ĐÁNH GIÁ HIỆU NĂNG MỘT SỐ MẠNG HỌC SÂU TRONG NHẬN DẠNG HÌNH ẢNH KHUÔN MẶT THẬT GIẢ

Trần Quý Nam
Học Viện Công Nghệ Bưu Chính Viễn Thông

Tóm tắt: Nghiên cứu này thử nghiệm một số mô hình học sâu để đưa ra đánh giá hiệu năng mô hình phù hợp nhất cho bài toán nhận dạng ảnh khuôn mặt thật giả. Các kỹ thuật học chuyển đổi, mô hình lai ghép, mô hình Vision Transformer được áp dụng trên cùng một tập dữ liệu ảnh khuôn mặt thật giả để so sánh. Trong đó, kỹ thuật học chuyển đổi được áp dụng cho 8 mô hình InceptionResNetV2, EfficientNetV2, VGG16, ResNet50, EfficientNetB7, MobileNet, ResNet101, DenseNet201, kỹ thuật lai ghép áp dụng cho 3 mô hình VGG16-XGBoost, VGG16-CatBoost, VGG16-LightGBM. Kết quả thử nghiệm cho thấy các mô hình lai ghép và mô hình Vision Transformer cho kết quả khá tốt, nhưng mô hình ResNet50 cho kết quả tốt nhất.

Từ khóa: Ảnh khuôn mặt, thật giả, học sâu, chuyển đổi, lai ghép.

I. GIỚI THIỆU

Hiện nay, sự phát triển mạnh mẽ của Internet và mạng xã hội dẫn đến hiện tượng lan tràn các thông tin, hình ảnh thật và giả trên các mạng xã hội (Facebook, Twitter, TikTok...), trên các website, các luồng thông tin trên mạng Internet, trang thông tin điện tử, blogs,… Sự hình thành các hình ảnh giả mạo bắt nguồn từ những năm 1990, nhưng đến năm 2017, khi trí tuệ nhân tạo phát triển mạnh với điển hình là sáng tạo ra mạng đối nghịch (GAN - Generative Adversarial Network) đã giả mạo các hình ảnh rất khó phân biệt đâu là ảnh thật, đâu là ảnh giả mạo. Điển hình là từ việc sao chép khuôn mặt của người nổi tiếng để quảng cáo sản phẩm của họ cho đến gài bẫy các chính trị gia và người nổi tiếng trên mạng Internet. Ngày nay, ảnh giả mạo bị lạm dụng nhiều trong tội phạm mạng để đánh cắp danh tính, tống tiền trên mạng, tin tức giả mạo, gian lận tài chính, video khiêu dâm giả mạo của người nổi tiếng để tống tiền. Trong an toàn thông tin, các hệ thống định danh bằng khuôn mặt cũng bị giả mạo, trộm cắp danh tính, tống tiền. Tội phạm mạng sử dụng các công nghệ tạo ảnh giả mạo như thật để truy cập tài khoản ngân hàng, truy cập các hệ thống chính phủ điện tử, các hệ thống điểm danh hoặc bảo mật an ninh bằng xác thực ảnh khuôn mặt, gây khó khăn trong xác định danh tính, xác thực nhân vật. Nhiều quốc gia trên thế giới như Vương quốc Anh, Hoa Kỳ, Canada, Ấn Độ, Trung Quốc và Hàn Quốc đã ban hành các đạo luật để kiểm chế sự lan tràn các tin tức, hình ảnh, video giả mạo trên mạng Internet. Đồng thời các quốc gia này cũng tăng cường đầu tư phát triển các công nghệ tự động phát hiện tin tức, hình ảnh, video giả mạo để cảnh báo người dùng. Trong đó, phát hiện hình ảnh khuôn mặt giả mạo là một thách thức lớn và có nhu cầu ứng dụng cao về công nghệ nhận dạng ảnh số. Tại Việt Nam, các nghệ sỹ thường bị giả mạo hình ảnh. Ví dụ như ca sĩ Phương Mỹ Chi bị lan truyền ảnh giả mạo trên mạng Internet năm 2023 và đã gửi đơn trình báo đến cơ quan chức năng của Bộ Công an và đã được minh oan. Bộ Thông tin và Truyền thông đã xuất bản "Cẩm nang phòng chống tin giả, tin sai sự thật trên không gian mạng". Tuy vậy, vấn nạn hình ảnh, tin tức giả mạo vẫn còn là vấn đề lớn ở Việt Nam. Mặc dù Việt Nam đã có các nghiên cứu về vấn đề này nhưng chủ yếu mang tính định tính về tin tức giả mạo nói chung, chưa có nhiều nghiên cứu vận dụng trí tuệ nhận tạo để tự động xác định nội dung giả mạo, đặc biệt còn ít nghiên cứu về giả mạo ảnh khuôn mặt, một vấn đề quan trọng hiện nay.

Như vậy, bài toán nhận dạng hình ảnh các khuôn mặt thật và ảnh giả mạo một cách tự động có thể cung cấp căn cứ cho phát triển các phần mềm quét tự động, cảnh báo người dùng. Vì vậy, nghiên cứu này đề xuất thử nghiệm một số mô hình sử dụng mô hình học sâu trong nhận dạng hình ảnh khuôn mặt thật giả. Kết quả thực nghiệm được so sánh giữa các mô hình với nhau và được trình bày trong các phần tiếp theo.

II. CÁC NGHIÊN CỨU LIÊN QUAN

Thực tế đã có rất nhiều nghiên cứu sử dụng các mô hình học máy, học sâu để nhận dạng các hình ảnh khuôn mặt thật giả trên thế giới. Một trong số đó là nghiên cứu do Raza và cộng sự (2022) đã nghiên cứu các công nghệ học máy để bảo vệ nạn nhân khỏi bị tống tiền bằng cách phát hiện các hình ảnh khuôn mặt giả mạo. Mục đích

Tác giả liên hệ: Trần Quý Nam,
Email: namtq@ptit.edu.vn
Đến tòa soạn: 12/2024, chỉnh sửa: 01/2025, chấp nhận đăng: 02/2025.

chính của nghiên cứu là phát hiện phương tiện deepfake bằng cách sử dụng sự kết hợp giữa VGG16 và kiến trúc mạng thần kinh tích chập CNN. Bộ dữ liệu deepfake dựa trên khuôn mặt thật và giả được sử dụng để xây dựng các mạng nơ ron Xception, NAS-Net, Mobile Net và VGG16 với kỹ thuật học chuyển đổi được sử dụng để so sánh. Phương pháp nghiên cứu của Raza và cộng sự giúp các chuyên gia an ninh mạng khắc phục tội phạm mạng liên quan đến deepfake bằng cách phát hiện chính xác nội dung deepfake và cứu nạn nhân deepfake khỏi bị tống tiền [1]. Năm 2023, Chen và cộng sự (2023) đã nghiên cứu một số các mô hình hiện có để phát hiện deepfake, trong đó đều tập trung vào hình ảnh các khuôn mặt gốc, được xem xét trong các ứng dụng thực tế để phát hiện ảnh khuôn mặt giả mạo và dữ liệu đầu vào có thể chứa thông tin riêng tư và nhạy cảm. Do đó, các tác giả đã xây dựng một mô hình bảo vệ quyền riêng tư có tên là mạng phát hiện deepfake an toàn (SecDFDNet). Trong đó, SecDFDNet sử dụng phương pháp chia sẻ bí mật bổ sung để phát hiện khuôn mặt bị giả mạo một cách an toàn. Cụ thể, một số giao thức tương tác an toàn nhiều bên được thiết kế cho các chức năng kích hoạt phi tuyến tính, ví dụ SecReLU thay cho chức năng ReLU, SecSigm thay cho chức năng sigmoid, SecSpatial cho chức năng chú ý không gian và SecChannel cho chức năng chú ý kênh. Kết quả thử nghiệm cho thấy SecDFDNet có thể phát hiện các khuôn mặt giả mạo mà mà không tiết lộ bất kỳ thông tin đầu vào riêng tư nào của cá nhân cần xác định [2].

Rafique và cộng sự (2023) nghiên cứu phương pháp tự động để phân loại hình ảnh khuôn mặt thật giả bằng cách sử dụng các phương pháp dựa trên Deep Learning và Machine Learning. Trong đó, các tác giả phân tích các hệ thống dựa trên Machine Learning (ML) truyền thống sử dụng trích xuất đặc trưng thủ công không thể trích xuất được các mẫu ảnh phức tạp, ảnh có nội dung khó hiểu hoặc khó biểu diễn bằng các tính năng đơn giản. Các hệ thống này rất nhạy cảm với nhiễu hoặc các biến thể trong dữ liệu, điều này có thể làm giảm hiệu suất của hệ thống học máy. Các tác giả ban đầu thực hiện phân tích mức độ lỗi của hình ảnh để xác định xem hình ảnh đã được sửa đổi hay chưa. Hình ảnh này sau đó được cung cấp cho một mạng nơ ron chuyển đổi để trích xuất các đặc trưng của ảnh. Các vectơ đặc trưng tổng hợp sau đó được phân loại thông qua thuật toán máy vectơ hỗ trợ (SVM) và K láng giềng gần nhất (KNN) bằng cách thực hiện tối ưu hóa siêu tham số. Kết quả chứng minh tính hiệu quả và bền vững của các kỹ thuật đề xuất, do đó, có thể được sử dụng để phát hiện những hình ảnh giả khuôn mặt mạo và giảm thiểu mối đe dọa tiềm ẩn về hành vi lan truyền sai lệch thông tin trên mạng Internet [3]. Sangavi và cộng sự (2023) đã áp dụng tính năng tăng cường ảnh kết hợp với phương pháp học chuyển đổi được lựa chọn là mạng đã huấn luyện Mobilenet và Resnet với dự đoán mang lại độ chính xác cao hơn các phương pháp học sâu khác. Việc tinh chỉnh các siêu tham số đã được đánh giá bằng các số liệu hiệu suất như độ chính xác, độ nhạy và điểm F1-Score. Phương pháp đề xuất cho độ chính xác cao hơn gần 75% và 78% so với các phương pháp khác [4]. Nhóm tác giả Jaffar và cộng sự (2024) đã nghiên cứu nhận dạng các bức ảnh giả sâu bằng cách sử dụng học sâu với kỹ thuật học chuyển đổi. Bài báo này đã đề xuất mô hình mạng nơ ron tích chập (CNN) dựa trên các phương pháp học chuyển đổi trong đó bộ phân loại được thiết kế sử dụng mức trung bình gộp toàn cục (Global Average Pooling - GAP), có dropout, lớp FCL và hàm Softmax để thay thế cho lớp kết nối đầy đủ cuối cùng (FCL) trong các mô hình được huấn luyện trước. Các tác giả kết luận, mạng DenseNet201 tạo ra độ chính xác tốt nhất là 86,85% cho cả hai bộ dữ liệu hình ảnh thật và ảnh giả, trong khi Mobilenet tạo ra độ chính xác thấp hơn 82,78% [5].

Đối với nghiên cứu trong nước, trong lĩnh vực nghiên cứu có các tác giả Cường và Tùng (2021) đã thực hiện nghiên cứu và đánh giá một số phương pháp học máy ứng dụng trong phát hiện ảnh giả mạo. Trong đó, các tác giả thực hiện bài toán phân biệt các ảnh giả mạo nói chung, không phải là giải quyết bài toán chuyên sâu về nhận dạng ảnh khuôn mặt. Các tác giả thực hiện nghiên cứu và đánh giá ba mô hình học sâu phổ biến (CNN, GAN và VGG-16) trong ứng dụng trong phát hiện ảnh giả mạo. Với mỗi mô hình, các tác giả thiết kế các mạng để xác định các đặc trưng và từ đó phân loại được ảnh thật và ảnh giả. Việc thử nghiệm của cả ba phương pháp được thực hiện trên cùng bộ dữ liệu để có thể so sánh công bằng với nhau. Kết quả cho thấy cả ba mô hình học sâu trên đều có thể áp dụng để phát hiện ảnh số giả mạo, trong đó, mô hình sử dụng VGG-16 cho kết quả tốt nhất [6].

III. PHƯƠNG PHÁP VÀ KẾT QUẢ THỰC NGHIỆM

Phần nội dung này trình bày phương pháp, các mô hình và kết quả thử nghiệm tương ứng. Trong đó, các mô hình được lựa chọn căn cứ kết quả khảo sát các nghiên cứu liên quan ở phần II, một số mô hình lai ghép là các đề xuất thử nghiệm mới cũng như thử nghiệm cho mô hình Vision Transformer.

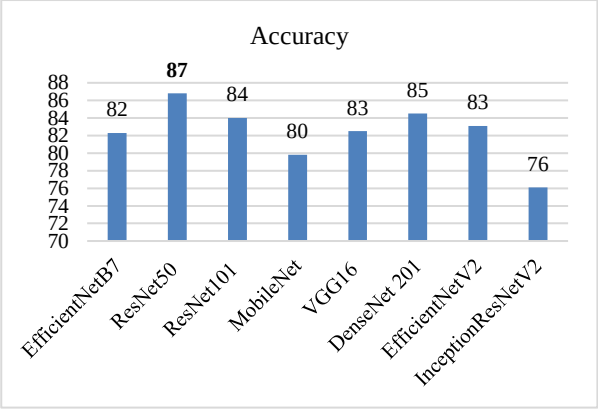
A. Tập dữ liệu thử nghiệm

Bộ dữ liệu thử nghiệm sử dụng trong nghiên cứu này do Xhlulu (2019) cung cấp chứa tổng cộng 140 nghìn ảnh khuôn mặt [7], nhưng do hạn chế về tài nguyên máy tính và bộ nhớ nên nghiên cứu đã trích xuất ngẫu nhiên 10 nghìn ảnh khuôn mặt với 5 nghìn ảnh khuôn mặt giả và 5 nghìn ảnh khuôn mặt thật. Tập dữ liệu này cung cấp các khuôn mặt thật từ tập dữ liệu Flickr do Nvidia thu thập và các khuôn mặt giả được lấy mẫu từ 1 triệu khuôn mặt giả do StyleGAN tạo ra. Các chuyên gia đã sử dụng bộ dữ liệu này để thử nghiệm đánh giá bằng cảm nhận dựa trên trực quan. Khi xem qua tập dữ liệu này và hầu hết các chuyên gia đều thấy không thể hoặc rất khó phân biệt được sự khác biệt giữa khuôn mặt thật và khuôn mặt giả

được đưa ra trong tập dữ liệu này là gì. Khi nhìn vào khuôn mặt thật và giả, cả hai đều trông giống nhau và các chuyên gia không thể tìm ra điểm khác biệt nào chỉ khi nhìn vào chúng.

B. Thử nghiệm kỹ thuật transfer learning

Đối với kỹ thuật học chuyển đổi (transfer learning), nghiên cứu này thử nghiệm mô hình học chuyển đổi dựa trên một số mạng học sâu CNN được đào tạo trước để phân loại hình ảnh khuôn mặt thật và giả. Có 8 mô hình được áp dụng thử nghiệm là EfficientNetB7, ResNet50, MobileNet, InceptionResNetV2, VGG16, ResNet101, EfficientNetV2, DenseNet201. Để có cơ sở so sánh các kết quả thử nghiệm, thực hiện cấu hình chung các tham số thử nghiệm một số mạng học sâu đã huấn luyện cho bài toán. Mỗi tập dữ liệu thử nghiệm được chia theo tỷ lệ 60:20:20, tức 60% ảnh cho phần huấn luyện (training), 20% ảnh cho phần xác thực (validation) và 20% cho phần kiểm tra (testing) mô hình học sâu để đánh giá hiệu quả nhận dạng hình ảnh khuôn mặt thật và ảnh giả. Chế độ màu (color mode) áp dụng chung là 'RGB', hàm phi tuyến sử dụng trong các mô hình là hàm Relu, trình tối ưu Adam, sử dụng chạy mô hình với 100 epoches có áp dụng dừng sớm (early stopping) dựa trên kết quả theo dõi giá trị hàm mất mát xác thực (theo dõi giá trị validation loss) với số epoch liên tục là 2 không cải thiện (không giảm) giá trị mất mát trên tập dữ liệu xác thực. Các siêu tham số tốc độ học (learning rate) lấy giá trị bằng 0.001 và kích thước tập dữ liệu (batch size) được lấy giá trị là 32. Kết quả thử nghiệm với 8 mô hình thể hiện trong hình 1 dưới đây.

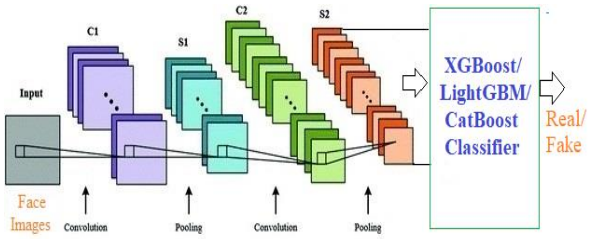


Hình 1. Kết quả 8 mô hình theo transfer learning

Kết quả áp dụng trên tập dữ liệu thử nghiệm đối với các mô hình mạng CNN đã huấn luyện: EfficientNetB7, ResNet50, MobileNet, InceptionResNetV2, VGG16, ResNet101, EfficientNetV2, DenseNet201 theo nguyên tắc áp dụng học chuyển đổi áp dụng cho bài toán nhận dạng ảnh khuôn mặt thật và ảnh khuôn mặt giả cho thấy mô hình mạng ResNet50 cho kết quả tốt nhất, với độ chính xác đạt 87%.

C. Mô hình thử nghiệm lai ghép

Đối với kỹ thuật lai ghép, nghiên cứu này xây dựng mô hình lai ghép (hybrid models) có 2 giai đoạn xử lý khác nhau. Giai đoạn thứ nhất là phần trích xuất đặc trưng hình ảnh, giai đoạn thứ 2 là phân loại hình ảnh. Trong giai đoạn 1, mô hình lai ghép sử dụng mạng nơ ron tích chập (Convolutional Neural Network - CNN) để trích xuất đặc trưng (features extraction). Sau đó trong giai đoạn 2, mô hình sử dụng một số thuật toán phân loại phổ biến như máy vec tơ hỗ trợ SVM, cây quyết định, rừng ngẫu nhiên, k láng giềng gần nhất… Trong nghiên cứu này, giai đoạn 2 sử dụng một số thuật tiên tiến mới ra đời trong vài năm trở lại đây, gồm có XGBoost, LightGBM và CatBoost để thực hiện phân loại, nhận dạng hình ảnh dựa trên các đặc trưng đã được trích xuất từ các mạng CNN trước đó. Hình 2 dưới đây mô tả kiến trúc mạng lai ghép này.



Hình 2. Mô tả kiến trúc mạng lai ghép

Trong mô hình lai ghép nêu trên, ở giai đoạn 1 trích xuất đặc trưng, nghiên cứu này sử dụng mạng nơ ron tích chập CNN đã được huấn luyện có tên là VGG16 để trích xuất đặc trưng. Năm 2015, các tác giả Karen và Andrew (2015) đã giới thiệu mạng VGG tại Hội nghị ICLR 2015. Kiến trúc của mô hình mạng này có nhiều biến thể khác nhau: 11 lớp, 13 lớp, 16 lớp và 19 lớp. VGG16 đề cập đến một kiến trúc mạng có 16 lớp, trong khi VGG19 đề cập đến một kiến trúc mạng với 19 lớp. Nguyên tắc thiết kế của mạng VGG nói chung bao gồm 2 hoặc 3 lớp Convolution (Conv) được nối tiếp theo sau bởi lớp 2D Max Pooling. Lớp cuối cùng là một lớp làm phẳng (Flatten layer) nhằm mục đích chuyển đổi ma trận 4 chiều thành ma trận 2 chiều. Tiếp theo là các lớp được kết nối đầy (Fully-Connected Layers) và 1 lớp Softmax [8].

Đối với giai đoạn 2 phân loại hình ảnh, nghiên cứu này thử nghiệm 3 thuật toán là XGBoost, LightGBM và CatBoost. Trong đó, thuật toán XGBoost viết tắt của cụm từ Extreme Gradient Boosting, một thuật toán học máy hiệu quả cao dựa trên sự kết hợp các kỹ thuật để điều chỉnh các trọng số lỗi trên các mô hình yếu hơn để tạo ra một mô hình mạnh hơn. Nguyên tắc thuật toán XGBoost dựa trên cây quyết định và kỹ thuật tăng cường độ dốc để đưa ra mô hình tối ưu. Các cây mới sinh ra tuần tự được giảm thiểu lỗi từ cây trước đó bằng cách học lại lỗi của cây trước đó, thực hiện sửa lỗi để được cây tốt hơn. XGBoost ban đầu được Chen và Guestrin (2016) giới thiệu để cải thiện hiệu suất và tốc độ của cây quyết định theo nguyên tắc tăng cường độ dốc (gradient-boosted) [9]. Đối với thuật toán LightGBM, là viết tắt của cụm từ Light

Gradient Boosting Machine, cũng là một thuật toán tăng cường độ dốc (Gradient Boosting) được phát triển bởi Microsoft. LightGBM ban đầu được giới thiệu bởi Guolin và cộng sự (2017) để cải thiện hiệu suất và tốc độ sử dụng phương pháp học tập dựa trên cây quyết định [10]. LightGBM là một thuật toán theo nguyên tắc tăng cường độ dốc phân tán và hiệu quả. Thuật toán này cũng dựa trên biểu đồ và đặt các giá trị liên tục vào các khoảng giá trị riêng biệt. Theo mô tả thuật toán LightGBM, có 2 ưu điểm của LightGBM, đó là lấy mẫu một mặt dựa trên gradient (GOSS - Gradient-based One-Side Sampling) và gói tính năng độc quyền (EFB - Exclusive Feature Bundling) [10]. Đối với thuật toán CatBoost, đây cũng là một thuật toán dựa trên nguyên lý gradient descent (Dorogush, 2018), nhưng có một số khác biệt nhỏ. CatBoost triển khai cây đối xứng giúp giảm thời gian dự đoán và nó cũng có độ sâu của cây nông hơn so với một số phương pháp khác. Trong bài báo mô tả thuật toán CatBoost, Dorogush và các tác giả (2018) đã giải thích đề xuất về nguyên tắc sắp xếp bắt nguồn từ việc tăng cường theo thứ tự, một sửa đổi của thuật toán tăng cường độ dốc tiêu chuẩn để xử lý các tính năng phân loại [11]. Kết quả thử nghiệm 3 mô hình lai ghép thể hiện như trong bảng 1 dưới đây.

Bảng 1: So sánh các mô hình lai ghép

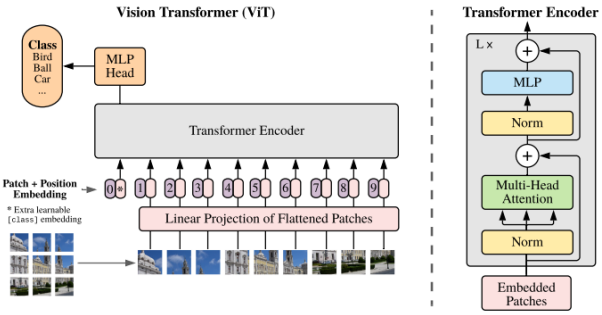
Hybrid models	*Acc.*	*Prec.*	*Rec.*	*F1-scr.*
VGG16-XGBoost	73,04	73,42	73,29	73,35
VGG16-LightGBM	75,05	74,87	75,12	74,99
VGG16-CatBoost	**77,45**	**76,34**	**77,02**	**76,68**

Theo dữ liệu tại bảng 1 so sánh kết quả phân loại của 3 mô hình lai ghép VGG16-XGBoost, VGG16-LightGBM và VGG16-CatBoost áp dụng trên cùng tập dữ liệu ảnh khuôn mặt thật giả thử nghiệm cho thấy mô hình mạng VGG16-CatBoost cho kết quả tốt nhất, với độ chính xác đạt trên 77%.

D. Mô hình thử nghiệm Vision Transformer

Năm 2021, các tác giả Dosovitskiy và cộng sự (2021) đã đề xuất một hướng tiếp cận khác trong các vấn đề về thị giác máy tính, lấy ý tưởng từ việc áp dụng kiến trúc Transformer để xử lý ngôn ngữ tự nhiên. Nhóm các nhà nghiên cứu từ Google Research đã giới thiệu kiến trúc Vision Transformer (xem hình 3) là phiên bản kiến trúc Transformer cho hình ảnh [12]. Trên thực tế, kiến trúc này đã đạt được rất nhiều kết quả nổi bật trong nhiều bài toán khác nhau. Kết quả nghiên cứu và ứng dụng mô hình Transformer trong xử lý ngôn ngữ tự nhiên đã đạt được rất ấn tượng. Tuy nhiên, về mặt thị giác máy tính, việc ứng dụng và nghiên cứu mô hình Transformer còn nhiều hạn chế và thiếu các nghiên cứu. Trong khoa học máy tính, khi gặp các bài toán về thị giác máy tính như phân loại, nhận dạng ảnh, phát hiện đối tượng, phân đoạn đối tượng… thì kiến trúc tích chập mà cụ thể là mạng CNN vẫn là kiến trúc quen thuộc mà chúng ta thường sử dụng.

Kiến trúc Transformer trong học máy là một mô hình học sâu sử dụng các cơ chế của cơ chế chú ý (cơ chế Attention), cân nhắc kỹ lưỡng tầm quan trọng của từng phần dữ liệu đầu vào. Transformer trong học máy bao gồm nhiều lớp với cơ chế Self-Attention, chủ yếu được sử dụng trong các lĩnh vực AI của xử lý ngôn ngữ tự nhiên (NLP) và thị giác máy tính (CV). Mô hình Transformer trong học máy thể hiện cách tiềm năng của một phương pháp học tổng quát. Transformer có thể được áp dụng cho các tập dữ liệu khác nhau trong thị giác máy tính, để đạt được độ chính xác tốt hơn với ít tham số hơn trong bối cảnh tài nguyên máy tính hạn chế (Dosovitskiy và cộng sự, 2021).



Hình 3. Kiến trúc Vision Transformer [12]

Kiến trúc Transformer truyền thống lấy đầu vào là một chuỗi mã thông báo nhúng 1D. Do đó, để xử lý đầu vào dưới dạng hình ảnh 2D, mô hình Vision Transformer chia nhỏ hình ảnh đầu vào thành các mảnh ghép ảnh riêng lẻ kích thước nhỏ được chia ra từ ảnh lớn hay còn gọi là gói ảnh (bản vá) có kích thước cố định giống như trình tự nhúng từ được sử dụng trong mô hình Transformer truyền thống được sử dụng cho văn bản. Kết quả huấn luyện và thử nghiệm mô hình Vision Transformer kích thước 16x16 áp dụng trên cùng tập dữ liệu ảnh khuôn mặt thật giả, cho độ chính xác của mô hình Vision Transformer đạt 0.7705 (tức 77,05%).

Quá trình nghiên cứu thử nghiệm các mô hình cho thấy thời gian huấn luyện của mô hình Vision Transformer là lâu nhất, sau đó là các mô hình lai ghép và cuối cùng là các mô hình học chuyển đổi. Thời gian xử lý để dự đoán, nhận dạng một hình nhả khuôn mặt thật và giả giữa các mô hình hầu như không có sự khác biệt lớn. Tốc độ phân tích, nhận dạng một hình ảnh hoặc một tập hợp các hình ảnh là tương đối nhanh, không có nhiều sự khác biệt giữa các mô hình. Lý do là các mô hình sau khi huấn luyện thì các tham số đã được hình thành, cố định trong quá trình nhận dạng nên thời gian xử lý được nhanh, không tốn thời gian tính toán như giai đoạn huấn luyện.

Kết quả thực nghiệm cho thấy mạng ResNet 50 cho kết quả tốt nhất với độ chính xác là 87%. Lý do ResNet 50 có kết quả tốt nhất là một trong những đặc điểm chính

của ResNet 50 là cấu trúc rất sâu, bao gồm 50 lớp mạng. Các mạng sâu hơn đã được hiển thị để nắm bắt các tính năng và phân cấp phức tạp hơn trong hình ảnh, cho phép học các đặc trưng tốt hơn. Độ sâu này cho phép ResNet 50 tìm hiểu các mẫu và chi tiết phức tạp trong ảnh danh sách, do đó nâng cao độ chính xác của quy trình phân loại. Ngoài ra, ResNet 50 sử dụng các kết nối tắt (shortcut connections) cho phép mạng giảm bớt vấn đề độ dốc biến mất thường gặp phải trong các mạng rất sâu. Bằng cách truyền các gradient hiệu quả hơn, các kết nối còn lại tạo điều kiện thuận lợi cho việc đào tạo các kiến trúc sâu hơn, dẫn đến hiệu suất được cải thiện. Hơn nữa, ResNet 50 được đào tạo trước trên tập dữ liệu quy mô lớn, như ImageNet nên độ chính xác cao hơn.

Đối với các trường hợp dự đoán sai, qua nghiên cứu chủ yếu do các đặc trưng ảnh thật và ảnh giả không có sự khác biệt rõ ràng. Như đã trình bày trong mục 2, phần 1 mô tả tập dữ liệu, nghiên cứu này sử dụng tập dữ liệu Flickr do Nvidia thu thập và do StyleGAN tạo ra. Các chuyên gia đã sử dụng bộ dữ liệu này để thử nghiệm đánh giá bằng cảm nhận dựa trên trực quan đều thấy không thể hoặc rất khó phân biệt được sự khác biệt giữa khuôn mặt thật và khuôn mặt giả. Vì vậy, các trường hợp dự đoán sai là do cấu trúc đặc trưng khó phân biệt giữa ảnh thật và ảnh giả mạo.

IV. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Nghiên cứu này đã này thử nghiệm một số mô hình học sâu để đưa ra đánh giá hiệu năng mô hình phù hợp nhất cho bài toán nhận dạng ảnh khuôn mặt thật giả. Các kỹ thuật học chuyển đổi (transfer learning), mô hình lai ghép (hybrid models), mô hình Vision Transformer đã được áp dụng trên cùng một tập dữ liệu ảnh khuôn mặt thật giả để so sánh. Trong đó, kỹ thuật học chuyển đổi được áp dụng cho 8 mô hình InceptionResNetV2, EfficientNetV2, VGG16, ResNet50, EfficientNetB7, MobileNet, ResNet101, DenseNet201, kỹ thuật lai ghép áp dụng cho 3 mô hình VGG16-XGBoost, VGG16-CatBoost, VGG16-LightGBM. Kết quả thử nghiệm cho thấy các mô hình lai ghép và mô hình Vision Transformer cho kết quả khá tốt, nhưng mô hình ResNet50 cho kết quả tốt nhất đạt 87% độ chính xác.

Hướng phát triển các nghiên cứu tiếp theo có thể thử nghiệm các mô hình học sâu với các tập dữ liệu khác nhau, thử nghiệm tích hợp trên các website, ứng dụng di động… để cảnh báo người dùng khi tự động nhận dạng có hình ảnh giả mạo khi truy cập Internet.

TÀI LIỆU THAM KHẢO

[1] Raza A., Kashif M. and Mubarak A. (2022) "A Novel Deep Learning Approach for Deepfake Image Detection" Applied Sciences 12, no. 19: 9820. https://doi.org/10.3390/app12199820.

[2] Chen B., Xin L., Zhihua X., Guoying Z. (2023) "Privacy-preserving DeepFake face image detection, Digital Signal Processing, Volume 143, 104233, ISSN 1051-2004, https://doi.org/10.1016/j.dsp.2023.104233

[3] Rafique R., Gantassi R., Amin R. (2023) Deep fake detection and classification using error-level analysis and deep learning. Sci Rep 13, 7422, https://doi.org/10.1038/s41598-023-34629-3

[4] Sangavi N., K. Premalatha, V. Dhananjayan, V. R. Kiruthika and P. Ananthi, "Augmentation Applied with Transfer Learning Approach to Detect Real and Fake Faces," 2023 Third International Conference on Smart Technologies, Communication and Robotics (STCR), Sathyamangalam, India, 2023, pp. 1-5, doi: 10.1109/STCR59085.2023.10396919

[5] Jaffar A., Mohammad W., Dheeb Albashish, Elaf Aljaafrah, Ryan Alturki, Bandar Alshawi (2024) "Using Deep Learning to Recognize Fake Faces", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 15, No. 1, 2024

[6] Cường N. H và Tùng Đ. C (2021) "Nghiên cứu và đánh giá một số phương pháp học máy ứng dụng trong phát hiện ảnh giả mạo", Kỷ yếu Hội thảo Khoa học Quốc gia về Công nghệ thông tin và Ứng dụng trong các lĩnh vực lần thứ 10 - CITA 2021, Đại học Đà Nẵng, ISBN: 978-604-84-5998-7

[7] Xhlulu (2019) "140k Real and Fake Faces - 70k real faces from Flickr and 70k fake faces GAN-generated", Retrieved on 04 Nov. 2023 from https://www.kaggle.com/datasets/xhlulu /140k-real-and-fake-faces

[8] Karen S. & Andrew Z. (2015) "Very Deep Convolutional Networks for Large-Scale Image Recognition", ICLR 2015, Computer Vision and Pattern Recognition, https://doi.org/10.48550/arxiv.1409.1556

[9] Chen T. and Guestrin C. (2016) "XGBoost: A Scalable Tree Boosting System", KDD'16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 2016, San Francisco, USA, Pages 785–794, https://doi.org/10.1145/2939672.2939785

[10] Guolin K., Meng Q., Finley T., Wang T., Chen W., Ma W., Ye Q. & Liu T.Y. (2017) "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," in Proceedings of the 31st International Conference on Neural Information Processing, Long Beach California USA

[11] Dorogush A. V., Gulin A., Gusev G., Ostroumova L., Prokhorenkova, & Vorobev A. (2018) "Catboost: unbiased boosting with categorical features", Advances in Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada

[12] Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn, D., Zhai X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S. và cộng sự, (2021) "An image is worth 16×16 words: Transformers for image recognition at scale," Proceedings of the 9th International Conference on Learning Representations

PERFORMANCE EVALUATION OF SOME DEEP LEARNING NETWORKS IN RECOGNITION OF FAKE REAL FACE IMAGES

Abstract: This study tests several deep learning models to evaluate the performance of the most suitable model for the problem of recognizing real and fake facial images. Transfer learning techniques, hybrid models, and Vision Transformer models are applied on the same dataset of real and fake face images for comparison. In particular, transfer learning techniques are applied to 8 models InceptionResNetV2, EfficientNetV2, VGG16, ResNet50, EfficientNetB7, MobileNet, ResNet101,

DenseNet201, hybrid techniques are applied to 3 models VGG16-XGBoost, VGG16-CatBoost, VGG16 - LightGBM. Experimental results show that the hybrid models and Vision Transformer models give quite good results, but the ResNet50 model gives the best results.

Keywords: Face image, real and fake, deep learning, transfer learning, hybrid models

Trần Quý Nam tốt nghiệp Kỹ sư (năm 1999) và Thạc sỹ (năm 2005) ngành công nghệ thông tin hệ chính quy tại Đại học Bách Khoa Hà Nội, nhận bằng Tiến sỹ quản lý công nghệ thông tin tại Đại học Quốc gia Xê-Un Hàn Quốc năm 2010. Lĩnh vực nghiên cứu: ứng dụng học máy, học sâu, trí tuệ nhân tạo, kiến trúc chính phủ điện tử, chuyển đổi số, tối ưu hệ thống thông tin, phát triển ứng dụng đa phương tiện.