THÔNG TIN CHUNG CỦA NHÓM

- Link YouTube video của báo cáo (tối đa 5 phút):

- Link slides: https://github.com/nhanhuynh123/CS519.P11/blob/main/slide.pdf

<table>
<tr>
<td>

- Họ và Tên: Huỳnh Chí Nhân
- MSSV: 22520996

</td>
<td>

- Lớp: CS519.P11
- Tự đánh giá (điểm tổng kết môn): 9.5/10
- Số buổi vắng: 0
- Số câu hỏi QT cá nhân: 6
- Số câu hỏi QT của cả nhóm: 6
- Link Github: https://github.com/nhanhuynh123/CS519.P11/
- Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm:
    - Lên ý tưởng
    - Viết Proposal
    - Làm Slide

</td>
</tr>
<tr>
<td>

- Họ tên: Nguyễn Hồ Nam
- MSSV: 22520915

</td>
<td>

- Lớp: CS519.P11
- Tự đánh giá (điểm tổng kết môn): 7.5/10

</td>
</tr>
</table>

	• Số buổi vắng: 0
	• Số câu hỏi QT cá nhân: 6
	• Số câu hỏi QT của cả nhóm: 6
	• Link Github: https://github.com/mynameuit/CS519.P11/
	• Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm:
	  ○ Làm poster
	  ○ Làm video YouTube

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

BẢO VỆ HÌNH ẢNH KHỎI MÔ HÌNH MÁY HỌC BẰNG MIẾNG VÁ ĐỐI KHÁNG NGỤY TRANG TỰ NHIÊN.

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

PROTECTING IMAGES FROM MACHINE LEARNING MODELS USING NATURALLY CAMOUFLAGED ADVERSARIAL PATCHES.

TÓM TẮT *(Tối đa 400 từ)*

Deepfake, sự giao thoa giữa các mô hình học sâu trong thị giác máy tính và các mô hình tạo sinh (generative models), ngày càng phổ biến nhưng cũng mang đến nhiều thách thức lớn về bảo mật thông tin cá nhân. Các dữ liệu hình ảnh công khai trên không gian mạng thường trở thành nguồn khai thác chính cho Deepfake, đặt ra nguy cơ bị lợi dụng cho các mục đích phi pháp như giả mạo danh tính hoặc chiếm đoạt thông tin.

Đáng chú ý, sự nhạy cảm của các mô hình phát hiện và nhận diện đối với các thay đổi nhỏ trong dữ liệu đầu vào đã mở ra hướng nghiên cứu sử dụng các tấn công đối kháng (adversarial attacks) để bảo vệ dữ liệu hình ảnh. Trong bài báo này, chúng tôi phát triển một phương pháp mới dựa trên Adversarial Patches (một loại tấn công đối kháng) kết hợp với kỹ thuật Style Transfer. Phương pháp đề xuất không chỉ tập trung vào việc gây nhiễu dự đoán của các mô hình nhận diện mà còn chú trọng tính tự nhiên và thẩm mỹ của hình ảnh để giảm thiểu khả năng bị loại bỏ có chủ đích. Bằng cách thay đổi phong cách và nội dung của miếng vá dựa trên hình ảnh chỉ định, phương pháp này đảm bảo tính nguỵ trang hiệu quả, tăng độ tương đồng giữa ảnh gốc và các ví dụ đối kháng (Adversarial Examples).

Nghiên cứu kỳ vọng đóng góp một cách tiếp cận mới, kết hợp tính thẩm mỹ và hiệu quả bảo mật, nhằm đối phó với các nguy cơ tiềm tàng từ các hệ thống trí tuệ nhân tạo hiện đại

GIỚI THIỆU *(Tối đa 1 trang A4)*

Sự phát triển mạnh mẽ của mạng nơ-ron sâu (DNN), đặc biệt là Convolutional Neural Networks (CNN), đã thúc đẩy những tiến bộ vượt bậc trong các lĩnh vực như thị giác máy tính và xử lý ngôn ngữ. Một ứng dụng nổi bật của DNN là công nghệ Deepfake, cho phép tái

tạo hoặc thay đổi nội dung hình ảnh và video với độ chính xác cao. Mặc dù Deepfake có tiềm năng ứng dụng lớn trong giải trí và sáng tạo nội dung, công nghệ này cũng gây ra những lo ngại nghiêm trọng về bảo mật thông tin và quyền riêng tư. Hình ảnh cá nhân trên không gian mạng có thể bị khai thác để tạo ra các nội dung giả mạo, phục vụ cho các mục đích như lừa đảo, giả danh, hoặc phá hoại uy tín [1].

Một trong những giải pháp để bảo vệ dữ liệu hình ảnh cá nhân khỏi sự khai thác của các hệ thống học sâu như Deepfake là tận dụng điểm yếu của các mô hình DNN. Các nghiên cứu trước đây [2] đã chỉ ra rằng việc thêm các nhiễu đối kháng (adversarial noise) vào dữ liệu đầu vào có thể làm giảm đáng kể độ chính xác của các mô hình phát hiện và nhận diện, qua đó mở ra một hướng tiếp cận mới trong bảo mật dữ liệu.

Trong bài báo này, chúng tôi tập trung vào Digital Attacks, một loại Adversarial Attack tác động trực tiếp lên hình ảnh đầu vào nhằm gây nhiễu kết quả dự đoán của mô hình mục tiêu[4]. Một phương pháp phổ biến trong nhóm này là Adversarial Patches, nơi các miếng vá gây nhiễu được gắn lên hình ảnh để làm rối các hệ thống nhận diện, như YOLO và Faster-RCNN [3]. Tuy nhiên, các Adversarial Patches hiện tại thường có kích thước lớn, màu sắc và nội dung khác biệt rõ rệt so với ảnh gốc, làm giảm hiệu quả ngụy trang và dễ bị loại bỏ. Nhằm giải quyết vấn đề này, chúng tôi đề xuất áp dụng kỹ thuật Style Transfer để ngụy trang các Adversarial Patches. Kỹ thuật này không chỉ đảm bảo tính tự nhiên và sự tương đồng giữa hình ảnh gốc và các Adversarial Examples [4], mà còn nâng cao hiệu quả bảo vệ dữ liệu hình ảnh cá nhân trước các hệ thống học sâu như Deepfake.

Tóm lại, mục tiêu của nghiên cứu là đánh giá hiệu quả của Camouflaged Adversarial Patches trong việc bảo vệ dữ liệu hình ảnh khỏi các hệ thống nhận diện và phát hiện, đồng thời đề xuất hướng ứng dụng phương pháp này vào việc ngăn chặn các rủi ro bảo mật từ công nghệ Deepfake.

MỤC TIÊU *(Viết trong vòng 3 mục tiêu)*

1. Thử nghiệm và đánh giá hiệu quả của Adversarial Patches theo định hướng Digital Attack trên mô hình YOLO, với mục tiêu đạt được Attack Success Rate (ASR) tối thiểu 80%.
2. Cải thiện tính tự nhiên của Adversarial Patches bằng kỹ thuật Style Transfer, đồng thời duy trì ASR ổn định.

3. Thiết kế và đóng gói mô hình Camouflaged Adversarial Patches dưới dạng module, hướng đến ứng dụng trong các hệ thống bảo mật dữ liệu ảnh cá nhân và tổ chức.

NỘI DUNG VÀ PHƯƠNG PHÁP

I. Nội dung

1. Tìm hiểu về tấn công đối kháng

Adversarial Attacks được chia thành hai hướng chính: Digital Attack và Real World Attack.

- Digital Attack: Nhiễu được thêm trực tiếp vào ảnh đầu vào, giúp nghiên cứu và đánh giá hiệu năng của các phương pháp tấn công trong môi trường kiểm soát. Digital Attack được chọn trong nghiên cứu này vì phù hợp hơn với bối cảnh ảnh được sử dụng trong các hệ thống Deepfake, thường được lấy từ các nền tảng truyền thông (media) hoặc không gian mạng, thay vì môi trường thực.
- Real World Attack: Nhiễu được tích hợp vào vật thể 3D, chịu ảnh hưởng của môi trường thực như ánh sáng và góc máy.

2. Cài đặt mô hình Adversarial Patch

Phương pháp nghiên cứu:

- Để tối ưu hóa việc áp dụng Adversarial Patches, chúng tôi sử dụng kỹ thuật Grad-CAM++ [5] để tạo ra Saliency Map. Grad-CAM++ là một phương pháp trực quan hóa, làm nổi bật các khu vực trong ảnh đầu vào quan trọng nhất đối với dự đoán của mô hình, bằng cách sử dụng gradient từ lớp đầu ra đến các lớp convolution cuối cùng. Từ Saliency Map, một Mask được tạo ra để xác định vị trí Adversarial Patch, kích thước của Patch được giới hạn trong một Threshold.
- Miếng vá Adversarial Patch được gắn lên khu vực đã xác định trong ảnh đầu vào dựa trên Mask, tạo ra một Adversarial Example.
- Mô hình được tối ưu bằng hàm: L = Detection Loss. Hàm mất mát sẽ điều chỉnh miếng vá sao cho: bounding box bị lệch, class bị nhận diện sai hoặc không nhận diện được.

Kết quả dự kiến:

- Tạo thành công mô hình Adversarial Patches.

3. Tiến hành nguỵ trang patch sử dụng Style Transfer

Phương pháp nghiên cứu:

- Áp dụng kỹ thuật Style Transfer, một phương pháp tách biệt nội dung và phong cách từ các hình ảnh, để tạo ra miếng vá nguỵ trang tự nhiên. Quá trình tối ưu hóa là quá trình tối thiểu hàm L = Content Loss + Style Loss, trong đó:
  - *Content Loss*: Hàm loss này được tính dựa trên sự khác biệt giữa nội dung (cấu trúc, đặc trưng) của Content Image (phần ảnh gốc bị Patches phủ lên) và Patch.
  - *Style Loss:* Tái tạo phong cách (màu sắc, họa tiết) từ Style Image bằng cách tính toán sự khác biệt giữa ma trận Gram của đặc trưng trích xuất từ Style Image và ảnh đầu ra.

Kết quả dự kiến:

- Tạo thành công mô hình Camouflaged Adversarial Patches.
- Quan sát được sự tương đồng giữa ảnh gốc và Adversarial Examples.

4. Thử nghiệm, đánh giá hiệu năng của Camouflaged Adversarial Patches

Phương pháp nghiên cứu:

- Mô hình được huấn luyện trên các bộ dữ liệu CelebA. Mô hình mục tiêu để thử nghiệm độ hiệu quả của Camouflaged Adversarial Patches là YOLO. Để đánh giá hiệu quả của phương pháp, bài báo đánh giá hai khía cạnh: Attack Performance và Imperceptibility.
- Với Attack Performance, chúng tôi sử dụng ASR (Attack Success Rate). Với Imperceptibility, sử dụng SSIM (Structure Similarity Index Measure) và Patch Size, trong đó SSIM thể hiện độ tương đồng giữa ảnh gốc và ảnh sau khi thêm Adversarial Patch.

Kết quả dự kiến:

- Thử nghiệm thành công mô hình Camouflaged Adversarial Attack trên mô hình YOLO.
- Phát sinh được bảng số liệu đánh giá các tiêu chí ASR, SSIM.

KẾT QUẢ MONG ĐỢI

1. Huấn luyện thành công mô hình Adversarial Patch, ASR trung bình đạt trên 80% khi tấn công mô hình YOLO.

2. Phát sinh thành công mô hình Camouflaged Adversarial Patches. Thử nghiệm tấn công mô hình YOLO và đạt SSIM trên 0.8, ASR trên 70%

3. Mô hình ổn định và được đóng gói thành module.

TÀI LIỆU THAM KHẢO *(Định dạng DBLP)*

[1]. Gan Pei, Jiangning Zhang, Menghan Hu, Guangtao Zhai, Chengjie Wang, Zhenyu Zhang, Jian Yang, Chunhua Shen and Dacheng Tao, "Deepfake Generation and Detection: A Benchmark and Survey," CoRR, vol. abs/2403.17881, 2024.

[2]. Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras and Adrian Vladu, "Towards Deep Learning Models Resistant to Adversarial Attacks," CoRR, vol. abs/1706.06083, 2017.

[3]. Xin Liu, Huanrui Yang, Ziwei Liu, Linghao Song, Hai Li and Yiran Chen, "DPatch: An Adversarial Patch Attack on Object Detectors," in AAAI, 2019.

[4]. Ranjie Duan, Xingjun Ma, Yisen Wang, James Bailey, A. Kai Qin and Yun Yang, "Adversarial Camouflage: Hiding Physical-World Attacks With Natural Styles," in IEEE/CVF, 2020, pp. 997-1005.

[5]. Chattopadhyay, Aditya and Sarkar, Anirban and Howlader, Prantik and Balasubramanian and Vineeth, "Grad-CAM++: Generalized Gradient-based Visual Explanations for Deep Convolutional Networks," in WACV, 2018, pp. 839-847.