HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Vilayvone Phimsipasom

NGHIÊN CỨU, ỨNG DỤNG PHƯƠNG PHÁP HỌC SÂU VÀO PHÁT HIỆN ẢNH DEEPFAKE

Chuyên ngành: Khoa học máy tính
Mã số: 8.48.01.01

TÓM TẮT ĐỀ ÁN TỐT NGHIỆP THẠC SĨ

HÀ NỘI - NĂM 2025

Đề án tốt nghiệp được hoàn thành tại:

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Người hướng dẫn khoa học: **PGS.TSKH. Hoàng Đăng Hải**

Phản biện 1: …….

Phản biện 2: …….

Đề án tốt nghiệp sẽ được bảo vệ trước Hội đồng chấm đề án tốt nghiệp thạc sĩ tại Học viện Công nghệ Bưu chính Viễn thông

Vào lúc: ..... giờ ..... ngày ..... tháng ..... năm .....

Có thể tìm hiểu đề án tốt nghiệp tại:

- Thư viện của Học viện Công nghệ Bưu chính Viễn thông.

MỞ ĐẦU

Sự phát triển mạnh mẽ của trí tuệ nhân tạo (Artificial Intelligence – AI) và học sâu (Deep Learning – DL) đã mang lại những tiến bộ vượt bậc trong nhiều lĩnh vực khác nhau, đặc biệt là trong thị giác máy tính và xử lý hình ảnh. Một trong những ứng dụng nổi bật của học sâu là khả năng tạo ra các nội dung số chân thực đến mức khó phân biệt bằng mắt thường, điển hình là công nghệ Deepfake. Công nghệ này cho phép tạo ra các hình ảnh, video giả mạo với độ chân thực cao, có thể thay thế khuôn mặt của một người trong video bằng khuôn mặt của người khác, hoặc tạo ra các bức ảnh trông như thật nhưng hoàn toàn không có thực.

Ban đầu, Deepfake được phát triển nhằm phục vụ các mục đích tích cực như hỗ trợ trong ngành điện ảnh, sản xuất nội dung kỹ thuật số hoặc phục hồi hình ảnh lịch sử. Tuy nhiên, với sự phát triển ngày càng tinh vi của công nghệ này, Deepfake đang trở thành một công cụ có thể bị lợi dụng cho các mục đích xấu, chẳng hạn như lừa đảo, phát tán thông tin sai lệch, xâm phạm quyền riêng tư và gây ảnh hưởng đến danh dự cá nhân. Sự lan truyền nhanh chóng của nội dung Deepfake đặt ra những thách thức lớn đối với tính xác thực của thông tin, an ninh mạng, pháp lý và đạo đức xã hội.

Trước thực trạng này, việc nghiên cứu và phát triển các phương pháp phát hiện ảnh Deepfake trở thành một yêu cầu cấp thiết nhằm hạn chế tác động tiêu cực của công nghệ này đối với xã hội. Các phương pháp truyền thống như phân tích bằng mắt thường, kiểm tra siêu dữ liệu (metadata) hoặc sử dụng phần mềm chỉnh sửa ảnh hiện có không còn đủ hiệu quả trước sự phát triển không ngừng của các thuật toán tạo ra ảnh Deepfake.

Hiện nay, một số nghiên cứu đã tập trung vào phát hiện Deepfake dựa trên các đặc trưng bất thường trong hình ảnh, chẳng hạn như sai lệch ánh sáng, biến dạng khuôn mặt, thiếu nhất quán trong biểu cảm hoặc dấu vết chỉnh sửa trong các chi tiết nhỏ, ví dụ [1,3, 4-7]. Tuy nhiên, những phương pháp này vẫn còn nhiều hạn chế về độ chính xác và khả năng tổng quát hóa khi áp dụng trên các tập dữ liệu lớn và đa dạng. Vì vậy, cần có những phương pháp phát hiện tự động, chính xác và hiệu quả hơn, trong đó các phương pháp học sâu nổi lên như một giải pháp tiềm năng đầy hứa hẹn. Học sâu đã chứng minh được khả năng vượt trội trong việc xử lý và phân loại hình ảnh thông qua các kiến trúc mạng nơ-ron nhân tạo tiên tiến như Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs) và Transformers [11-13, 15-17]. Các mô hình học sâu có khả năng tự động trích xuất đặc trưng từ hình ảnh mà không cần phải xác định trước các đặc điểm quan trọng, từ đó giúp cải thiện đáng kể độ chính xác trong phát hiện ảnh Deepfake.

Các nghiên cứu gần đây [7, 19, 62] cho thấy, việc áp dụng mạng CNN và các phương pháp học sâu khác vào bài toán phát hiện ảnh Deepfake có thể giúp nhận diện các đặc trưng tinh vi mà mắt thường khó phân biệt, chẳng hạn như sự sai lệch trong kết cấu da, bóng đổ, ánh sáng hoặc sự bất thường trong biểu cảm khuôn mặt. Ngoài ra, các phương pháp dựa trên Transformer như Vision Transformers (ViTs) cũng đang được nghiên cứu và thử nghiệm để cải thiện khả năng tổng quát hóa của mô hình, giúp phát hiện ảnh Deepfake hiệu quả hơn trên nhiều loại dữ liệu khác nhau.

Tuy nhiên, vẫn còn nhiều thách thức trong việc phát triển mô hình học sâu hiệu quả để phát hiện ảnh Deepfake. Một trong những vấn đề chính là sự đa dạng của các thuật toán tạo

Deepfake, khiến mô hình cần phải có khả năng phát hiện cả những hình ảnh giả mạo chưa từng xuất hiện trong quá trình huấn luyện. Bên cạnh đó, việc cân bằng giữa độ chính xác và hiệu suất xử lý cũng là một yếu tố quan trọng cần được xem xét.

Xuất phát từ thực tiễn trên, đề án tốt nghiệp "**NGHIÊN CỨU, ỨNG DỤNG PHƯƠNG PHÁP HỌC SÂU VÀO PHÁT HIỆN ẢNH DEEPFAKE**" được thực hiện nhằm cung cấp một cái nhìn toàn diện về công nghệ Deepfake và các phương pháp phát hiện ảnh giả mạo, qua đó đánh giá việc áp dụng một số mô hình học sâu tối ưu nhằm nâng cao hiệu quả nhận diện ảnh Deepfake. Kết quả nghiên cứu có thể được ứng dụng trong nhiều lĩnh vực như an ninh mạng, báo chí, truyền thông, pháp lý và bảo vệ danh tính số, góp phần ngăn chặn những tác động tiêu cực của ảnh Deepfake đối với xã hội.

Bố cục đề án tốt nghiệp ngoài phần mở đầu, kết luận gồm 03 chương, như sau:

Chương 1: **Tổng quan nghiên cứu về học sâu và ảnh DeepFake**: cơ sở lý thuyết về công nghệ AI, học sâu, đặc trưng của ảnh DeepFake do AI tạo ra, các kỹ thuật tạo ảnh DeepFake phổ biến, yêu cầu đối với việc phát hiện ảnh DeepFake, một số nghiên cứu nổi bật liên quan đến phát hiện ảnh DeepFake.

Chương 2: **Giải pháp sử dụng học sâu trong phát hiện ảnh Deepfake**: đặc điểm của ảnh DeepFake do AI tạo ra; các phương pháp học sâu phổ biến trong phát hiện ảnh DeepFake bao gồm kiến trúc CNNs và Transformers; mô hình học sâu cải tiến cho phát hiện ảnh DeepFake; lựa chọn giải pháp sử dụng học sâu cho cho phát hiện ảnh DeepFake.

Chương 3: Xây dựng mô hình phát hiện ảnh Deepfake: đề xuất mô hình tổng quát; thu thập dữ liệu; mô tả các công cụ, thư viện sử dụng để thực hiện mô hình; triển khai

áp dụng các thuật toán học máy cho mô hình; thực hiện thử nghiệm; so sánh giữa các mô hình để đánh giá kết quả phát hiện.

CHƯƠNG 1. TỔNG QUAN NGHIÊN CỨU VỀ HỌC SÂU VÀ ẢNH DEEPFAKE

1.1. Khái quát về học sâu và công nghệ AI

1.2. Ứng dụng của AI, học sâu trong phát hiện ảnh giả mạo

1.2.1. Ứng dụng của AI và học sâu trong các lĩnh vực hiện nay

1.2.2. AI và học sâu trong nhận diện hình ảnh và phát hiện DeepFake

Một trong những ứng dụng nổi bật và mang tính chất thời sự nhất của AI và học sâu trong nhận diện hình ảnh là *nhận diện nội dung hình ảnh và video giả mạo – thường được gọi là phát hiện Deepfake*. Đây là một vấn đề không chỉ mang tính kỹ thuật mà còn liên quan trực tiếp đến an ninh thông tin, truyền thông đại chúng và pháp y kỹ thuật số.

Deepfake là sản phẩm của các mô hình tự sinh – đặc biệt là Generative Adversarial Networks (GAN), trong đó một mô hình tự sinh (generator) học cách tạo dữ liệu giả (hình ảnh hoặc video) sao cho không thể phân biệt được với dữ liệu thật, trong khi mô hình phân biệt (discriminator) cố gắng phát hiện ra sự khác biệt. Quá trình này tạo ra nội dung giả mạo có độ chân thực cao, đặc biệt là khuôn mặt, biểu cảm, giọng nói [68].

Nhiều nghiên cứu đã đề xuất các phương pháp phát hiện Deepfake dựa trên học sâu như:

- **CNN**: Khai thác đặc trưng vi mô và hiện tượng nhiễu [3].
- **Xception, EfficientNet**: Mô hình nhẹ, hiệu quả trong nhận diện giả mạo ảnh/video [15, 65].

- **Vision Transformer (ViT)**: Học đặc trưng toàn cục và không gian ảnh tốt [77].
- **Hybrid CNN–RNN hoặc CNN–LSTM**: Phân tích chuỗi thời gian trong video DeepFake [2].
- **Ensemble Learning**: Kết hợp nhiều mô hình và nguồn dữ liệu để nâng cao độ chính xác [68].

Mặc dù công nghệ phát hiện Deepfake đã có những bước tiến, vẫn còn nhiều thách thức như:

- Deepfake ngày càng tinh vi, khó bị phát hiện bởi mô hình truyền thống [70, 76].
- Thiếu dữ liệu thật–giả đa dạng, đặc biệt là trong môi trường thực tế [64].
- Vấn đề tổng quát hóa giữa các tập dữ liệu khác nhau và sự thay đổi liên tục của kỹ thuật làm giả [1, 7, 57].
- Chi phí tính toán cao, đòi hỏi hạ tầng GPU mạnh để huấn luyện mô hình phức tạp [5].

1.3. Các kỹ thuật tạo ảnh Deepfake phổ biến bằng công nghệ AI

1.4. Ảnh DeepFake và những vấn đề đặt ra

1.4.1. Giới thiệu về ảnh DeepFake do AI tạo ra

1.4.2. Mặt trái của việc tạo ảnh bằng công nghệ AI và tác động xã hội

1.4.3. Yêu cầu nhận diện, phát hiện ảnh Deepfake do AI tạo ra

Để tăng cường hiệu quả khả năng nhận dạng, phát hiện ảnh Deepfake, các hệ thống nhận diện và phát hiện ảnh Deepfake cần tập trung nghiên cứu vào các vấn đề:

- ✓ *Độ chính xác và độ nhạy cao*
- ✓ *Tính thời gian thực và khả năng triển khai*
- ✓ *Khả năng giải thích và minh bạch*
- ✓ *Đảm bảo quyền riêng tư và đạo đức*

Tuy nhiên, một trong những thách thức lớn nhất trong việc phát hiện ảnh Deepfake là sự thiếu hụt dữ liệu huấn luyện chất lượng cao, do các mẫu Deepfake liên tục thay đổi và trở nên ngày càng tinh vi hơn. Do đó, việc xây dựng các tập dữ liệu lớn và đa dạng về Deepfake để huấn luyện mô hình phát hiện là một yêu cầu cấp thiết, giúp tăng cường khả năng nhận diện và phân biệt giữa nội dung thật và giả.

1.5. Một số nghiên cứu nổi bật liên quan đến phát hiện ảnh Deepfake

1.6. Kết luận chương

Trong chương này, đề án tốt nghiệp đã trình bày tổng quan nền tảng về trí tuệ nhân tạo, học máy và đặc biệt là học sâu – lĩnh vực đóng vai trò cốt lõi trong các hệ thống xử lý và phân tích hình ảnh hiện đại. Học sâu, với các kiến trúc mạng nơ-ron sâu như CNN, Transformer hay Diffusion Models, đã tạo nên bước ngoặt trong khả năng sinh và phân tích dữ liệu hình ảnh.

Bên cạnh những thành tựu vượt trội, học sâu cũng mở ra nhiều thách thức mới, điển hình là sự xuất hiện của ảnh Deepfake – sản phẩm của các mô hình sinh tạo có khả năng giả mạo khuôn mặt, biểu cảm và danh tính một cách tinh vi. Các ứng dụng Deepfake đang đặt ra những nguy cơ nghiêm trọng về đạo đức, an ninh thông tin, và độ tin cậy của truyền thông số.

Trong bối cảnh đó, việc phát triển các hệ thống có khả năng phát hiện ảnh Deepfake một cách chính xác và đáng tin cậy

đang trở thành một hướng nghiên cứu cấp thiết. Những nội dung trình bày trong chương này là cơ sở lý luận quan trọng để từ đó đi sâu vào phân tích các phương pháp nhận diện ảnh Deepfake và đề xuất mô hình phù hợp trong các chương tiếp theo.

CHƯƠNG 2. GIẢI PHÁP SỬ DỤNG HỌC SÂU TRONG PHÁT HIỆN ẢNH DEEPFAKE

2.1. Đặc điểm của ảnh DeepFake

2.2. Các phương pháp học sâu ứng dụng trong phát hiện ảnh DeepFake

2.3. Mô hình học sâu cải tiến cho phát hiện ảnh DeepFake

2.4. Giải pháp sử dụng học sâu phát hiện ảnh DeepFake

2.5. Kết luận chương

Trong chương 2, đề án đã trình bày các nội dung liên quan đến giải pháp sử dụng học sâu trong phát hiện ảnh DeepFake. Các nội dung đã đề cập gồm: đặc điểm của ảnh DeepFake, các dấu hiệu đặc trưng và cách nhận biết; các phương pháp học sâu sử dụng trong phát hiện ảnh DeepFake bao gồm: các phương pháp học máy truyền thống, mô hình CNN và các biến thể, các mô hình Transformer và các cải tiến.

Bốn mô hình CNN được đánh giá cao trong phát hiện ảnh DeepFake gồm: Xception, MesoNet, ResNet-50 và EfficientNet-B0. Bốn mô hình Transformer gồm: Vision Transformer (ViT), Swin Transformer, TimeSpace Transformer, Dettection Transformer (DETR). Hai mô hình học sâu cải tiến được trình bày gồm: DTN (Distilled Transformers with Locally Enhanced Global Representations) và mô hình kết hợp CNN và Vision Transformer.

Qua khảo sát của học viên, các mô hình CNN và Transformer đều có các ưu và nhược điểm trong bài toán phát hiện ảnh DeepFake. Để lựa chọn một giải pháp sử dụng học sâu cho phát hiện ảnh DeepFake và tính đa dạng của ảnh DeepFake,

học viên đặt ra bốn tiêu chí kỹ thuật cần quan tâm gồm: Mô hình cần phù hợp với tính thiếu đặc trưng rõ ràng và ổn định của dữ liệu; Mô hình cần có khả năng tổng quát hóa với các tập dữ liệu; Mô hình cần đạt cân bằng giữa hiệu năng và độ phức tạp; Mô hình cần có khả năng chịu lỗi và thích ứng với kỹ thuật mới, điển hình là các kỹ thuật tạo ảnh giả ngày càng đa dạng với công nghệ AI.

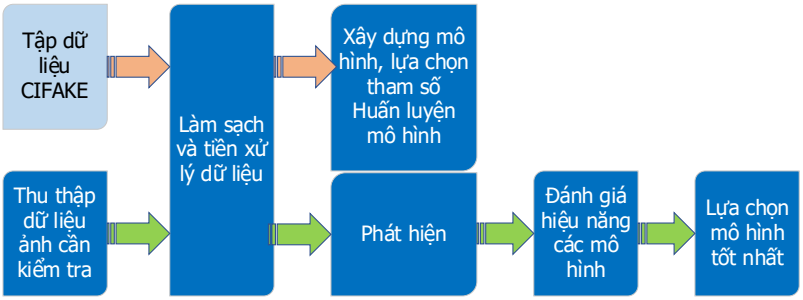
Chương 2 đã đưa ra đề xuất giải pháp học sâu gồm: 1) Lựa chọn tập dữ liệu phù hợp để thực hiện đánh giá các mô hình học sâu, cụ thể là tập CIFAKE; 2) Lựa chọn ba mô hình học sâu phù hợp với các yêu cầu đã nêu so sánh đánh giá theo các tiêu chí hiệu năng để chọn ra mô hình tốt nhất. Ba mô hình khác nhau được lựa chọn xuất phát từ phân tích lý thuyết ở mục 2.2, cụ thể là hai mô hình CNN là ResNet-50 và EfficientNet-B0, một mô hình Transformer là Swin Transformer được so sánh, đánh giá trên cùng một tập dữ liệu ảnh giả CIFAKE.

Trong chương tiếp theo, đề án sẽ trình bày chi tiết về việc triển khai giải pháp, cách thức tiền xử lý dữ liệu ảnh, chuẩn hóa dữ liệu ảnh, phương pháp huấn luyện mô hình, tiến trình huấn luyện, phương pháp đánh giá các mô hình theo các tiêu chí hiệu năng và thực hiện thử nghiệm, đánh giá kết quả.

CHƯƠNG 3. THỰC HIỆN MÔ HÌNH HỌC SÂU TRONG PHÁT HIỆN ẢNH DEEPFAKE

3.1. Sơ đồ khối mô hình học sâu phát hiện ảnh DeepFake

Hình 3.1 là sơ đồ khối mô hình học sâu cho phát hiện ảnh Deepfake được đề xuất trong đề án tốt nghiệp.



Hình 0.1: Các bước thực hiện mô hình phát hiện ảnh DeepFake

Mô hình phát hiện ảnh Deepfake được đề xuất cơ bản tuân thủ đúng 5 bước thực hiện ở trên. Bước 1 được gồm 2 thành phần:

- Tập dữ liệu CIFAKE chứa cả ảnh thật và ảnh giả (do GAN và StyleGAN2 tạo), đã được gán nhãn rõ ràng (real, fake). Tập dữ liệu CIFAKE gồm các ảnh có kích thước 32x32, RGB, theo định dạng png hoặc jpg. Tổng số ảnh khoảng 120.000 ảnh được chia theo tỷ lệ 50% ảnh thật, 50% ảnh giả. Tỷ lệ cân bằng này giúp cho việc huấn luyện mô hình tốt hơn, không bị lệch về ảnh giả hay ảnh thật. Tập dữ liệu CIFAKE sử dụng tỷ lệ chia dữ liệu sẵn có cho huấn luyện là 80%, nghĩa là 96.000 ảnh và cho kiểm tra là 20%, nghĩa là 24.000 ảnh.

- Dữ liệu ảnh cần được kiểm tra thật / giả được thu thập từ thực tế. Dữ liệu này cần được tiền xử lý, chuẩn hóa tương tự như ảnh từ tập CIFAKE để bảo đảm sự tương đồng của dữ liệu.

  Trong phạm vi đề án, dữ liệu ảnh để kiểm thử mô hình cũng được lấy từ tập CIFAKE.

Khối làm sạch và tiền xử lý dữ liệu thực hiện các nhiệm vụ: hiệu chỉnh kích thước ảnh (đối với ảnh thu thập từ thực tế), chuẩn hóa ảnh, tăng cường chất lượng ảnh.

Khối mô hình học sâu: có thể chọn một trong ba mô hình là: ResNet-50, EfficientNet-B0 hoặc Swin Transformer như đã đề cập ở chương 2. Module huấn luyên thực hiện huấn luyện theo tập dữ liệu huấn luyện (48000 ảnh) để có khả năng phân loại ảnh thật /giả. Module phát hiện nhận ảnh đầu vào chưa rõ thật /giả đã qua bước tiền xử lý để nhận diện, phát hiện ảnh thật /giá với mô hình học sâu đã huấn luyện.

Đầu ra của khối phát hiện trả về kết quả phân loại ảnh thật / giả kèm theo các tham số phục vụ cho đánh giá hiệu năng cho các mô hình. Căn cứ vào kết quả đánh giá hiệu năng theo các tiêu chí, có thể lựa chọn mô hình cho kết quả tốt nhất đối với ảnh cần nhận diện, phát hiện.

3.2. Thu thập và mô tả dữ liệu

3.3. Tiền xử lý và làm sạch dữ liệu

3.4. Xây dựng mô hình phát hiện ảnh Deepfake

3.5. Đánh giá hiệu năng của các mô hình

3.5.1. Độ chính xác (Accuracy)

Độ chính xác được định nghĩa là tỷ lệ giữa tổng số giá trị dự đoán đúng so với tổng số dự đoán. Độ chính xác thường được sử dụng để đánh giá các mô hình phân loại. Độ chính xác của mô hình học máy cho biết tổng số lần mô hình dự đoán đúng trên toàn bộ tập dữ liệu. Tuy nhiên, accuracy có thể không phản ánh đầy đủ hiệu suất trong trường hợp dữ liệu không cân bằng (ví dụ, nhiều ảnh thật hơn ảnh giả hoặc ngược lại) [60].

Công thức tính:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \qquad (3.1)$$

trong đó:

- TP (True Positive) – Dương tính thật: Số lượng ảnh giả được dự đoán đúng là giả.
- TN (True Negative) – Âm tính thật: Số lượng ảnh thật được dự đoán đúng là thật.
- FP (False Positive) – Dương tính giả: Số lượng ảnh thật bị dự đoán sai là giả.
- FN (False Negative) - Âm tính giả: Số lượng ảnh giả bị dự đoán sai là thật.

3.5.2. Tỷ lệ trúng (Precision)

Precision đo lường tỷ lệ các dự đoán dương tính thật (ảnh giả được dự đoán là giả) trên tổng số dự đoán dương tính. Precision phản ánh khả năng của mô hình trong việc không tạo ra quá nhiều kết quả dương tính giả. Chỉ số này rất quan trọng trong các bài toán mà việc dự đoán sai (ảnh thật bị nhận diện là giả) có thể gây ra hậu quả nghiêm trọng. Precision cao cho thấy mô hình ít có xu hướng nhầm lẫn ảnh thật thành ảnh giả [81-60].

Công thức tính:

$$\text{Precision} = \frac{TP}{TP+FP} \qquad (3.2)$$

trong đó:

- TP (True Positive): Số lượng ảnh giả được dự đoán đúng là giả.
- FP (False Positive): Số lượng ảnh thật bị dự đoán sai là giả.

3.5.3. Độ nhạy (Recall)

Recall, hay còn gọi là độ nhạy hoặc độ thu hồi, đo lường tỷ lệ các dự đoán dương tính thật trên tổng số mẫu dương tính thực tế (ảnh giả được nhận diện là giả trên tổng số ảnh giả thực sự). Recall cho biết mô hình có thể nhận diện bao nhiêu phần trăm ảnh giả trong tổng số ảnh giả thực tế. Đây là chỉ số quan trọng để đánh giá hiệu suất của mô hình trong việc phát hiện ảnh giả, đặc biệt là khi việc bỏ sót các ảnh giả (FN) cần được hạn chế tối đa. Recall cao cho thấy mô hình ít bỏ sót ảnh giả [60]

Công thức tính:

$$\text{Recall} = \frac{TP}{TP+FN} \qquad (3.3)$$

trong đó:

- TP (True Positive): Số lượng ảnh giả được dự đoán đúng là giả.
- FN (False Negative): Số lượng ảnh giả bị dự đoán sai là thật.

3.5.4. F1 Score

F1 Score là độ đo kết hợp giữa Precision và Recall, cung cấp cái nhìn cân bằng hơn trong các trường hợp có sự đánh đổi giữa precision và recall. F1 Score sẽ cao khi cả Precision và Recall đều cao. F1 Score là chỉ số rất hữu ích khi chúng ta muốn một cân bằng giữa Precision và Recall, nhất là trong các trường hợp dữ liệu không cân bằng. F1 Score cao cho thấy mô hình có

thể vừa nhận diện chính xác ảnh giả (Precision) vừa không bỏ sót ảnh giả nào (Recall).

Công thức tính:

$$\text{F1 Score} = 2\ x\ \frac{Precision\ x\ Recall}{Precision + Recall} \qquad (3.4)$$

trong đó:

- Precision : Tỷ lệ trúng
- Recall: Độ nhạy

3.6. Môi trường thử nghiệm

3.6.1. Ngôn ngữ lập trình và thư viện

Phần thực nghiệm của luận văn này được thực hiện bằng ngôn ngữ lập trình Python (Python 3.6.6 :: Anaconda custom 64-bit) và các thư viện đã được phát triển cho nó. Python là một ngôn ngữ mã nguồn mở cấp cao [29], đã trở thành một trong những ngôn ngữ được sử dụng nhiều nhất trong học máy và khoa học máy tính. Nó có nhiều thư viện mã nguồn mở được phát triển mà một số trong đó được sử dụng trong luận văn này. Các thư viện được sử dụng bao gồm: numpy, scipy, pandas, matplotlib, scikit-learning [53], TensorFlow, xgboost, imblearn, plotly, Keras và seaborn. Các thư viện này cung cấp các công cụ để tiền xử lý, các thuật toán cho học máy và cách thức vẽ biểu đồ dữ liệu.

3.6.2. Cấu hình máy tính thử nghiệm

- Hệ điều hành: Windows 10 Pro
- CPU: Intel(R) Core(TM) i7-4720HQ CPU @2.60GHz
- RAM: 8GB

3.6.3. Cấu trúc tập dữ liệu thử nghiệm

Tập dữ liệu CIFAKE gồm 2 loại ảnh:

- Ảnh thật: Lấy trực tiếp từ tập dữ liệu CIFAR-10, gán nhãn Real
- Ảnh giả: Sinh ra từ mô hình StyleGAN2 huấn luyện trên tập ảnh CIFAR-10, gán nhãn Fake.
- Kích thước ảnh: 32x32 pixel, RGB (3 màu)
- Định dạng ảnh: png, jpg.
- Số lượng ảnh: 120.000 ảnh
- Tỷ lệ ảnh thật: 60.000 ảnh (50%). Tỷ lệ ảnh giả: 60.000 ảnh (50%)
- Tỷ lệ tập huấn luyện: 96.000 ảnh (80%).
- Tỷ lệ tập kiểm tra: 24.000 ảnh (20%).

3.6.4. Các tham số cơ bản của các mô hình học sâu

- Input_size: (32, 32, 3)
- Num_classes: 2 (real/fake)
- Learning Rate: 1e-3
- Batch_size: 64 (mô hình Swin Transformer sử dụng batch nhỏ = 16 để tránh quá tải)
- Epochs: 50
- Loss function: Binary Cross Entropy.

3.6.5. Tiêu chí đánh giá hiệu năng cho các mô hình học sâu

Các tham số dùng để đánh giá hiệu năng của các mô hình bao gồm:

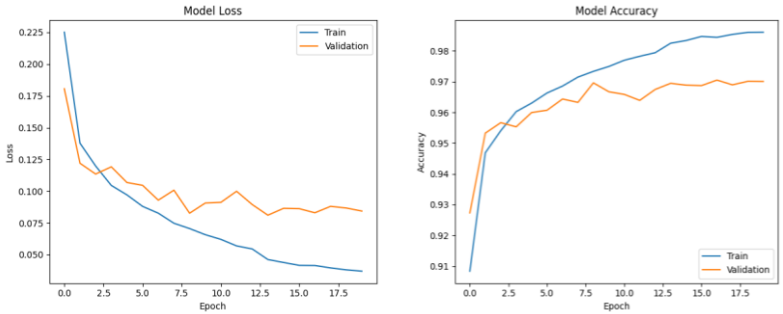
- Độ chính xác (Accuracy)
- Độ dự đoán chính xác (Precision)
- Độ nhạy (Recall)
- F1-score

3.7. Kết quả thử nghiệm

3.7.1. Kết quả huấn luyện và tối ưu tham số các mô hình

a. Mô hình Resnet-50

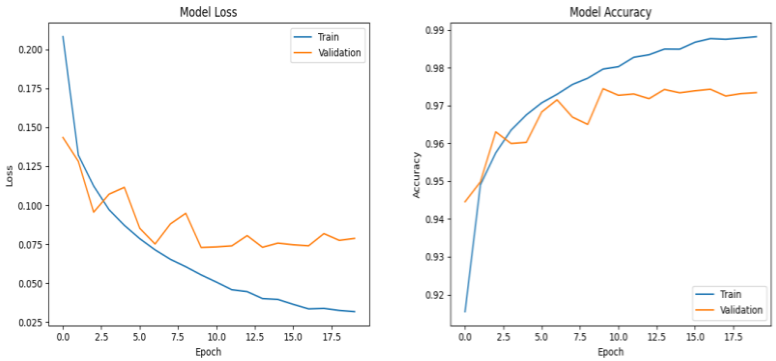
(a) (b)



Hình 0.2: Kết quả huấn luyện và tối ưu tham số của mô hình ResNet-50

b. Mô hình EficientNet-B0

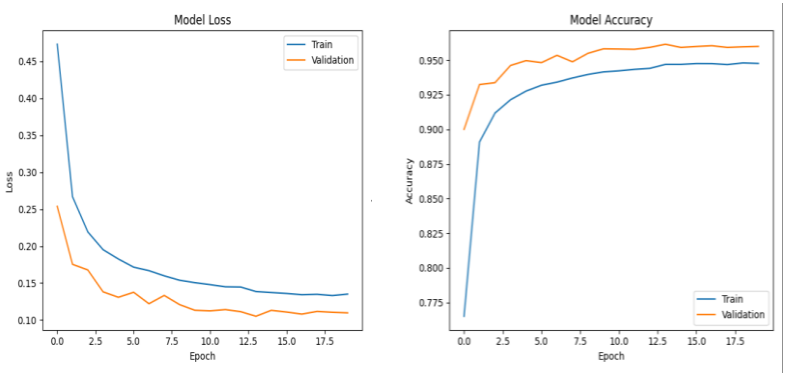
(a) (b)



Hình 0.3: Kết quả huấn luyện và tối ưu tham số của mô hình EficientNet-B0

c. Mô hình Swin

(a) (b)



Hình 0.4: Kết quả huấn luyện và tối ưu tham số của mô hình Swin

Tổng quan, cả ba mô hình đều phù hợp cho bài toán phát hiện ảnh Deepfake, trong đó Swin Transformer thể hiện sự ổn định và khả năng tổng quát hóa vượt trội, EfficientNet-B0 đạt hiệu suất rất cao với chi phí tính toán thấp, còn ResNet-50 vẫn là một mô hình nền tảng mạnh với khả năng học sâu và hiệu suất cao, mặc dù cần điều chỉnh để giảm thiểu overfitting.

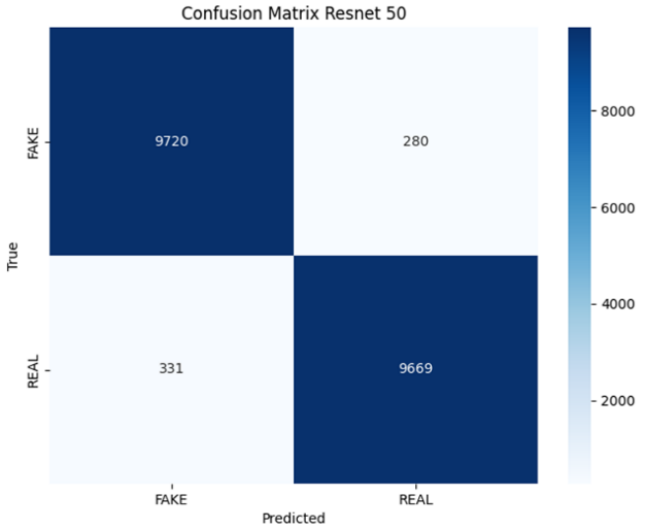
3.7.2. Kết quả đánh giá hiệu năng của các mô hình

Bảng 0.1: Phân tích đánh giá hiệu năng của các mô hình ResNet-50, EfficientNet-B0 và Swin Transformer

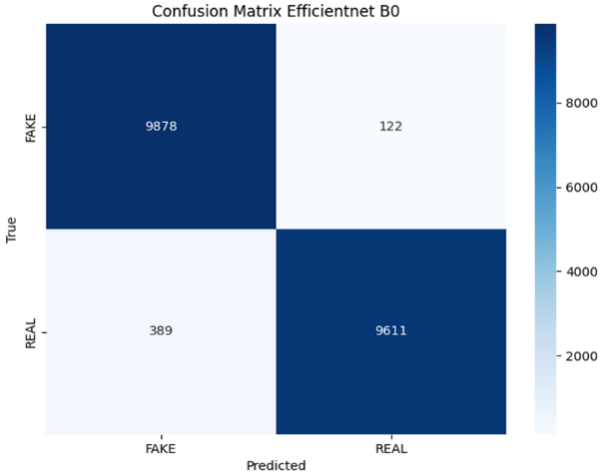
Mô hình	Độ chính xác (Accuracy)	Tỷ lệ trúng (Precision)	Độ nhạy (Recall)	F1-Score
ResNet-50	0,9695	0,9719	0,9669	0,9694
EfficientNet-B0	0,9745	0,9875	0,9611	0,9741

Swin Transformer	0,9615	0,9733	0,9489	0,9610

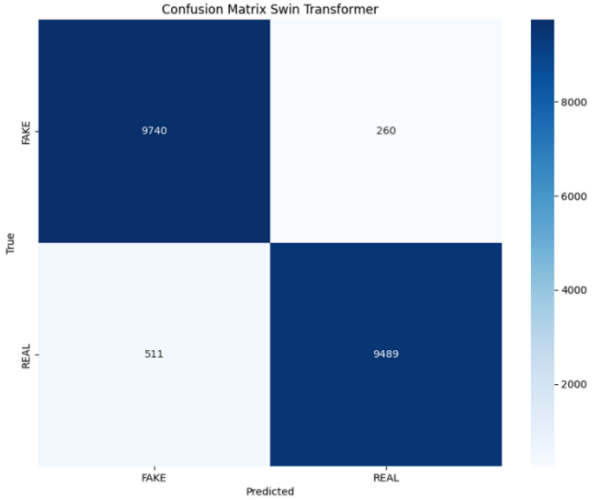
Một chỉ số quan trọng khác được sử dụng trong nghiên cứu này để đánh giá các mô hình là ma trận nhầm lẫn, được trình bày trong Hình 3.9, 3.10 và 3.11 tương ứng với các mô hình ResNet-50, EfficientNet-B0 và Swin Transformer.



Hình 0.5: Phân tích ma trận nhầm lẫn của mô hình ResNet-50



Hình 0.6: Phân tích ma trận nhầm lẫn của mô hình EfficientNet-B0



Hình 0.7: Phân tích ma trận nhầm lẫn của mô hình Swin Transformer

Kết quả từ ma trận nhầm lẫn cho thấy EfficientNet-B0 là mô hình có hiệu suất phân loại theo lớp tốt nhất, với số lượng lỗi thấp nhất trong ba mô hình khảo sát. Mô hình này chỉ ghi nhận 122 ảnh Deepfake bị phân loại nhầm là thật, và 389 ảnh thật bị phân loại nhầm là giả. ResNet-50 có tổng số lỗi ở mức trung bình, trong khi Swin Transformer ghi nhận số lượng lỗi cao nhất, đặc biệt là tỷ lệ ảnh thật bị nhầm thành Deepfake (511 trường hợp) – yếu tố cần được lưu ý trong các hệ thống yêu cầu độ tin cậy cao. Những kết quả này khẳng định hiệu quả nhận diện của EfficientNet-B0, đặc biệt trong việc giảm thiểu false negative, điều rất quan trọng đối với bài toán phát hiện ảnh giả mạo.

3.8. Kết luận chương

Trong Chương 3, đề án đã trình bày các nội dung về triển khai thực hiện mô hình học sâu trong phát hiện ảnh DeepFake. Nội dung chương đã trình bày sơ đồ các khối trong mô hình học sâu, mô tả các thành phần trong mô hình: thu thập dữ liệu, trình bày ba mô hình học sâu áp dụng gồm: ResNet-50, EfficientNet-B0, Swin Transformer. Tiếp đó, bài đã trình bày các tiêu chí đánh giá hiệu năng cho các mô hình, môi trường và các tham số dùng cho thử nghiệm. Các kết quả thử nghiệm cho ba mô hình cho thấy, với tập dữ liệu CIFAKE sử dụng, mô hình EfficientNet cho độ chính xác cao hơn so với hai mô hình còn lại. Kết quả từ ma trận nhầm lẫn cho thấy EfficientNet-B0 là mô hình có hiệu suất phân loại theo lớp tốt nhất, với số lượng lỗi thấp nhất trong ba mô hình khảo sát. Ngoài ra, đề án cũng minh họa kết quả xác thực ảnh thật/giả của ứng dụng phát hiện ảnh Deepfake - ứng dụng xây dựng dựa trên kết quả thực nghiệm của đề án.

KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN TIẾP

Kết luận

Phát hiện ảnh Deepfake là một nhu cầu thực tiễn và là vấn đề kỹ thuật khó, có nhiều thách thức do các ảnh DeepFake do AI tạo ra rất đa dạng. Ảnh Deepfake đang bị lợi dụng cho các mục đích xấu, như lừa đảo, phát tán thông tin sai lệch, xâm phạm quyền riêng tư và gây ảnh hưởng đến danh dự cá nhân. Sự lan truyền nhanh chóng của ảnh Deepfake đặt ra những thách thức lớn đối với tính xác thực của thông tin, an ninh mạng, pháp lý và đạo đức xã hội. Việc nghiên cứu, phát triển các phương pháp và xây dựng giải pháp phát hiện ảnh Deepfake trở thành một yêu cầu cấp thiết hiện nay.

Các phương pháp và mô hình học sâu tỏ ra rất hiệu quả trong việc xử lý và phân loại hình ảnh thông qua các kiến trúc mạng nơ-ron nhân tạo tiên tiến như Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs) và Transformers [11-13, 15-17]. Các mô hình học sâu có khả năng tự động trích xuất đặc trưng từ hình ảnh mà không cần phải xác định trước các đặc điểm quan trọng, từ đó giúp cải thiện đáng kể độ chính xác trong phát hiện ảnh Deepfake. Trên cơ sở đó, đề án tốt nghiệp hướng tới mục tiêu: nghiên cứu, ứng dụng phương pháp học sâu vào xây dựng giải pháp phát hiện ảnh Deepfake.

Các kết quả chính đã đạt được trong bài gồm:

- Nghiên cứu về công nghệ AI sử dụng tạo ảnh DeepFake; Các mô hình học máy và học sâu; Các đặc trưng của ảnh DeepFake do AI tạo ra; Phương pháp nhận diện và phát hiện ảnh DeepFake; Các phương pháp học sâu trong phát hiện ảnh DeepFake.

- Đưa ra được một mô hình áp dụng phương pháp học sâu vào phát hiện DeepFake do AI tạo ra.
- Thực hiện thử nghiệm mô hình, đánh giá mô hình bằng các chỉ số Accuracy, Precision, Recall, F1-score và so sánh đánh giá ba mô hình tiêu biểu là ResNet-50, EfficientNet-B0 và Swin Transformer.

Từ kết quả thực nghiệm và phân tích hiệu năng, có thể thấy rằng cả ba mô hình ResNet-50, EfficientNet-B0 và Swin Transformer đều đạt hiệu quả cao trong bài toán phát hiện ảnh Deepfake. Tuy nhiên, xét trên các tiêu chí toàn diện như độ chính xác, F1-score và đặc biệt là khả năng giảm thiểu sai sót trong ma trận nhầm lẫn, EfficientNet-B0 là mô hình tối ưu nhất. Mô hình này không chỉ đạt độ chính xác và F1-score cao nhất mà còn có số lượng ảnh deepfake bị nhầm là ảnh thật thấp nhất, đồng thời giữ mức sai lệch tổng thể nhỏ hơn so với các mô hình còn lại. Đây là yếu tố then chốt trong các hệ thống phát hiện nội dung giả mạo, nơi việc bỏ sót deepfake (false negative) có thể dẫn đến hậu quả nghiêm trọng. Tuy nhiên, kết quả cũng cho thấy rằng chưa có mô hình nào đạt hiệu suất tuyệt đối, đặc biệt là trong việc duy trì độ nhạy (recall) và hạn chế false positive ở mức tối thiểu.

Hướng phát triển tiếp

Mặc dù bài đã đạt được một số kết quả khả quan, tuy nhiên vẫn còn một số vấn đề cần tiếp tục giải quyết trong thời gian tới như sau:

- Tăng cường kỹ thuật tiền xử lý ảnh, đặc biệt là các phương pháp làm nổi bật chi tiết khuôn mặt hoặc phát hiện hiện vật bất thường (artifacts) trong ảnh deepfake.

- Kết hợp nhiều mô hình (ensemble learning) hoặc sử dụng kiến trúc lai giữa CNN và Transformer để khai thác ưu điểm của từng loại mô hình.
- Phân tích sâu về lỗi sai, đặc biệt là false positive đối với ảnh thật, nhằm giảm thiểu các cảnh báo sai trong ứng dụng thực tiễn.
- Khám phá thêm các tập dữ liệu mới, có tính đa dạng cao hơn về nguồn ảnh, phong cách sinh ảnh (GAN, diffusion), để nâng cao tính tổng quát hóa cho mô hình.

Tóm lại, EfficientNet-B0 là lựa chọn tối ưu trong bối cảnh nghiên cứu hiện tại, nhưng việc cải tiến mô hình và mở rộng phạm vi thử nghiệm là cần thiết để hướng tới các hệ thống phát hiện ảnh Deepfake có độ tin cậy cao hơn trong môi trường thực tế.