ĐẠI HỌC QUỐC GIA HÀ NỘI
**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ**

**NGUYỄN ĐỨC ANH**

# CÁC PHƯƠNG PHÁP ĐẢM BẢO TÍNH CHẮC CHẮN CHO MỘT SỐ MÔ HÌNH HỌC SÂU

LUẬN ÁN TIẾN SĨ KỸ THUẬT PHẦN MỀM

**Hà Nội - 2023**

**Công trình được hoàn thành tại:**
Trường Đại học Công nghệ, Đại học Quốc gia Hà Nội.

Người hướng dẫn khoa học: 1. PGS. TS. Phạm Ngọc Hùng
2. PGS. TS. Nguyễn Lê Minh

Phản biện 1: PGS. TS. Đặng Văn Đức

Phản biện 2: PGS. TS. Nguyễn Mạnh Hùng

Phản biện 3: PGS. TS. Lê Hồng Phương

Luận án sẽ được bảo vệ tại Hội đồng chấm luận án cấp Đại học Quốc Gia họp tại: Trường Đại học Công nghệ – Đại học Quốc gia Hà Nội vào hồi ... giờ ngày ... tháng ... năm ....

**Có thể tìm hiểu luận án tại:**

- Thư viện Quốc gia Việt Nam,
- Trung tâm thông tin - Thư viện, Đại học Quốc gia Hà Nội.

# TÓM TẮT

Để đảm bảo chất lượng của mô hình học sâu, nhiều độ đo đã được đề xuất. Tuy nhiên, dù mô hình được kiểm thử kĩ càng bởi các độ đo này, nhiều nghiên cứu gần đây cho thấy mô hình có thể dễ dàng bị tấn công đối kháng. Tính chắc chắn của mô hình học sâu là khả năng mô hình nhận diện được chính xác nhãn của ảnh đầu vào khi ảnh này được thêm nhiễu đối kháng. Cụ thể, luận án đã đạt được bốn kết quả chính.

Thứ nhất, luận án đề xuất phương pháp HA4FNN để cải thiện tỉ lệ thành công và hiệu năng thấp của DeepCheck khi kiểm thử mô hình nơ-ron truyền thẳng. HA4FNN sử dụng bộ giải phỏng đoán thay vì bộ giải SMT và loại bỏ việc duy trì trạng thái kích hoạt nơ-ron. Thực nghiệm trên MNIST, Fashion-MNIST và bộ chữ cái viết tay cho thấy phương pháp HA4FNN có hiệu năng và tỉ lệ thành công vượt trội so với DeepCheck.

Thứ hai, luận án đề xuất phương pháp PatternAttack để cải thiện tính đa dạng và chất lượng ảnh đối kháng sinh bởi ATN. Tư tưởng chính của PatternAttack là xây dựng ATN khái quát có kiến trúc mô hình mã hóa tự động để thêm nhiễu đối kháng vào ảnh đầu vào theo các mẫu thêm nhiễu khác nhau, từ đó làm tăng tính đa dạng của ảnh đối kháng. Thực nghiệm trên MNIST và CIFAR-10 cho thấy ATN khái quát có thể tấn công mô hình học sâu với tỉ lệ thành công cao và thuật toán tham lam có khả năng cải thiện chất lượng ảnh đối kháng với tỉ lệ giảm nhiễu tốt.

Thứ ba, luận án đề xuất phương pháp QI4AE để nâng cao chất lượng ảnh đối kháng sinh bởi các phương pháp tấn công đối kháng. Độ đo chất lượng ảnh đối kháng là $L_0$ và $L_2$. Thực nghiệm trên MNIST và CIFAR-10 cho thấy phương pháp QI4AE có thể cải thiện chất lượng ảnh đối kháng đáng kể với chi phí tính toán thấp.

Cuối cùng, để nâng cao tính chắc chắn của mô hình học sâu, luận án đề xuất phương pháp SCADefender để loại bỏ nhiễu đối kháng khỏi ảnh đối kháng. Thực nghiệm trên MNIST, CIFAR-10 và Fashion-MNIST cho thấy SCADefender có thể loại bỏ nhiễu đối kháng khỏi ảnh đối kháng khá tốt.

**Từ khóa:** Tính chắc chắn, Kiểm thử hộp trắng, Tấn công đối kháng, Phân tích chương trình, Phòng thủ đối kháng.

# Chương 1

# Giới thiệu

## 1.1 Đặt vấn đề

Trong bài toán phân loại ảnh, với đầu vào là tập học gồm các ảnh và nhãn tương ứng, lập trình viên sẽ định nghĩa kiến trúc mô hình học sâu, rồi chọn các siêu tham số phù hợp như tốc độ học, số lần lặp, v.v. để xây dựng mô hình. Để đánh giá chất lượng mô hình học sâu, các độ đo được sử dụng phổ biến gồm độ chính xác, độ chuẩn xác và điểm số F1. Tuy nhiên, dù mô hình học sâu phân loại ảnh đạt được kết quả tốt với các độ đo nêu trên, mô hình học sâu vẫn có thể có tính chắc chắn chưa đủ tốt. Tính chắc chắn của mô hình học sâu là khả năng mô hình nhận diện được chính xác nhãn của ảnh đầu vào khi ảnh này được thêm nhiễu đối kháng.

Quá trình kẻ tấn công cố tình thêm nhiễu đối kháng vào ảnh đã dự đoán đúng để đánh lừa mô hình gọi là tấn công đối kháng. Ảnh trước khi thêm nhiễu đối kháng và được dự đoán đúng bởi mô hình học sâu gọi là ảnh dự đoán đúng. Ảnh sau khi thêm nhiễu đối kháng gọi là ảnh đối kháng. Trong đó, nhiễu đối kháng được tính dựa trên những điểm ảnh khác nhau giữa ảnh dự đoán đúng và ảnh đối kháng.

Để đánh giá được tính chắc chắn của mô hình học sâu, có hai hướng nghiên cứu chính gồm chứng minh tính chắc chắn của mô hình học sâu và sinh ảnh đối kháng. Đối với hướng sinh ảnh đối kháng, các phương pháp theo hướng này sinh các ảnh đối kháng và coi đó là bằng chứng thể hiện tính chắc chắn của mô hình học sâu. Ưu điểm của hướng này là dễ dàng áp dụng cho các mô hình học sâu phức tạp nên được sử dụng phổ biến.

Theo hướng sinh ảnh đối kháng, hai tiêu chí phổ biến để đánh giá chất lượng phương pháp tấn công đối kháng gồm chất lượng ảnh đối kháng và

tỉ lệ thành công. Công thức đánh giá chất lượng ảnh đối kháng có hai đầu vào chính gồm ảnh dự đoán đúng và ảnh đối kháng tương ứng. Các công thức phổ biến là sử dụng độ đo khoảng cách $L_p$, độ đo cấu trúc như SSIM và các độ đo khác như PSNR. Đối với tỉ lệ thành công, tiêu chí này thể hiện tỉ lệ ảnh dự đoán đúng được thêm nhiễu đối kháng thành công để sinh ảnh đối kháng.

Hai hướng chính để sinh ảnh đối kháng là kiểm thử hộp đen và kiểm thử hộp trắng. Trong kiểm thử hộp trắng, kiểm thử viên có thể truy cập kiến trúc và trọng số của mô hình kiểm thử. Chi phí của kiểm thử hộp trắng thường cao hơn hộp đen do thường phải tính toán đạo hàm hàm mục tiêu của mô hình kiểm thử. Do kiểm thử viên biết được kiến trúc mô hình nên tỉ lệ thành công của kiểm thử hộp trắng thường cao hơn so với kiểm thử hộp đen.

Trong hướng kiểm thử hộp trắng, tấn công đối kháng có hai hướng chính gồm tấn công đối kháng có định hướng và tấn công đối kháng không định hướng. Điểm chung của hai hướng này là thực hiện thêm nhiễu đối kháng vào ảnh dự đoán đúng để sinh ảnh đối kháng có nhãn khác nhãn của ảnh dự đoán đúng. Điểm khác biệt chính giữa hai hướng này là nhãn của ảnh đối kháng. Hướng tấn công đối kháng không định hướng cho mô hình nơ-ron truyền thẳng sử dụng thực thi tượng trưng được đề xuất lần đầu tiên trong DeepCheck. Tuy nhiên, thực nghiệm cho thấy phương pháp này có tỉ lệ thành công và hiệu năng chưa đủ tốt.

Trong hướng tấn công đối kháng có định hướng cho mô hình học sâu, nhiều phương pháp đã đề xuất thiếu tính khái quát hóa. Tính khái quát hóa là khả năng một phương pháp có thể học được cách thêm nhiễu đối kháng vào ảnh dự đoán đúng để sinh ảnh đối kháng và áp dụng tri thức này để thêm nhiễu đối kháng vào ảnh đầu vào mới trong tương lai. Để giải quyết vấn đề thiếu tính khái quát hóa của các phương pháp này, ATN đã được đề xuất để thêm nhiễu đối kháng vào ảnh dự đoán đúng theo độ đo khoảng cách $L_2$. Tuy nhiên, ảnh đối kháng sinh bởi ATN gặp hai vấn đề gồm chất lượng ảnh đối kháng và tính đa dạng của ảnh đối kháng.

Sau khi đã hiểu được bản chất của các phương pháp tấn công đối kháng, nhiệm vụ kế tiếp là chống lại các cuộc tấn công như vậy. Đây là bài toán cải thiện tính chắc chắn. Hướng loại bỏ nhiễu đối kháng khỏi ảnh đầu

vào là một hướng phổ biến. Cụ thể, ảnh đầu vào được đi qua một mô hình loại bỏ nhiễu đối kháng, ví dụ như mô hình mã hóa tự động. Ảnh sau khi loại bỏ nhiễu đối kháng sẽ được đẩy vào mô hình kiểm thử để lấy kết quả. Theo hướng này, các phương pháp kinh điển có thể kể đến MagNet, PuVAE và Defense-VAE. Tuy nhiên, ba phương pháp này chưa loại bỏ nhiễu đối kháng đủ tốt đối với ảnh đối kháng có nhiễu đối kháng đa dạng.

Từ các phân tích trên, luận án hướng tới giải quyết các vấn đề sau. Vấn đề thứ nhất là nghiên cứu phương pháp cải thiện tỉ lệ thành công và hiệu năng của DeepCheck. Vấn đề thứ hai là đề xuất phương pháp cải thiện ATN để sinh ảnh đối kháng có nhiễu đối kháng đa dạng. Vấn đề thứ ba là nghiên cứu phương pháp loại bỏ nhiễu dư thừa khỏi ảnh đối kháng, hay nói cách khác khoảng cách $L_0$ hoặc $L_2$ giữ ảnh dự đoán đúng và ảnh đối kháng càng nhỏ càng tốt. Vấn đề thứ bốn là kết hợp các kết quả nghiên cứu về phương pháp tấn công đối kháng trước đó để xây dựng phương pháp cải thiện tính chắc chắn.

## 1.2 Phạm vi nghiên cứu

Thứ nhất, luận án tập trung vào đánh giá chất lượng các mô hình học sâu phân loại ảnh có kích thước nhỏ. Trong đó, hai loại ảnh được nghiên cứu gồm ảnh xám và ảnh màu. Thứ hai, luận án tập trung vào đề xuất phương pháp sinh các ảnh đối kháng để đánh giá tính chắc chắn của mô hình học sâu. Thứ ba, luận án áp dụng phương pháp kiểm thử hộp trắng để sinh ảnh đối kháng.

## 1.3 Các đóng góp chính của luận án

Thứ nhất, luận án cải thiện tỉ lệ thành công và hiệu năng thấp của phương pháp DeepCheck. Luận án đề xuất phương pháp HA4FNN. Tư tưởng của phương pháp HA4FNN là sử dụng bộ giải phỏng đoán và loại bỏ việc duy trì trạng thái kích hoạt nơ-ron để sinh ảnh đối kháng. Mô hình kiểm thử là mô hình nơ-ron truyền thẳng.

Thứ hai, luận án cải thiện phương pháp ATN để sinh ảnh đối kháng có nhiễu đối kháng đa dạng cho mô hình học sâu bằng cách sử dụng mẫu

thêm nhiễu. Ngoài ra, luận án đề xuất thuật toán tham lam để cải thiện chất lượng ảnh đối kháng theo độ đo $L_0$ và $L_2$. Hai kết quả này được cài đặt trong phương pháp PatternAttack.

Thứ ba, luận án kết hợp thuật toán tham lam và sử dụng mô hình mã hóa tự động để nâng cao hiệu năng của quá trình cải thiện chất lượng ảnh đối kháng, gọi là QI4AE. Đề xuất này là cải tiến của thuật toán tham lam trình bày trong phương pháp PatternAttack.

Thứ bốn, luận án đề xuất phương pháp cải thiện tính chắc chắn, gọi là SCADefender, để loại bỏ nhiễu đối kháng khỏi ảnh đầu vào. Trong khi ba đề xuất trên liên quan đến tấn công đối kháng, phương pháp SCADefender hướng đến chống lại các phương pháp tấn công đối kháng.

## 1.4 Cây nghiên cứu

Để có một cái nhìn rõ hơn về mối tương quan giữa phương pháp đề xuất và các phương pháp so sánh, phần này trình bày cây nghiên cứu liên quan.

## 1.5 Mối quan hệ giữa các chương

Bố cục luận án gồm bảy chương. Phần này trình bày sự kết nối giữa bảy chương này. Chương 2 trình bày kiến thức nền tảng như khái niệm mô hình học sâu, các phương pháp tấn công đối kháng và cải thiện tính chắc chắn, các tiêu chí để đánh giá chất lượng tấn công đối kháng, chất lượng cải thiện tính chắc chắn và bộ giải SMT. Từ Chương 3 đến Chương 6 trình bày bốn đóng góp chính. Kết luận được trình bày trong Chương 7. Chương này tóm tắt lại các kết quả chính của luận án. Sau đó, luận án trình bày những hạn chế còn tồn tại và đề xuất phương hướng giải quyết các hạn chế này.

# Chương 2

# Kiến thức nền tảng

## 2.1 Mô hình học sâu cho bài toán phân loại ảnh

Phần này trình bày mô hình học sâu và hai loại phổ biến gồm mô hình nơ-ron truyền thẳng và mô hình tích chập. Sau đó, phần này trình bày quy trình học mô hình học sâu cho bài toán phân loại ảnh.

## 2.2 Mô hình mã hóa tự động

Phần này trình bày các mô hình mã hóa tự động gồm mô hình mã hóa tự động thưa, mô hình mã hóa tự động xếp chồng và mô hình mã hóa tự động tích chập xếp chồng.

## 2.3 Tấn công đối kháng

### 2.3.1 Hai loại tấn công đối kháng phổ biến

Tấn công đối kháng là một hướng phổ biến để đánh giá tính chắc chắn của mô hình học sâu. Phần này trình bày hai loại tấn công đối kháng phổ biến gồm tấn công đối kháng có định hướng và tấn công đối kháng không định hướng.

### 2.3.2 Tính chắc chắn

Đối với hướng sinh ảnh đối kháng, tính chắc chắn được đánh giá với một phương pháp tấn công đối kháng cụ thể. Các phương pháp tấn công

đối kháng khác nhau sẽ có các kĩ thuật thêm nhiễu đối kháng khác nhau. Mô hình học sâu có tính chắc chắn cao khi phương pháp tấn công đối kháng đó *(i)* khó thêm nhiễu đối kháng nhỏ vào ảnh dự đoán đúng và *(ii)* số lượng ảnh dự đoán đúng thêm nhiễu đối kháng thành công là nhỏ nhất.

### 2.3.3 Phân loại ảnh

Giá trị các điểm ảnh có thể thuộc khoảng số nguyên từ 0 đến 255 hoặc số thực từ 0 đến 1. Nếu không nói gì thêm, luận án mặc định các giá trị điểm ảnh thuộc khoảng [0, 1]. Luận án phân loại ảnh thuộc các loại như sau gồm ảnh đầu vào, ảnh dự đoán đúng, mô hình kiểm thử và ảnh đối kháng.

### 2.3.4 Tính chất nhiễu

Mục đích của quá trình tấn công đối kháng là tìm nhiễu đối kháng để thêm vào ảnh dự đoán đúng. Nhiễu đối kháng có hai tính chất chính gồm tính đa dạng và tính bất định. Phần này trình bày về hai tính chất này.

#### 2.3.4.1 Tiêu chí chất lượng ảnh đối kháng

Ta cần sinh ảnh đối kháng trông giống ảnh dự đoán đúng hết mức có thể. Với tiêu chí này, độ đo khoảng cách $L_p$ thường được sử dụng. Phần này trình bày ba độ đô $L_p$ phổ biến gồm $L_0$, $L_1$ và $L_\infty$.

#### 2.3.4.2 Tiêu chí tỉ lệ thành công

Tỉ lệ thành công là một tiêu chí phổ biến để đánh giá tính chắc chắn của mô hình học sâu trước một phương pháp tấn công đối kháng. Tiêu chí tỉ lệ thành công phản ánh khả năng thêm nhiễu đối kháng vào ảnh dự đoán đúng để sinh ảnh đối kháng thành công. Phần này trình bày tỉ lệ thành công của tấn công đối kháng có định hướng và tấn công đối kháng không định hướng.

#### 2.3.4.3 Tiêu chí tỉ lệ giảm nhiễu

Các phương pháp tấn công đối kháng sinh ảnh đối kháng có thể chứa nhiễu dư thừa. Nếu loại bỏ những nhiễu dư thừa này thì chất lượng ảnh đối kháng sẽ tăng lên. Tiêu chí tỉ lệ giảm nhiễu được sử dụng để đánh giá chất lượng của phương pháp cải thiện chất lượng ảnh đối kháng.

### 2.3.5 Các phương pháp tấn công đối kháng không định hướng

Phần này trình bày các phương pháp tấn công đối kháng không định hướng gồm DeepCheck, FGSM và BIM.

### 2.3.6 Các phương pháp tấn công đối kháng có định hướng

Phần này trình bày các phương pháp tấn công đối kháng có định hướng gồm L-BFGS, FGSM, ATN và CW $L_2$.

## 2.4 Các phương pháp phòng thủ sử dụng mô hình mã hóa tự động

Phần này trình bày các phương pháp phòng thủ sử dụng mô hình mã hóa tự động gồm PuVAE, MagNet và Defense-VAE. Sau đó, phần này trình bày tiêu chí tỉ lệ phát hiện để đánh giá chất lượng mô hình mã hóa tự động phòng thủ.

## 2.5 Các bộ dữ liệu sử dụng trong thực nghiệm

Phần này trình bày về các bộ dữ liệu được sử dụng trong thực nghiệm gồm MNIST, Fashion-MNIST, CIFAR-10 và bộ chữ cái viết tay.

# Chương 3

# Phương pháp sử dụng bộ giải phỏng đoán để tấn công đối kháng không định hướng mô hình nơ-ron truyền thẳng

## 3.1 Giới thiệu

DeepCheck là một kĩ thuật mới theo hướng tấn công đối kháng không định hướng. DeepCheck hỗ trợ kiểm thử mô hình nơ-ron truyền thẳng sử dụng hàm kích hoạt ReLU ở tầng ẩn và softmax ở tầng đầu ra. Thực nghiệm cho thấy DeepCheck có tỉ lệ thành công và hiệu năng khá thấp. Hai nguyên nhân chính cho vấn đề này gồm sự hạn chế của việc duy trì trạng thái kích hoạt nơ-ron và sự hạn chế của bộ giải SMT.

Do đó, để cải thiện hai vấn đề nêu trên của DeepCheck, luận án đề xuất HA4FNN. HA4FNN sử dụng bộ giải phỏng đoán thay vì bộ giải SMT và đơn giản hóa hệ ràng buộc bằng cách loại bỏ tiêu chí duy trì trạng thái kích hoạt nơ-ron.

## 3.2 Các nghiên cứu liên quan

Phần này trình bày các nghiên cứu liên quan đến HA4FNN gồm các phương pháp sinh ảnh đối kháng cho mô hình học sâu phân loại ảnh hiện nay, tổng quan về thực thi tượng trưng, hướng sử dụng bộ giải SMT trong bài toán sinh ảnh đối kháng.

## 3.3 Phương pháp HA4FNN

Phần này trình bày tổng quan HA4FNN. Đầu vào gồm mô hình nơ-ron truyền thẳng **M**, ảnh dự đoán đúng **x** và độ nhạy $k <= 0$. Đầu ra là một tập ảnh đối kháng được kí hiệu bởi *advs*.

### 3.3.1 Sinh mã nguồn từ mô hình & Chèn câu lệnh đánh dấu

#### 3.3.1.1 Sinh mã nguồn từ mô hình

Phần này phân tích mô hình nơ-ron truyền thẳng và chuyển đổi sang mã nguồn viết bằng ngôn ngữ C. Mục đích là thực thi mã nguồn này với đầu vào là ảnh dự đoán đúng để thu thập được đường thi hành.

#### 3.3.1.2 Chèn câu lệnh đánh dấu

Việc biên dịch và thực thi mã nguồn sẽ không trả về danh sách câu lệnh hoặc nhánh đã viếng thăm. Do đó, mã nguồn sẽ được thêm các câu lệnh đánh dấu để tạo thành mã nguồn đánh dấu.

### 3.3.2 Thực thi tượng trưng

Ý tưởng chính của là phân tích sự thay đổi của các nơ-ron ẩn khi thực thi các câu lệnh từ câu lệnh đầu tiên đến câu lệnh cuối cùng của đường thi hành.

### 3.3.3 Bộ giải phỏng đoán

Điểm khác biệt giữa HA4FNN và DeepCheck là cách sử dụng bộ giải để tìm nghiệm của hệ ràng buộc. Nghiệm tìm thấy tương ứng với một ảnh đối kháng. Tuy nhiên, DeepCheck gặp hai vấn đề. Vấn đề đầu tiên là DeepCheck có tỉ lệ thành công thấp. Vấn đề thứ hai là hiệu năng của DeepCheck khi thêm nhiễu đối kháng vào nhiều điểm ảnh trên ảnh dự đoán đúng không tốt. Để giảm thiểu hai vấn đề này, HA4FNN sử dụng một bộ giải phỏng đoán.

## 3.4 Thực nghiệm

Thực nghiệm trả lời ba câu hỏi nghiên cứu gồm **RQ1 - Sửa một điểm ảnh**, **RQ2 - Sửa nhiều điểm ảnh** và **RQ3 - Hiệu năng**.

### 3.4.1 Cấu hình

Đầu tiên, phần này trình bày thông tin về các mô hình kiểm thử. Các mô hình này được học sử dụng bộ thư viện Keras [1].

Từ 500 ảnh đầu tiên của tập học trên từng bộ dữ liệu, thực nghiệm chọn những ảnh được dự đoán đúng bởi mô hình kiểm thử.

Để đảm bảo tính công bằng, DeepCheck và HA4FNN sử dụng các mô-đun giống hệt nhau ngoại trừ mô-đun giải hệ ràng buộc. Các mô-đun dùng chung gồm mô-đun chuyển đổi mô hình kiểm thử thành mã nguồn, mô-đun chèn câu lệnh vào mã nguồn và mô-đun thực thi tượng trưng.

### 3.4.2 Kết quả

#### 3.4.2.1 RQ1 - Sửa một điểm ảnh

Phần thực nghiệm này đánh giá hiệu quả của HA4FNN khi thêm nhiễu đối kháng vào một điểm ảnh trong ảnh dự đoán đúng để sinh ảnh đối kháng.

**Ảnh hưởng của độ nhạy** $k$: Trước khi đi sâu vào so sánh chi tiết, thực nghiệm cần chọn giá trị độ nhạy $k$ phù hợp. Giá trị độ nhạy $k$ trong thực nghiệm này được chọn bằng 0.

**Tỉ lệ thành công**: Với độ nhạy $k = 0$, tỉ lệ thành công của phương pháp HA4FNN cao hơn hẳn DeepCheck.

#### 3.4.2.2 RQ2 - Sửa nhiều điểm ảnh

Trong thực tế, kẻ tấn công có thể muốn sinh ra nhiều ảnh đối kháng nhất có thể và không quan tâm đến tiêu chí chỉ thêm nhiễu đối kháng vào duy nhất một điểm ảnh. Khoảng cách $L_0$ nhỏ nhất giữa ảnh đối kháng và

1. https://keras.io

ảnh dự đoán đúng bằng một. Thực nghiệm cho thấy tỉ lệ thành công của phương pháp HA4FNN cao hơn hẳn tỉ lệ thành công của DeepCheck.

#### 3.4.2.3 RQ3 - Hiệu năng của thực nghiệm thêm nhiễu đối kháng vào nhiều điểm ảnh

Trong khi bộ giải đề xuất cần trung bình 0.40 giây để thêm nhiễu đối kháng vào một ảnh dự đoán đúng, SMTInterpol@1 và SMTInterpol@20 cần trung bình 0.51 giây và 1.07 giây. Đối với Z3@1, Z3@20, Z3@40 và SMTInterpol@40, thời gian chạy trung bình vượt ngưỡng 40 phút cho tấn công một ảnh dự đoán đúng.

## 3.5 Thảo luận

Phần này thảo luận về ảnh hưởng của hàm kích hoạt, việc sử dụng GPU và CPU, mô hình học sâu sử dụng dropout và thực thi tượng trưng.

## 3.6 Tổng kết

Luận án đã trình bày cải tiến trong DeepCheck và cài đặt trong công cụ HA4FNN. Mô hình kiểm thử là mô hình nơ-ron truyền thẳng sử dụng hàm kích hoạt ReLU. Thực nghiệm được tiến hành trên ba mô hình nơ-ron truyền thẳng với ba bộ dữ liệu MNIST, Fashion-MNIST và chữ cái viết tay. Kết quả đầu tiên cho thấy tỉ lệ thành công trung bình của HA4FNN tốt hơn DeepCheck. Kết quả thứ hai cho thấy bộ giải phỏng đoán hoạt động tốt hơn bộ giải SMT. Khi số điểm ảnh được thêm nhiễu đối kháng tăng lên, bộ giải SMT tiêu tốn nhiều thời gian hơn để tìm nghiệm do độ phức tạp của hệ ràng buộc. Kết quả chương này được công bố tại tạp chí Automated Software Engineering (Q2) [anh3].

# Chương 4

# Phương pháp sử dụng mô hình mã hóa tự động để tấn công đối kháng có định hướng mô hình tích chập

## 4.1 Giới thiệu

Sử dụng mô hình mã hóa tự động để thực hiện tấn công đối kháng có định hướng mô hình tích chập được đề xuất lần đầu trong *mô hình chuyển đổi đối kháng* (ATN). Tuy nhiên, ATN có hai hạn chế bao gồm tính đa dạng của ảnh đối kháng và chất lượng ảnh đối kháng.

Đối với vấn đề tính đa dạng của ảnh đối kháng, ý tưởng của ATN là thêm nhiễu đối kháng vào tất cả điểm ảnh trong ảnh dự đoán đúng. Tuy nhiên, ảnh đối kháng có thể sinh ra khi chỉ cần thêm nhiễu đối kháng vào một vài vùng ảnh thay vì toàn bộ ảnh dự đoán đúng. Việc sinh ảnh đối kháng bằng cách thêm nhiễu đối kháng vào ảnh dự đoán đúng theo nhiều tiêu chí thêm nhiễu đối kháng khác nhau sẽ giúp tạo thêm nhiều bằng chứng về tính chắc chắn.

Đối với vấn đề chất lượng ảnh đối kháng, bởi vì ATN thêm nhiễu đối kháng vào mọi điểm ảnh trong ảnh dự đoán đúng nên ảnh đối kháng và ảnh dự đoán đúng có thể trông rất khác nhau. Nếu mô hình mã hóa tự động sinh bởi ATN học chưa đủ tốt thì ảnh đối kháng sinh ra có thể sẽ rất khác ảnh dự đoán đúng. Một trong những nguyên nhân phổ biến là do kiểm thử viên chưa tìm được thông số học mô hình phù hợp.

Để giảm thiểu hai vấn đề của ATN, luận án đề xuất phương pháp PatternAttack. Đối với vấn đề đầu tiên, PatternAttack đề xuất ATN khái quát để sinh ảnh đối kháng có nhiễu đa dạng. Đối với vấn đề thứ hai,

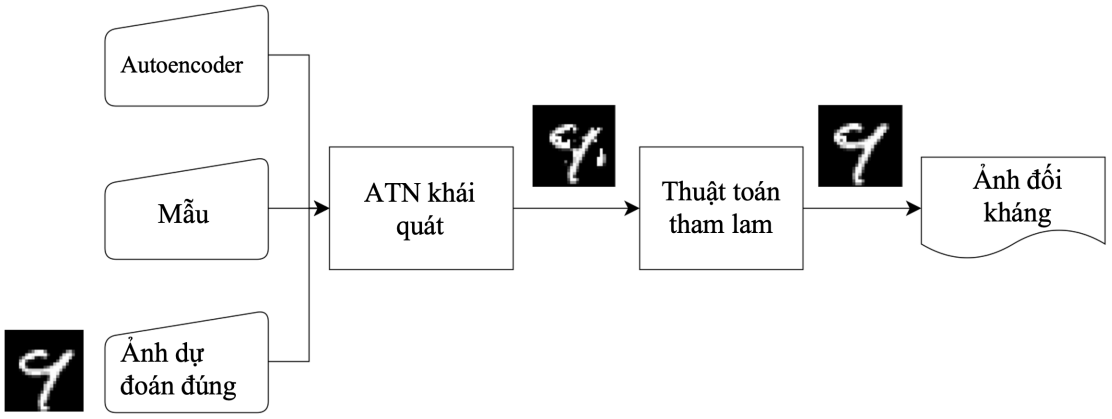
PatternAttack đề xuất thuật toán tham lam để giảm thiểu khoảng cách $L_0$ và $L_2$ giữa ảnh đối kháng và ảnh dự đoán đúng.

## 4.2 Các nghiên cứu liên quan

Phần này trình bày các nghiên cứu liên quan đến đề xuất trong chương. Đầu tiên, luận án sẽ trình bày về các phương pháp tấn công đối kháng cho mô hình tích chập. Sau đó, luận án trình bày tổng quan nghiên cứu xoay quanh bản đồ nổi bật và giải thích tại sao PatternAttack sử dụng bản đồ nổi bật là một mẫu thêm nhiễu.

## 4.3 Phương pháp PatternAttack

Để sinh ảnh đối kháng có chất lượng tốt và đa dạng, luận án đề xuất PatternAttack với hai cải tiến gồm ATN khái quát và thuật toán tham lam. Tổng quan PatternAttack được mô tả trong Hình 4.1.




**Hình 4.1 :** Tổng quan phương pháp PatternAttack.

### 4.3.1 ATN khái quát

Trong pha đầu tiên của PatternAttack, ATN khái quát được áp dụng để sinh tập ảnh đối kháng theo mẫu thêm nhiễu. Kiến trúc của ATN khái quát là kiến trúc của mô hình mã hóa tự động tích chập xếp chồng. Nghiên cứu này sử dụng ba mẫu thêm nhiễu gồm mẫu sửa mọi điểm ảnh, mẫu sửa điểm ảnh ở biên đối tượng và mẫu bản đồ nổi bật.

### 4.3.2 Cải thiện chất lượng ảnh đối kháng

Để cải thiện chất lượng ảnh đối kháng, luận án cần nhận diện nhiễu không có ảnh hưởng hoặc ảnh hưởng rất nhỏ vào kết quả phân lớp. Luận án đề xuất thuật toán tham lam để cải thiện chất lượng ảnh đối kháng.

## 4.4 Thực nghiệm

Thực nghiệm trả lời ba câu hỏi nghiên cứu gồm **RQ1 - Cải thiện tính đa dạng**, **RQ2 - Cải thiện chất lượng** và **RQ3 - Hiệu năng**.

### 4.4.1 Cấu hình

#### 4.4.1.1 Mô hình kiểm thử

Thực nghiệm xây dựng bốn mô hình kiểm thử gồm hai mô hình được học trên MNIST và hai mô hình khác học trên CIFAR-10. Kiến trúc hai mô hình là LeNet-5 và AlexNet.

#### 4.4.1.2 Cấu hình PatternAttack

Phần này trình bày cấu hình ATN khái quát và thuật toán tham lam.

#### 4.4.1.3 Cấu hình các phương pháp so sánh

Các phương pháp so sánh gồm FGSM, L-BFGS và CW $L_2$. Giá trị trọng số của FGSM và L-BFGS giống ATN khái quát. Đối với CW $L_2$, phương pháp này không cần chọn trọng số thủ công vì sử dụng thuật toán nhị phân để tìm trọng số tốt.

### 4.4.2 Kết quả

#### 4.4.2.1 RQ1 - Cải thiện tính đa dạng ảnh đối kháng

Phần này đánh giá khả năng của ATN khái quát để cải thiện tính đa dạng của ảnh đối kháng.

#### 4.4.2.2 RQ2 - Cải thiện chất lượng ảnh đối kháng

Thực nghiệm đánh giá mức độ cải thiện chất lượng ảnh đối kháng của thuật toán tham lam. Đối với độ đo $L_0$, thực nghiệm cho thấy thuật toán tham lam có thể khiến ảnh dự đoán đúng chỉ thêm nhiễu đối kháng vào một điểm ảnh để sinh ảnh đối kháng.

#### 4.4.2.3 RQ3 - Hiệu năng

Thực nghiệm cần đánh giá hiệu năng của PatternAttack khi áp dụng trong thực tế. Thực nghiệm cho thấy PatternAttack có thể sinh ảnh đối kháng từ bộ dữ liệu mới với hiệu năng tốt.

## 4.5 Tổng kết

Chương này đã trình bày phương pháp tấn công đối kháng có định hướng cho mô hình tích chập. Phương pháp PatternAttack có hai đóng góp chính gồm ATN khái quát và thuật toán cải thiện chất lượng ảnh đối kháng. Để đánh giá hiệu quả của PatternAttack, thực nghiệm được tiến hành trên MNIST và CIFAR-10. Kết quả nghiên cứu đã được công bố tại tạp chí Soft Computing (Q2) [anh2] và hội nghị quốc tế RIVF (bài báo xuất sắc nhất) [anh5].

# Chương 5

# Phương pháp sử dụng mô hình mã hóa tự động kết hợp thuật toán tham lam để cải thiện chất lượng ảnh phản ví dụ

## 5.1 Giới thiệu

Để giải quyết vấn đề nhiễu dư thừa của các phương pháp tấn công đối kháng, luận án đã đề xuất thuật toán tham lam. Mặc dù thuật toán tham lam có thể cải thiện chất lượng ảnh đối kháng, thuật toán này có hai hạn chế. Hạn chế thứ nhất là không hỗ trợ tính khái quát hóa. Hạn chế thứ hai là hiệu năng chưa đủ tốt.

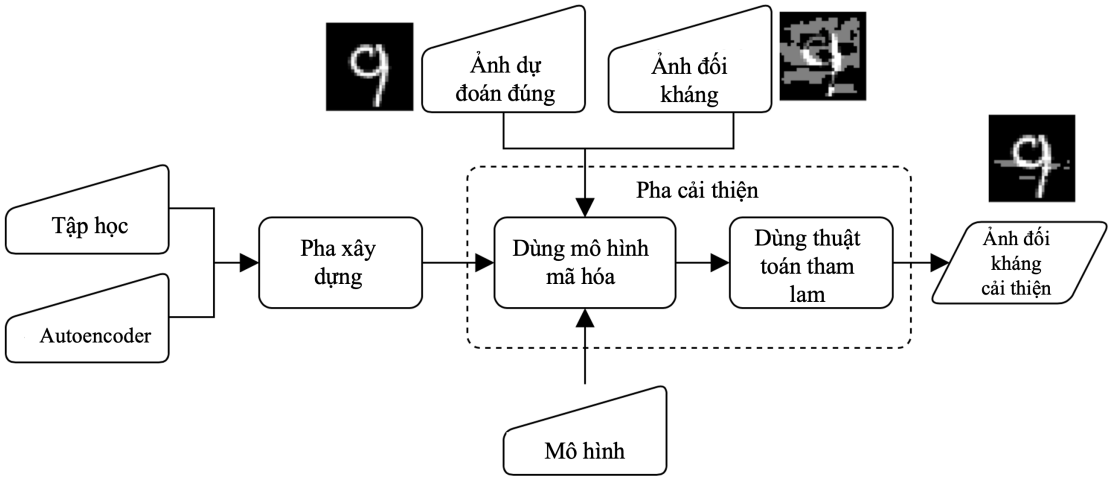
Do đó, chương này trình bày cải tiến của thuật toán tham lam, gọi là QI4AE. Cụ thể, QI4AE gồm hai pha chính gọi là pha xây dựng và pha cải thiện. Trong pha xây dựng, QI4AE xây dựng một mô hình mã hóa tự động có khả năng nhận diện các điểm ảnh chứa nhiễu dư thừa trong ảnh đối kháng. Trong pha cải thiện, ảnh đối kháng sẽ được đẩy vào mô hình mã hóa tự động này để sinh phiên bản ảnh đối kháng cải thiện mức thô. QI4AE áp dụng thuật toán tham lam để loại bỏ nhiễu dư thừa trên ảnh đối kháng cải thiện mức thô để sinh ảnh đối kháng cải thiện mức tinh chế.

## 5.2 Các nghiên cứu liên quan

Các phương pháp tấn công đối kháng cố gắng sinh ảnh đối kháng giống ảnh dự đoán đúng hết mức có thể. Ngoài ra, nhãn của ảnh đối kháng bị phân loại sai. Hai nhóm phương pháp chính là cải thiện chất lượng ảnh đối kháng dựa theo đạo hàm và dựa theo bộ giải.

## 5.3 Phương pháp QI4AE

Tổng quan QI4AE được trình bày trong Hình 5.1. Phương pháp gồm hai pha chính gọi là pha xây dựng và pha cải thiện.




**Hình 5.1 :** Tổng quan phương pháp QI4AE.

## 5.4 Thực nghiệm

Để chứng minh hiệu quả của QI4AE, thực nghiệm trả lời ba câu hỏi gồm **RQ1 - Tỉ lệ thành công của mô hình mã hóa tự động**, **RQ2 - Chất lượng** và **RQ3 - Hiệu năng**.

### 5.4.1 Cấu hình

Phần này trình bày cấu hình của mô hình kiểm thử, phương pháp sinh ảnh đối kháng, phương pháp QI4AE, và phương pháp cơ sở.

### 5.4.2 Kết quả

#### 5.4.2.1 RQ1 - Tỉ lệ thành công của mô hình mã hóa tự động

Thực nghiệm cho thấy các mô hình mã hóa tự động có thể cải thiện chất lượng của hầu hết  ảnh đối kháng chưa cải thiện trong tập $\mathbf{X}_{train}$, từ

khoảng 87.3% đến khoảng 97.2%. Đối với tập $\mathbf{X}_{test}$, tỉ lệ thành công từ 81% tới khoảng 92.2%.

#### 5.4.2.2 RQ2 - Khả năng cải thiện chất lượng

Thực nghiệm chọn giá trị $\delta$ sử dụng trong RQ1 (0.93, 0.96 và 0.99) để so sánh với thuật toán tham lam. Nói chung, pha cải thiện có kết quả tốt hơn thuật toán tham lam trong 8/12 trường hợp. Đối với $L_0$, tỉ lệ giảm nhiễu trung bình từ khoảng 82.38% đến khoảng 95.20%. Đối với $L_2$, tỉ lệ giảm nhiễu trung bình trong khoảng 59.67% đến xấp xỉ 81.07%.

#### 5.4.2.3 RQ3 - Hiệu năng

Phần này đánh giá hiệu năng của pha cải thiện khi gặp ảnh mới. Hiệu năng của pha cải thiện tốt hơn hẳn thuật toán tham lam. Trên MNIST, pha cải thiện thường cần khoảng 14.2 giây đến khoảng 28.8 giây, nhanh hơn khoảng ba lần đến sáu lần so với thuật toán tham lam. Trên CIFAR-10, hiệu năng của pha cải thiện nhanh gấp khoảng bốn lần đến tám lần so với thuật toán tham lam.

### 5.5 Tổng kết

Chương này đã trình bày phương pháp để cải thiện chất lượng ảnh đối kháng. Phương pháp này là cải tiến của thuật toán tham lam. Phương pháp QI4AE có hai pha chính gồm pha xây dựng và pha cải thiện. Thực nghiệm trên MNIST và CIFAR-10 cho thấy QI4AE có thể cải thiện chất lượng ảnh đối kháng theo tiêu chí $L_0$ và $L_2$ đáng kể. Ngoài ra, chi phí khi triển khai QI4AE thấp hơn nhiều so với thuật toán tham lam. QI4AE có thể áp dụng như một pha hậu xử lý với nhiều phương pháp tấn công đối kháng khác. Kết quả nghiên cứu được công bố tại hội nghị quốc tế ICAART [anh4].

# Chương 6

# Phương pháp sử dụng mô hình mã hóa tự động để cải thiện tính chắc chắn của mô hình tích chập

## 6.1 Giới thiệu

Để đánh giá chất lượng của phương pháp cải thiện tính chắc chắn, tỉ lệ phát hiện được sử dụng phổ biến. Các hướng cải thiện tính chắc chắn cho mô hình tích chập phổ biến gồm *(i)* xây dựng lại mô hình kiểm thử, *(ii)* xây dựng một mô hình phân lớp để nhận diện ảnh dự đoán đúng và ảnh đối kháng, *(iii)* loại bỏ nhiễu đối kháng khỏi ảnh đầu vào. Đối với cách tiếp cận *(iii)*, các phương pháp kinh điển có thể kể đến MagNet, PuVAE và Defense-VAE.

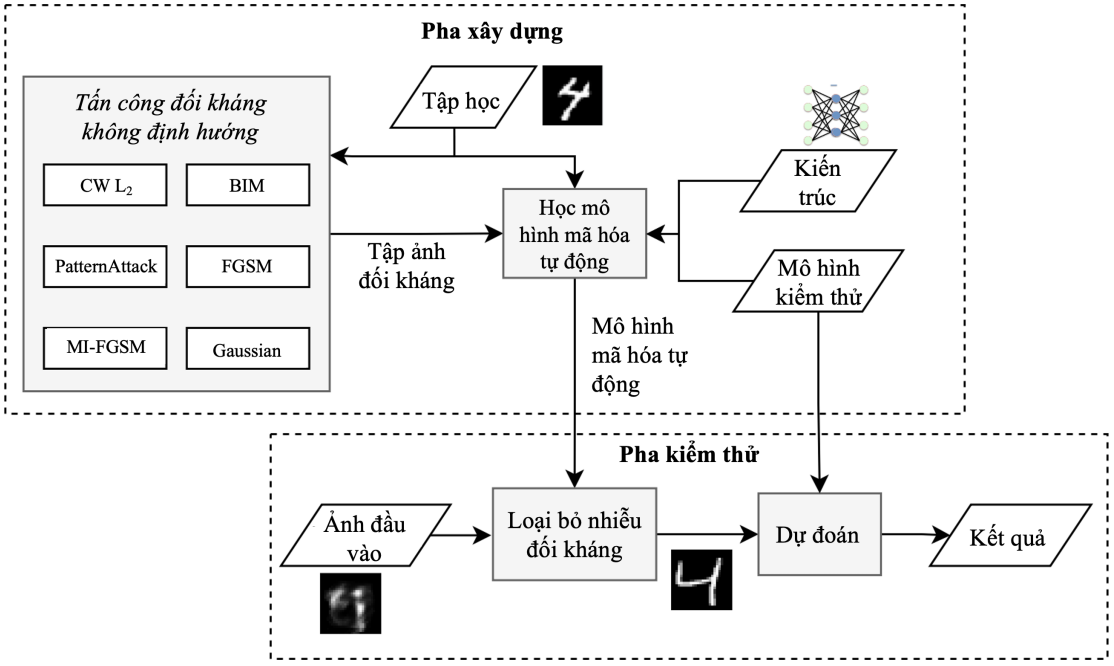
Ba phương pháp trên chưa loại bỏ nhiễu đối kháng đủ tốt đối với ảnh đầu vào có nhiễu đối kháng đa dạng. Trong thực tế, nhiễu đối kháng rất khó có loại phân phối cụ thể đối với mọi phương pháp tấn công đối kháng không định hướng. Cụ thể, một ảnh dự đoán đúng có nhiều cách thêm nhiễu đối kháng để tạo thành ảnh đối kháng. Từ quan sát về tính bất định của nhiễu đối kháng, điểm yếu của MagNet, PuVAE và Defense-VAE được bộc lộ. Ba phương pháp này có thể dễ dàng nhận diện sai nhãn của các ảnh đầu vào chứa nhiễu đối kháng không phải Gaussian. Để giảm thiểu vấn đề nêu trên, nghiên cứu này đề xuất phương pháp SCADefender để loại bỏ nhiễu đối kháng khỏi ảnh đầu vào.

## 6.2 Các nghiên cứu liên quan

Phần này trình bày hai hướng phòng thủ phổ biến gồm hướng sử dụng mô hình mã hóa tự động phòng thủ và hướng học đối kháng.

## 6.3 Phương pháp SCADefender

Hình 6.1 mô tả tổng quan SCADefender. Phương pháp gồm hai pha là pha xây dựng và pha kiểm thử.




**Hình 6.1 :** Tổng quan phương pháp SCADefender.

### 6.3.1 Sinh tập ảnh đối kháng

Bước đầu tiên trong pha xây dựng là sinh tập ảnh đối kháng để làm đầu vào cho quá trình học mô hình mã hóa tự động phòng thủ. Phương pháp SCADefender sử dụng hai loại nhiễu gồm tự nhiên và nhân tạo.

### 6.3.2 Xây dựng mô hình mã hóa tự động

Từ tập ảnh đối kháng và bộ dữ liệu sạch, SCADefender xây dựng một mô hình mã hóa tự động.

## 6.4 Thực nghiệm

Thực nghiệm trả lời ba câu hỏi gồm **RQ1 - Tỉ lệ phát hiện của ảnh không có nhiễu**, **RQ2 - Tỉ lệ phát hiện của ảnh đối kháng** và **RQ3 - Hiệu năng**.

### 6.4.1 Cấu hình

Phần này trình bày cấu hình mô hình kiểm thử, cách sinh tập kiểm thử và cấu hình các phương pháp so sánh.

### 6.4.2 Kết quả

#### 6.4.2.1 RQ1 - Tỉ lệ phát hiện ảnh không có nhiễu

Thực nghiệm này đánh giá hiệu quả của SCADefender khi xử lý ảnh không có nhiễu. Tất cả mọi ảnh đều thuộc tập kiểm thử của MNIST, CIFAR-10 và Fashion-MNIST. Có thể thấy tỉ lệ phát hiện của SCADefender tốt hơn các phương pháp đang so sánh.

#### 6.4.2.2 RQ2 - Tỉ lệ phát hiện ảnh đối kháng

Tổng quan cho thấy SCADefender đạt kết quả tốt nhất trong 17/21 tấn công và kết quả xếp hạng nhì trong 3/21 tấn công.

#### 6.4.2.3 RQ3 - Hiệu năng

Thực nghiệm này đánh giá hiệu năng của SCADefender khi sử dụng trong thực tế. Đối với học đối kháng, thực nghiệm không tính thời gian học lại mô hình kiểm thử với bộ dữ liệu bổ sung. Thực nghiệm cho thấy phương pháp đề xuất có hiệu năng ổn định hơn các phương pháp khác.

## 6.5 Tổng kết

Luận án đề xuất một cải thiện tính chắc chắn đơn giản nhưng hiệu quả có tên là SCADefender sử dụng mô hình mã hóa tự động tích chập xếp chồng để loại bỏ nhiễu đối kháng khỏi ảnh đầu vào. Các thực nghiệm trên MNIST, Fashion-MNIST và CIFAR-10 chứng minh rằng tỉ lệ phát hiện của SCADefender tốt hơn so với các cải thiện tính chắc chắn tương tự bao gồm MagNet và PuVAE. Kết quả nghiên cứu đã được chấp thuận đăng tại tạp chí International Journal of Pattern Recognition and Artificial Intelligence (Q3) [anh1].

# Chương 7

# Kết luận và hướng phát triển

## 7.1 Các kết quả đạt được

Luận án đề xuất các phương pháp để cải thiện tính chắc chắn của mô hình học sâu. Luận án đã giải quyết được bốn vấn đề. Vấn đề đầu tiên là chất lượng tấn công của DeepCheck đối với mô hình nơ-ron truyền thẳng chưa đủ tốt. Vấn đề thứ hai là ảnh đối kháng sinh bởi ATN cho mô hình tích chập chưa đủ đa dạng. Vấn đề thứ ba là ảnh đối kháng sinh bởi nhiều phương pháp tấn công đối kháng chứa nhiều nhiễu dư thừa. Vấn đề thứ bốn là khả năng phòng thủ tấn công đối kháng theo hướng mô hình mã hóa tự động của các phương pháp chưa đủ tốt với ảnh đối kháng có nhiễu đa dạng.

Các nghiên cứu được trình bày trong luận án không những có ý nghĩa về mặt lý thuyết mà còn góp phần làm phương pháp kiểm thử tính chắc chắn cho mô hình học sâu dễ dàng được áp dụng hơn trong thực tiễn. Điều này đặc biệt có ý nghĩa với những mô hình học sâu yêu cầu cao về chất lượng, có khả năng chống lại tấn công từ bên ngoài, trong đó có tấn công đối kháng. Ngoài ra, các công cụ của luận án đã được triển khai sử dụng tại TSDV Việt nam và nhận được những phản hồi tích cực.

## 7.2 Hướng phát triển tiếp theo

Mặc dù các kết quả nghiên cứu đã có những đóng góp cụ thể như đã trình bày nêu trên, các kết quả này còn có những hạn chế cần khắc phục. Nghiên cứu tiếp theo của luận án hướng đến giải quyết các hạn chế này. Cụ thể, các hạn chế và hướng nghiên cứu tiếp theo của luận án như sau.

Trong nghiên cứu thứ nhất, HA4FNN có hai hạn chế. Hạn chế thứ nhất là HA4FNN chưa hỗ trợ tấn công đối kháng có định hướng cho mô hình nơ-ron truyền thẳng. Trong HA4FNN, bộ giải phỏng đoán sẽ thêm nhiễu đối kháng vào ảnh dự đoán đúng để mô hình kiểm thử nhận diện sai. Một hạn chế của bộ giải phỏng đoán là chưa thêm nhiễu đối kháng được ảnh dự đoán đúng để sinh ảnh đối kháng được phân loại là một nhãn cụ thể. Hạn chế thứ hai là HA4FNN chưa hỗ trợ tấn công mô hình tích chập. Mặc dù PatternAttack được đề xuất để tấn công mô hình tích chập và coi là một giải pháp tốt hơn HA4FNN, cách tiếp cận hai phương pháp này khác nhau.

Trong nghiên cứu thứ hai, tuy PatternAttack có thể sinh ảnh đối kháng với chất lượng tốt và có tính đa dạng, phương pháp này có ba hạn chế sau đây. Thứ nhất, PatternAttack chưa được thực nghiệm trên ảnh có kích thước lớn như ImageNet. Thứ hai, việc lựa chọn kiến trúc mô hình mã hóa tự động là một thách thức. Kiến trúc mô hình mã hóa tự động quyết định khả năng mô hình liệu có thêm nhiễu đối kháng vào ảnh dự đoán đúng một cách phù hợp. Thứ ba, việc lựa chọn trọng số giữa các  thành phần có ảnh hưởng đến tốc độ hội tụ của quá trình học mô hình mã hóa tự động.

Trong nghiên cứu thứ bốn, tuy SCADefender có thể loại bỏ được nhiều nhiễu đối kháng của ảnh đối kháng, phương pháp này có hai hạn chế. Thứ nhất, SCADefender có thể biến ảnh đầu vào đang nhận diện đúng thành sai. Nguyên nhân bởi vì khi triển khai trong thực tế, SCADefender chưa phân biệt được ảnh thực sự có nhiễu đối kháng. Thứ hai, ảnh hưởng của kiến trúc mô hình mã hóa tự động chưa được khảo sát triệt để trong thực nghiệm. Thứ ba, bộ dữ liệu để học mô hình mã hóa tự động có thể khá lớn nếu sử dụng ảnh đối kháng sinh bởi nhiều tấn công đối kháng không định hướng, từ đó khiến cho quá trình học càng ngày càng phức tạp khi ngày càng nhiều tấn công đối kháng mới được đề xuất.

# DANH MỤC CÁC CÔNG TRÌNH KHOA HỌC CỦA TÁC GIẢ LIÊN QUAN TỚI LUẬN ÁN

1. *[anh1]* **Duc-Anh Nguyen**, Kha Do Minh, Ngoc Nguyen Nhu, Pham Ngoc Hung (2023). *SCADefender: An Autoencoder-based Defense for CNN-based Image Classifiers*. In International Journal of Pattern Recognition and Artificial Intelligence (Q3)
2. *[anh2]* **Duc-Anh Nguyen**, Kha Do Minh, Khoi Nguyen Le, Minh Nguyen Le, Pham Ngoc Hung (2022). *Improving Diversity and Quality of Adversarial Examples in Adversarial Transformation Network*. In Soft Computing (Q2)
3. *[anh3]* **Duc-Anh Nguyen**, Kha Do Minh, Pham Ngoc Hung, Nguyen Le Minh (2022). *A Symbolic Execution-based Method to Perform Untargeted Attack on Feed-forward Neural Networks*. In Automated Software Engineering (Q2)
4. *[anh4]* **Duc-Anh Nguyen**, Kha Do Minh, Duc-Anh Pham, Pham Ngoc Hung (2022). *Method for Improving Quality of Adversarial Examples*. In the 14th International Conference on Agents and Artificial Intelligence (ICAART - rank B)
5. *[anh5]* **Duc-Anh Nguyen**, Do Minh Kha, Pham Thi To Nga, Pham Ngoc Hung (2021). *An Autoencoder-based Method for Targeted Attack on Deep Neural Network Models*. In the 15th IEEE-RIVF International Conference on Computing and Communication Technologies (RIVF - **The best paper award**)

Danh mục này gồm 05 công trình.