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

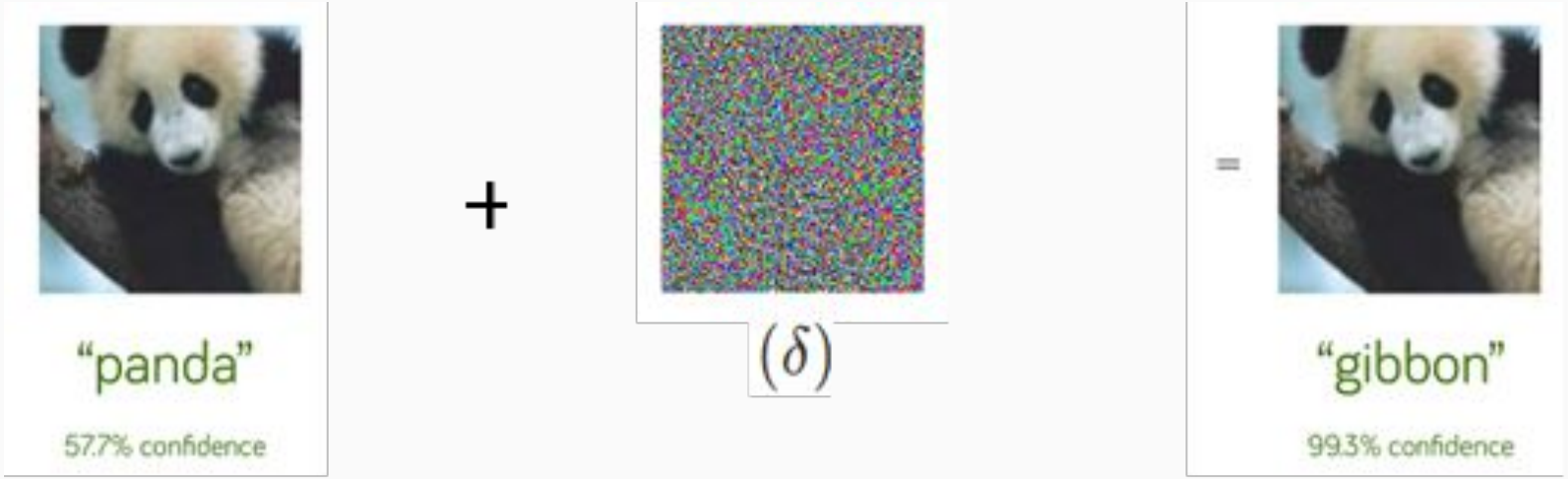
Phan Trường Trí - 21520117
Hà Trọng Tài - 21520436

Tóm tắt

- Họ và Tên: Hà Trọng Tài
- MSSV: 21520436

- Họ và Tên: Phan Trường Trí
- MSSV: 21520117

Giới thiệu



Giới thiệu

- Bảo vệ mô hình trước các cuộc tấn công là rất cần thiết.
- Chưa có nhiều nghiên cứu dành cho việc tấn công (cũng như bảo vệ) các mô hình.
- Đặc biệt là mô hình hộp đen thuộc lĩnh vực Thị giác máy tính.
- Các phương pháp hiện nay chủ yếu là **transfer-based attack** và **query-based attack.**
- Tỉ lệ tấn công thành công của các phương pháp đó khá thấp.
- Nghiên cứu này mong muốn tạo ra một phương pháp tấn công là sự kết hợp cả hai phương pháp trên với kì vọng tỉ lệ thành công khi tấn công sẽ cao hơn.

Mục tiêu

- Tìm ra phương pháp tấn công hộp đen có sự kết hợp giữa transfer-based attack và query-based attack
- Đánh giá mô hình đề xuất thông qua các mô hình của hai bài toán object detection và segmentation.
- Tìm ra nguyên nhân khiến các mô hình bị đánh lừa bởi các cuộc tấn công, từ đó đưa ra các ý tưởng để các mô hình hạn chế việc bị tấn công.

Nội dung và Phương pháp

Nội dung 1: Tìm hiểu các phương pháp hiện có **Phương pháp:** Thông qua các nguồn thông tin uy tín. Đặc biệt chú ý đến transfer-based attack và query-based attack	**Nội dung 2:** Xây dựng phương pháp tấn công **Phương pháp:** Xây dựng một phương pháp tấn công mới: kết hợp giữa transfer-based attack và query-based attack	**Nội dung 3:** Đánh giá mô hình tấn công **Phương pháp:** Đánh giá mô hình tấn công thông qua các mô hình và các bộ dữ liệu phổ biến. So sánh kết quả đó với các phương pháp tấn công khác

Nội dung và Phương pháp

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

Phương pháp: Phân tích kết quả thực nghiệm để tìm nguyên nhân khiến các mô hình bị tấn công, từ đó tìm cách hạn chế việc bị tấn công.

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 dự kiến

- Mô hình tấn công hộp đen đề xuất có hiệu suất khả quan (xấp xỉ các mô hình hiện tại) khi 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.

Tài liệu tham khảo

[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.