

HỆ THỐNG NHẬN DẠNG ẢNH MẶT NGƯỜI TẠO RA BỞI TRÍ TUỆ NHÂN TẠO A SYSTEM FOR DETECTING AI GENERATED HUMAN FACES IMAGES

SVTH: Cao Kiều Văn Mạnh, Nguyễn Tuấn Hưng, Võ Hoàng Bảo

Lớp 20TCLC_KHDL, Khoa Công Nghệ Thông Tin, Trường Đại Học Bách Khoa – Đại Học Đà Nẵng;

Email: caomanh.qng2019@gmail.com, tuanhungg26@gmail.com, hoangbao022002dn@gmail.com

GVHD: TS. Ninh Khánh Duy

Khoa Công Nghệ Thông Tin, Trường Đại Học Bách Khoa – Đại Học Đà Nẵng; Email: nkduy@dut.udn.vn

Tóm tắt

Với sự phát triển mạnh mẽ của trí tuệ nhân tạo (AI), theo sau đó là các mô hình trí tuệ nhân tạo tạo sinh ảnh, việc phân biệt hình ảnh do AI tạo ra và hình ảnh thật ngày càng khó khăn. Sự phát triển của công nghệ trên đã đặt ra nhiều mối nguy hại về độ tin cậy và an toàn trong nhiều lĩnh vực như bảo mật thông tin, quyền riêng tư, quyền sở hữu trí tuệ,... Điều này còn đặc biệt quan trọng đối với các dữ liệu liên quan đến danh danh như hình ảnh khuôn mặt con người. Hiện nay, tuy đã có nhiều nghiên cứu nhằm phân biệt hình ảnh thực tế và hình ảnh do trí tuệ nhân tạo sinh ra, tuy nhiên các mô hình vẫn chưa đạt được kết quả quá cao với dữ liệu từ các mô hình sinh ảnh mới, chưa đủ tập trung vào mảng dữ liệu khuôn mặt con người, cũng như vẫn chưa được áp dụng vào trong một hệ thống thực tế. Do đó, trong nghiên cứu này, nhóm chúng tôi tiến hành xây dựng một bộ dữ liệu ảnh mặt người do AI tạo ra từ các mô hình sinh ảnh mới nhất. Bên cạnh đó, chúng tôi cũng đề xuất một số mô hình học sâu dựa trên các kiến trúc Convolutional Neural Network và các kỹ thuật Transfer Learning - Ensemble Learning nhằm phát hiện ảnh mặt người thực tế và ảnh mặt người do trí tuệ nhân tạo tạo ra. Ngoài ra, chúng tôi cũng tiến hành xây dựng hệ thống website để tích hợp kết quả nghiên cứu, giúp người dùng có thể dễ dàng tương tác và sử dụng. Qua quá trình thử nghiệm và đánh giá trên tập dữ liệu nhóm xây dựng, kết quả cho thấy sự hiệu quả của các mô hình với độ chính xác cao nhất lên đến 97.91% cùng với hệ thống có tốc độ xử lý nhanh, trung bình xử lý 7 - 8s/ảnh cho quá trình phân loại và khoảng 70 - 74s/ảnh cho quá trình giải thích kết quả phân loại.

Từ khóa: Trí tuệ nhân tạo tạo sinh; Ảnh do trí tuệ nhân tạo sinh ra; Học sâu; Mô hình phân loại hình ảnh.

1. Giới thiệu

Các mô hình sinh ảnh ảo sử dụng các mô hình học sâu ngày càng phát triển mạnh mẽ với chất lượng ảnh vô cùng chân thực và gần như không thể xác định bằng mắt thường. Điều này là vô cùng nguy hiểm đối với đang dữ liệu định danh như khuôn mặt con người. Một vài ví dụ tiêu biểu có thể nhắc đến là các mô hình dựa trên kiến trúc GANs hay các mô hình Diffusions đã trở thành một thách thức lớn trong các lĩnh vực như xác minh danh tính trực tuyến và bảo vệ quyền riêng tư, quyền sở hữu trí tuệ,...

Hầu hết các phương pháp phân loại ảnh mặt người do AI tạo ra hiện nay đều sử dụng các kiến trúc mạng nơ-ron học sâu để tự động trích xuất các đặc trưng khác nhau từ

Abstract

With the rapid development of Artificial Intelligence (AI), followed by the emergence of AI-generated images, distinguishing between images created by AI and real images has become increasingly challenging. Technological advancements have raised concerns about reliability and safety in various domains such as information security, privacy, copyrights, etc. This is particularly crucial for identity-related data, such as human facial images. Currently, despite numerous studies focusing on building deep learning models to differentiate between real and AI-generated images, these models have not yet achieved consistently high accuracy with data from newer image generation models. There is also a limited focus on human facial data, and existing research has not been widely applied in real-world systems. In this study, we undertook the construction of a dataset consisting of human facial images generated by the latest AI image generation models. We propose several deep learning models based on Convolutional Neural Network architectures and Transfer Learning - Ensemble Learning techniques to detect AI-generated human facial images and real human facial images. Additionally, we develop a website system to integrate research results, allowing users easy interaction and utilization. Through testing and evaluation on the constructed dataset, the results demonstrate the effectiveness of the models, with the highest accuracy reaching up to 97.91%. The system also exhibits fast processing speeds, averaging 7-8 seconds per image for the classification process and around 70 - 74 seconds per image for result explanation.

Key words: Generative AI, AI-Synthesized images, Deep Learning; Image classification model.

mặt người và đưa vào một bộ phân loại nhị phân để phân biệt giữa ảnh thực và ảnh ảo. Các mô hình này cho hiệu quả tốt trên dữ liệu của các mô hình sinh ảnh đã được huấn luyện, tuy nhiên độ hiệu quả của các phương pháp này có thể giảm khi áp dụng vào các ứng dụng thực tế, với dữ liệu từ các nguồn hay các mô hình chưa biết. Để cải thiện những vấn đề trên, trong nghiên cứu này chúng tôi tiến hành thu thập một bộ dữ liệu ảnh mặt người do AI tạo ra từ các mô hình sinh ảnh mới và có chất lượng cao nhất hiện nay. Sau đó, nhóm sẽ tiến hành thử nghiệm xây dựng các mô hình phân loại ảnh giả dựa trên các kiến trúc mạng CNN và các kỹ thuật như Transfer learning, Ensemble learning, data augmentation,... Ngoài ra, nhóm

còn tiến hành xây dựng thêm một hệ thống tích hợp mô hình và cung cấp dịch vụ phân loại ảnh do AI tạo ra. Với hệ thống này và dữ liệu được xác thực bởi người dùng, nhóm có thể tiếp tục cải thiện độ chính xác của mô hình.

Bên cạnh đó, chúng tôi cũng đã nghiên cứu và tiến hành tích hợp các kỹ thuật giải thích kết quả của các mô hình học sâu với hai phương pháp là Gradient-weighted Class Activation Mapping (Grad-CAM) và Locally Interpretable Model-Agnostic Explanations (LIME).

Những đóng góp chính của nghiên cứu:

- (1) Thu thập và xây dựng bộ dữ liệu từ các nguồn khác nhau, tạo nên bộ dữ liệu chứa lên đến 70,000 ảnh mặt người thật và hơn 50,000 ảnh mặt người do AI sinh ra.
- (2) Thủ nghiệm các mô hình học sâu phổ biến hiện nay như VGG, MobileNet, EfficientNet, ResNet, DenseNet, RegNet,... và kết hợp với các kỹ thuật Transfer Learning - Ensemble Learning để xây dựng mô hình phân loại ảnh giả với độ chính xác cao.
- (3) Triển khai tích hợp mô hình vào hệ thống và cung cấp dịch vụ giúp người dùng dễ dàng tương tác và sử dụng, đồng thời kết hợp các kỹ thuật để giải thích kết quả dự đoán của mô hình.

2. Nghiên cứu liên quan

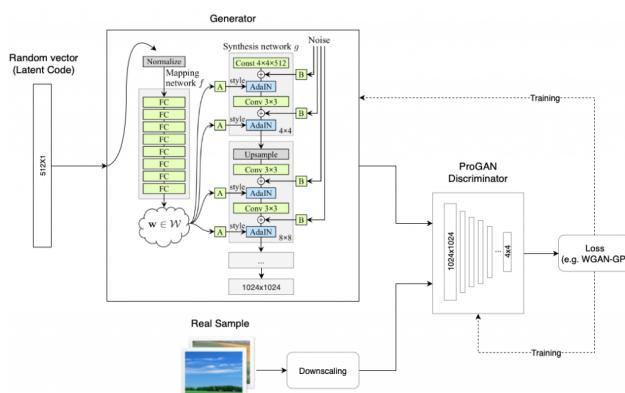
Trong phần này, nhóm chúng tôi tiến hành khảo sát các nghiên cứu liên quan đến các mô hình sinh ảnh hiện nay cũng như các nghiên cứu về kỹ thuật nhận diện ảnh do trí tuệ nhân tạo sinh ra. Ngoài ra, nhóm còn tiến hành khảo sát các phương pháp giúp giải thích kết quả dự đoán mà mô hình học sâu đưa ra.

2.1. Các mô hình sinh ảnh hiện nay

Hầu hết các mô hình sinh ảnh hiện nay đều được xây dựng dựa trên Kiến trúc GANs. Nhìn chung, kiến trúc GANs bao gồm 2 mạng chính:

- The generator network với mục tiêu là sản xuất ra các ảnh khó phân biệt so với dữ liệu thực tế.
- The discriminator network với mục tiêu cô gắng phân biệt giữa dữ liệu thật được đưa vào huấn luyện và dữ liệu được sinh ra bởi generator.

Một số biến thể của GAN có thể kể đến là: DCGANs, CGANs, CycleGANs, StyleGANs,... Trong đó, StyleGANs [13] là mô hình sinh ảnh có chất lượng chi tiết tốt nhất và là mô hình phổ biến nhất được sử dụng để sinh ảnh mặt người giả. Bên cạnh đó, StyleGAN cũng thường được dùng để cải thiện ảnh mặt người dựa trên hình ảnh ban đầu sinh ra từ các mô hình khác như Diffusion.



Hình 1. Kiến trúc StyleGAN [13]

2.2. Các phương pháp nhận diện ảnh do trí tuệ nhân tạo sinh ra

Các nghiên cứu dùng để nhận diện ảnh do Trí tuệ nhân tạo sinh ra đã và đang được phát triển bằng nhiều phương pháp khác nhau và được chia thành ba hướng nghiên cứu chính: dựa trên các kỹ thuật thị giác máy tính (Computer Vision), dựa trên các mạng học sâu CNN và dựa trên các kiến trúc học sâu riêng biệt của một số phương pháp mới gần đây.

2.2.1. Các phương pháp sử dụng Computer Vision

Với phương pháp dựa trên các kỹ thuật Computer Vision sẽ chủ yếu sử dụng các đặc trưng nông của ảnh như màu sắc, tần số và sử dụng các kỹ thuật Computer Vision như Histogram Color,... để tiến hành phân loại.

Trong nghiên cứu [8], tác giả đã tiến hành phân tích thống kê thành phần màu của ảnh ảo và phân biệt với ảnh thật. Hay như trong nghiên cứu [9], tác giả nhận diện dựa trên màu sắc và độ bão hòa. Họ cho rằng dựa trên màu sắc, ảnh do AI sinh ra sẽ có độ tương quan giữa các pixel với nhau trong không gian sắc độ khác với ảnh thật.

Điểm mạnh của các phương pháp trên là mô hình nhỏ, nhẹ, tốc độ xử lý nhanh và đạt được kết quả phân loại tốt với các mô hình sinh ảnh đời đầu. Tuy nhiên, với các mô hình, kiến trúc sinh ảnh mới gần đây thì các phương pháp chỉ dựa trên Computer Vision gần như không thể xử lý tốt.

2.2.2. Các phương pháp sử dụng kiến trúc CNN

Các mạng học sâu CNN đã được áp dụng thành công trong nhiều lĩnh vực, bao gồm cả bài toán phân loại ảnh thật và ảnh do AI sinh ra. Các phương pháp này thường sử dụng các mô hình CNN sâu để học các đặc trưng phức tạp của ảnh, từ đó có thể phân biệt được ảnh giả.

Trong [10], các tác giả đã sử dụng mô hình học máy phân loại dựa trên ResNet18 để phát hiện hình ảnh được tạo bằng Stable diffusion với prompt (chuyển văn bản thành hình ảnh). Cụ thể, họ đã thay đổi một số thành phần của ResNet18 để cải thiện khả năng phát hiện hình ảnh giả.

Các phương pháp sử dụng kiến trúc CNN thường có độ chính xác cao hơn các phương pháp chỉ sử dụng Computer Vision. Tuy nhiên, các phương pháp này thường yêu cầu lượng dữ liệu huấn luyện lớn và được cập nhật thường xuyên để đảm bảo độ chính xác cao trên các mô hình sinh ảnh mới.

2.2.3. Các phương pháp gần đây

Gần đây, một số phương pháp mới đã được đề xuất để giải quyết bài toán phân loại ảnh thật và ảnh do AI sinh ra. Các phương pháp này thường sử dụng các kỹ thuật mới như học sâu, học máy tự động, v.v... để cải thiện độ chính xác và hiệu quả của hệ thống.

Cụ thể, trong [1], tác giả đã kết hợp đặc trưng lớp convolutional nông và sâu của một bức ảnh (nhánh global) cùng với các đặc trưng của các patch được chia nhỏ từ bức ảnh (nhánh local). Ở nhánh global, tác giả sử dụng kỹ thuật "Attention-based Multi-Scale Feature Fusion (AMSSFF)" để trích xuất các đặc trưng tổng thể như: màu sắc chung của ảnh, tổng cộng độ sáng, tỷ lệ khung hình,... Tiếp theo, tác giả sử dụng kỹ thuật Patch Selection Module để chọn lọc các khung ảnh quan trọng để trích xuất đặc trưng cục bộ như sự mất cân đối hoặc sai lệch của các chi tiết. Đặc trưng của cả hai nhánh sau cùng được kết hợp lại bởi Attentional Feature Fusion Module

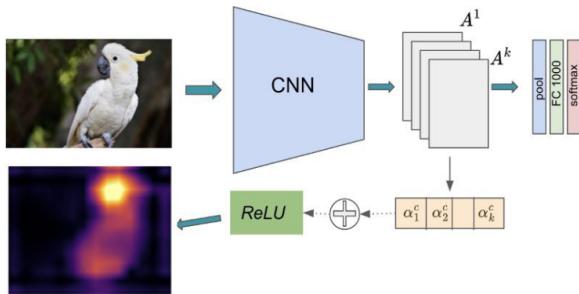
và được đưa vào hệ thống phân loại nhị phân để phân biệt ảnh giả và ảnh thật.

Các phương pháp gần đây đã đạt được những thành tựu đáng kể trong việc cải thiện độ chính xác và hiệu quả của hệ thống phân loại ảnh thật và ảnh do AI sinh ra. Tuy nhiên, đây vẫn là một bài toán khó và cần tiếp tục được nghiên cứu để đạt được kết quả tốt hơn.

2.3. Các phương pháp giải thích kết quả của mô hình học sâu

2.3.1. Phương pháp GRAD - CAM

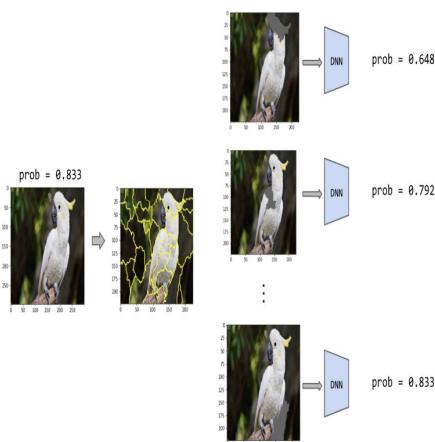
Grad-CAM là một trong những kỹ thuật giải thích đầu tiên được phát triển cho các mô hình xử lý ảnh và có thể được áp dụng lên các bất kỳ mạng CNN nào. Grad-CAM là một kỹ thuật giải thích hậu xử lý và không yêu cầu bất kỳ thay đổi nào về kiến trúc hay huấn luyện [7]. Thay vào đó, Grad-CAM truy cập vào lớp tích chập bên trong của mô hình để xác định khu vực có ảnh hưởng cao nhất đến sự dự đoán của mô hình. Bởi vì Grad-CAM chỉ dựa vào các bước chuyển tiếp qua mô hình, không có lan truyền ngược nên nó cũng có hiệu quả về mặt tính toán. Kết quả của Grad-CAM là một heatmap đánh dấu khu vực ảnh hưởng đến lớp phân loại tương ứng.



Hình 2. Sơ đồ tổng quát quá trình hoạt động của phương pháp Grad-CAM [7]

2.3.2. Phương pháp LIME

LIME là một trong những kỹ thuật giải thích phổ biến nhất. LIME được sử dụng sau khi đã huấn luyện xong mô hình. Về bản chất, LIME coi mô hình đã huấn luyện như một API, lấy các mẫu và tạo ra các giá trị dự đoán [7]. LIME giải thích bằng sự nhiễu loạn xảy ra ở cấp độ đặc trưng của đầu vào. Bằng cách này, những pixel và vùng pixel có ảnh hưởng cao nhất với sự dự đoán mô hình được đánh dấu ràng tăng hay giảm sự dự đoán của mô hình với đầu vào đã cho. Kết quả cuối cùng thể hiện các vùng có ảnh hưởng cao nhất đối với lớp phân loại được chọn.



Hình 3. Sơ đồ mô tả quá trình hoạt động của phương pháp LIME [7]

3. Phương pháp

Các mục tiêu chính của dự án lần này bao gồm: (1) xây dựng được bộ dữ liệu ảnh mặt người thật và ảnh mặt người do AI tạo ra, (2) thử nghiệm xây dựng các mô hình học sâu nhằm phân loại ảnh mặt người thật và giả và (3) xây dựng hệ thống tích hợp mô hình cung cấp dịch vụ cho người dùng.

3.1. Thu thập dữ liệu

3.1.1. Thu thập ảnh mặt người thật

Với dữ liệu ảnh khuôn mặt người thật, nhóm sử dụng bộ dữ liệu Flickr-Faces-HQ Dataset (FFHQ) [12]. FFHQ là bộ dữ liệu mặt người do NVIDIA công bố vào năm 2019, bao gồm khoảng 70.000 ảnh mặt người có kích thước 1024 x 1024 với chất lượng cao và đa dạng về quốc tịch, độ tuổi, nền ảnh, nhiều loại phụ kiện như mũ, kính mát và kính râm trong ảnh.

Kể từ khi phát hành, bộ dữ liệu này đã trở thành bộ dữ liệu khuôn mặt được sử dụng rộng rãi nhất cho nhiều ứng dụng nghiên cứu và thương mại khác nhau, từ nhận dạng khuôn mặt đến nhận dạng giới tính và đặc biệt là dùng để huấn luyện các mô hình sinh ảnh mặt người giả.



Hình 4. Một số ảnh từ bộ FFHQ [12]

3.1.2. Thu thập ảnh mặt người giả

Sau quá trình tìm hiểu, các nguồn ảnh mặt người do AI tạo ra có chất lượng tốt mà nhóm có thể tiếp cận được bao gồm:

- *thispersondoesnotexist.com*: website này sẽ trả về một ảnh mặt người giả được sinh ra từ StyleGAN2 sau mỗi lần làm mới.
- *SFHQ phần 1* [11]: ảnh được tạo ra bằng cách đưa các tranh vẽ, tranh 3D vào mô hình StyleGAN2 để tăng độ chân thật.
- *SFHQ phần 2* [11]: lấy các ảnh 3D và các ảnh sinh ra từ Stable Diffusion 1.4 và tiến hành cải thiện độ chân thật với StyleGAN2.
- *SFHQ phần 3* [11]: chứa các ảnh được sinh trực tiếp từ StyleGAN2.
- *SFHQ phần 4* [11]: dùng Stable Diffusion 2.1 để chuyển văn bản thành hình ảnh và dùng StyleGAN để cải thiện chất lượng.
- *generated.photos*: website cung cấp nhiều hình ảnh mặt người giả được tạo ra từ StyleGAN.
- *Stable Diffusion*: được sinh bằng cách sử dụng prompt và mô hình Stable Diffusion XL 1.0.

Trong các nguồn ảnh trên, nhóm tiến hành tự thu thập dữ liệu từ các nguồn *thispersondoesnotexist.com*, *generated.photos*. Với Stable Diffusion, nhóm xây dựng một bộ prompt đa dạng và tiến hành sinh ảnh sử dụng mô hình Stable Diffusion XL 1.0.

Đối với dữ liệu từ bộ SFHQ nhóm sử dụng bộ dữ liệu có sẵn trên Kaggle, với mỗi phần, nhóm tiến hành lấy ra 5000 ảnh để sử dụng.

Bảng 1. Danh sách các bộ dữ liệu ảnh mặt người giả

Dataset/ Data source	Số lượng ảnh thu thập/ sử dụng	Kích thước gốc của ảnh	Độ chân thật của hình ảnh
thispersondoesnotexist.com	20,000	1024 x 1024	Cao
SFHQ phần 1	5,000	1024 x 1024	Thấp
SFHQ phần 2	5,000	1024 x 1024	Thấp
SFHQ phần 3	5,000	1024 x 1024	Trung bình
SFHQ phần 4	5,000	1024 x 1024	Cao
generated.photos	10,000	256 x 256	Cao
Stable Diffusion	1,000	1024 x 1024	Trung bình

3.1.3. Tiền xử lý dữ liệu

Trong quá trình khảo sát, nhóm nhận thấy dữ liệu thu thập từ website generated.photos có nền chủ yếu là các màu đơn sắc như trắng, xám, nâu,... Điều này có thể dẫn đến mô hình bị thiên kiêng với các ảnh có nền đơn sắc. Bên cạnh đó, tuy có chất lượng sinh ảnh giả cao nhưng độ phân giải của ảnh thu tập được trên website này khá thấp (256x256). Để giải quyết vấn đề trên, nhóm tiến hành xử lý bằng cách sử dụng ảnh thật từ bộ FFHQ và tiến hành swap/blend khuôn mặt giả từ bộ ảnh generated.photos vào các ảnh thật này. Lúc này ta có một ảnh mới với background ảnh là thật và khuôn mặt giả mạo.

Theo đánh giá chủ quan, phương pháp này có thể giúp đa dạng hóa thêm bộ dữ liệu và khắc phục vấn đề bias của mô hình như đã đề cập cũng như cải thiện điểm yếu của bộ dữ liệu generated.photos là độ phân giải thấp. Trong nghiên cứu này, nhóm sử dụng API Face Dancer của tác giả Felix Rosberg [2] vì API này có tốc độ xử lý khá nhanh, chất lượng swap cao và miễn phí.

**Hình 5. Xử lý nền cho các ảnh giả từ generated.photos**

Trước khi sử dụng cho quá trình huấn luyện, nhóm tiến hành một số bước tiền xử lý dữ liệu sau:

- Toàn bộ dataset sẽ được resize về kích thước chung 512x512 và nén về định dạng JPEG nhằm giảm kích thước dữ liệu cho phù hợp với tài nguyên huấn luyện hiện có của nhóm.
- Chuẩn hóa dữ liệu: đưa giá trị các pixel về phạm vi [0, 1]. (Với một số kiến trúc CNN mà nhóm thử nghiệm như kiến trúc MobileNet, EfficientNet,... có thể bỏ qua bước này vì trong kiến trúc đã có sẵn lớp chuẩn hóa dữ liệu).
- Tiếp theo, nhóm tiến hành làm giàu dữ liệu nhằm mô phỏng các thay đổi trên ảnh mà có thể xảy ra trong thực tế như xoay, lật, tăng giảm độ sáng, tương phản,... (Chỉ thực hiện tăng cường dữ liệu trên tập huấn luyện).

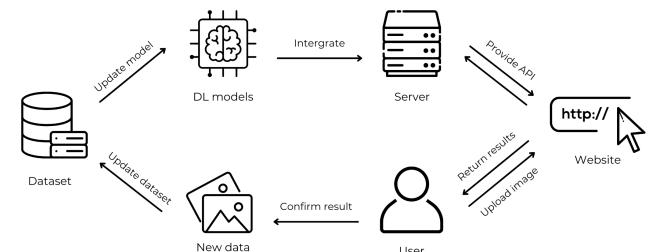
Sau quá trình tiền xử lý, dữ liệu sẽ được chia thành 3 bộ train/test/validate với tỷ lệ chia 6/2/2 và dữ liệu ảnh giả sẽ được chia đều cho các nguồn ảnh giả khác nhau.

3.2. Xây dựng hệ thống

3.2.1. Cấu trúc hệ thống và dịch vụ

Hệ thống cung cấp dịch vụ bao gồm các thành phần chính sau:

- Server: được xây dựng dựa trên Framework FastAPI.
- Website clients: được thiết kế sử dụng HTML, CSS và JavaScript.
- Mô hình phân loại ảnh do AI tạo ra.

**Hình 6. Sơ đồ hoạt động của hệ thống dịch vụ**

3.2.2. Face detection

Vì mô hình phân loại chỉ hỗ trợ nhận diện ảnh có đúng một mặt người. Do đó, nhóm tiến hành tích hợp thêm kỹ thuật **Haar Cascade OpenCV Python** nhằm phát hiện khuôn mặt và loại bỏ hai trường hợp: ảnh không có mặt người và ảnh có nhiều hơn một mặt người.

3.2.3. Tính năng crop image (cắt ảnh)

Qua quá trình khảo sát thực tế, có hai vấn đề về ảnh đầu vào gây ảnh hưởng tới kết quả dự đoán của mô hình như ảnh có khuôn mặt nhỏ hơn nhiều so với kích thước ảnh tổng thể hoặc ảnh hình chữ nhật khi tiến hành thay đổi kích thước trước khi đưa vào mô hình sẽ làm ảnh sẽ bị biến dạng đáng kể hoặc khiến ảnh chứa các dữ liệu không mong muốn.

Giải pháp nhóm đưa ra cho hai vấn đề này đó là tính năng crop image (cắt ảnh) trên giao diện người dùng. Trang này sẽ cho phép người dùng phóng to, thu nhỏ và cắt ảnh gốc thành ảnh mới có kích thước vuông. Với phương pháp này, người dùng có thể dễ dàng điều chỉnh vùng ảnh muốn phân loại mà không ảnh hưởng đến kết quả dự đoán của mô hình.

3.2.4. Lưu lại feedback của người dùng

Nhóm tiến hành lưu lại kết quả trong quá trình hoạt động của hệ thống. Kết quả lưu lại bao gồm: ảnh, kết quả mô hình dự đoán và kết quả xác nhận từ người dùng về kết quả dự đoán. Từ những feedback và dữ liệu từ người dùng trong quá trình hoạt động của hệ thống. Nhóm có thể tiếp tục tiến hành cải thiện kết quả của mô hình.

3.3. Giải pháp phân biệt ảnh mặt người thật và ảnh do AI tạo ra sử dụng deep learning

3.3.1. Transfer Learning và Ensemble Learning

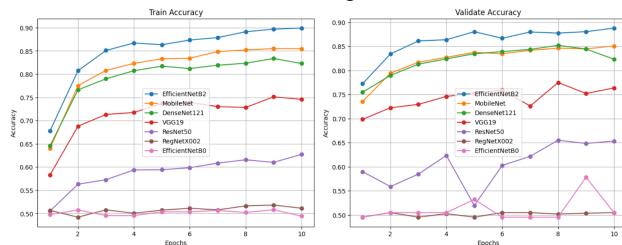
Transfer learning là kỹ thuật huấn luyện tái sử dụng một mô hình đã được huấn luyện như điểm khởi đầu cho một mô hình mới với một nhiệm vụ khác. Ứng dụng transfer learning có thể giúp cải thiện độ chính xác của mô hình và đồng thời giảm thiểu thời gian huấn luyện.

Ensemble learning là phương pháp giúp tăng độ chính xác trên tập dữ liệu bằng cách kết hợp một số mô hình với nhau. Có nhiều kỹ thuật ensemble learning khác nhau và trong nghiên cứu này, nhóm tiến hành thử nghiệm hai kỹ thuật Concatenation và Average Ensemble.

3.3.2. Các kiến trúc CNN phổ biến cho bài toán Image Classification

Nhóm tiến hành lựa chọn các mô hình CNN phổ biến

với bài toán phân loại ảnh bao gồm: VGG19, MobileNetV3, EfficientNetB0, EfficientNetB2, ResNet50, DenseNet121, RegNetX002 và thử nghiệm với kỹ thuật Transfer Learning. Nhóm sử dụng một tập dữ liệu nhỏ để thử nghiệm bao gồm 7000 ảnh mỗi nhãn ảnh thật và ảnh do AI tạo ra với 10 epochs.



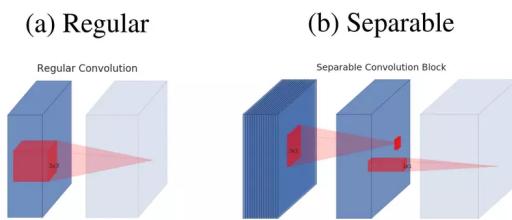
Hình 7. Kết quả thử nghiệm mô hình

Việc transfer Learning với các kiến trúc CNN bước đầu cho kết quả khả quan. Nhóm tiếp tục lựa chọn hai mô hình có kết quả tốt nhất ở đây là MobileNetV3 và EfficientNetB2 để tiếp tục phát triển cho nghiên cứu.

3.3.3. Xây dựng mô hình hệ thống

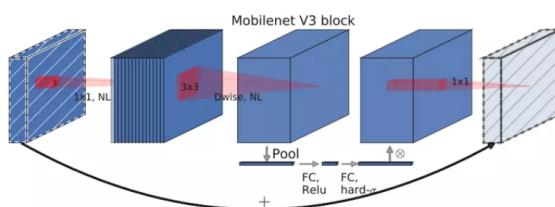
❖ Kiến trúc mô hình MobileNetV3

MobileNet là mô hình CNN được thiết kế với mục đích trở nên gọn nhẹ để ứng dụng vào các thiết bị di động và thiết bị nhúng.



Hình 8. Sự khác biệt của Separable Convolution [4]

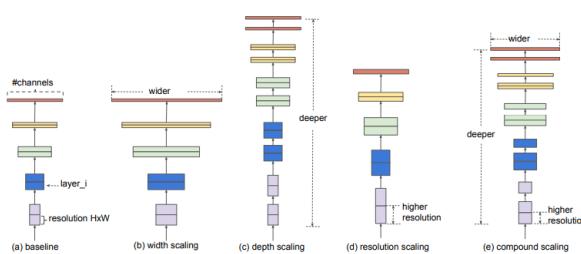
Depthwise separable convolution ở MobileNet đã giảm khối lượng tính toán và giảm số lượng tham số [3], đồng thời có thể thực hiện trích xuất đặc trưng một cách riêng biệt trên từng channel. MobileNetV2 để xuất thêm Linear Bottlenecks giúp giảm kích thước input và Inverted Residual Block giúp tăng độ chính xác của mô hình mà không cần đến chi phí lớn [4]. MobileNetV3 tiếp tục cải tiến bằng việc dùng Squeeze and Excite nhằm tăng lượng thông tin giữa các kênh [5].



Hình 9. MobileNet V3 [5]

❖ Kiến trúc mô hình EfficientNet

EfficientNet xoay quanh khái niệm thu phóng mô hình (Model Scaling) với các chiều sâu, rộng, phân giải.



Hình 10. Model Scaling [6]

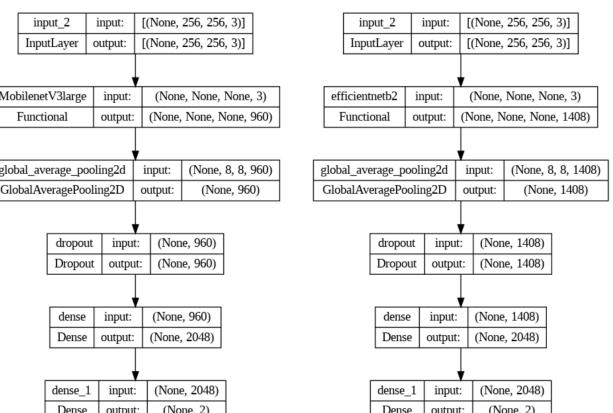
Mỗi việc thu phóng riêng lẻ có thể không tối ưu với đa số các bài toán mà còn mang hiệu ứng ngược. Vì vậy để đạt được độ chính xác và hiệu quả tốt hơn, điều quan trọng là phải cân bằng tất cả các kích thước của chiều rộng, chiều sâu và độ phân giải mạng trong quá trình thu phóng quy mô mạng nơ ron tích chập [6]. Từ đó các mô hình phiên bản của EfficientNet được phát triển bằng cách thu phóng các mô hình phổ biến như MobileNet hay ResNet để cải thiện độ chính xác của các mô hình này.

❖ Các mô hình đề xuất

Sử dụng backbone là kiến trúc MobileNetV3 và EfficientNet B2, nhóm tiến hành xây dựng 4 mô hình sử dụng hai kỹ thuật chính là Transfer Learning và Ensemble Learning.

Với kỹ thuật Transfer Learning, nhóm sử dụng mô hình pretrained của hai kiến trúc trên làm bộ trích xuất đặc trưng. Sau đó nhóm tiến hành thêm các lớp Global Average Pool, Dropout, Fully Connected và Softmax vào sau bộ trích xuất đặc trưng. Từ đó nhóm xây dựng được hai mô hình:

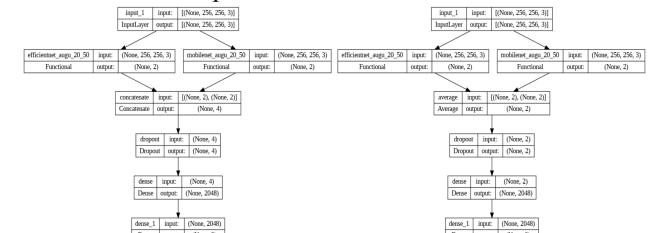
- **Mô hình 1:** sử dụng kỹ thuật Transfer Learning với backbone MobileNetV3.
- **Mô hình 2:** sử dụng kỹ thuật Transfer Learning với backbone EfficientNet B2.



Hình 11. Cấu trúc mô hình 1 (trái) và mô hình 2 (phải)

Sau khi huấn luyện hai mô hình 1 và mô hình 2, nhóm sử dụng thêm hai kỹ thuật Concatenation Ensemble và Average Ensemble để tiến hành kết hợp hai mô hình trên. Từ đó, nhóm xây dựng thêm hai mô hình.

- **Mô hình 3:** sử dụng kỹ thuật Concatenation Ensemble để kết hợp hai mô hình 1 và mô hình 2.
- **Mô hình 4:** sử dụng kỹ thuật Average Ensemble để kết hợp hai mô hình 1 và mô hình 2.



Hình 12. Cấu trúc mô hình 3 (trái) và mô hình 4 (phải)

3.3.4. Huấn luyện mô hình

Hai mô hình 1 và mô hình 2 được huấn luyện với các thông số sau:

- Kích thước ảnh đầu vào: 256x256 với batch size được sử dụng là 256. Huấn luyện trong 50 epochs và finetuning trong 10 epochs.

- Learning rate khởi tạo cho quá trình huấn luyện là 0.01 và cho quá trình finetuning là 0.001. Trong quá trình training, giảm learning rate sau mỗi 5 epochs, và sau mỗi 3 epochs trong quá trình finetuning.
- Optimizer được sử dụng là Adam optimizer. Hàm loss được sử dụng là Binary Crossentropy.
- Dừng huấn luyện khi mô hình không cải thiện trong 10 epochs, dừng finetune khi mô hình không cải thiện trong 6 epochs.
- Với mô hình 1 nhóm tiến hành finetune trên 63 layers của backbone MobileNetV3, với mô hình 2 nhóm tiến hành finetune trên 40 layers của backbone EfficientNetB2.

Hai mô hình 3 và mô hình 4 được huấn luyện chủ yếu để cập nhật các tham số của các lớp phân loại ở cuối của mô hình, do đó nhóm chỉ huấn luyện mô hình trong 10 epochs với learning rate khởi tạo là 0.001. Giảm learning rate sau mỗi 2 epochs và dừng huấn luyện khi mô hình không cải thiện trong 4 epochs.

Cả 4 mô hình trên đều được huấn luyện trên cấu hình chung như sau: GPU NVIDIA V100 - 16GB vRAM; CPU Intel(R) Xeon(R) - 2.00GHz; RAM: 51 GB.

4. Thực nghiệm và đánh giá kết quả

Trong phần này, nghiên cứu sẽ trình bày các thông tin về kết quả thu thập dữ liệu, kiến trúc mô hình đã thử nghiệm, thực hiện đánh giá và so sánh nhằm lựa chọn kiến trúc phù hợp nhất cho bài toán nhận diện hình ảnh mặt người thật và do AI tạo ra.

4.1. Kết quả thu thập và xử lý dữ liệu

Tổng kết, nhóm đã xây dựng được bộ dữ liệu bao gồm 120,959 ảnh thuộc về hai lớp là ảnh thật (70,000 ảnh) và ảnh do AI tạo ra (50,959) ảnh. Bộ dữ liệu trên được chia thành 3 tập train/test/validation theo tỷ lệ 6/2/2, các lớp ảnh giả được chia đều cho mỗi lớp train/test/validate.

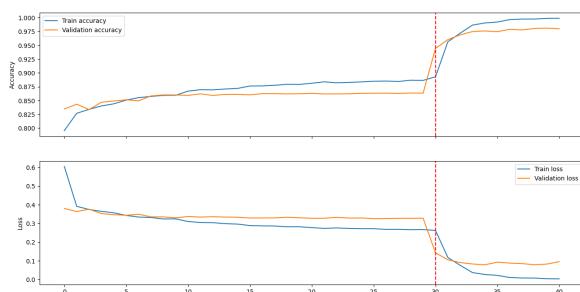
Bảng 2. Bảng thống kê số lượng dữ liệu

Lớp	Bộ dữ liệu	Số lượng ảnh
Ảnh thật	FFHQ	70,000
Ảnh do AI sinh ra	thispersondoesnotexist.com	20,499
	SFHQ	20,000
	Swap face	9,429
	Stable Diffusion	1,031

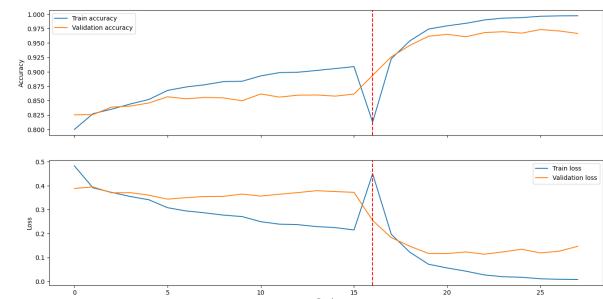
4.2. Kết quả huấn luyện và kiểm thử mô hình

4.2.1. Kết quả huấn luyện mô hình

Sau đây là kết quả huấn luyện của hai mô hình 1 và mô hình 2. Đối với hai mô hình 3 và mô hình 4, nhóm sẽ không trình bày kết quả huấn luyện ở đây vì quá trình huấn luyện của hai kỹ thuật này chỉ chủ yếu để cập nhật tham số của hai lớp phân loại cuối cùng.



Hình 13. Kết quả huấn luyện mô hình 1



Hình 14. Kết quả huấn luyện mô hình 2

4.2.2. Kết quả kiểm thử mô hình

Nhóm tiến hành đánh giá mô hình trên tập kiểm thử dựa trên các metrics sau: accuracy, precision, F1 và recall.

Bảng 3. Bảng kết quả kiểm thử của 4 mô hình

Mô hình	Class	Precision	Recall	F1	Accuracy (%)
Mô hình 1	Real	0.97	0.99	0.98	97.91
	AI	0.99	0.96	0.97	
Mô hình 2	Real	0.98	0.96	0.97	96.58
	AI	0.95	0.97	0.96	
Mô hình 3	Real	0.97	0.99	0.98	97.86
	AI	0.99	0.96	0.97	
Mô hình 4	Real	0.97	0.99	0.98	97.85
	AI	0.99	0.96	0.97	

Bảng 4. Đánh giá kích thước và thời gian chạy

Mô hình	Kích thước mô hình (MB)	Inference time trên tập kiểm thử (s)
Mô hình 1	50.3	15
Mô hình 2	90.8	28
Mô hình 3	60.6	36
Mô hình 4	60.6	36

4.2.3. Nhận xét và đánh giá kết quả

Cả 4 mô hình nhóm tiến hành thử nghiệm đều cho kết quả khá cao (độ chính xác đều trên 96%). Trong đó mô hình 1 (sử dụng kỹ thuật Transfer Learning với backbone MobileNetV3) cho kết quả tốt nhất trên hầu hết các metric đánh giá. Bên cạnh đó mô hình 1 cũng là mô hình có kích thước nhỏ nhất cũng như thời gian chạy nhanh nhất.



Giải thích kết quả với LIME



Giải thích kết quả với GRAD-CAM

Hình 15. Giải thích kết quả dự đoán

4.3. Đánh giá hệ thống trong sử dụng thực tế

Nhóm tiến hành đánh giá thời gian xử lý của hệ thống trong thực tế với cấu hình server:

- CPU: Intel(R) Core(TM) i3-1005G1 - 1.20GHz.
- RAM: 4.00 GB.

❖ *Đánh giá thời gian kiểm tra một ảnh có hợp lệ (chứa đúng một mặt người) hay không.*

Nhóm tiến hành đánh giá hệ thống với 10 ảnh trong đó không chứa mặt người hoặc chứa nhiều hơn một mặt người. Thời gian kiểm tra trung bình là xấp xỉ **1.24s/ảnh**.

❖ *Đánh giá thời gian chạy thực tế của 4 mô hình:*

Nhóm tiến hành đánh giá thời gian chạy của 4 mô hình khi tích hợp vào hệ thống với 10 ảnh. Lưu ý, thời gian đưa ra dự đoán của mô hình đã bao gồm thời gian kiểm tra ảnh có hợp lệ (chứa đúng một mặt người) hay không.

Bảng 5. Đánh giá thời gian chạy trong thực tế

Mô hình	Thời gian trung bình để đưa kết quả dự đoán (s)	Thời gian trung bình để giải thích kết quả dự đoán (s)
Mô hình 1	7	74
Mô hình 2	13	160
Mô hình 3	21	x
Mô hình 4	20	x

Với hai mô hình 3 và mô hình 4, vì cấu trúc mô hình phức tạp nên chưa áp dụng được hai kỹ thuật Grad-CAM và LIME để giải thích kết quả nên chưa đánh giá được thời gian trung bình để giải thích kết quả dự đoán của hai mô hình này.

5. Kết luận

Qua nghiên cứu này, nhóm đã xây dựng thành công được một bộ dữ liệu đa dạng về ảnh mặt người do AI tạo ra cũng như đã tiến hành thử nghiệm và đề xuất việc sử dụng các mô hình học sâu CNN cùng các kỹ thuật Transfer Learning - Ensemble Learning để xây dựng mô hình phân loại ảnh mặt người do AI tạo ra. Bên cạnh đó, nhóm cũng đã tiến hành đánh giá hiệu quả của các mô hình khi áp dụng vào hệ thống thực tế, từ đó xây dựng thành công hệ thống dịch vụ với độ chính xác cao, tốc độ xử lý nhanh và có khả năng mở rộng về sau.

Mặc dù đã đạt được một số kết quả tích cực, tuy nhiên nghiên cứu của nhóm vẫn còn một số điểm hạn chế và khó khăn.

Về mặt dữ liệu: tuy bộ dữ liệu của nhóm có chất lượng khá tốt và độ đa dạng cao, nhưng vì những hạn chế về tài nguyên tính toán dẫn đến việc nhóm phải giảm chất lượng và kích thước của bộ dữ liệu huấn luyện. Bên cạnh đó, việc thực hiện các kỹ thuật tăng cường dữ liệu cũng bị hạn chế vì lý do trên.

Về mặt mô hình: tuy các mô hình đều đạt kết quả khá cao, những cách tiếp cận về mặt mô hình của nhóm hiện tại đang phụ thuộc khá nhiều về dữ liệu. Do đó, trong tương lai, nhóm dự định sẽ nghiên cứu và phát triển thêm các mô hình khác để giảm sự phụ thuộc vào dữ liệu và tăng cường tính tổng quát hóa của hệ thống.

Tài liệu tham khảo

- [1] Yan Ju, Shan Jia, Jialing Cai, Haiying Guan, Siwei Lyu, "GLFF: Global and Local Feature Fusion for AI-synthesized Image Detection" (2023), <https://arxiv.org/pdf/2211.08615.pdf>
- [2] Rosberg, Felix and Aksoy, Eren Erdal and Alonso-Fernandez, Fernando and Englund, Cristofer, "FaceDancer: Pose- and Occlusion-Aware High Fidelity Face Swapping" (2023), <https://arxiv.org/pdf/2210.10473.pdf>
- [3] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, Hartwig Adam, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications" (2017), <https://arxiv.org/pdf/1704.04861.pdf>
- [4] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, Liang-Chieh Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks" (2018), <https://arxiv.org/pdf/1801.04381v4.pdf>
- [5] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, Hartwig Adam, "Searching for MobileNetV3" (2019), <https://arxiv.org/pdf/1905.02244.pdf>
- [6] Mingxing Tan, Quoc V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks" (2020), <https://arxiv.org/pdf/1905.11946.pdf>
- [7] Michael Munn, David Pitman, "Explainable AI for Practitioners" (2022)
- [8] Haodong Li, Bin Li, Shunquan Tan, Jiwu Huang, "Identification of Deep Network Generated Images Using Disparities in Color Components" (2020), <https://arxiv.org/pdf/1808.07276.pdf>
- [9] Scott McCloskey, Michael Albright, "Detecting GAN-generated Imagery using Color Cues" (2018), <https://arxiv.org/pdf/1812.08247.pdf>
- [10] Zeyang Sha, Zheng Li, Ning Yu, Yang Zhang, "DE-FAKE: Detection and Attribution of Fake Images Generated by Text-to-Image Generation Models", (2023) <https://arxiv.org/pdf/2210.06998.pdf>
- [11] David Beniaguev, "Synthetic Faces High Quality (SFHQ) dataset" (2022), <https://github.com/SelfishGene/SFHQ-dataset>
- [12] Tero Karras, Janne Hellsten, "Flickr-Faces-HQ Dataset (FFHQ)", <https://github.com/NVlabs/ffhq-dataset>
- [13] Tero Karras, Samuli Laine, Timo Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks" [2019], <https://arxiv.org/pdf/1812.04948.pdf>