BUILDING THE FACE MASK WARNING SYSTEM IN PUBLIC PLACE BASED ON CONVOLUTIONAL NEURAL NETWORK

XÂY DỰNG HỆ THỐNG CẢNH BÁO ĐEO KHẨU TRANG NƠI CÔNG CỘNG DỰA TRÊN MẠNG NƠ-RON HỌC SÂU

Nguyễn Duy Linh
Trường Đại học Quảng Bình

ABSTRACT: *Covid-19 is one of the most serious epidemics ever to ever happen in human history. This pandemic originated from Wuhan - China and rapidly spread around the world. Its influence on all aspects of social life such as health, economics, politics. In response to Covid-19, a lot of countries have enacted many strict regulations on the people and rushed to produce vaccines to prevent and fight the epidemic. In Vietnam, the Ministry of Health and local leaders have regulated and implemented many solutions to ensure the health of the people and avoid the spread of epidemics. In which, wearing a mask in public is the mandatory issue in the 5K Message. With the goal of improving people's awareness of wearing masks, this paper proposes a face mask warning system used in public based on deep neural networks. This application can be deployed in several places with a high concentration of people such as supermarkets, schools, pedestrian streets, offices, ... with available, low-cost equipment.*

Keywords: *Covid-19, convolutional neural network, face mask detection, face mask recognition, machine learning, deep learning, intelligent system.*

TÓM TẮT: *Covid-19 là một trong những dịch bệnh nghiêm trọng nhất từng xảy ra với lịch sử loài người. Đại dịch này xuất phát từ Vũ Hán - Trung Quốc và lan rộng khắp thế giới với mức độ lây lan nhanh chóng. Mức độ ảnh hưởng của nó đến mọi mặt của đời sống xã hội như sức khỏe, kinh tế, chính trị. Để đối phó với Covid-19, các quốc gia đã ban hành nhiều quy định bặt buộc đối với người dân và chạy đua sản xuất vắc-xin (vaccine) phòng và chống dịch bệnh. Ở Việt Nam, Bộ Y tế và lãnh đạo các địa phương đã quy định và thực hiện nhiều biện pháp nhằm đảm bảo sức khỏe cho người dân, tránh lây lan dịch bệnh. Trong đó việc đeo khẩu trang nơi công cộng là một nội dung bắt buộc trong Thông điệp 5K. Với mục đích nâng cao ý thức đeo khẩu trang của người dân, bài báo đề xuất hệ thống cảnh báo việc đeo khẩu trang nơi công cộng sử dụng mạng nơ-ron học sâu. Ứng dụng này có thể được triển khai ở một số nơi tập trung đông người như siêu thị, trường học, phố đi bộ, công sở, ... với các thiết bị sẵn có, chi phí thấp.*

Từ khoá*: Covid-19, mạng nơ-ron tích chập, phát hiện khẩu trang, nhận diện khẩu trang, máy học, học sâu, hệ thống thông minh.*

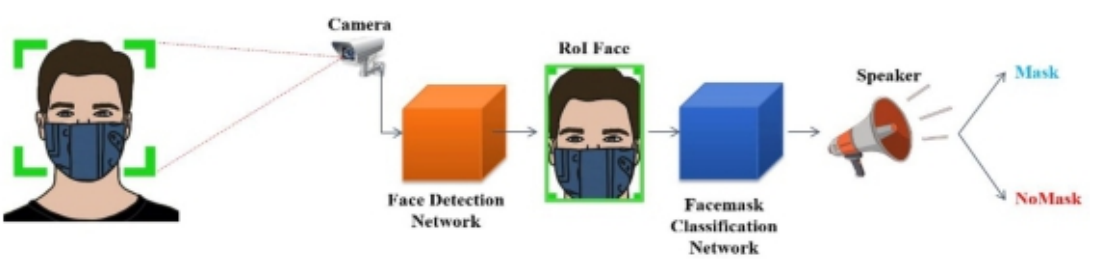
1. ĐẶT VẤN ĐỀ

Covid-19 là một đại dịch bệnh truyền nhiễm với tác nhân là virus SARS-CoV-2 bắt nguồn từ Vũ Hán - Trung Quốc. Theo thống kê của Tổ chức Y tế Thế giới tính đến ngày 15 tháng 3 năm 2021, thế giới có khoảng 119,603,761 ca lây nhiễm, 2,649,722 người chết do Covid-19 và con

số này đang tăng lên từng ngày [1]. Đây có thể coi là đại dịch lớn nhất xảy ra với con người, gây ảnh hưởng nghiêm trọng tới kinh tế, chính trị, đời sống xã hội của hầu hết các quốc gia trên thế giới. Để hạn chế sự lây nhiễm, chính phủ các nước đã đề ra nhiều sáng kiến và quy định bắt buộc đối với người dân và đặc biệt là bào chế, thử nghiệm các loại vắc-xin. Bên cạnh đó, việc phát triển các thiết bị công nghệ nhằm hỗ trợ cho công tác phòng dịch bệnh càng trở nên cấp thiết. Các hệ thống thông minh và sản phẩm công nghệ đã được nhiều nước triển khai như máy đo thân nhiệt từ xa, máy rửa tay sát khuẩn tự động, robot hỗ trợ công tác y tế, hệ thống khám chữa bệnh trực tuyến. Tuy nhiên, việc nghiên cứu sản xuất và chế tạo các loại máy này còn đang hạn chế do nguồn lực và kinh phí thực hiện cao. Với thông điệp 5K của Bộ Y tế Việt Nam, thì một số thao tác đơn giản người dân có thể chủ động để ngăn ngừa sự lây lan của dịch bệnh trong đó việc sử dụng khẩu trang được đặt lên hàng đầu. Từ những nghiên cứu đó, bài báo này đề xuất xây dựng hệ thống cảnh báo đeo khẩu trang nơi công cộng dựa trên mạng nơ-ron học sâu. Ứng dụng được triển khai chủ yếu dựa trên hai mạng nơ-ron học sâu hay còn được gọi là mạng nơ-ron tích chập (Convulotional Neural Network - CNN) được đề xuất là mạng phát hiện khuôn mặt (Face Detection Network) và mạng phân loại đeo hoặc không đeo khẩu trang (Face Mask Classification Network). Toàn bộ hệ thống được thực hiện dựa trên thư viện Pytorch sử dụng ngôn ngữ lập trình Python. Kết quả thử nghiệm trên máy tính cá nhân sử dụng vi xử lý Intel Core I7-4770 CPU @ 3.40 GHz, 8GB RAM kết nối với camara. Từ các đánh giá thử nghiệm, hệ thống này có thể được triển khai tại các địa điểm công cộng để góp phần phòng chống dịch bệnh với các thiết bị đơn giản và chi phí thấp.

2. PHÂN TÍCH VÀ XÂY DỰNG ỨNG DỤNG

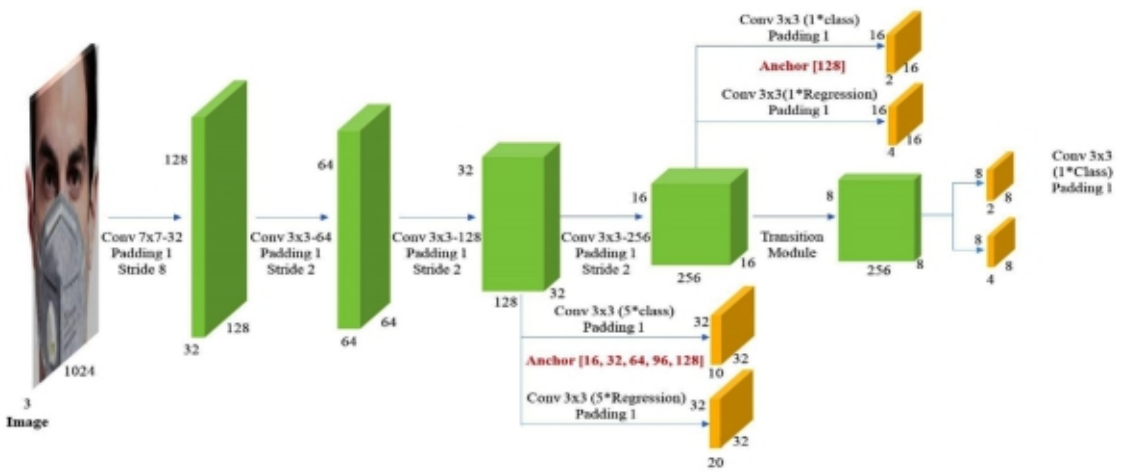
Tổng quan hệ thống cảnh báo việc sử dụng khẩu trang được mô tả chi tiết trong Hình 1. Hệ thống được chia thành hai mô-đun chính là phát hiện khuôn mặt (Face Detection Network) và phân loại đeo hoặc không đeo khẩu trang (Face Mask Classification Network). Trong mô-đun phát hiện khuôn mặt, bài báo đề xuất một mạng nơ-ron học sâu nhỏ gọn (Light-weight) để trích xuất vùng mặt (Face RoI). Đầu ra này được chuyển đến mô-đun phân loại đeo hoặc không đeo khẩu trang (Face Mask Classification Network). Ngoài ra, hệ thống còn được tích hợp camera để thu hình ảnh từ các địa điểm công cộng và hệ thống loa để phát thông tin cảnh báo.



Hình 1. Tổng quan mô hình hệ thống cảnh báo đeo khẩu trang nơi công cộng

2.1. Mạng phát hiện khuôn mặt (Face Detection Network)

Lấy cảm hứng từ hệ thống phân cấp tính năng kim tự tháp (Pyramid feature) [2] với một giai đoạn để phát hiện đối tượng, kiến trúc mạng CNN được sử dụng trong phần này được mô tả như trong Hình 2. Mạng này bao gồm một mạng xương sống (Backbone network) theo sau là một mô-đun chuyển tiếp (Transition Module) [3] và cuối cùng là ba bộ phát hiện (Detector) sử dụng để phân loại và hồi quy hộp giới hạn (Bounding box). Đầu ra của mạng này là vùng chứa khuôn mặt (RoI Face).

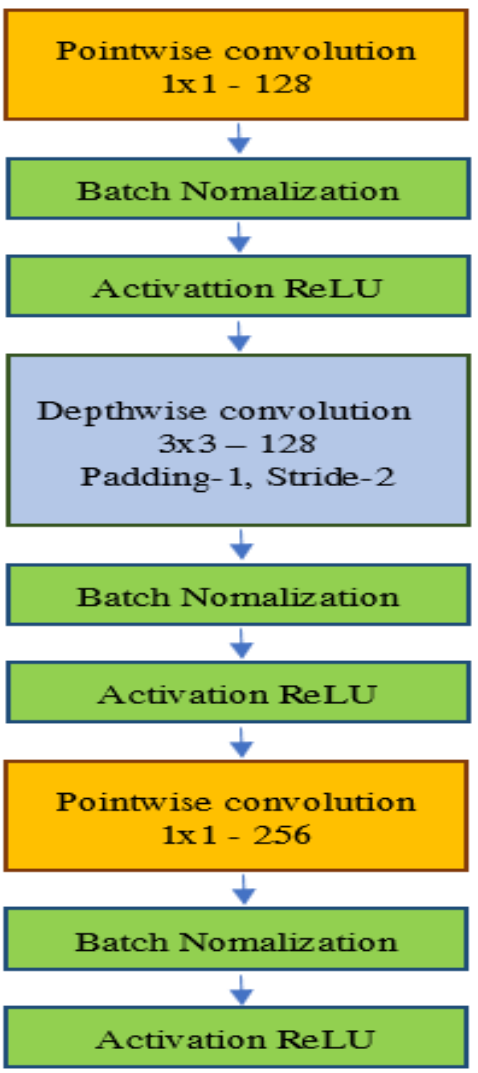


Hình 2. Mạng phát hiện khuôn mặt được đề xuất (Face Detection Network)

Mạng xương sống (Backbone network) bao gồm bốn phép toán tích chập (Convolution operations). Phép toán tích chập đầu tiên sử dụng kích thước bộ lọc 7 x 7 để giảm nhanh kích thước của bản đồ đặc trưng (Feature map). Mặc dù, bước này có thể mất nhiều thông tin từ ảnh gốc nhưng nó vẫn đảm bảo trích xuất các đặc điểm có lợi của kích thước khuôn mặt. Ba lớp tích chập tuần tự sử dụng bộ lọc kích thước 3 x 3 để giảm 64 lần kích thước đầu vào, đầu ra với 16 x 16 là kích thước của bản đồ đặc trưng cuối cùng. Nhóm tích chập với bộ lọc kích thước 3 x 3 có thể nhanh chóng giảm kích thước của bản đồ đối tượng và duy trì thông tin hữu ích so với việc sử dụng các lớp Max-pooling nhưng chúng sẽ tăng chi phí tính toán. Các bộ lọc được sử dụng tuần tự để lấy thông tin đặc trưng cơ bản như màu sắc, hình dạng và thông tin ngữ cảnh dựa trên độ sâu của lớp mạng. Ở các lớp cấp thấp, mạng chỉ có thể lọc thông tin cơ bản về hình dạng, cạnh và góc. Trong khi đó, các cấp cao hơn có thể cung cấp thông tin về các đặc tính cơ bản của khuôn mặt như mắt, mũi, miệng, cằm, má và da.

Tiếp theo sau mạng xương sống là mô-đun chuyển tiếp (Transition module), được phát triển dựa trên việc sử dụng phép tích chập theo chiều sâu (Depthwise convolution) [4]. Mô-đun này được sử dụng để giúp mạng phát hiện khuôn mặt hoạt động nhanh với các phép toán tích chập nhỏ gọn. Ngoài ra, sử dụng các phép

toán tích chập trước và sau phép tích chập theo chiều sâu để thay đổi số kênh trong mỗi bước. Cấu trúc của mô-đun này được thể hiện trong Hình 3.



Hình 3. Mô-đun chuyển tiếp (Transition module)

Mạng phát hiện khuôn mặt sử dụng ba bộ phát hiện, mỗi bộ sử dụng hai phép tích chập giống nhau 3 x 3 để phân loại và hồi quy hộp giới hạn (Bounding box). Các lớp này áp dụng tại các bản đồ đặc trưng có kích thước 32 x 32, 16 x 16 và 8 x 8. Bộ phát hiện sử dụng các hộp neo (Anchor box) hình vuông với nhiều kích cỡ khác nhau để dự đoán vị trí của khuôn mặt tương ứng trong ảnh gốc. Trong trường hợp này, nó sử dụng năm hộp neo với kích thước 16, 32, 64, 96 và 128 cho khuôn mặt cỡ nhỏ, một hộp neo có kích thước 256 cho khuôn mặt cỡ vừa và một hộp neo với kích thước 512 cho khuôn mặt cỡ lớn. Cuối cùng, bộ phát hiện tạo ra một vectơ bốn chiều (x, y, w, h) như độ lệch (Offset) vị trí khôn mặt và một vectơ hai chiều (Khuôn mặt hoặc không phải khuôn mặt) cho phân loại nhãn.

Bộ phát hiện sử dụng MultiBox loss [2] để tính toán mất mát (Loss funtion) trong mỗi lần lặp trong suốt giai đoạn huấn luyện. Sự mất cân bằng giữa hồi quy và phân loại sẽ được điều chỉnh bằng hệ số cân bằng mất mát, được mô tả như sau:

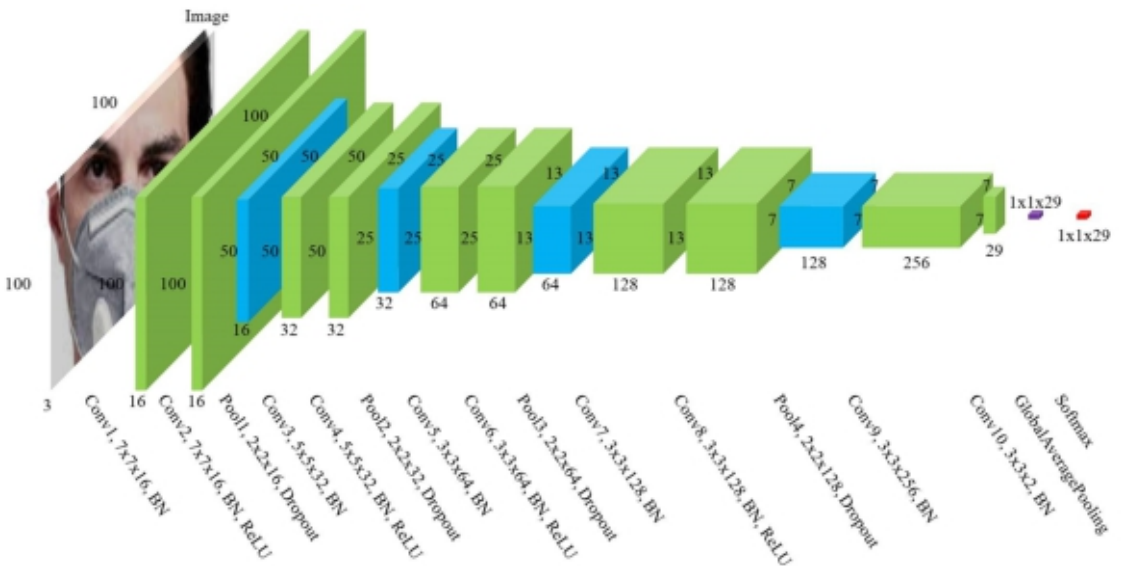
$$L(c_i, r_i) = \frac{2}{P}\sum_i L_{cls}\,(c_i, c_i^*) + \frac{1}{P}\sum_i L_{reg}\,(r_i, r_i^*),$$

trong đó P là các hộp phù hợp (Matching box), $L_{reg}(r_i, r_i^*)$ là mất mát hồi quy sử dụng Smooth $L1$ loss, $L_{cls}(c_i, c_i^*)$ là mất mát phân loại sử dụng Softmax-loss, r_i là tọa độ điểm trung tâm và kích thước của hộp dự đoán, c_i là các lớp được dự đoán, r_i^* là tọa độ thực của hộp giới hạn và c_i^* là nhãn đã cho.

2.2. Mạng phân loại việc sử dụng khẩu trang (Face Mask Classification Network)

Mô tả chi tiết về kiến trúc mạng của mô-đun này, như trong Hình 4. Tương tự như CNN dùng cho phân loại, mạng này được xây dựng bởi các khối tuần tự bao

gồm các lớp tích chập, lớp Average Pooling và hàm Softmax để phân loại dữ liệu. Sự khác biệt so với CNN được phân loại phổ biến là từ bản đồ đối tượng cuối cùng, nó sử dụng Global Average Pooling [5] thay vì các lớp được kết nối đầy đủ (Fully conneted layers).



Hình 4. Mạng phân loại việc đeo khẩu trang được đề xuất (Face Mask Classification Network)

Kiến trúc mạng này sử dụng một khối tích chập (Convolutional block) gồm hai lớp tích chập với bộ lọc kích thước 7 x 7, một khối tích chập với bộ lọc kích thước 5 x 5, hai khối tích chập với bộ lọc kích thước 3 x 3. Mỗi khối tích chập được theo sau bởi một lớp Average Pooling và một hàm kích hoạt ReLU (Rectified linear activation function). Bộ trích xuất đặc trưng kết thúc bằng một lớp chập với bộ lọc kích thước 3 x 3. Kích thước không gian của bản đồ đặc trưng được giảm từ 100 x 100 xuống còn 7 x 7. Sau đó, lớp Global Average Pooling được áp dụng để giảm thêm kích thước của bản đồ đặc trưng xuống còn 1 x 1. Cuối cùng, mạng sử dụng hàm Softmax để tính toán xác suất dự đoán của mỗi lớp (hai lớp tương ứng với 2 trạng thái đeo hoặc không đeo khẩu trang). Global Average Pooling có thể giảm thiểu khả năng overfitting bằng cách giảm tổng số tham số trong mạng. Mặt khác, để tránh tình trạng overfitting, phương pháp tối ưu Batch Normalization cũng được sử dụng. Bộ phân loại sử dụng hàm mất mát Cross-Entropy để tính toán tổn thất trong quá trình huấn luyện.

3. KẾT QUẢ THỬ NGHIỆM

3.1. Chuẩn bị tập dữ liệu (Dataset)

Mạng phát hiện khuôn mặt được đào tạo trên tập dữ liệu WIDER FACE [6] và được đánh giá trên tập dữ liệu Annotated Faces in the Wild (AFW) và tập dữ liệu về khuôn mặt PASCAL face. Các hình ảnh trong tập dữ liệu WIDER FACE được chọn từ tập dữ liệu

WIDE. Nó cung cấp 32,203 hình ảnh và gán nhãn 393,703 khuôn mặt với mức độ thay đổi về tỷ lệ và tư thế. Tập dữ liệu AFW [7] chứa 473 nhãn khuôn mặt với 203 hình ảnh. Bộ dữ liệu PASCAL face [8] cung cấp 1,335 nhãn khuôn mặt với 851 hình ảnh khuôn mặt người lớn và trẻ em được chụp ở các vị trí trong nhà và ngoài trời.

Mạng phân loại được đào tạo và đánh giá trên tập dữ liệu về khẩu trang được thu thập từ Internet với 8,019 hình ảnh trong đó 4,408 hình ảnh chứa khuôn mặt người đeo khẩu trang và 3,611 hình ảnh khuôn mặt không có khẩu trang. Tập dữ liệu này được chia thành 75% cho quá trình đào tạo và 25% cho quá trình đánh giá.

3.2. Cài đặt thử nghiệm

Các thí nghiệm trong bài báo này được đào tạo trên GPU Titan RTX và thử nghiệm trên CPU Intel Core I7-4770 @ 3.40 GHz, RAM 8GB. Để đào tạo mạng phát hiện khuôn mặt, nhiều tham số cấu hình đã được sử dụng để cải thiện chất lượng phát hiện. Phương pháp tối ưu Stochastic Gradient Descent đã được sử dụng, batch size là 16, weight decay là 5.10^{-4}, momentum là 0.9, learning rate từ 10^{-6} đến 10^{-3}. Ngưỡng IoU (Intersection over Union threshold) được chọn là 0.5 để tạo ra hộp giới hạn tốt nhất. Mạng phân loại sử dụng một số cấu hình cơ bản để phân loại hình ảnh như phương pháp tối ưu hóa Adam, batch size là 16, learning rate là 10^{-4}.

3.3. Kết quả

Mỗi mạng trong hệ thống được đào tạo, thử nghiệm riêng trên tập dữ liệu hình ảnh và mạng tổng thể được thử nghiệm trên hệ thống thời gian thực sử dụng camera kết nối PC (máy tính cá nhân) sử dụng CPU. Với bộ dữ liệu hình ảnh, mạng nhận diện khuôn mặt đạt được 95.37% AP (Average Presicion) trên tập dữ liệu AFW và 90.81 % AP trên tập dữ liệu PASCAL face. Kết quả của mạng phân loại trên bộ dữ liệu khẩu trang được thể hiện trong Bảng 1. Mạng phân loại được đề xuất có thể đạt độ chính xác là 99.50%. Kết quả này tốt hơn so với các mạng phân loại phổ biến như ResNet50, AlexNet, LeNet và tương đương với mạng VGG13 trong khi nó chứa số lượng tham số (Number of parameters - mục tiêu để xây dựng các mạng Light-weight CNN) rất nhỏ chỉ bằng 632,978.

Bảng 1. Kết quả so sánh phân loại việc đeo hoặc không đeo khẩu trang giữa mạng đề xuất và các kiến trúc mạng phổ biến

Mạng (Network)	**Độ chính xác (Accuracy) Đơn vị: %**	**Số lượng tham số (Number of parameters)**
Mạng đề xuất	**99.50**	**632,978**
VGG13 [9]	99.50	7,0527738
ResNet50 [10]	99.44	23,591,810
AlexNet [11]	99.38	67,735,938
LeNet [12]	99.19	15,653,072

Cuối cùng, toàn bộ hệ thống được thử nghiệm trên một camera được kết nối với hệ thống PC chạy trên CPU. Người tham gia đứng trước camera và thực hiện việc đeo hoặc không đeo khẩu trang. Hình ảnh từ camera sẽ được hệ thống phân loại, xuất ra các nhãn dự đoán trên màn hình được chỉ định bởi các nhãn trong tập dữ liệu và đồng thời xuất tính hiệu cảnh báo ra loa nếu có người không đeo khẩu trang đúng cách.

Hệ thống trong bài báo này đề xuất có thể đạt 46.93 FPS (frames per second) khi thử nghiệm trên camera và hoạt động tốt trong điều kiện bình thường mà không có độ trễ. Tuy nhiên, trong các điều kiện nhiễu như ánh sáng, tốc độ di chuyển của con người và khoảng cách giữa người với camera có thể làm giảm đáng kể độ chính xác của hệ thống. Những yếu tố này cần được cân bằng để có được RoI của khuôn mặt tốt nhất nhằm phân loại việc đeo khẩu trang một cách chính xác. Hình 5 thể hiện kết quả thử nghiệm của hệ thống với CPU và camera thông thường. Trong Hình 5, ở hàng thứ nhất thể hiện kết quả khi thử nghiệm với người tham gia đeo khẩu trang và không đeo khẩu trang. Hàng thứ hai là kết quả thử nghiệm với người tham gia đeo khẩu trang với các tư thế đầu khác nhau. Hàng thứ ba là kết quả thử nghiệm với nhiều người tham gia có đeo và không đeo khẩu trang. Nhãn hiển thị trên ảnh với Mask là chú thích cho việc có đeo khẩu trang và No_Mask là không đeo khẩu trang.

Hình 5. Kết quả thử nghiệm của hệ thống trên PC sử dụng CPU có kết nối với camera thông thường

4. KẾT LUẬN

Bài báo này đã đề xuất một hệ thống cảnh báo đeo khẩu trang với hai mạng nơ-ron học sâu nhỏ gọn (Light-weight). Mạng nhận diện khuôn mặt bao gồm mạng xương sống (Backbone network), mô-đun chuyển tiếp (Transition module) và ba bộ phát hiện (Detector). Mạng phân loại việc đeo hoặc không đeo khẩu trang là một mạng nơ-ron tích chập đơn giản bao gồm các lớp tích chập xen kẽ với các lớp Average Pooling, kết thúc bằng lớp Global Average Pooling và hàm Softmax. Số lượng tham số và tính toán đã được tối ưu giúp nó có thể được triển khai trên các thiết bị di động và máy tính chạy trên CPU. Trong tương lai, hệ thống này có thể tiếp tục được tối ưu hóa và tăng khả năng ứng dụng trong các hệ thống thời gian thực. Ngoài ra, tập dữ liệu cần được thu thập và chú thích trong nhiều điều kiện khác nhau để đảm bảo hệ thống này hoạt động chính xác nhất.

TÀI LIỆU THAM KHẢO

[1] Website của Tổ chức Y tế Thế giới (WTO): https://covid19.who.int, truy cập ngày 16/03/2021.

[2] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C. Fu, and A. C.Berg, "SSD: single shot multibox detector,"CoRR, vol. abs/1512.02325,2015.

[3] G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand,M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neuralnetworks for mobile vision applications,"CoRR, vol. abs/1704.04861,2017.

[4] Chollet, "Xception: Deep learning with depthwise separable convo-lutions,"CoRR, vol. abs/1610.02357, 2016.

[5] M. Lin, Q. Chen, and S. Yan, "Network in network,"2013.citearxiv:1312.4400Com ment: 10 pages, 4 figures, for iclr2014.

[6] Yang, Shuo and Luo, Ping and Loy, Chen Change and Tang, Xiaoou,"WIDER FACE: A Face Detection Benchmark", IEEE Conference onComputer Vision and Pattern Recognition (CVPR), 2016.

[7] X. Zhu, D. Ramanan, "Face detection, pose estimation, and landmarklocalization in the wild", in Proceedings of IEEE International Con-ference on Computer Vision Pattern Recognition (CVPR), IEEE, pp.2879-2886, 2012.

[8] Everingham, M. and Van Gool, L. and Williams, C. K. I. and Winn, J.and Zisserman, A, "The Pascal Visual Object Classes (VOC) Challenge",International Journal of Computer Vision, volume 88, 303-338, 2010.

[9] Simonyan, Karen and Andrew Zisserman. "Very Deep ConvolutionalNetworks for Large-Scale Image Recognition.", CoRR abs/1409.1556,2014.

[10] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for ImageRecognition", 2016 IEEE Conference on Computer Vision and PatternRecognition (CVPR), Las Vegas, NV, pp. 770-778, 2016.

[11] Krizhevsky, Alex, Ilya Sutskever and Geoffrey E. Hinton. "ImageNetClassification with Deep Convolutional Neural Networks.", NIPS (2012).

[12] Y. Lecun, L. Bottou, Y. Bengio and P. Haffner, "Gradient-based learningapplied to document recognition", in Proceedings of the IEEE, vol. 86,no. 11, pp. 2278-2324, Nov. 1998.

Liên hệ:
NCS.ThS. Nguyễn Duy Linh
Khoa Kỹ thuật - Công nghệ thông tin, Trường Đại học Quảng Bình
Địa chỉ: 312 Lý Thường Kiệt, TP Đồng Hới, tỉnh Quảng Bình
Email: nguyenduylinhqbu@gmail.com

Ngày nhận bài: 7/4/2021
Ngày gửi phản biện: 8/4/2021
Ngày duyệt đăng: 20/5/2021