.1, trên Google Colab Notebook với GPU T4 và 16 GB VRAM. Bộ dữ liệu sử dụng trong thực nghiệm là Hybrid CIFAR-10 [17], gồm 158.498 ảnh màu có kích thước 32×32 điểm ảnh, chia thành 10 phân lớp, mô phỏng theo tập dữ liệu CIFAR-10 [27] với 72.249 ảnh gốc và 72.249 ảnh đối kháng. Hybrid CIFAR-10 được tạo theo phương pháp GAN do Trung và cộng sự đề xuất (xem trong [28]), gồm 2 bước chính:

Bước 1. Các tác giả sử dụng phương pháp của mô hình GAN để nhằm tạo mặt nạ nhiễu loạn đánh lừa mô hình học máy để xác định sai lớp của hình ảnh đầu vào.

Bước 2. Kết hợp mặt nạ nhiễu loạn với hình ảnh gốc bằng cách sử dụng hệ số nhiễu loạn $k \in [0,1; 1]$ với mỗi lần tăng k thêm 0,1 đơn vị theo (4).

$$x_{adv} = x + (k \times \Delta), \quad (4)$$

trong đó, x_{adv} là ảnh đối nghịch, x là ảnh gốc và Δ là mặt nạ nhiễu loạn được tạo ra bởi mô hình dựa theo GAN được Trung và cộng sự [28] đề xuất.

Phương pháp tiếp cận của Trung và cộng sự nhằm điều chỉnh gia tăng hệ số nhiễu loạn, cho phép tác giả đánh giá độ khác nhau của ảnh đối nghịch so với ảnh gốc (chi tiết cụ thể về mô hình và các bước trong [28]). Một số hình ảnh của tập dữ liệu Hybrid CIFAR-10 hiển thị trong Hình 3.

Hình 3. *Một số hình ảnh về tập dữ liệu Hybrid CIFAR-10* [24] *được biểu diễn theo từng cặp trái - phải tương ứng ảnh thật - ảnh đối kháng với cường độ nhiễu đối kháng tăng dần từ trên xuống dưới*

Chúng tôi đánh giá phương pháp đề xuất thông qua sử dụng 5 mô hình học sâu hiện đại phổ biến như ResNet-56, MobileNetV2, VGG19_bn, ShuffleNetV2, và RepVGG_a2. Các mô hình này đạt hiệu quả cao trên tập dữ liệu CIFAR-10 (đã được huấn luyện, kiểm tra và công bố trên nền tảng Pytorch). Năm mô hình học sâu được sử dụng để phân lớp các hình ảnh sau khi thử nghiệm thuật toán, sau đó tiến hành đánh giá và thống kê kết quả. Thử nghiệm với các giá trị T khác nhau, chúng tôi chọn T = 20, thu được n = 250 hình ảnh sau F = 5000 vòng lặp cho kết quả tái tạo tốt nhất.

3.2. Kết quả thực nghiệm và thảo luận

Ảnh đối kháng được chọn ngẫu nhiên từ tập dữ liệu Hybrid CIFAR-10 (số lượng 500 ảnh) đánh lừa thành công trên 5 mô hình học sâu đã trình bày. Hình ảnh đối kháng được chọn chia đều cho 10 phân lớp, với các độ nhiễu khác nhau. Kết quả từ Bảng 1 cho thấy, phương pháp đề xuất hiệu quả trong việc loại bỏ nhiễu đối kháng bằng khối G được xây dựng dựa trên kiến trúc Encoder-Decoder, giúp loại bỏ thông tin không quan trọng mà kẻ tấn công thêm vào khi tạo ảnh đối kháng.

Tuy nhiên, kết quả thực nghiệm còn hạn chế do một số ảnh thuộc phân lớp Car, Bird, Cat vẫn khó loại bỏ nhiễu, vì mang đặc trưng chưa rõ ràng. Đặc biệt thông qua kết quả thực nghiệm cũng cho thấy MobileNetV2 dù là mô hình nhẹ và hiệu quả dành cho thiết bị di động và nhúng, nhưng gặp khó khăn khi xử lý các tác vụ phức tạp như phân loại ảnh chi tiết hoặc cần thông tin ngữ

cảnh. Điều này xuất phát từ thiết kế tối ưu hóa nhẹ với các khối trong thiết kế giúp giảm tham số và phép tính nhưng hạn chế khả năng biểu diễn. Do đó dẫn đến kết quả của MobileNetV2 trong Bảng 1 thấp hơn so với 4 mô hình học sâu còn lại.

Nghiên cứu này chỉ đánh giá tỉ lệ chính xác của mô hình trên dữ liệu đã loại bỏ nhiễu đối kháng (theo Bảng 1) mà chưa xem xét các chỉ số khác. Thực tế, việc xác định hình ảnh nguyên bản x của ảnh đối kháng x_{adv} để so sánh, đo lường và đánh giá độ trung thực của ảnh đối kháng sau khi loại bỏ nhiễu vẫn còn nhiều khó khăn. Ngoài ra, các thử nghiệm mới chỉ thực hiện trên hình ảnh kích thước nhỏ (32×32 điểm ảnh) và chưa thực nghiệm trên nhiều tập dữ liệu khác nhau nên chưa đánh giá được toàn diện, phần nào hạn chế tính ứng dụng thực tiễn của phương pháp đề xuất.

Bảng 1. *Tỉ lệ chính xác (%) phân lớp hình ảnh sau khi khử nhiễu bằng mô hình tạo sinh G*

STT	Phân lớp	Số lượng	Mô hình học sâu dùng trong phân lớp hình ảnh				
			ResNet-56	**MobileNetV2**	**VGG19_bn**	**ShuffleNetV2**	**RepVGG_a2**
1	Airplane	10	90	100	100	90	100
2	**Car**	10	0	0	0	0	0
3	**Bird**	10	0	0	0	0	0
4	**Cat**	10	80	0	80	80	90
5	Deer	10	100	100	90	100	100
6	Dog	10	100	100	100	100	80
7	Frog	10	100	100	100	100	90
8	Horse	10	100	100	100	100	90
9	Ship	10	100	100	100	100	90
10	Truck	10	100	100	100	100	80
Trung bình		**100**	**77**	**70**	**77**	**77**	**72**

So với các nghiên cứu trước [24] – [26], phương pháp đề xuất giảm thiểu năng lực tính toán bằng cách khôi phục trực tiếp trên từng ảnh riêng lẻ, và đạt hiệu quả cao hơn khi phương pháp HGD [24] giảm tỷ lệ sai sót do tấn công đối kháng xuống 10-20% so với không sử dụng phòng thủ. Trên CIFAR-10, PixelDefend [25] giảm sai sót khoảng 10% với các tấn công như FGSM và BIM, trong khi mô hình không phòng thủ có sai sót gần 50%. Sau khi áp dụng Defense-GAN [26], độ chính xác phục hồi lên 70-80%, cho thấy khả năng bảo vệ tốt. Khi không có phòng thủ, các mô hình như ResNet hoặc VGG bị giảm độ chính xác dưới 20-30% với các tấn công FGSM hoặc PGD.

4. Kết luận

Trong bài báo này, nhóm tác giả đã khảo sát lý thuyết về tấn công và phòng thủ đối kháng, đồng thời chỉ ra hạn chế của các phương pháp phòng thủ hiện tại. Nhóm đề xuất một phương pháp khôi phục hình ảnh đầu vào bằng cách loại bỏ nhiễu đối kháng thông qua bộ sinh dữ liệu, học các đặc trưng từ chính hình ảnh. Kết quả cho thấy ảnh sau khi khôi phục vẫn giữ được hầu hết các đặc trưng cơ bản, giúp các mô hình học sâu nhận dạng chính xác phân lớp của ảnh gốc.

Trong tương lai, nhóm tác giả sẽ thử nghiệm phương pháp trên các tập dữ liệu với kích thước hình ảnh lớn hơn không chỉ dừng ở 32×32 điểm ảnh và số lượng hình ảnh nhiều hơn. Hướng phát triển tiếp theo là tích hợp thuật toán vào bước tiền xử lý của mô hình học sâu để cải thiện hiệu quả xử lý ảnh đối kháng, từ đó nâng cao độ chính xác và tính an toàn cho mô hình học máy.

TÀI LIỆU THAM KHẢO/ REFERENCES

[1] L. Li, "Application of deep learning in image recognition," *Journal of Physics: Conference Series,* vol. 1693, no. 1, 2020, Art. no. 012128.

[2] N. Xu, "The application of deep learning in image processing is studied based on the reel neural network model," *Journal of Physics: Conference Series*, vol. 1881, no. 3, 2021, Art. no. 032096.

[3] J. Yang, Y. Sheng, Y. Zhang, W. Jiang, and L. Yang., "On-device unsupervised image segmentation," *2023 60th ACM/IEEE Design Automation Conference (DAC)*, IEEE, 2023, pp.1-6.

[4] J. Ma, P. Liang, W. Yu, C. Chen, X. Guo, J. Wu, and J. Jiang, "Infrared and visible image fusion via detail preserving adversarial learning," *Information Fusion*, vol. 54, pp. 85-98, 2020.
[5] Y. Shi, Y. Han, Q. Zhang, and X. Kuang, "Adaptive iterative attack towards explainable adversarial robustness," *Pattern recognition*, vol. 105, 2020, Art. no. 107309.
[6] Y. Xiao, C. M. Pun, and B. Liu, "Fooling deep neural detection networks with adaptive object-oriented adversarial perturbation," *Pattern Recognition*, vol. 115, 2021, Art. no. 107903.
[7] M. O. K. Mendonça, J. Maroto, P. Frossard, and P. S. R. Diniz, "Adversarial training with informed data selection," in *2022 30th European Signal Processing Conference (EUSIPCO)*, IEEE, 2022, pp. 608-612.
[8] E. C. Yeats, Y. Chen, and H. Li, "Improving gradient regularization using complex-valued neural networks," in *International Conference on Machine Learning*, 2021, pp. 11953-11963.
[9] Z. Liu, Q. Liu, T. Liu, N. Xu, X. Lin, Y. Wang, and W. Wen, "Feature distillation: DNN-oriented jpeg compression against adversarial examples," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2019, pp. 860-868.
[10] X. Jia, X. Wei, X. Cao, and H. Foroosh, "Comdefend: An efficient image compression model to defend adversarial examples," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 6084-6092.
[11] F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, "Ensemble adversarial training: Attacks and defenses," *arXiv preprint arXiv:1705.07204*, 2017.
[12] C. Xie, Y. Wu, L. Maaten, A. L. Yuille, and K. He, "Feature denoising for improving adversarial robustness," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 501-509.
[13] J. Chen, X. Zhang, R. Zhang, C. Wang, and L. Liu, "De-pois: An attack-agnostic defense against data poisoning attacks," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 3412-3425, 2021.
[14] Y. Bai, Y. Feng, Y. Wang, T. Dai, S. T. Xia, and Y. Jiang, "Hilbert-based generative defense for adversarial examples," in *Proceedings of the IEEE/CVF International conference on computer vision*, 2019, pp. 4784-4793.
[15] A. Shukla, P. Turaga, and S. Anand, "Gracias: Grassmannian of corrupted images for adversarial security," *arXiv preprint arXiv:2005.02936*, 2020.
[16] C. Guo, M. Rana, M. Cisse, and L V. D. Maaten, "Countering adversarial images using input transformations," *arXiv preprint arXiv:1711.00117*, 2017.
[17] P. H. Truong, C. T. Nguyen, N. M. Pham, D. T. Pham, and T. L. Bui, "A novel Hybrid CIFAR-10 dataset for Adversarial training to enhance the Robustness of Deep learning models," in *The XXVII National Conference "Some Selected Issues on Information and Communication Technology"*, 2024, pp. 27-32.
[18] I. J. Goodfellow, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.
[19] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Artificial intelligence safety and security*, Chapman and Hall/CRC, 2018, pp. 99-112.
[20] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li, "Boosting adversarial attacks with momentum," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 9185-9193.
[21] F. Nesti, A. Biondi, and G. Buttazzo, "Detecting adversarial examples by input transformations, defense perturbations, and voting," *IEEE transactions on neural networks and learning systems*, vol. 34, no. 3, pp. 1329-1341, 2021.
[22] W. Zhang, "Generating adversarial examples in one shot with image-to-image translation gan," *IEEE Access*, vol. 7, pp. 151103-151119, 2019.
[23] J. J. Bird and A. Lotfi, "Cifake: Image classification and explainable identification of ai-generated synthetic images," *IEEE Access*, vol. 12, pp. 15642-15650, 2024.
[24] F. Liao, M. Liang, Y. Dong, T. Pang, X. Hu, and J. Zhu, "Defense against adversarial attacks using high-level representation guided denoiser," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1778-1787.
[25] Y. Song, T. Kim, S. Nowozin, S. Ermon, and N. Kushman, "Pixeldefend: Leveraging generative models to understand and defend against adversarial examples," *arXiv preprint arXiv:1710.10766*, 2017.
[26] P. Samangouei, "Defense-gan: protecting classifiers against adversarial attacks using generative models," *arXiv preprint arXiv:1805.06605*, 2018.
[27] Y. Abouelnaga, O. S. Ali, H. Rady, and M. Moustafa, "Cifar-10: Knn-based ensemble of classifiers," in *2016 International Conference on Computational Science and Computational Intelligence (CSCI)*, IEEE, 2016, pp. 1192-1195.
[28] D. T. Pham, C. T. Nguyen, P. H. Truong, and N. H. Nguyen, "Automated generation of adaptive perturbed images based on GAN for motivated adversaries on deep learning models," in *Proceedings of the 12th International Symposium on Information and Communication Technology*, 2023, pp. 808-815.