BUILDING A DEEP LEARNING MODEL FOR THE IDENTIFICATION OF HUMANS WITHOUT A MASK

Ma Thi Hong Thu[1*], Nguyen Thi Ngoc Anh[2]
[1] *Tan Trao University*
[2]*TNU - School of Foreign Languages*

ARTICLE INFO

Received: 10/3/2022
Revised: 23/5/2022
Published: 25/5/2022

KEYWORDS

Deep learning
Identify human without masks
Convolutional Neural Network
Covid-19
Face Recognition

ABSTRACT

Nowadays, the Covid-19 epidemic has been causing significant impacts on health, economy, and society in many countries around the world as well as in Vietnam. This is the top concern of WHO and quarantine centers of countries. Therefore, identifying people who are not wearing masks is one of the prerequisites to prevent the spread of the virus. In this paper, we present a real-time maskless person recognition system based on deep learning. Our system consists of two main models that are the RetinaFace and lightweight CNN models. The RetinaFace model is responsible for extracting faces from the camera input data. The proposed lightweight CNN model to identify people without masks from faces that extracted from the RetinaFace model. The experiment results show that the lightweight CNN model achieves 96.88% accuracy on the test data set. Besides, compared with existing systems in practice, our proposed system has more advantages in terms of accuracy, analysis time, construction, and maintenance costs of the system.

XÂY DỰNG MÔ HÌNH HỌC SÂU NHẬN DẠNG NGƯỜI KHÔNG ĐEO KHẨU TRANG

Ma Thị Hồng Thu[1*], Nguyễn Thị Ngọc Anh[2]
[1]*Trường Đại học Tân Trào*
[2]*Trường Ngoại Ngữ - ĐH Thái Nguyên*

THÔNG TIN BÀI BÁO

Ngày nhận bài: 10/3/2022
Ngày hoàn thiện: 23/5/2022
Ngày đăng: 25/5/2022

TỪ KHÓA

Học sâu
Nhận dạng người không đeo khẩu trang
Mạng tích chập
Covid-19
Nhận dạng khuôn mặt

TÓM TẮT

Hiện nay, dịch Covid-19 đã và đang gây ra những ảnh hưởng không nhỏ đến sức khỏe, kinh tế và xã hội ở nhiều nước trên thế giới cũng như Việt Nam. Đây là mối quan tâm hàng đầu của WHO cũng như các trung tâm kiểm dịch của các quốc gia. Do đó, nhận dạng người không đeo khẩu trang là một trong những yếu tố tiên quyết để phòng chống sự lây lan của virus. Trong bài báo này chúng tôi trình bày một hệ thống nhận dạng người không đeo khẩu trang trong thời gian thực dựa trên học sâu. Hệ thống của chúng tôi bao gồm hai mô hình chính là mô hình RetinaFace và CNN nhẹ. Mô hình RetinaFace có nhiệm vụ trích xuất ra khuôn mặt từ dữ liệu đầu vào là camera. Mô hình CNN nhẹ được đề xuất nhằm nhận dạng người không đeo khẩu trang từ khuôn mặt được trích xuất từ mô hình RetinaFace. Kết quả thử nghiệm cho thấy mô hình CNN nhẹ đạt hiệu suất 96,88% độ chính xác trên tập dữ liệu thử nghiệm. Bên cạnh đó, so sánh với các hệ thống hiện có trên thực tế, hệ thống được chúng tôi đề xuất có nhiều ưu điểm hơn về độ chính xác, thời gian phân tích và trả kết quả, chi phí xây dựng, bảo trì hệ thống.

DOI: https://doi.org/10.34238/tnu-jst.5667

[*] Corresponding author. *Email: thutq7@gmail.com*

1. Giới thiệu

Đại dịch Covid-19 với những diễn biến khó lường đang diễn ra tại Việt Nam nói riêng và thế giới nói chung. Chỉ trong khoảng thời gian ngắn, nước ta đã ghi nhận hơn 4 triệu trường hợp mắc Covid-19 với các biến thể khác nhau. Số lượng ca bệnh đang gia tăng nhanh chóng trong cộng đồng, đặc biệt là tại các nơi tập trung đông các khu công nghiệp.

Đứng trước đại dịch và nhận thấy mức độ nguy hiểm, Bộ Y tế đã đưa ra các quy định để bảo đảm sức khỏe cho người dân gồm: Đeo khẩu trang thường xuyên; rửa tay bằng dung dịch sát khuẩn; giữ khoảng cách khi tiếp xúc với người khác; đo thân nhiệt,… Không những thế, khi có các triệu chứng mắc Covid-19 như ho, sốt, khó thở, mất khứu giác cần liên lạc với cơ quan Y tế để được tư vấn hỗ trợ [1].

Để hỗ trợ cho công tác phòng dịch bệnh Covid-19, đã có một số trường đại học, nhóm nghiên cứu thành công trong việc tạo ra các mô hình phát hiện người không đeo khẩu trang. Tuy nhiên, các mô hình đó vẫn còn có hạn chế như: tốc độ xử lý hình ảnh chậm, sử dụng các công nghệ truyền thống nên khả năng phát hiện các khuôn mặt trong cùng một thời điểm bị giới hạn, độ chính xác chưa cao,…[2], [3].

Ngày nay, các mô hình mạng nơron tích chập Convolutional Neural Network (CNN) ngày càng đạt được nhiều thành tựu nổi bật trong các bài toán về thị giác máy tính và xử lý ảnh như phân lớp ảnh [4], [5], phát hiện đối tượng trong ảnh [6], [7]. Các mô hình CNN nhẹ cũng được quan tâm với nhiều biến thể khác nhau như [8]-[11] nhằm mục đích cho phép các mô hình có thể triển khai trên các thiết bị di động, thiết bị nhúng trong thời gian thực. Trong bài báo này, chúng tôi giới thiệu một mô hình hiện đại dựa trên học sâu để giải quyết bài toán. Trong đó, chúng tôi đề xuất một mô hình CNN nhẹ nhằm thực hiện việc nhận dạng người không đeo khẩu trang, mạng của chúng tôi nhận đầu vào từ mô hình mạng RetinaFace [12] cho phép phát hiện khuôn mặt. Chúng tôi sử dụng mạng RetinaFace thay vì các mô hình nhận dạng khuôn mặt khác chẳng hạn như MTCNN vì RetinaFace cho độ chính xác cao hơn, có thể phát hiện khuôn mặt ngay cả trong những điều kiện ảnh khó khăn khi mà MTCNN không nhận dạng được, chẳng hạn như khuôn mặt nhỏ, ảnh thiếu sáng, góc chụp xa,... So sánh với một số phương pháp được đề xuất hiện nay [13], mô hình của chúng tôi cho phép nhận dạng người không đeo khẩu trang nhanh và chính xác hơn. Bên cạnh đó, phương pháp của chúng tôi cũng tốn ít tài nguyên và trang thiết bị hơn so với phương pháp sử dụng robot [14]. Các đóng góp của chúng tôi có thể tóm tắt như sau:

(1) Chúng tôi đề xuất một mô hình nhẹ nhằm nhận dạng người không đeo khẩu trang trong thời gian thực.

(2) Kết quả thử nghiệm cho thấy, mô hình của chúng tôi đạt hiệu suất vượt trội hơn so với một số mô hình hiện đại với chi phí về mặt bộ nhớ thấp và thời gian thực hiện nhanh.

(3) Kết hợp mô hình được đề xuất với mạng RetinaFace, chúng tôi xây dựng được một hệ thống phát hiện người không đeo khẩu trang hoàn chỉnh với chi phí thấp.

2. Nghiên cứu gần đây

Trong phần này, chúng tôi giới thiệu ngắn gọn một số phương pháp phát hiện khuôn mặt truyền thống qua thư viện OpenCV hay HOG. Bên cạnh đó, các mô hình học sâu hiện đại cũng được mô tả chẳng hạn như MTCNN hay RetinaFace.

2.1. Thư viện OpenCV

OpenCV (Open Source Computer Vision) [15] là thư viện thị giác máy tính phổ biến do Intel bắt đầu vào năm 1999. Thư viện đa nền tảng đặt trọng tâm vào xử lý hình ảnh thời gian thực và bao gồm các triển khai miễn phí bằng sáng chế của các thuật toán thị giác máy tính mới nhất. Vào năm 2008, Willow Garage tiếp nhận hỗ trợ và từ phiên bản OpenCV 2.3.1 hiện đi kèm với giao diện lập trình C, C ++, Python và Android. OpenCV được phát hành theo giấy phép BSD nên nó được sử dụng trong các dự án học thuật và các sản phẩm thương mại. OpenCV phiên bản

3.3 trở lên hiện đi kèm với nhiều lớp Face Recognition để nhận diện khuôn mặt chẳng hạn như Haar cascades [16].

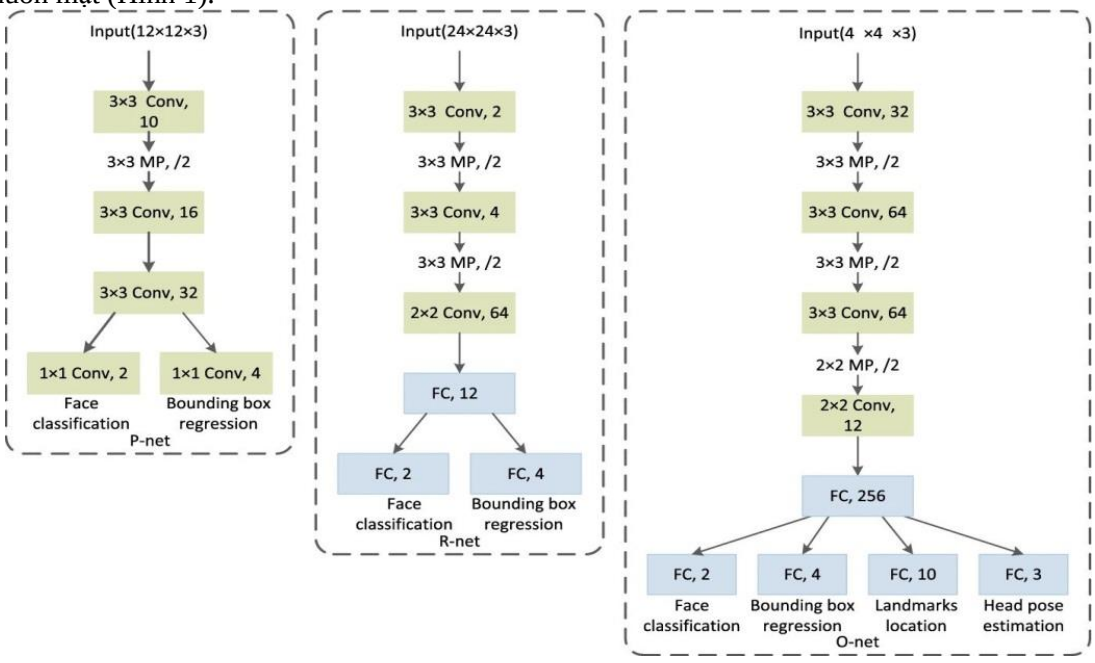
2.2. HOG + Linear SVM

HOG (Histogram of Oriented Gradients) là một trình mô tả tính năng đơn giản và mạnh mẽ. Nó không chỉ được sử dụng để nhận diện khuôn mặt mà còn được sử dụng rộng rãi để phát hiện đối tượng như ô tô, vật nuôi và trái cây. HOG mạnh mẽ để phát hiện đối tượng vì hình dạng đối tượng được đặc trưng bằng cách sử dụng phân bố gradient cường độ cục bộ và hướng cạnh. Để nhận dạng khuôn mặt thu được, một vectơ HOG các đặc điểm của khuôn mặt được trích xuất. Vectơ này sau đó được sử dụng trong mô hình SVM để xác định điểm phù hợp cho vectơ đầu vào với mỗi nhãn. SVM trả về nhãn có điểm tối đa, đại diện cho độ tin cậy đối với kết quả khớp gần nhất trong dữ liệu khuôn mặt được đào tạo. Nhược điểm duy nhất với tính năng nhận dạng khuôn mặt dựa trên HOG là nó không hoạt động trên khuôn mặt ở các góc lạ, nó chỉ hoạt động với khuôn mặt thẳng và mặt trước. Nó chỉ thực sự hữu ích nếu bạn sử dụng nó để phát hiện khuôn mặt từ các tài liệu được quét như giấy phép lái xe và hộ chiếu nhưng không phù hợp với video thời gian thực.

2.3. MTCNN

Một số phương pháp học sâu đã được phát triển và đề xuất để nhận dạng khuôn mặt trong những năm gần đây là "Mạng nơ-ron đa nhiệm xếp tầng" ("Multi-Task Cascaded Convolutional Neural Network") hay gọi tắt là MTCNN, được mô tả bởi Zhang, Kaipeng và cộng sự trong bài báo năm 2016 có tiêu đề "Phát hiện và căn chỉnh khuôn mặt chung bằng cách sử dụng mạng kết nối đa nhiệm xếp tầng" [17].

Mạng sử dụng cấu trúc tầng với ba mạng; đầu tiên hình ảnh được thay đổi tỷ lệ thành nhiều kích thước khác nhau (được gọi là kim tự tháp hình ảnh), sau đó mô hình đầu tiên (Mạng đề xuất hoặc P-Net) đưa ra các vùng khuôn mặt ứng viên, mô hình thứ hai (Mạng tinh chỉnh hoặc R-Net) lọc các bounding boxes và mô hình thứ ba (Mạng đầu ra hoặc O-Net) đề xuất các điểm mốc trên khuôn mặt (Hình 1).



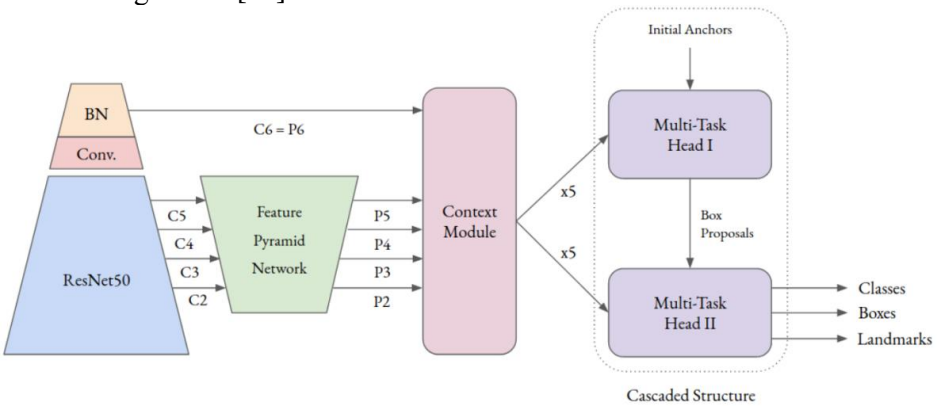
Hình 1. *Kiến trúc mạng MTCNN. Trong đó, Conv là phép tích chập covolution, FC là tầng kết nối đầy đủ Fully Connected, MP là tầng MaxPooling*

Quá trình hoạt động của mạng MTCCN được mô tả như sau: Trong giai đoạn đầu, nó tạo ra các cửa sổ ứng viên một cách nhanh chóng thông qua P-Net. Sau đó, nó tinh chỉnh các cửa sổ để loại bỏ một số lượng lớn các cửa sổ không có mặt thông qua một mạng CNN phức tạp hơn (R-Net). Cuối cùng, nó sử dụng một mạng CNN thứ ba (O-Net) mạnh hơn để tinh chỉnh kết quả và xuất ra các vị trí mốc trên khuôn mặt.

Ưu điểm của mạng MTCNN là độ chính xác cao hơn so với Haar cascades từ OpenCV và HOG+SVM. Do được huấn luyện trên tập dữ liệu đa dạng về khuôn mặt, góc chụp, độ sáng,... nên MTCNN phù hợp hơn đối với việc ứng dụng nhận dạng khuôn mặt trong các ứng dụng thực tế. Nhược điểm của MTCNN là khó phát hiện các khuôn mặt có độ phân giải thấp trong hình ảnh (các khuôn mặt ở xa so với máy ảnh khi được chụp).

2.4. RetinaFace

Một hệ thống nhận dạng khuôn mặt hiện đại bao gồm 4 giai đoạn phổ biến: phát hiện, căn chỉnh, đại diện và xác minh. Phát hiện và căn chỉnh là giai đoạn đầu và rất quan trọng. RetinaFace xử lý những giai đoạn đầu đó với điểm số tin cậy cao. RetinaFace chủ yếu dựa trên một nghiên cứu học thuật: "RetinaFace: Single-stage Dense Face Localisation in the Wild" được đăng trên hội nghị quốc tế CVPR 2020. Thiết kế mô hình dựa trên các kim tự tháp đặc trưng được mô tả như trong hình 2 [12].



Hình 2. *Kiến trúc mô hình RetinaFace. Trong đó, ResNet50 là kiến trúc mạng ResNet-50 trong [4]. Conv là phép tích chập Convolution, BN là BatchNormalization.*

Các thí nghiệm cho thấy RetinaFace có thể phát hiện khuôn mặt ngay cả trong những điều kiện ảnh khó khăn nhất (ví dụ: kích thước khuôn mặt nhỏ, thiếu ánh sáng, mặt bị che khuất một phần,...). Chính vì vậy, mạng RetinaFace cho độ chính xác cao hơn nhiều so với mạng MTCNN (chi tiết về so sánh được phân tích trong [12]). Do đó, sử dụng RetinaFace có thể xây dựng được một hệ thống nhận dạng khuôn mặt mạnh mẽ. Một số ví dụ về kết quả nhận dạng khuôn mặt của mô hình RetinaFace được thể hiện như trong hình 3.

Hình 3. *Ví dụ về kết quả phát hiện khuôn mặt của mô hình mạng RetinaFace (Hình ảnh được chụp lại từ bài báo gốc* [12]*)*

3. Mô hình đề xuất

3.1. Kiến trúc mạng đề xuất

Kiến trúc mạng CNN để nhận dạng người không đeo khẩu trang được thể hiện trong bảng 1, trong đó, Conv là tầng tích chập, BatchNorm là tầng Batch normalization [18], ReLU [19] và sigmoid là các hàm kích hoạt. FC là tầng kết nối đầy đủ Fully Connected, tầng Global Average Pooling là phép lấy trung bình toàn cục ma trận đầu vào. Mạng CNN được đề xuất dựa trên kiến trúc của mạng VGG-Net [20], tuy nhiên đã được tinh chỉnh nhằm làm giảm số tầng và độ phức tạp tính toán trong mạng, do đó phù hợp hơn với bài toán nhận dạng người không đeo khẩu trang trong thời gian thực. Đầu vào của mô hình này là các ảnh khuôn mặt nhận được sau khi áp dụng mô hình phát hiện khuôn mặt (chẳng hạn MTCNN hoặc RetinaFace). Các bức ảnh này được resize về kích thước 56x56 và được đưa vào mô hình. Đầu ra của mô hình bao gồm một giá trị tương ứng với:

1 - Thể hiện rằng ảnh khuôn mặt đầu vào có đeo khẩu trang.

0 - Thể hiện rằng ảnh khuôn mặt đầu vào không đeo khẩu trang.

Bảng 1. *Kiến trúc mạng CNN nhận dạng người không đeo khẩu trang*

STT	Tầng	Kích thước đầu ra
1	Input	56x56x3
2	Conv (7x7, stride = 2) BatchNorm ReLU	28x28x64
3	Conv (3x3, stride = 2) BatchNorm ReLU	14x14x128
4	Conv (3x3, stride = 2) BatchNorm ReLU	7x7x256

5	Global Average Pooling	256
6	FC Sigmoid	1

Cụ thể, mô hình được đề xuất của chúng tôi sẽ nhận đầu vào là các khuôn mặt được trích xuất từ mạng RetinaFace, sau đó sẽ dự đoán xem khuôn mặt đầu vào đó có đeo khẩu trang hay không. Sau đó, một hệ thống cảnh báo được xây dựng để tiến hành xử lý và đưa ra thông báo nhắc nhở đối với những người không đeo khẩu trang. Để thực hiện được điều đó, mô hình cần được huấn luyện trước trên bộ dữ liệu gồm các khuôn mặt có và không đeo khẩu trang.

3.2. Hàm mất mát

Binary Cross Entropy tính toán độ chênh lệch giữa 2 phân phối xác suất của dự đoán và của nhãn. Đầu ra của mô hình là phân phối xác suất (p, 1 - p) biểu diễn xác suất xảy ra của một trong hai nhãn. Phân phối xác suất của nhãn có dạng (y, 1 - y) với y nhận giá trị 0 hoặc 1 được thể hiện như trong công thức (1).

$$\mathcal{L} = -\sum_{i=1}^{N}(y_i \log a_i + (1 - y_i)\log(1 - a_i)) \quad (1)$$

Trong đó, y_i là nhãn chân lý và a_i là xác suất dự đoán của mô hình. N được dùng để thể hiện số điểm dữ liệu trong tập training. Cụ thể, đầu ra dự đoán của điểm dữ liệu thứ i đó là $a_i = sigmoid(F)$ là xác suất để điểm đó rơi vào class thứ nhất (với F là tầng cuối cùng của mô hình). Xác suất để điểm đó rơi vào class thứ hai có thể được dễ dàng suy ra là $(1 - a_i)$.

4. Thử nghiệm

4.1. Bộ dữ liệu và chi tiết cài đặt

4.1.1. Bộ dữ liệu

Bộ dữ liệu được sử dụng để huấn luyện mô hình nhận dạng người không đeo khẩu trang bao gồm hơn 92 nghìn hình ảnh. Dưới đây là địa chỉ tải bộ cơ sở dữ liệu: https://tinyurl.com/y2hajjvz

Trong bộ dữ liệu này, các ảnh khuôn mặt được chia thành 2 lớp: có đeo khẩu trang (with mask) và không đeo khẩu trang (without mask). Có 90.468 hình ảnh thuộc lớp "không đeo khẩu trang" và 2.203 hình ảnh thuộc lớp "có đeo khẩu trang". Một số ví dụ về các hình ảnh có và không đeo khẩu trang được mô tả qua hình 4.

Có đeo khẩu trang (with mask)

Không đeo khẩu trang (Without mask)

Hình 4. *Mô tả một số hình ảnh khuôn mặt thuộc cả hai lớp trong bộ dữ liệu*

Với bộ dữ liệu này, để cân bằng số lượng mẫu (hình ảnh) giữa lớp "có đeo khẩu trang" và "không đeo khẩu trang", chúng tôi chọn ngẫu nhiên 3.000 ảnh trong số 90.468 ảnh người không đeo khẩu trang và 2203 ảnh người có đeo khẩu trang để tiến hành huấn luyện mô hình nhận dạng người không đeo khẩu trang. Đối với bộ dữ liệu sau khi lựa chọn ngẫu nhiên này, chúng tôi sử dụng 90% tổng số ảnh để huấn luyện và 10% ảnh còn lại để đánh giá mô hình.

4.1.2. Chi tiết cài đặt

Chúng tôi huấn luyện mô hình mạng CNN nhận dạng người không đeo khẩu trang (bảng 1) trên máy tính sử dụng GPU 2070S, kích thước lô (batch size) được thiết lập là 64. Chúng tôi sử dụng trình tối ưu hóa SGD (Stochastic Gradient Descent) với động lượng (momentum) 0,9 và tốc độ học (learning rate) ban đầu là 0,001. Quá trình đào tạo được thực hiện trong 200 kỷ nguyên (epochs). Tỷ lệ học tập sẽ giảm 10 lần nếu sau 10 kỷ nguyên mà độ chính xác trên tập dữ liệu xác thực không được cải thiện. Tất cả các tham số trên được thiết lập dựa trên quá trình huấn luyện các mạng phổ biến như: ResNet [4], VGG-Net [20], Mobile-Net [10], ShuffleNet [11].

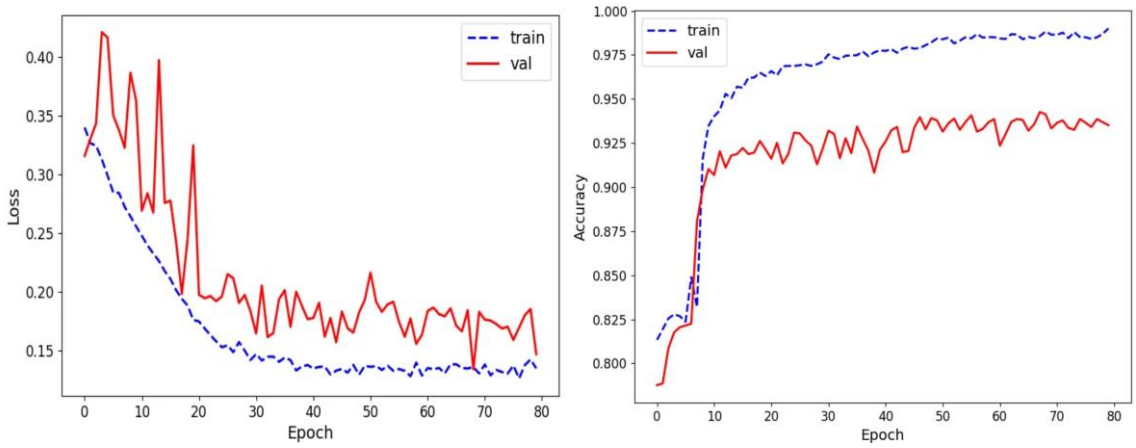
4.2. Kết quả thử nghiệm

Bảng 2 mô tả kết quả độ chính xác của mô hình nhận dạng người không đeo khẩu trang với bộ dữ liệu đã được mô tả ở trên. Kết thúc quá trình huấn luyện, chúng tôi thu được kết quả độ chính xác trên tập dữ liệu huấn luyện là 99,53% và độ chính xác trên tập dữ liệu đánh giá là 96,88%.

Bảng 2. *Kết quả và độ chính xác của mô hình*

Phương pháp	Độ chính xác	
	Tập huấn luyện	Tập đánh giá
Mô hình nhận dạng người không đeo khẩu trang từ khuôn mặt	99,53%	96,88%

Hình 5 mô tả trực quan hóa quá trình huấn luyện với hàm loss và độ chính xác trong 100 epoch đầu. Nhìn hình 5 ta có thể thấy rằng mô hình hội tụ sau khoảng epoch 50.



Hình 5. *Trực quan hóa hàm mất mát (loss) và độ chính xác (accuracy) của mô hình trên tập huấn luyện và tập đánh giá*

Tiếp theo, chúng tôi tiến hành so sánh mô hình mà chúng tôi đề xuất với các mô hình đã có trong thực tế (chẳng hạn sử dụng robot phát hiện khuôn mặt [6]) để thấy rõ ưu nhược điểm của từng loại mô hình trong bảng 3.

Bảng 3. *Đánh giá chung mô hình nhận dạng người không đeo khẩu trang*

Mô hình đã có trong thực tế	Mô hình tác giả đề xuất
Thuật toán áp dụng chưa tối ưu nên chưa nhận dạng và phân biệt được nhiều khuôn mặt trong cùng một thời điểm.	Nhận dạng và phân biệt được nhiều khuôn mặt trong cùng một thời điểm.
Phải đứng trong khoảng cách nhất định và mặt hướng thẳng vào Robot thì Robot mới phát hiện và đưa ra được cảnh báo.	Độ chính xác cao khi có thể phát hiện khuôn mặt không đeo khẩu trang và cảnh báo ngay cả trong những điều kiện khó khăn nhất như kích thước khuôn mặt nhỏ, thiếu ánh sáng, mặt bị che khuất một phần,…
Hệ thống xử lý gồm 4 thành phần chính gồm: camera, máy tính, loa và thân vỏ Robot dẫn đến bị cồng kềnh.	Hệ thống cần 3 thiết bị chính gồm máy tính, camera và loa.
Không thuận tiện trong trường hợp cần di chuyển.	Thuận tiện trong trường hợp cần di chuyển.
Tốn nhiều chi phí bảo trì, bảo dưỡng.	Tốn ít chi phí bảo trì, bảo dưỡng.

Cuối cùng, chúng tôi cung cấp một video demo về hệ thống nhận dạng người không đeo khẩu trang sử dụng mạng RetinaFace và mô hình CNN được chúng tôi đề xuất nhằm phát hiện và nhận dạng khuôn mặt đeo và không đeo khẩu trang tại địa chỉ: https://youtu.be/gXGq9FoLKGI.

5. Kết luận

Bài báo này trình bày một hệ thống nhận dạng người không đeo khẩu trang trong thời gian thực dựa trên học sâu. Trong đó, hệ thống bao gồm ba quá trình: (1) sử dụng mạng RetinaFace để phát hiện khuôn mặt, (2) đề xuất mô hình mạng CNN nhẹ nhận dạng người không đeo khẩu trang dựa trên các khuôn mặt đã được phát hiện từ mạng RetinaFace, (3) đưa ra cảnh báo đối với những người không đeo khẩu trang. Kết quả thử nghiệm mạng CNN nhẹ nhận dạng người không đeo khẩu trang cho thấy mô hình đạt hiệu suất 96,88% độ chính xác trên tập dữ liệu thử nghiệm. Bên cạnh đó, so sánh với một số hệ thống đã có trong thực tế như sử dụng robot phát hiện, hệ thống của chúng tôi tỏ ra ưu việt hơn ở nhiều khía cạnh như độ chính xác, thời gian phân tích và trả kết quả, chi phí xây dựng, bảo trì hệ thống,…

Lời cám ơn

Nghiên cứu này là kết quả của đề tài cấp cơ sở mã số 2021.2.01, được hỗ trợ bởi Trường Đại học Tân Trào, Tuyên Quang, Việt Nam.

TÀI LIỆU THAM KHẢO/ REFERENCES

[1] T. Phong, P. Phuong, T. My. "Cough reftlex: reason and mechanism" 2020-06-02. [Online]. Available: https://www.dieutri.vn/trieuchungnoi/phan-xa-ho-tai-sao-va-co-che-hinh-thanh. [Accessed Jan. 07, 2022].

[2] L. Li *et al.*, "A review of face recognition technology," *Proceedings of the IEEE/CVF International Conference on Computer Vision*, access 8, 2020, pp. 139110-139120.

[3] S. -H. Lin, "An introduction to face recognition technology," *Informing Sci. Int. J. an Emerg. Transdiscipl*, vol. 3, pp. 1-7, 2000.

[4] K. He *et al.*, "Deep residual learning for image recognition," *Proceedings of the IEEE/CVPR Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770-778.

[5] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," *International conference on machine learning*, PMLR, 2019, pp. 6105-6114.

[6] J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger," *Proceedings of the IEEE/CVPR Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7263-7271.

[7] T. Y. Lin *et al.*, "Focal loss for dense object detection," *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2017, pp. 2980-2988.

[8] D. Q. Vu, N. Le, and J. C. Wang, "Teaching yourself: A self-knowledge distillation approach to action recognition," *IEEE Access*, vol. 9, pp. 105711-105723, 2021.

[9] D. Q. Vu *et al.*, "A Novel Self-Knowledge Distillation Approach with Siamese Representation Learning for Action Recognition," *Proceedings of the IEEE/VCIP International Conference on Visual Communications and Image Processing*, 2021, pp. 1-5.

[10] A. Howard *et al.*, "Searching for mobilenetv3," *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1314-1324.

[11] N. Ma, Ningning *et al.*, "Shufflenet v2: Practical guidelines for efficient cnn architecture design," *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 116-131.

[12] J. Deng *et al.*, "Retinaface: Single-stage dense face localisation in the wild," *Proceedings of the IEEE/CVPR Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5203-5212.

[13] Jain, K. Anil, and S. Z. Li, *Handbook of face recognition*, vol. 1. New York: Springer, 2011.

[14] Q. Luu. "Using the robot to detect human without wearing mask", 2020. [Online]. Available: https://vnexpress.net/dung-robot-de-phat-hien-nguoi-khong-deo-khau-trang-4099618.html. [Accessed Jan. 12, 2022].

[15] "Open source library - OpenCV", 1999. [Online]. Available: https://opencv.org/. [Accessed Jan. 12, 2022].

[16] Sakai, T., Kanade, T., Nagao, M., & Ohta, Y. I. "Picture processing system using a computer complex", *Computer Graphics and Image Processing*, 2(3-4), 1973, pp. 207-215.

[17] K. Zhang *et al.*, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499-1503, 2016.

[18] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *International conference on machine learning*, PMLR, 2015, pp. 448-456.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 84-90, 2012.

[20] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Proceedings of the International Conference on Learning Representations*, 2015, pp.1-14.