CẢI THIỆN ĐỘ BỀN VỮNG ĐỐI KHÁNG CỦA CÁC MÔ HÌNH PHÂN TÍCH ẢNH Y KHOA THÔNG QUA PHÒNG THỦ HỘP ĐEN KẾT HỢP TÁI CẤU TRÚC HÌNH ẢNH.

IMPROVING ADVERSARIAL ROBUSTNESS OF MEDICAL IMAGE ANALYSIS MODELS VIA BLACK-BOX DEFENSE INTEGRATING IMAGE RECONSTRUCTION.

Thông tin nhóm

Trần Nhật Khoa
2250691

Lý Nguyên Thùy Linh
22520766

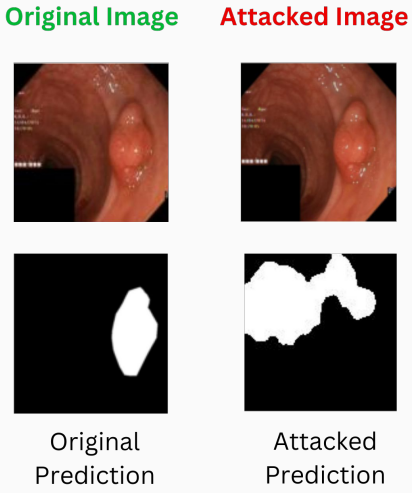
Lê Trần Quốc Khánh
2250638

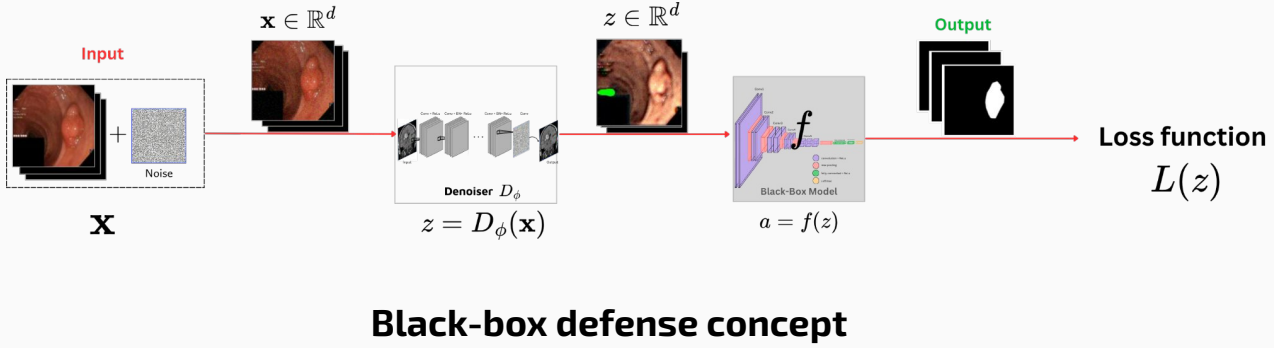
- Link Github của nhóm: github.com/khoa16122004/CS519.O21.KHTN
- Link YouTube video:

Giới thiệu

Adversarial Attack là một mối đe dọa lớn đối với các mô hình học máy, đặc biệt là trong lĩnh vực hình ảnh y tế sử dụng mạng học sâu (DNNs).

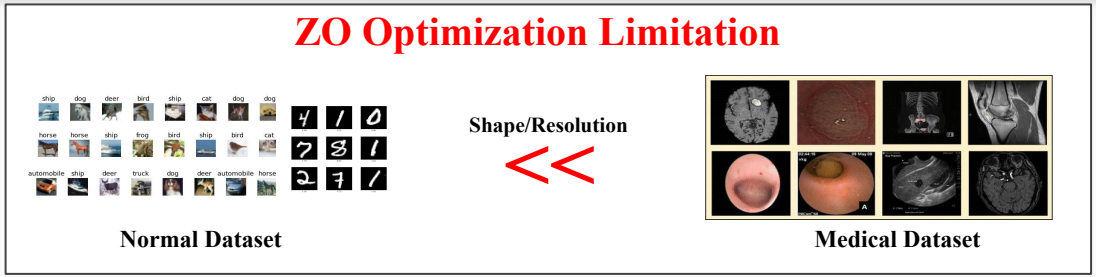
→ **Black-box Defense**: Một phương pháp sử dụng **Denoised Smoothing (DS)** kết hợp **Zeroth-order (ZO)** có thể giúp bảo vệ mô hình học sâu chỉ cần truy cập vào **đầu vào** và **đầu ra** của mô hình, **đặc biệt phù hợp với mô hình liên quan đến hình ảnh y khoa**.





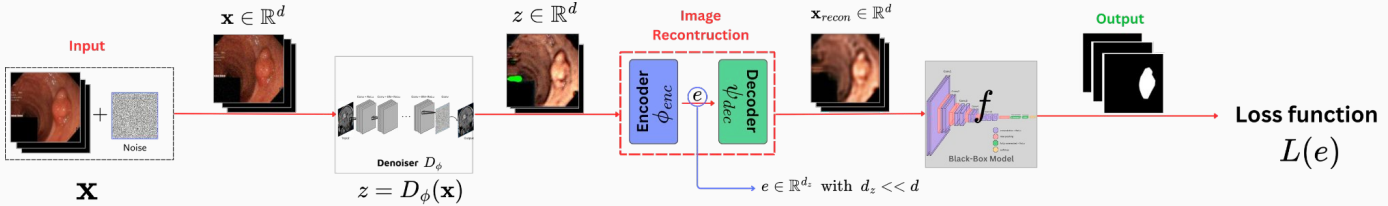
Black-box defense concept

Giới thiệu



→ nén hình ảnh hoặc vector đầu vào xuống không gian có số chiều thấp hơn.

Làm thế nào để có thể huấn luyện DS sử dụng ZO optimization xấp xỉ gradient của vector hình ảnh đã được giảm chiều mà vẫn có thể đưa vector đó vào mô hình y học ?



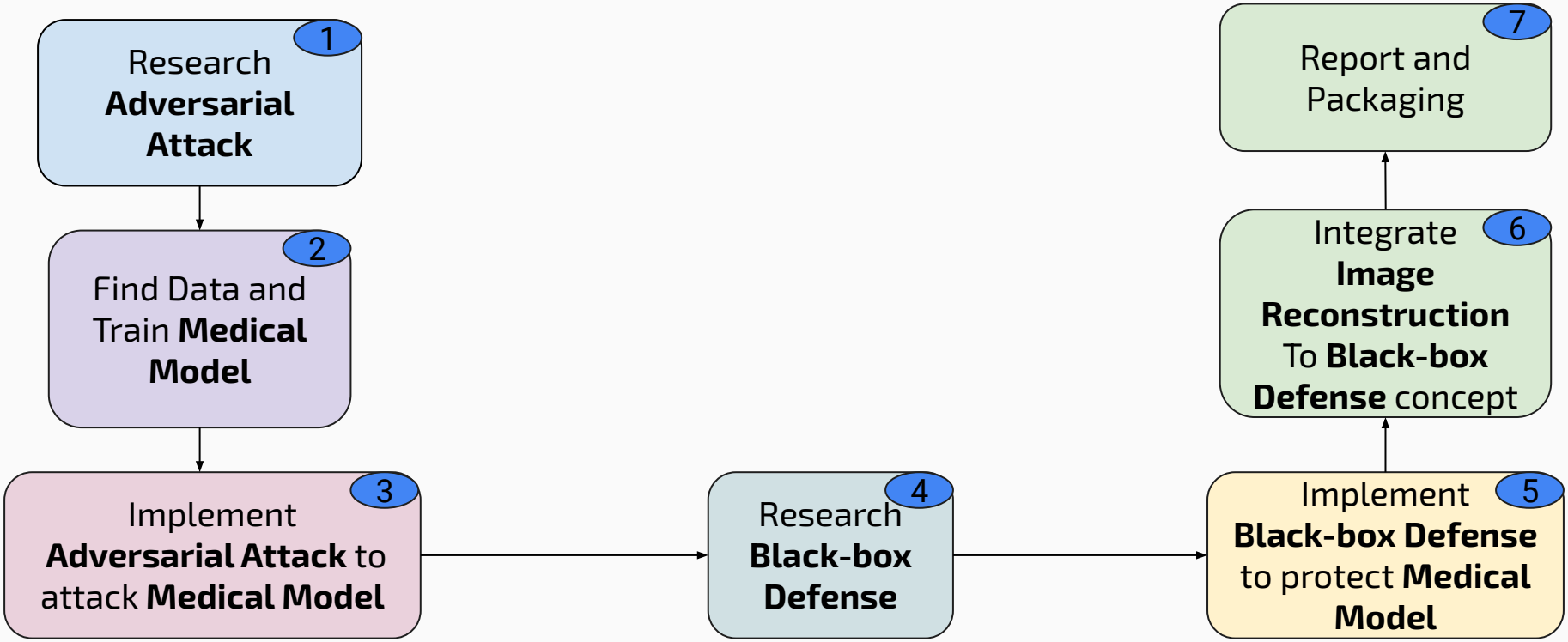
Pipeline: Black-box Defense integrating Image Reconstruction

→ Kết hợp các kỹ thuật **Image Reconstruction** vào Black-box Defense.

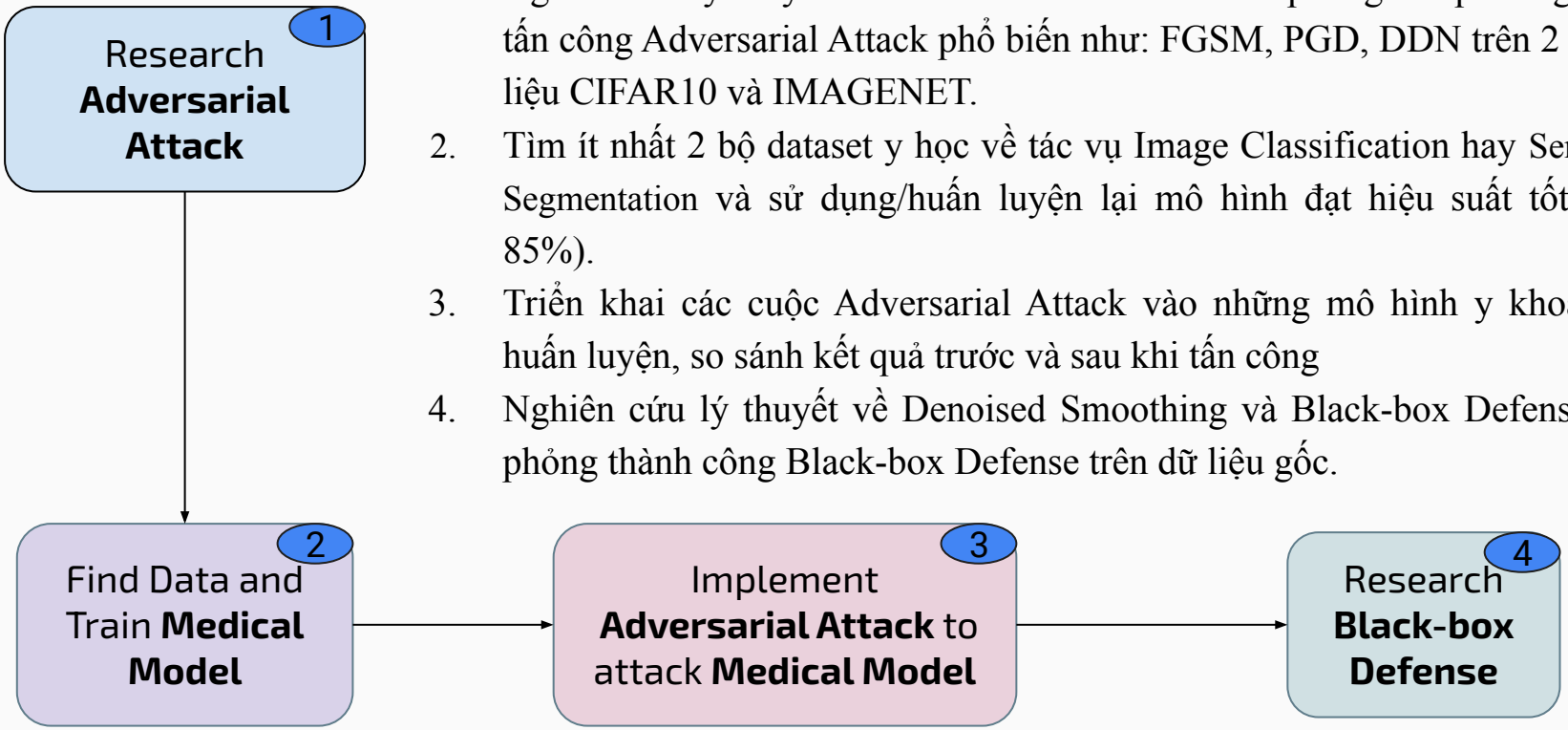
Mục tiêu

- **Giả lập Adversarial Attack** trên dữ liệu y khoa với các tác vụ thị giác máy tính sử dụng DNNs.
- **Cải thiện độ bền vững đối kháng**: Áp dụng Black-box defense kết hợp với Image Reconstruction trên các bộ dữ liệu y học để chống lại thành công sự ảnh hưởng từ Adversarial Attack.
- **Cung cấp kiến thức và đánh giá** bằng chứng minh toán học và thực nghiệm về phương pháp phòng thủ Black-box defense kết hợp Image Reconstruction.

Nội dung và Phương pháp



Nội dung và Phương pháp



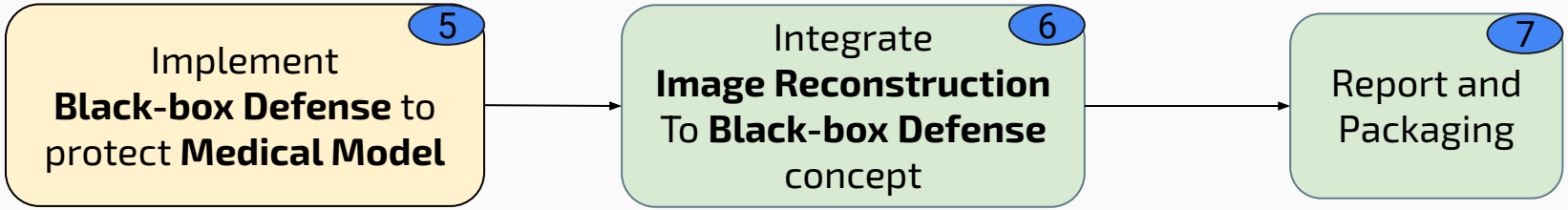
1. Nghiên cứu lý thuyết về Adversarial Attack và mô phỏng các phương pháp tấn công Adversarial Attack phổ biến như: FGSM, PGD, DDN trên 2 bộ dữ liệu CIFAR10 và IMAGENET.
2. Tìm ít nhất 2 bộ dataset y học về tác vụ Image Classification hay Semantic Segmentation và sử dụng/huấn luyện lại mô hình đạt hiệu suất tốt (trên 85%).
3. Triển khai các cuộc Adversarial Attack vào những mô hình y khoa vừa huấn luyện, so sánh kết quả trước và sau khi tấn công
4. Nghiên cứu lý thuyết về Denoised Smoothing và Black-box Defense, mô phỏng thành công Black-box Defense trên dữ liệu gốc.

Nội dung và Phương pháp

5. Thực hiện Black-Box Defense trên mô hình mục tiêu (**nội dung 3**) với hai tác vụ: Image Classification và Semantic Segmentation.

6. Tìm hiểu lý thuyết về các cấu trúc Image Reconstruction tiên tiến, đặc biệt là các mô hình có thể ứng dụng cho ảnh y học (PCA, CNN AutoEncoder, Transformer, …). Tiến hành thử nghiệm và tích hợp các cấu trúc Image Reconstruction khác nhau vào Black-Box Defense và so sánh kết quả.

7. Viết báo cáo chi tiết về nghiên cứu và đóng gói chương trình thực nghiệm.



Kết quả dự kiến

- Kết quả bảo vệ của Black-box Defense phải tốt hơn ban đầu khi mô hình bị tấn công trong hai tác vụ Image Classification và Semantic Segmentation với các bộ dữ liệu y học.
- Tích hợp thành công Image Reconstruction vào trong phương pháp Black-Box Defense.
- Có tối thiểu một phương pháp, cấu trúc cho ra kết quả tốt hơn các cấu trúc hiện tại về thời gian chạy nhưng độ chính xác có thể giảm trong khoảng chấp nhận được (5-8%).
- Bảng kết quả, so sánh và nhận ra ưu nhược điểm của những cấu trúc, phương pháp qua quá trình thực nghiệm.

Tài liệu tham khảo

[1]. I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” in ICLR, 2015.

[2]. H. Salman, M. Sun, G. Yang, A. Kapoor, and J. Z. Kolter, “Denoised smoothing: A provable defense for pretrained classifiers,” in NeurIPS, 2020.

[3]. Y. Zhang, Y. Yao, J. Jia, J. Yi, M. Hong, S. Chang, and S. Liu, “How to robustify black-box ML models? A zeroth-order optimization perspective,” in ICLR. OpenReview.net, 2022

[4]. S. Liu, B. Kailkhura, P. Chen, P. Ting, S. Chang, and L. Amini, “Zeroth- order stochastic variance reduction for nonconvex optimization,” CoRR, vol. abs/1805.10367, 2018

[5]. A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, “Towards deep learning models resistant to adversarial attacks,” in ICLR. OpenReview.net, 2018.

[6]. J. Rony, L. G. Hafemann, L. S. Oliveira, I. B. Ayed, R. Sabourin, and E. Granger, “Decoupling direction and norm for efficient gradient-based L2 adversarial attacks and defenses,” in CVPR, 2019, pp. 4322–4330.