


THÔNG TIN CHUNG CỦA NHÓM

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

<https://www.youtube.com/watch?v=xBcMlxgMs48>

- Link slides:

<https://github.com/Ly-Lynn/CS519.O21.KHTN/blob/main/slide.pdf>

<ul style="list-style-type: none">● Họ và Tên: Trần Nhật Khoa● MSSV: 22520691 	<ul style="list-style-type: none">● Lớp: CS519.O21.KHTN● Tự đánh giá (điểm tổng kết môn): 9.5/10● Số buổi vắng: 1● Số câu hỏi QT cá nhân: 5● Số câu hỏi QT của cả nhóm: 15● Link Github: https://github.com/khoa16122004/CS519.O21.KHTN● Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm:<ul style="list-style-type: none">○ Tìm hiểu đề tài, lên ý tưởng○ Viết phần Tóm tắt và Giới thiệu○ Góp ý phần Kết quả mong đợi○ Làm nội dung và thiết kế slide○ Chỉnh sửa poster
<ul style="list-style-type: none">● Họ và Tên: Lý Nguyên Thùy Linh● MSSV: 22520766	<ul style="list-style-type: none">● Lớp: CS519.O21.KHTN● Tự đánh giá (điểm tổng kết môn): 9.5/10● Số buổi vắng: 0● Số câu hỏi QT cá nhân: 5● Số câu hỏi QT của cả nhóm: 15● Link Github: https://github.com/Ly-Lynn/CS519.O21.KHTN



- Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm:
 - Đóng góp ý tưởng cho đề tài
 - Viết đề cương phần Nội dung và Phương pháp
 - Chỉnh sửa đề cương phần Tóm tắt và Giới thiệu
 - Thuyết trình video Youtube
 - Làm poster

- Họ và Tên: Lê Trần Quốc Khánh
- MSSV: 22520638



- Lớp: CS519.O21.KHTN
- Tự đánh giá (điểm tổng kết môn): 9.5/10
- Số buổi vắng: 0
- Số câu hỏi QT cá nhân: 5
- Số câu hỏi QT của cả nhóm: 15
- Link Github:
<https://github.com/LeeKhanhs/CS519.O21.KHTN/>
- Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm:
 - Đóng góp ý tưởng đề tài
 - Viết phần Giới thiệu và Kết quả mong đợi
 - Góp ý, chỉnh sửa phần Nội dung, phương pháp
 - Làm slide
 - Chỉnh sửa video

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI

CẢI THIẾN 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.

TÊN ĐỀ TÀI TIẾNG ANH

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

TÓM TẮT

Những mô hình học máy sử dụng mạng học sâu (Deep Neural Networks) đã được chứng minh rằng rất nhạy cảm trước *tấn công đối kháng (Adversarial Attack)* [1] - những cuộc tấn công thay đổi nhỏ dữ liệu đầu vào để đánh lừa mô hình. Việc tăng cường tính ổn định trước các cuộc tấn công này là rất quan trọng để đảm bảo độ tin cậy và an toàn của các mô hình học máy, đặc biệt trong các ứng dụng liên quan đến hình ảnh y tế sử dụng mạng học sâu.

Denoised Smoothing (DS) [2] thực hiện phương pháp phòng thủ bằng cách thêm một mạng khử nhiễu ngay trước mô hình mục tiêu đã được huấn luyện trước, giúp bảo vệ mô hình trước tấn công mà không cần thay đổi tham số bên trong mô hình. Tuy nhiên, việc huấn luyện mạng cần sử dụng các thuật toán *First-order (FO) optimization*, yêu cầu mô hình mục tiêu cho phép truy cập vào cấu trúc, tham số để tính toán đạo hàm (gradient) dựa vào lan truyền ngược (backpropagation). Việc truy cập như vậy thường bất khả thi trong các hệ thống y tế do vấn đề bảo mật. *Black-box Defense* [3] giải quyết hạn chế này bằng cách kết hợp DS và *Zeroth-order (ZO) optimization* [4] để xấp xỉ gradient của mô hình, giúp ta có thể huấn luyện được mạng khử nhiễu mà không cần phải biết chi tiết cấu trúc cũng như tham số của mô hình bị tấn công.

Tuy nhiên, dữ liệu hình ảnh y khoa thường có **độ phân giải và kích thước rất cao**, gây ra thách thức lớn về độ chính xác khi xấp xỉ gradient bằng *ZO Optimization*, một kỹ thuật chỉ khả thi khi không gian hay kích thước của vector nhỏ.

Do đó, nghiên cứu này đề xuất kết hợp các kỹ thuật **Image Reconstruction** tiên tiến, có tác dụng giảm chiều và tái cấu trúc hình ảnh, vào phương pháp *Black-box Defense* nhằm cải thiện độ chính xác của thuật toán *ZO Optimization* và tăng cường độ bền vững của các mô hình học sâu phân tích hình ảnh y khoa trước các cuộc tấn công đối kháng phổ biến.

GIỚI THIỆU

DNNs, đặt biệt là Convolutional Neural Networks (CNNs) đã rất thành công trong việc giải quyết các tác vụ liên quan đến hình ảnh y học. Tuy nhiên, những mô hình này thường thiếu độ bền vững, cụ thể các mô hình sử dụng mạng học sâu rất nhạy cảm với thay đổi của đầu vào. Các cuộc tấn công lợi dụng nhược điểm trên, được gọi là *Adversarial Attack*. Và đa phần các cuộc tấn công đối kháng thường không dễ thấy bằng mắt thường, gây ra rủi ro lớn trong các ứng dụng hình ảnh y tế sử dụng các mô hình mạng học sâu.

Huấn luyện đối kháng (Adversarial Training) [5] đã được chứng minh là một phương pháp phòng thủ hiệu quả bằng cách huấn luyện lại mô hình với cả dữ liệu gốc và dữ liệu đã bị tấn công. Tuy nhiên, phương pháp này tốn nhiều thời gian và tài nguyên, đặc biệt không phù hợp với các mô hình y tế do chi phí huấn luyện lại rất cao.

Do đó, Salman đã đề xuất DS [2], kỹ thuật sử dụng một mạng khử nhiễu được thêm vào trước mô hình mục tiêu. Bằng cách đóng băng tham số và sử dụng gradient của mô hình mục tiêu để thực hiện backpropagation, mạng khử nhiễu được huấn luyện mà không thay đổi bất kỳ thông số nào của mô hình gốc. Tuy nhiên, trong lĩnh vực y tế, các chủ sở hữu mô hình hoặc bệnh viện thường **không tiết lộ cấu trúc cũng như tham số** của mô hình do các vấn đề bảo mật. Vì vậy Zhang đề xuất *Black-box Defense* [3], huấn luyện mạng khử nhiễu bằng cách sử dụng *ZO Optimization* để xấp xỉ gradient của mô hình mục tiêu chỉ bằng đầu vào và đầu ra.

Tuy nhiên, nghiên cứu này chỉ tập trung vào **phân loại hình ảnh trên các bộ dữ liệu đơn giản** như MNIST hay CIFAR (hình ảnh có kích thước 32x32x3 hoặc 64x64). Trong khi đó, các bộ dữ liệu hình ảnh y khoa thường có **độ phân giải và kích thước lớn** (1024x1024x3), khiến cho việc xấp xỉ gradient bằng *ZO Optimization* một cách chính xác là không khả thi do ta **cần phải phát sinh rất nhiều vector trong không gian đó**, gây ra rào cản lớn về tài nguyên.

Giải pháp tự nhiên là nén hình ảnh hoặc vector đầu vào xuống không gian có số chiều

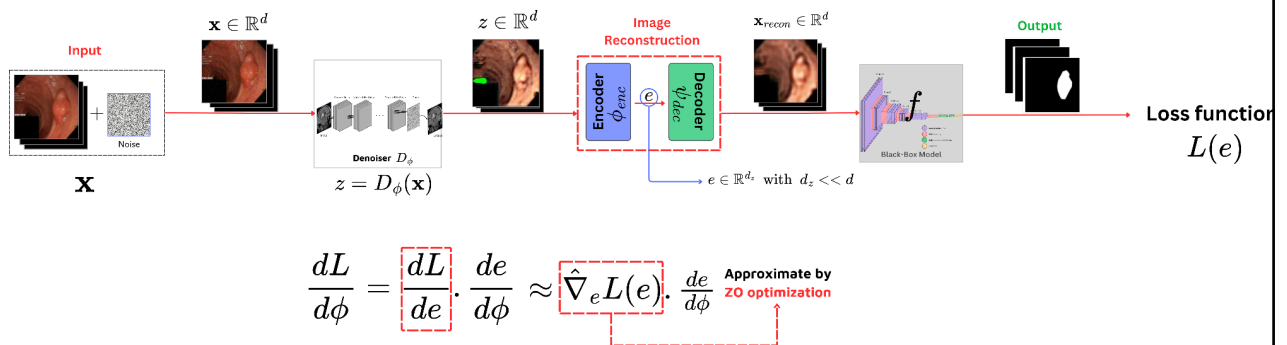
thấp hơn, giúp cải thiện hiệu quả của *ZO Optimization*. Tuy nhiên, vector được nén không thể trực tiếp đưa vào mô hình ban đầu do không tương thích kích thước. Vậy câu hỏi đặt ra là:

“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?”

Nghiên cứu về Black-Box Defense của Zhang sử dụng **AutoEncoder** để giải quyết vấn đề này, cụ thể sau khi được nén vào không gian có số chiều thấp, ta sẽ tái cấu trúc lại hình ảnh vào trong không gian ban đầu trước khi đưa vào mô hình, khi đó ta chỉ cần **xấp xỉ gradient của mô hình theo vector có số chiều thấp**, giải quyết được câu hỏi được đặt ra ban đầu.

Nhưng các cấu trúc này khá **đơn giản** và chỉ hiệu quả trên những bộ dữ liệu quen thuộc như MNIST, CIFAR10 - những hình ảnh có kích thước rất nhỏ và dễ phục hồi từ không gian vector có số chiều thấp. Đối với hình ảnh y học có độ phân giải và số chiều cao, phương pháp này có thể xấp xỉ tốt gradient của mô hình, giúp ta huấn luyện được mạng khử nhiễu, tuy nhiên lại khiến hình ảnh bị giảm chất lượng so với ban đầu, dẫn đến khi đưa vào mô hình lại không dự đoán tốt được như ta mong muốn.

Vì vậy, chúng tôi đề xuất áp dụng các kỹ thuật **Image Reconstruction** tiên tiến, từ một vector có kích thước nhỏ có thể tái tạo tốt lại ảnh gốc. Ngoài ra có thể sử dụng những kỹ thuật chuyên biệt dành riêng cho phục hồi ảnh y khoa, để có thể thành công thực hiện *Black-Box Defense* trên các mô hình phân tích ảnh y khoa.



Hình 1: Minh họa pipeline của Black-box defense kết hợp Image Reconstruction cho bài toán Polyp Segmentation.

Đóng góp chính và tính mới của nghiên cứu này bao gồm:

1. Ứng dụng *Black-Box Defense* để cải thiện độ bền vững trước các cuộc tấn công đối kháng phổ biến đối các mô hình phân tích hình ảnh y khoa, mở rộng với tác vụ khác

không chỉ riêng với phân loại hình ảnh.

2. Kết hợp và thử nghiệm các kỹ thuật *Image Reconstruction* tiên tiến để cải thiện hiệu quả của thuật toán *ZO Optimization* trong phương pháp *Black-Box Defense* trên các hình ảnh y học có độ phân giải lớn.

MỤC TIÊU

Chúng tôi đề ra mục tiêu của nghiên cứu này gồm:

- Giả lập thành công các thuật toán Adversarial Attack khác nhau trên những bộ dữ liệu y học trong các tác vụ thị giác máy tính sử dụng mạng học sâu (Image Classification, Semantic Segmentation): thử nghiệm các loại tấn công khác nhau và thành công làm giảm hiệu suất của mô hình.
- Cải thiện độ bền vững đối kháng: áp dụng Black-box defense trên những bộ dữ liệu đó, đồng thời tích hợp các kỹ thuật Image Reconstruction.
- Cung cấp các kiến thức, đánh giá bằng chứng minh toán học và thực nghiệm của phương pháp phòng thủ Black-box defense trước sự tấn công đối kháng trên lên các mô hình phân tích ảnh y khoa.

NỘI DUNG VÀ PHƯƠNG PHÁP

Nội dung 1: Tìm hiểu tổng quan về Adversarial Attack

Phương pháp thực hiện:

- Tiến hành nghiên cứu và tìm hiểu về lý thuyết về các phương pháp tấn công phổ biến vào các mô hình học sâu. Các mô hình học sâu này đã được huấn luyện trên các bộ dữ liệu quen thuộc như CIFAR10, IMAGENET với kết quả ở tập test cao (trên 95%).
- Các phương pháp tấn công dự kiến tìm hiểu: Fast Gradient-Sign Method (FGSM) [1], Projected Gradient Descent (PGD) [5], Decoupled Direction and Norm Attack (DDN) [6].

Kết quả dự kiến: Tóm tắt lý thuyết Adversarial Attack và mô phỏng thành công được các phương pháp tấn công đã tìm hiểu trên các bộ dữ liệu cơ bản (CIFAR10, IMAGENET).

Nội dung 2: Tìm kiếm dữ liệu và huấn luyện mô hình mục tiêu

Phương pháp thực hiện:

- Tìm kiếm các bộ dataset liên quan đến hình ảnh y học trên các tác vụ thị giác máy

tính (Classification và Semantic Segmentation).

- Tìm kiếm các cấu trúc, hoặc phương pháp sử dụng học sâu cho kết quả tốt trên các bộ dữ liệu này (để đánh giá tác động của Adversarial Attack và hiệu quả của phương pháp phòng thủ thì mô hình được chọn phải có kết quả tốt, thậm chí là State-of-The-art trên một hay nhiều Benchmark cho tác vụ đó).
- Thực hiện huấn luyện từ đầu, hoặc sử dụng pretrained model (nếu có) để đánh giá kết quả của những mô hình mục tiêu.

Kết quả dự kiến:

- Tìm kiếm ít nhất 2 bộ dataset thỏa mãn yêu cầu.
- Đạt được hiệu suất huấn luyện mô hình mục tiêu có độ chính xác tương đối so với các nghiên cứu trước đó.

Nội dung 3: Ứng dụng các phương pháp tấn công đã tìm hiểu vào các mô hình phân tích ảnh y khoa

Phương pháp thực hiện:

- Thực hiện Adversarial Attack trên các bộ dữ liệu và mô hình phân tích y học (đã huấn luyện ở **nội dung 2**).
- So sánh kết quả trước và sau khi tấn công.

Kết quả dự kiến:

- Mô phỏng thành công các cuộc tấn công trên mô hình y học
- Đạt được kết quả hiệu suất mô hình mục tiêu giảm rõ rệt so với trước khi bị tấn công.

Nội dung 4: Tìm hiểu tổng quan về Black-Box Defense

Phương pháp thực hiện:

- Tiến hành nghiên cứu và tìm hiểu về lý thuyết về Denoised Smoothing [2] và Black-box Defense [3].

Kết quả dự kiến:

- Tóm tắt được lý thuyết về Denoised Smoothing.
- Hiểu được phương pháp và quy trình thực hiện Black-Box Defense.
- Mô phỏng thành công Black-Box Defense trên bộ dữ liệu gốc của nghiên cứu [3].

Nội dung 5: Thực hiện Black-Box Defense trên mô hình mục tiêu đã bị tấn công

Phương pháp thực hiện:

- Thực hiện Black-Box Defense trên mô hình mục tiêu (**nội dung 3**) với hai tác vụ: Image Classification và Semantic Segmentation.
- So sánh kết quả trước và sau khi tấn công khi có phòng thủ và khi không có phòng thủ.

Kết quả dự kiến:

- Kết quả phòng thủ mô hình mục tiêu trước các thuật toán tấn công (**nội dung 1**).
- Bảng tóm tắt, kết quả so sánh trên.

Nội dung 6: Thay đổi cấu trúc Black-box defense

Phương pháp thực hiện:

- 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.

Kết quả dự kiến:

- Tóm tắt lý thuyết của các cấu trúc Image Reconstruction được sử dụng.
- Kết quả phòng thủ của các cấu trúc đó.
- Bảng tóm tắt, kết quả so sánh giữa các cấu trúc.

Nội dung 7: Tóm tắt các nội dung, viết báo cáo nghiên cứu và đóng gói chương trình

Phương pháp thực hiện:

- Tiến hành tổng hợp quá trình nghiên cứu.
- Chỉnh sửa và đóng gói chương trình.

Kết quả dự kiến:

- Bản báo cáo nghiên cứu.
- Chương trình đã được đóng gói.

KẾT QUẢ MONG ĐỢI

- 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.
- Cung cấp các kiến thức, đánh giá bằng chứng minh toán học và thực nghiệm của phương pháp phòng thủ Black-box Defense trước sự tấn công đối kháng trên lên các mô hình phân tích ảnh y khoa.
- Chương trình được đóng gói, dễ dàng chạy lại để giúp cho những người muốn nghiên cứu dễ dàng tham khảo.
- Dựa vào kết quả nghiên cứu, đưa ra được kết luận và góc nhìn khách quan về tính bảo mật và độ bền vững của những mô hình phân tích ảnh y khoa hiện nay.

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.