



THÔNG TIN CHUNG CỦA NHÓM

- Link YouTube video của báo cáo: <https://youtu.be/5ghcLJypVCc>
- Link slides: <https://github.com/hatrontai/CS519.O11/blob/main/Slide.pdf>

<ul style="list-style-type: none"> • Họ và Tên: Hà Trọng Tài • MSSV: 21520436 	<ul style="list-style-type: none"> • Lớp: CS519.O11 • Tự đánh giá (điểm tổng kết môn): 9/10 • Số buổi vắng: 2 • Số câu hỏi QT cá nhân: 5 • Số câu hỏi QT của cả nhóm: 15 • Link Github: https://github.com/hatrontai/CS519.O11 • Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm: <ul style="list-style-type: none"> ◦ Đưa ra ý tưởng cho bài toán ◦ Viết phần tóm tắt, giới thiệu, mục tiêu, kết quả mong đợi và tài liệu tham khảo ◦ Viết poster
<ul style="list-style-type: none"> • Họ và Tên: Phan Trường Trí • MSSV: 21520117 	<ul style="list-style-type: none"> • Lớp: CS519.O11 • Tự đánh giá (điểm tổng kết môn): 9/10 • Số buổi vắng: 2 • Số câu hỏi QT cá nhân: 5 • Số câu hỏi QT của cả nhóm: 15 • Link Github: https://github.com/triphan2k3/CS519.O11 • Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm: <ul style="list-style-type: none"> ◦ Đưa ra ý tưởng cho bài toán ◦ Viết phần giới thiệu, phương pháp ◦ Làm slide ◦ Làm video YouTube

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

TẤN CÔNG HỘP ĐEN DỰA TRÊN TẬP HỢP CÁC MÔ HÌNH DỰ ĐOÁN DÀY ĐẶC

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

ENSEMBLE-BASED BLACKBOX ATTACKS ON DENSE PREDICTION

TÓM TẮT

Hiện nay hầu hết các nghiên cứu đều hướng đến việc tăng hiệu suất cho mô hình về ở các khía cạnh như độ chính xác, thời gian chạy hoặc lượng bộ nhớ sử dụng mà thường bỏ qua việc đánh giá độ an toàn của hình trong khi an toàn cũng là một yếu tố quan trọng đối với một mô hình. Việc đánh giá độ an toàn của một mô hình có thể được thực hiện thông qua quá trình tấn công mô hình đó. Hiện nay đã có nhiều nghiên cứu về tấn công mô hình tuy nhiên phần lớn các nghiên cứu đó hướng đến việc tấn công vào các mô hình mà ta đã biết trước kiến trúc (whitebox). Nhưng trên thực tế không phải lúc nào ta cũng biết trước kiến trúc của mô hình cần tấn công. Và thông thường thì việc tấn công có mục tiêu thường khó thực hiện hơn tấn công không mục tiêu. Do đó trong nghiên cứu này, chúng tôi đề xuất một phương pháp tấn công có mục tiêu vào các mô hình hộp đen (blackbox), tức là kiến trúc mô hình không được biết trước của bài toán dự đoán dày đặc (dense prediction) để từ đó đánh giá mức độ an toàn của các mô hình đó. Về phương pháp, chúng tôi sẽ chọn ra một tập các mô hình đã biết (whitebox) và điều chỉnh các trọng số cho từng mô hình riêng lẻ để tạo ra lớp nhiễu tối ưu cho việc tấn công vào mô hình mục tiêu. Độ hiệu quả của phương pháp sẽ được đánh giá thông qua các thực nghiệm được thực hiện trên các mô hình thuộc lớp bài toán object detection và segmentation.

GIỚI THIỆU

Hiện nay, có rất nhiều công việc liên quan đến phân tích đã được tự động hóa được xử lý bởi trí tuệ nhân tạo (AI). Và cũng có nhiều đối tượng tận dụng sơ hở của các mô hình AI để từ đó sử dụng các kỹ thuật nhằm đánh lừa mô hình, tạo ra các cuộc tấn công nhắm vào sự chính xác của mô hình. Hiện nhiên, nếu một mô hình có ảnh hưởng lớn hay được dùng để giải quyết một công việc quan trọng thì việc nó bị tấn công sẽ tạo ra thiệt hại nặng nề cho chủ sử hữu. Và ta cũng có thể thấy rằng việc đảm bảo an toàn cho một mô hình AI là rất quan trọng nhất là các mô hình dễ bị "tổn thương", nhưng hiện nay chủ đề này chưa nhận được quá nhiều sự quan tâm.

Mà như đã biết thì các mô hình thị giác máy tính rất dễ bị tổn thương bởi các mẫu đối nghịch được tạo ra một cách có chủ đích. Việc tạo các mẫu đối nghịch để tấn công model whitebox thường đơn giản hơn khi tấn công các mô hình blackbox vì ta đã được biết toàn bộ kiến trúc. Đối với trường hợp mô hình là blackbox, tức là ta không biết thông tin gì về kiến trúc, việc tạo ra các mẫu đối nghịch gặp rất nhiều thử thách. Các nghiên cứu gần đây về tấn công hộp đen hầu hết đều là tấn công trên các mô hình phân loại và tập trung vào tấn công không mục tiêu. Còn đối với tấn công hộp đen cho các mô hình dự đoán dày đặc như object detection và segmentation thì sự số lượng công trình còn khá hạn chế. Hơn nữa, các phương pháp đó chủ yếu dựa trên **transfer attacks** (tấn công mô hình hộp đen dựa trên việc tạo nhiễu từ một mô hình whitebox) và **query-based attack** (dựa trên feedback của các truy vấn). Nhưng các phương pháp như vậy thường có tỷ lệ tấn công thành công khá thấp, đặc biệt đối với các cuộc tấn công có mục tiêu.

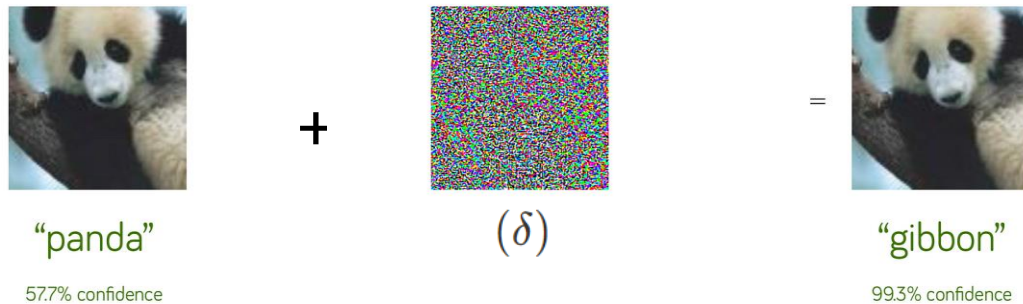
Trong nghiên cứu này chúng tôi sẽ tìm một phương pháp tấn công khác vào các mô hình hộp đen cho bài toán dự đoán dày đặc bằng cách kết hợp hai phương pháp **transfer-based attack** và **query-based attack** có sử dụng nhiều mô hình whitebox, và hiệu suất sẽ được đánh giá thông qua các mô hình của hai bài toán object detection và segmentation.

Ý tưởng được lấy từ 3 quan sát chính: (1) các cuộc tấn công dựa được tạo bởi một mô hình whitebox đơn có tỷ lệ thành công thấp; (2) các cuộc tấn công được tạo bởi nhiều mô hình whitebox sẽ thành

công nếu sự đóng góp từ các mô hình được chuẩn hóa đúng cách; (3) kết quả tấn công trên một mô hình có thể được cải thiện bằng cách điều chỉnh mức độ đóng góp của từng mô hình whitebox. Xác định bài toán: Tạo ra một lớp nhiễu phủ lên ảnh đầu vào nhằm mục đích tấn công, đánh lừa các mô hình của bài toán dự đoán dày đặc

Input: một tấm ảnh

Output: ảnh sau khi thêm nhiễu



MỤC TIÊU

- Tìm ra phương pháp tấn công hộp đen có sự kết hợp giữa transfer-based attack và query-based attack.
- Kết quả đánh giá mô hình đề xuất (thông qua các mô hình của hai bài toán object detection và segmentation) khả quan khi so sánh với các phương pháp hiện có.
- Tìm ra nguyên nhân các mô hình bị đánh lừa bởi các cuộc tấn công, từ đó đưa ra các ý tưởng để các mô hình có thể tránh bị tấn công.

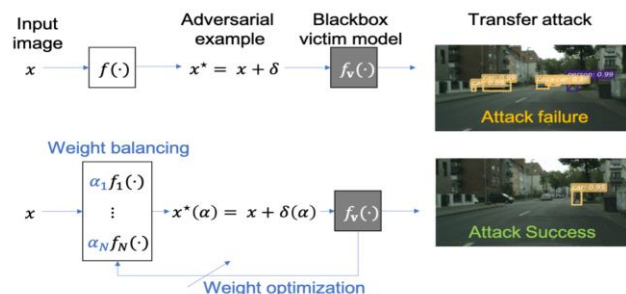
NỘI DUNG VÀ PHƯƠNG PHÁP

Nội dung 1: Tìm hiểu bài toán

Phương pháp: Tìm hiểu về các nghiên cứu liên quan đến bài toán tấn công hộp đen, đặc biệt là các phương pháp về **transfer-based attack** và **query-based attack** từ các nguồn đáng tin cậy như các hội nghị, tạp chí uy tín.

Nội dung 2: Xây dựng mô hình

Phương pháp: Xây dựng mô hình Ensemble-based attack, có ý tưởng chính là sự kết hợp giữa transfer-based attack và query-based attack.



- Transfer based attack: phương pháp tấn công bằng việc tạo nhiễu từ một model whitebox biết trước sau đó tấn công vào victim model. Phương pháp này khá dễ thực hiện nhưng đổi lại tỷ lệ tấn công thành công khá là thấp.

- Query based attack: phương pháp sử dụng query thả vào victim model để tìm ra nhiều có thể tấn công thành công vào victim model. Tuy nhiên nó yêu cầu lượng lớn query (có thể hàng trăm đến hàng ngàn).
- Ensemble based attack: đây cũng là phương pháp đề xuất của chúng tôi, được lấy ý tưởng từ Transfer based attack chúng tôi tiến hành tạo nhiều bằng cách sử dụng một tập whitebox model nhằm tăng tính bao phủ khi tấn công vào victim model. Bù lại thì số lượng tính toán sẽ tăng lên do sử dụng nhiều model whitebox hơn.

Nội dung 3: Đánh giá hiệu quả tấn công của mô hình

Phương pháp: Thực nghiệm tấn công trên các models object detection sử dụng dataset COCO và các models segmentation sử dụng dataset Pascal VOC 2012. Đánh giá phương pháp dựa trên tỷ lệ tấn công thành công vào victim model đồng thời đánh giá mức độ ổn định chống bị tấn công của victim model; so sánh kết quả thực nghiệm với các phương pháp khác. Từ đó rút ra nhận xét về ưu điểm và khuyết điểm của mô hình.

Nội dung 4: Đề xuất hướng đi chống lại các cuộc tấn công

Phương pháp: Phân tích kết quả của mô hình, để rút ra đặc điểm chung của các điểm dữ liệu tấn công thành công. Từ đó rút ra kết luận về nguyên nhân các mô hình dự đoán dày đặc bị lừa để đưa ra hướng giải quyết, hạn chế nguy cơ bị tấn công bởi các attack models.

Nội dung 5: Viết báo cáo cho nghiên cứu

Phương pháp: Viết lại kết quả nghiên cứu thành một bài báo khoa học

KẾT QUẢ MONG ĐỢI

- Mô hình tấn công hộp đen đề xuất có kết quả khả quan khi cho tấn công vào các models dự đoán dày đặc.
- Một bài báo khoa học viết về nghiên cứu được công bố ở hội nghị khoa học.

TÀI LIỆU THAM KHẢO

- [1]. Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.
- [2]. Nicholas A. Lord, Romain Mueller, and Luca Bertinetto. Attacking deep networks with surrogate-based adversarial black-box methods is easy. In *International Conference on Learning Representations*, 2022.
- [3]. Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9627–9636, 2019.
- [4]. Chun-Chen Tu, Paishun Ting, Pin-Yu Chen, Sijia Liu, Huan Zhang, Jinfeng Yi, Cho-Jui Hsieh, and Shin-Ming Cheng. Autozoom: Autoencoder-based zeroth order optimization method for attacking black-box neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 742–749, 2019.
- [5]. Siyuan Liang, Longkang Li, Yanbo Fan, Xiaojun Jia, Jingzhi Li, Baoyuan Wu, and Xiaochun Cao. A large-scale multiple-objective method for black-box attack against object detection. In *European Conference on Computer Vision*, pages 619–636. Springer, 2022.

- [6]. Anurag Arnab, Ondrej Miksik, and Philip HS Torr. On the robustness of semantic segmentation models to adversarial attacks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 888–897, 2018.
- [7]. Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. *International Conference on Machine Learning*, pages 2137–2146, 2018.