ỨNG DỤNG MÔ HÌNH FACENET TRONG VIỆC XÂY DỰNG VÀ PHÁT TRIỂN HỆ THỐNG NHẬN DIỆN KHUÔN MẶT TẠI TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP HÀ NỘI

FACENET MODEL APPLICATION IN THE CONSTRUCTION AND DEVELOPMENT OF FACE RECOGNITION SYSTEM AT HANOI UNIVERSITY OF INDUSTRY

Phạm Việt Anh[1,*], Lê Xuân Hải[1], Vương Trung Hiếu[1]

TÓM TẮT

Hệ thống nhận diện khuôn mặt là một trong những ứng dụng dựa trên nền tảng về xử lý ảnh và các phương pháp về học máy, nó giúp máy tính tự động xác định hoặc nhận dạng một người nào đó từ một bức ảnh hay một khung hình video. Có rất nhiều những thuật toán đã được đề cập và một trong số chúng có thể kể tới là việc so sánh các đặc điểm khuôn mặt được trích chọn từ hình ảnh với một cơ sở dữ liệu về các khuôn mặt đã được thu thập từ trước đó (one-to-many matching) [1]. Tuy nhiên, nếu chỉ sử dụng các thuật toán đơn thuần như vậy thì ngay cả với một cơ sở dữ liệu ảnh nhỏ, hệ thống nhận diện cũng sẽ tốn rất nhiều tài nguyên và thời gian trong việc tính toán mà vẫn chỉ đưa ra một dự đoán có độ chính xác rất thấp. Trong những năm gần đây, với sự phát triển mạnh mẽ về học sâu mà đặc biệt là sự phát triển của những mạng neural tích chập thì các hệ thống nhân diện đã được chú trọng và cải tiến đáng kể hơn bao giờ hết. Mô hình Facenet ra mắt vào năm 2015 và được ứng dụng vào hầu hết trong các hệ thống nhận diện cho tới nay khi mang một ưu điểm nổi trội từ việc phát triển kiến trúc mạng Siamese kết hợp với việc sử dụng một hàm mất mát linh hoạt để huấn luyện trên bộ dữ liệu ảnh lớn. Trong bài báo này, nhóm tác giả sẽ phân tích cũng như đưa ra một số phương pháp cải tiến cho mô hình Facenet để ứng dụng trong việc xây dựng và phát triển một hệ thống nhận diện đáp ứng được với số lượng lớn sinh viên phục vụ cho việc điểm danh và quản lý sinh viên tại trường Đại học Công nghiệp Hà Nội.

Từ khóa: *Mạng neural tích chập, học sâu - nhận diện khuôn mặt.*

ABSTRACT

The face recognition system is one of applications, based on the foundation of photography editing and machine learning methodology, which assists computers in confirming and recognising someone from a picture or a video frame. There have been a lot of algorithms mentioned and one of them can be listed as the comparison of facial characteristics determined from pictures with a database of faces collected previously (one-to-many matching) [1]. However, the fact that using those common algorithms solely, even with a small image database, can lead to the waste of resources and time for the recognition system in calculations while the accuracy rate of a prediction remains low. In recent years, the significant development of deep learning, especially the development of convolution neural networks, has contributed to the focus and enhance more than ever of the recognition systems. The Facenet model was introduced in 2015, which has been applied to almost all recognition systems until now, having a remarkable advantage in the development of Siamese network architecture, co-operated with the utilization of a flexible loss function for the training in large image databases. In this article, the authorities will analyse as well as provide methodologies to enhance Facenet model for the application in constructing and developing a suitable recognition system meeting the requirement of large numbers of students in taking attendance and managing students at Hanoi University of Industry.

***Keywords*:** *Convolutional network neurral, Deep Learning, Face recognition.*

[1]Trường Đại học Công nghiệp Hà Nội

[*]Email: anhpv@haui.edu.vn

Ngày nhận bài: 10/01/2021

Ngày nhận bài sửa sau phản biện: 15/3/2021

Ngày chấp nhận đăng: 25/10/2021

1. GIỚI THIỆU

Hệ thống nhận diện khuôn mặt được tích hợp rất nhiều trong các hệ thống an ninh, thực thi luật, chăm sóc sức khỏe, giải trí,… Hệ thống nhận diện khuôn mặt là một trong những ứng dụng dựa trên nền tảng về xử lý ảnh và học máy, nó giúp máy tính tự động xác định hoặc nhận dạng một người nào đó từ một bức ảnh hay một khung hình video. Một hệ thống nhận diện được mong muốn là nó có khả năng tự động nhận diện và kiểm chứng các cá nhân trong những video hoặc hình ảnh. Bài toán về nhận diện khuôn mặt đã được nghiên cứu từ rất lâu và có rất nhiều thuật toán đã được đưa ra để thực hiện điều này, một trong số chúng có thể kể tới là việc so sánh các đặc điểm khuôn mặt được trích chọn từ hình ảnh với một cơ sở dữ liệu về các khuôn mặt đã được thu thập (one-to-many matching) [1]. Tuy nhiên, nếu chỉ sử dụng các thuật toán đơn thuần như vậy thì ngay cả với một cơ sở dữ liệu ảnh nhỏ, hệ thống

nhận diện cũng sẽ tốn rất nhiều tài nguyên và thời gian trong việc tính toán mà vẫn chỉ đưa ra một dự đoán có độ chính xác rất thấp. Trong quá trình phát triển, các nhà nghiên cứu cũng đưa ra rất nhiều những thư viện để hỗ trợ cho việc xây dựng các ứng dụng về nhận diện khuôn mặt. Việc sử dụng các thư viện sẵn có sẽ tiết kiệm về thời gian cài đặt, thời gian thu thập dữ liệu nhưng lại có những hạn chế trong việc ứng dụng tại những nơi có số lượng người lớn do việc không đảm bảo tính ổn định của việc dự đoán khi cơ sở dữ liệu ảnh gia tăng một cách đáng kể.

Trong những năm gần đây, với sự phát triển mạnh mẽ về học sâu mà đặc biệt là sự phát triển của những mạng neural tích chập mà các hệ thống nhân diện đã được chú trọng và cải tiến đáng kể hơn bao giờ hết. Thuật toán về nhận diện khuôn mặt dựa trên các mô hình học sâu được đề xuất tại [2, 3] đã đạt được hiệu suất tốt về thời gian xử lý và có độ chính xác rất cao [4-6]. Mô hình Facenet ra mắt vào năm 2015 và được ứng dụng vào hầu hết trong các hệ thống nhận diện cho tới nay khi mang một ưu điểm nổi trội từ việc phát triển kiến trúc mạng Siamese kết hợp với việc sử dụng một hàm mất mát linh hoạt để huấn luyện trên bộ dữ liệu ảnh lớn. Kiến trúc mạng Siamese của mô hình Facenet dựa trên nền tảng là một mạng neural tích chập được loại bỏ đi lớp kết nối đầy đủ (fully connected), đầu vào là một bộ ba ảnh trong tập dữ liệu với hai ảnh thuộc cùng một lớp để huấn luyện dựa trên một hàm mất mát có khả năng học được đồng thời sự tương đồng giữa các bức ảnh thuộc cùng một lớp và sự khác biệt giữa các bức ảnh không thuộc cùng một lớp [7]. Kết quả cuối cùng của việc huấn luyện sẽ tạo ra một bộ mã hóa ảnh thành một vector 128 chiều. Các vector này (vector embedding) mang những đặc điểm riêng biệt của khuôn mặt một người và có sự khác biệt so với các vector không thuộc cùng nhóm với nó. Như vậy, việc nhận diện khuôn mặt của một người chỉ là việc phân lớp vector từ ảnh khuôn mặt của người đó so với các lớp vector đã được mã hóa trong toàn bộ cơ sở dữ liệu ảnh. Ưu điểm của Facenet khá nổi trội khi mô hình chỉ cần thu thập dữ liệu ảnh được căn chỉnh tối ưu về vùng cắt của mặt và luôn đáp ứng một hiệu suất vô cùng ổn định với một số lượng dữ liệu ảnh rất lớn.

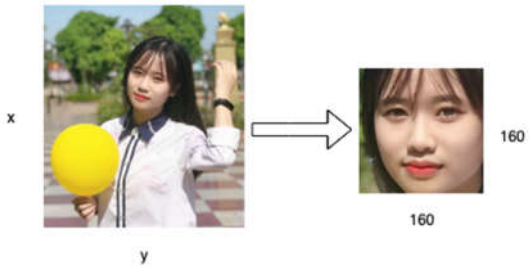
Trong bài báo này, nhóm nghiên cứu sẽ trình bày về mô hình Facenet và những cải tiến giúp nâng cao hơn hiệu suất của mô hình để từ đó ứng dụng vào việc xây dựng một hệ thống nhận diện khuôn mặt phục vụ trong việc quản lý, giám sát và điểm danh sinh viên tại trường Đại học Công nghiệp Hà Nội.

2. QUÁ TRÌNH TIỀN XỬ LÝ ẢNH

2.1. Xác định vị trí khuôn mặt trên ảnh

Trong giai đoạn tiền xử lý, việc xác định vị trí khuôn mặt trên ảnh là bước đầu tiên và được thực hiện dựa trên mô hình MTCNN (Multi-task Cascaded Convolutional Neural Networks) là một sự phát triển dựa trên mạng neural tích chập [8]. Lý thuyết về MTCNN đã được trình bày chi tiết tại [9] với việc dựa trên sự hoạt động của ba mạng neural tích chập là P-Net, R-Net và O-Net được thực hiện trên ba giai đoạn. Mỗi ảnh đầu vào được sao chép và thay đổi kích thước theo các tỷ lệ khác nhau. Trong giai đoạn đầu, P-Net hoạt động trong việc sử dụng một cửa sổ trượt có kích thước 12x12 chạy qua mỗi bức hình để tìm kiếm khuôn mặt. Sau lớp tích chập thứ ba, mạng được chia thành hai lớp nhỏ, một lớp đưa ra xác suất mà một khuôn mặt nằm trong miền xác định và lớp còn lại cung cấp các tọa độ của miền xác định. Trong giai đoạn tiếp theo, R-Net hoạt động tương tự như P-Net nhưng số lớp của R-Net nhiều hơn nhằm mục đích tinh chỉnh lại tọa độ của miền xác định từ P-Net. Cuối cùng, O-Net sẽ lấy các miền xác định từ R-Net làm đầu vào và đưa ra ba kết quả đầu ra bao gồm: xác suất của khuôn mặt nằm trong miền xác định, tọa độ được tinh chỉnh cuối cùng của miền xác định và tọa độ các bộ phận trên khuôn mặt.

Khi xác định được vị trí khuôn mặt, hình ảnh gốc trên tập dữ liệu có kích thước là x pixels × y pixels sẽ được thực hiện bằng cách cắt theo vùng khuôn mặt và đưa về kích thước 160×160. Kết quả được thực hiện qua hình 1.



Hình 1. Quá trình xác định khuôn mặt trên ảnh dựa trên phương pháp MTCNN

2.2. Tăng cường dữ liệu ảnh

Khi học sâu [10] đã trở nên phổ biến thì dữ liệu huấn luyện càng trở nên quan trọng hơn bao giờ hết [12]. Một mô hình học sâu cần có lượng dữ liệu rất lớn để có thể hoạt động tốt [11]. Về bản chất, mô hình Facenet cũng là một mạng học sâu nên để nâng cao độ chính xác cũng cần cung cấp một số lượng dữ liệu ảnh lớn. Trong nghiên cứu này, nhóm tác giả sẽ sử dụng một số kỹ thuật để có thể tăng cường dữ liệu của ảnh, phục vụ cho quá trình huấn luyện dữ liệu của mô hình học sâu.

Kỹ thuật tăng giảm độ sáng: Quá trình tăng giảm độ sáng của dữ liệu ảnh sẽ cho phép mô hình có khả năng dự đoán tốt với những điều kiện ánh sáng khác nhau trong thực tế. Tuy nhiên việc tăng giảm độ sáng cũng cần phải lưu ý với những giá trị hợp lý, tránh giảm ảnh quá tối hoặc quá sáng khiến mô hình phải học những dữ liệu không tốt.

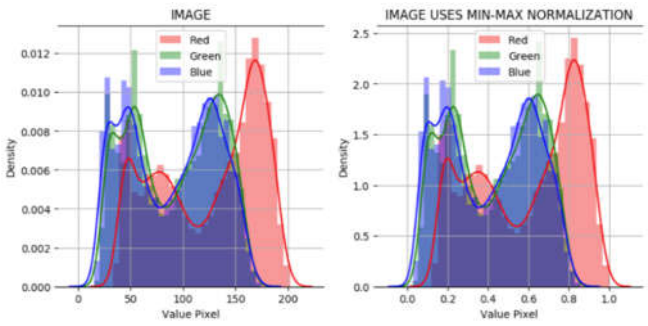
Hình 2. Tăng cường dữ liệu ảnh với kỹ thuật tăng giảm độ sáng

Kỹ thuật xoay ảnh: Phép xoay ảnh sẽ xoay hình ảnh một cách ngẫu nhiên theo kim đồng hồ với một góc nhất định từ trong khoảng 0 tới 360 độ. Phép xoay ảnh được sử dụng rất nhiều trong việc tạo ra dữ liệu khuôn mặt bởi lẽ trong thực tế, việc tính toán được thực hiện trong rất nhiều những góc độ khác nhau của khuôn mặt.

Hình 3. Tăng cường dữ liệu ảnh với kỹ thuật xoay ảnh

2.3 Chuẩn hóa dữ liệu ảnh

Xét với dữ liệu của bài toán là dữ liệu ảnh màu được đọc từ máy tính và được biểu diễn dưới dạng một ma trận ba chiều với các giá trị điểm ảnh là các số nguyên và nằm trong khoảng từ 0 tới 255. Nhận thấy rằng, miền giá trị của điểm ảnh có sự trải dài và có sự chênh lệch rõ rệt về giá trị điểm ảnh lớn nhất với điểm ảnh nhỏ nhất. Nếu sử dụng các thuật toán học sâu hay học máy với miền giá trị như vậy sẽ gây ra hai vấn đề, thứ nhất đó là việc thuật toán phải xử lý và làm việc với các dữ liệu có giá trị lớn, điều này sẽ làm cho việc tính toán mất nhiều thời gian, không ổn định và khó hội tụ. Thứ hai, các dữ liệu đầu vào trước khi đưa vào mạng học sâu để huấn luyện sẽ được lưu trữ vào bộ nhớ trong (RAM - Random Access Memory), với việc huấn luyện nhiều dữ liệu với giá trị cao rất dễ xảy ra hiện tượng tràn bộ nhớ và ảnh hưởng tới quá trình tính toán. Để khắc phục vấn đề này, nhóm nghiên cứu sử dụng phương pháp chuẩn hóa dữ liệu để điều chỉnh giá trị của dữ liệu về cùng một tỉ lệ trên một miền giá trị mới.



Hình 4. Kết quả sử dụng phương pháp chuẩn hóa min-max cho ảnh

Chuẩn hóa min-max có thể coi là phương pháp đơn giản nhất trong việc ánh xạ giá trị về phạm vi $[0,1]$. Công thức của phương pháp chuẩn hóa min-max:

$$x_i' = \frac{x_i - \min(x)}{\max(x) - \min(x)}$$

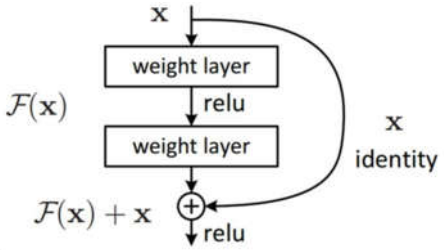
Với x_i là giá trị ban đầu, x_i' là giá trị mới sau khi được chuẩn hóa, $\min(x)$ là giá trị nhỏ nhất của đặc trưng và $\max(x)$ là giá trị lớn nhất của đặc trưng. Chuẩn hóa min-max bảo đảm mối quan hệ giữa các giá trị của dữ liệu gốc, nó sẽ phát hiện ra được các lỗi vượt quá giới hạn nếu như có giá trị đầu vào vượt quá khoảng giá trị cho phép. Ở đây, giá trị đặc trưng là điểm ảnh nên $\min(x) = 0$ và $\max(x) = 255$. Kết quả sử dụng phương pháp chuẩn hóa min-max được thể hiện ở hình 4.

3. MÔ HÌNH FACENET VÀ QUÁ TRÌNH HUẤN LUYỆN DỮ LIỆU

3.1. Lựa chọn mạng tích chập

Với việc được xây dựng dựa trên những ưu điểm của Inception module và Residual block được trình bày tại [13] mà Inception Resnet V1 là mạng sẽ được lựa chọn để phục vụ trong việc huấn luyện các dữ liệu hình ảnh khuôn mặt được thu thập. Residual block [14] sẽ giúp cho việc huấn luyện của mạng dễ dàng hơn rất nhiều khi tạo ra các kết quả tốt nhất. Mỗi Residual block chỉ cần thêm đầu vào của block (x) tới đầu ra của các lớp $\mathcal{F}(x)$ để thu được một kết quả $G(x)$ như công thức:

$$G(x) = \mathcal{F}(x) + x \quad (1)$$



Hình 5. Residual block

Khái niệm về Inception module đã được đề cập tại [15] khi nhóm nghiên cứu của Google đã phát triển và công bố mạng GoogLeNet. Inception module là một mạng tích chập giúp mạng huấn luyện được sâu và nhanh hơn thay vì việc phải tạo ra nhiều lớp sẽ rất dễ dẫn tới trường hợp mô hình bị overfitting (khái niệm về overfitting được trình bày tại [23]) và gia tăng số lượng về tham số [16]. Inception module sẽ tính toán trên các kernel có kích thước khác nhau từ một đầu vào của lớp trước đó và sau đó sẽ nối các đầu ra lại với nhau để tạo thành một đầu ra mới. Ưu điểm của các kernel 1×1 là để giảm số chiều và số lượng các tham số tính toán.

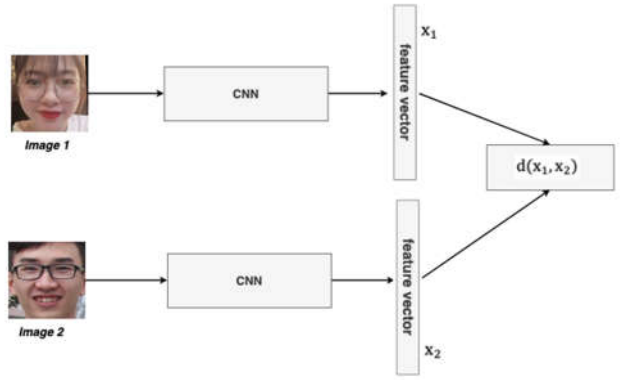
Về cơ bản thì Inception Resnet V1 cũng như các mạng tích chập khác gồm có 2 thành phần chính. Thành phần thứ nhất là khối chứa các lớp tích chập (hidden layers), thành phần thứ hai là khối chứa các lớp phân lớp. Tại thành phần thứ nhất, mạng thực hiện hàng loạt các phép tích chập và pooling để phát hiện ra các đặc trưng quan trọng của ảnh. Các pooling trong mạng tích chập còn có mục đích đạt được sự bất biến đối với việc thay đổi vị trí hoặc độ sáng của ảnh và tổng hợp một kết quả đầu ra dựa trên các giá trị nằm trong vùng mà kernel ánh xạ [17]. Tại thành phần thứ hai mỗi lớp với các liên kết sẽ đóng vai trò như một bộ phân lớp các đặc trưng đã được rút trích từ trước đó. Đầu ra cuối cùng của mạng sẽ đưa ra một xác suất của đối tượng tương ứng với ảnh đầu vào.

Bảng 1. Cấu trúc mạng Inception Resnet V1

Layer	Size-in	Size-out	Kernel	Stride, Padding	Params	ReLU	Scale
ConV_BN_ReLU	$160 \times 160 \times 3$	$79 \times 79 \times 32$	$3 \times 3 \times 3$	2,0	$32 \times 3 \times 3 \times 3$	True	–
ConV_BN_ReLU	$79 \times 79 \times 32$	$77 \times 77 \times 32$	$3 \times 3 \times 32$	1,0	$32 \times 3 \times 3 \times 32$	True	–
ConV_BN_ReLU	$77 \times 77 \times 32$	$77 \times 77 \times 64$	$3 \times 3 \times 32$	1,1	$64 \times 3 \times 3 \times 32$	True	–
MaxPool2D	$77 \times 77 \times 64$	$38 \times 38 \times 64$	3×3	2, –	0	True	–
ConV_BN_ReLU	$38 \times 38 \times 64$	$38 \times 38 \times 80$	$1 \times 1 \times 64$	1.0	$80 \times 1 \times 1 \times 64$	True	–
ConV_BN_ReLU	$38 \times 38 \times 80$	$36 \times 36 \times 192$	$3 \times 3 \times 80$	1,0	$192 \times 3 \times 3 \times 80$	True	–
ConV_BN_ReLU	$36 \times 36 \times 192$	$17 \times 17 \times 256$	$3 \times 3 \times 192$	2,0	$256 \times 3 \times 3 \times 192$	True	–
5×Inception A	$17 \times 17 \times 256$	$17 \times 17 \times 256$	Inception A	–	–	True	0,17
Reduction A	$17 \times 17 \times 256$	$8 \times 8 \times 896$	Reduction A	–	–	True	–
10×Inception B	$8 \times 8 \times 896$	$8 \times 8 \times 896$	Inception B	–	–	True	0,1
Reduction B	$8 \times 8 \times 896$	$3 \times 3 \times 1792$	Redution B	–	–	True	–
5×Inception C	$3 \times 3 \times 1792$	$3 \times 3 \times 1792$	Inception C	–	–	True	0,2
Inception C	$3 \times 3 \times 1792$	$3 \times 3 \times 1792$	Inception C	–	–	False	1,0
AvgPool2D	$3 \times 3 \times 1792$	$1 \times 1 \times 1792$	3×3	1, –	0	–	–
Flatten	$1 \times 1 \times 1792$	$1 \times 1 \times 1792$	–	–	–	–	–
Fully Connected	$1 \times 1 \times 128$	$1 \times 1 \times 128$	–	–	–	–	–
L2	$1 \times 1 \times 128$	$1 \times 1 \times 128$	–	–	–	–	–

3.2. Sử dụng hàm mất mát Triplet

Như đã trình bày ở phần trên, mô hình Facenet có một ưu điểm là việc phát triển kiến trúc mạng Siamese kết hợp với việc sử dụng một hàm mất mát linh hoạt để huấn luyện trên bộ dữ liệu ảnh lớn. Số lượng đầu ra của mạng neural tích chập trong (3.1) chính là số lượng lớp khuôn mặt trong cơ sở dữ liệu ảnh. Như vậy, nếu số lượng người cần dự đoán tăng lên một cách đáng kể thì lớp cuối của mạng sẽ chứa rất nhiều neural, điều này sẽ làm cho quá trình tính toán và huấn luyện trở nên phức tạp hơn chưa kể việc phải huấn luyện lại toàn bộ mạng khi có một lớp mới được tạo ra. Kiến trúc Siamese được tạo ra để giải quyết vấn đề này.



Hình 6. Kiến trúc của mạng Siemese

Kiến trúc của Siamese dựa trên nền tảng là một mạng tích chập được loại bỏ đi lớp đầu ra và chỉ được sử dụng để mã hóa ảnh thành một vector gọi là vector embedding. Đầu vào của mạng Siamese là hai bức ảnh bất kỳ được lựa chọn ngẫu nhiên từ dữ liệu ảnh và đầu ra của mạng là 2 vector embedding tương ứng với 2 ảnh từ đầu vào của mạng. Hai vector này thể hiện cho những đặc trưng của mỗi ảnh do quá trình được tính toán qua rất nhiều lớp tích chập trong mạng. Cuối cùng, hai vector sẽ được đưa vào một hàm mất mát (loss function) để đo lường sự khác biệt giữa chúng. Thông thường, hàm mất mát được sử dụng là một hàm norm chuẩn bậc 2.

Trong hình 6, mô hình đưa ra 2 vector là x_1 và x_2 biểu diễn lần lượt cho ảnh 1 và ảnh 2. Gọi $f(x)$ là hàm có tác dụng tương tự như một phép biến đổi qua lớp fully connected trong mạng neural để tạo phi tuyến và giảm chiều dữ liệu về các kích thước nhỏ. Khi đó nếu x_1, x_2 là cùng một người thì:

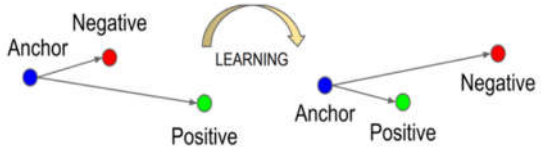
$$d(x_1, x_2) = \|f(x_1) - f(x_2)\|_2^2 \text{ đạt min} \quad (2)$$

và nếu x_1, x_2 là hai người khác nhau thì:

$$d(x_1, x_2) = \|f(x_1) - f(x_2)\|_2^2 \text{ đạt max} \quad (3)$$

Mục tiêu chính của Siamese là việc tìm ra cách ánh xạ một ảnh về một không gian vector n chiều cho nên cũng không nhất thiết phải lựa chọn hàm mất mát là binary cross entropy [18] như các bài toán phân loại nhị phân khác. Mô hình Facenet là một dạng Siemese với tác dụng là biểu diễn một ảnh về một không gian vector n chiều sao cho khoảng cách giữa các vector embedding càng nhỏ thì mức độ thuộc cùng một lớp của các ảnh tương ứng càng lớn. Việc ánh xạ ảnh như vậy có mục đích quan trọng là giảm chiều dữ liệu, giúp tăng tốc khả năng tính toán của thuật toán và hơn cả là vẫn giữ được độ chính xác khi nhận diện. Thông thường, số phần tử của vector embedding là 128 tương ứng với 128 điểm được trích chọn trên khuôn mặt.

Đối với hàm mất mát thông thường sẽ chỉ tính toán khoảng cách giữa hai bức ảnh, do vậy thì với một lần huấn luyện mô hình sẽ chỉ học được một trong hai khả năng đó là sự giống nhau nếu hai ảnh cùng một lớp hoặc sự khác nhau nếu hai ảnh khác lớp mà không thể học được cùng một lúc hai việc đó trong một lượt huấn luyện. Mô hình Facenet khắc phục điều này khi đưa ra hàm mất mát là Triplet với đầu vào là bộ ba ảnh anchor, positive và negative ký hiệu lần lượt là A, P và N. Ý tưởng chính của hàm mất mát này là đảm bảo với ảnh anchor A (là ảnh chỉ định đang xét) sẽ gần hơn với tất cả các ảnh positive P (là toàn bộ các ảnh của cùng người đó) so với các ảnh negative N là các ảnh không phải của người đó [19].



Hình 7. Quá trình huấn luyện của Facenet [19]

Khoảng cách giữa ảnh anchor tới positive nhỏ hơn so với ảnh anchor tới negative, nên:

$$d(A, P) < d(A, N) \quad (4)$$

$$\rightarrow \|f(A) - f(P)\|_2^2 + \alpha < \|f(A) - f(N)\|_2^2 \quad (5)$$

$$\forall (f(A), f(P), f(N)) \in \mathcal{T}$$

$$\rightarrow \|f(A) - f(P)\|_2^2 - \|f(A) - f(N)\|_2^2 + \alpha < 0 \quad (6)$$

Trong đó, hệ số $\alpha > 0$ có giá trị rất nhỏ được thêm vào để tạo ra lề giữa khoảng cách các các cặp ảnh positive và negative. $\mathcal{T}$ là tập hợp tất cả bộ ba trong tập huấn luyện. Hàm mất mát Triplet được viết đầy đủ:

$$\mathcal{L}(A, P, N) = \sum_{i=0}^{n} \|f(A) - f(P)\|_2^2 - \|f(A) - f(N)\|_2^2 + \alpha \quad (7)$$

Trong công thức (7), n là tổng số bộ ba trong tập huấn luyện của mô hình. Mục tiêu của Triplet loss về bản chất vẫn là giảm thiểu các trường hợp mô hình nhận diện sai ảnh negative thành positive nhất có thể và loại bỏ đi ảnh hưởng của các trường hợp mà mô hình nhận diện đúng negative và positive lên hàm mất mát. Để thể hiện chính xác mục tiêu, hàm Triplet trong (7) sẽ được điều chỉnh về dạng sau:

$$\mathcal{L}(A, P, N) = \sum_{i=0}^{n} \max(\|f(A) - f(P)\|_2^2 - \|f(A) - f(N)\|_2^2 + \alpha, 0) \quad (8)$$

3.3. Lựa chọn bộ ba ảnh đầu vào

Việc lựa chọn bộ ba ảnh đầu vào sẽ có ảnh hưởng rất lớn tới chất lượng của mô hình Facenet do mô hình sẽ hội tụ nhanh hơn và đồng thời đưa ra kết quả dự báo tốt hơn. Việc lựa chọn ngẫu nhiên bộ ba ảnh đầu vào về bản chất là cũng có thể được thực hiện do xác suất tỉ lệ chọn cặp ảnh ngẫu nhiên đều thuộc cùng một lớp là rất nhỏ do cơ sở dữ liệu ảnh của hệ thống nhận diện là rất lớn. Tuy nhiên việc này sẽ dẫn tới sự khó hội tụ của mô hình và là điều không mong muốn khi việc cải thiện mô hình là điều luôn được hướng tới.

Trong [19] đã đưa ra một chiến lược lựa chọn bộ ba ảnh đầu vào là Hard Triplets. Với mỗi ảnh A cần xác định ảnh P sao cho có khoảng cách là xa nhất với ảnh A tức là phải tìm nghiệm của:

$$\text{argmax}_{p_i} \|f(A_i) - f(P_i)\|_2^2 \quad (9)$$

Tương tự cũng xác định ảnh N sao cho có khoảng cách là gần nhất với ảnh A:

$$\text{argmin}_{n_j} \|f(A_i) - f(N_j)\|_2^2 \quad (10)$$

Trong đó, i, j là các nhãn của ảnh và các ảnh P và N lúc này được gọi lần lượt là hard positive và hard negative. Tuy nhiên, trong thực tế không thể tính toán được argmin và argmax trên toàn bộ tập huấn luyện và có thể dẫn tới việc mô hình huấn luyện kém do ảnh những khuôn mặt được gán nhãn sai và có chất lượng kém sẽ nhiều hơn hard positive và hard negative. Trong [19] cũng trình bày hai cách để giải quyết vấn đề này:

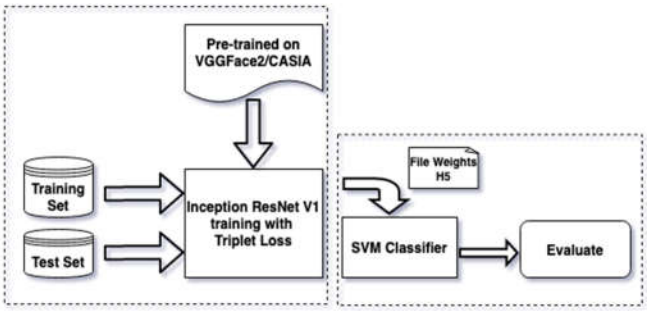
- Tạo bộ ba ảnh offline sau n bước, tính toán hard positive và hard negative và lưu vào checkpoint trên mỗi tập con dữ liệu.

- Tạo bộ ba ảnh online bằng cách chọn các mẫu hard positive và hard negative trên mỗi mini-batch.

4. HUẤN LUYỆN MÔ HÌNH VÀ ĐÁNH GIÁ KẾT QUẢ

4.1. Huấn luyện mô hình

Như đã trình bày trong phần trước, quá trình huấn luyện dữ liệu toàn mạng sẽ được thực hiện dựa trên kiến trúc của mạng Inception ResNet V1 với tập dữ liệu của hơn 200 người đã được thu thập và gán nhãn kết hợp với quá trình sử dụng pre-trained trên bộ dữ liệu là CASIA-WebFace.



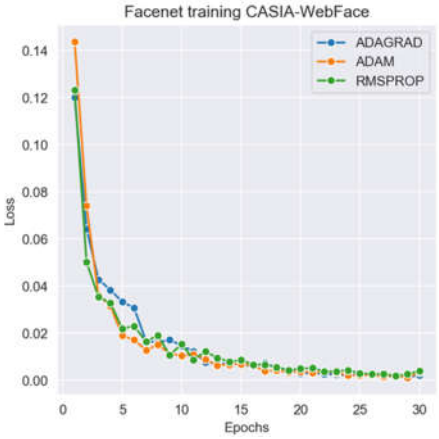
Hình 8. Quy trình huấn luyện và đánh giá mô hình

Việc tiến hành huấn luyện dữ liệu sẽ được thực thi trên máy chủ tính toán hiệu năng cao (HPC) được trang bị card đồ họa GPU NVIDIA TesLa P100 16GB.

Kiến trúc của Siamese dựa trên nền tảng là một mạng tích chập được loại bỏ đi lớp đầu ra và chỉ được sử dụng để mã hóa ảnh thành một vector gọi là vector embedding. Đầu vào của mạng Siamese là hai bức ảnh bất kỳ được lựa chọn ngẫu nhiên từ dữ liệu ảnh và đầu ra của mạng là 2 vector embedding tương ứng với 2 ảnh từ đầu vào của mạng. Hai vector này thể hiện cho những đặc trưng của mỗi ảnh do quá trình được tính toán qua rất nhiều lớp tích chập trong mạng. Cuối cùng, hai vector sẽ được đưa vào một hàm mất mát (loss function) để đo lường sự khác biệt giữa chúng. Thông thường, hàm mất mát được sử dụng là một hàm norm chuẩn bậc 2.

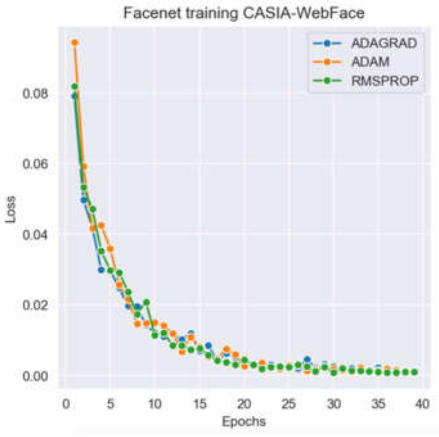
Bảng 2. Quá trình huấn luyện với tốc độ học 0,1

Epoch	Learning-rate	Optimizer	Number Image
39	0,1	ADAGRAD	453.953
39	0,1	ADAM	453.953
39	0,1	RMSPROP	453.953



Bảng 3. Quá trình huấn luyện với tốc độ học 0,01

Epoch	Learning-rate	Optimizer	Number Image
50	0,01	ADAGRAD	453.953
50	0,01	ADAM	453.953
50	0,01	RMSPROP	453.953



Quá trình huấn luyện sử dụng thuật toán lan truyền ngược tại [20] dựa trên hàm mất mát kết hợp với các giải thuật về tối ưu khác như Adam, Adagrad, Rmsprop được trình bày chi tiết tại [21] để tìm các trọng số tốt nhất cho mạng neural. Quá trình học được thể hiện rõ khi giá trị của hàm mất mát giảm dần và hội tụ về 0 sau mỗi lần học. Ngoài ra, nhóm nghiên cứu còn tiến hành huấn luyện dựa trên các giá trị về tốc độ học, số lượng epochs,... Kết quả cuối cùng của quá trình huấn luyện sẽ trả về một file trọng số của mạng neural.

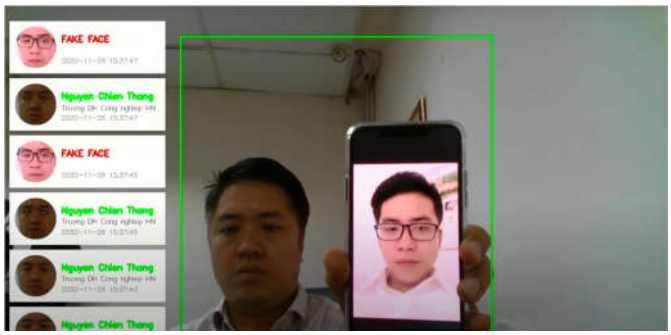
4.2. Đánh giá kết quả nhận diện

Dựa trên file trọng số đã tính toán, nhóm nghiên cứu tiến hành mã hóa ảnh trong cơ sở dữ liệu thành các vector 128 chiều. Số lượng ảnh thu thập được bao gồm 280 ảnh của 140 sinh viên trường Đại học Công nghiệp Hà Nội, trong đó 140 ảnh sẽ được đưa vào để mã hóa và huấn luyện, 140 ảnh còn lại sẽ được sử dụng để đánh giá độ chính xác. Nhóm nghiên cứu tiến hành đánh giá dựa trên giải thuật SVM (Support Vector Machine)[22] và thư viện FAISS (Facebook AI Similarity Search) của Facebook.

Bảng 4. Kết quả dự đoán dựa trên các phương pháp phân lớp vector

Method	Normalization	Face alignment	Processed time	Result
FAISS	False	False	9,98223	5,10%
SVM	False	False	7,89074	5,10%
FAISS	False	True	9,04781	8,02%
SVM	False	True	8,06475	6,56%
FAISS	True	False	11,0754	85,4%
SVM	True	False	7,83187	80,2%
FAISS	***True***	***True***	***8,81791***	***89,7%***
SVM	True	True	8,15443	88,3%

Từ những kết quả đã được nghiên cứu và phát triển, nhóm tác giả đã thực hiện thử nghiệm và xây dựng nên một hệ thống nhận diện và điểm danh khuôn mặt tại phòng lab mô phỏng và tính toán hiệu năng cao của Viện Công nghệ HaUI, Trường Đại học Công nghiệp Hà Nội. Kết quả được thể hiện trong hình 9.



Hình 9. Kết quả thử nghiệm được xây dựng trên hệ thống nhận diện

5. KẾT LUẬN

Bài báo đã trình bày về mô hình Facenet trong việc ứng dụng cho bài toán nhận diện khuôn mặt. Trong đó, những ưu điểm của mô hình đã được phân tích và thử nghiệm dựa trên bộ dữ liệu về sinh viên đã được thu thập. Các kết quả đánh giá cũng dựa trên nhiều phương pháp phân lớp khác nhau để đưa ra độ chính xác cao nhất.

Hướng phát triển tiếp theo của nghiên cứu là tối ưu các giải thuật để làm giảm thời gian tính toán cùng với đó là đề xuất các phương pháp chống giả mạo cho bài toán nhận diện và điểm danh sinh viên, cải thiện tốt hơn quá trình nhận diện và xây dựng một hệ thống điểm danh phục vụ cho trường Đại học Công nghiệp Hà Nội.

LỜI CẢM ƠN

Nghiên cứu này được thực hiện và thử nghiệm tại phòng Lab mô phỏng và tính toán hiệu năng cao thuộc Viện Công nghệ HaUI, Trường Đại học Công nghiệp Hà Nội trong đề tài cấp trường mã số 21-2020-RD/HĐ-ĐHCN.

TÀI LIỆU THAM KHẢO

[1]. Jiang X.D., Mandal B., Kot A., 2009. *Complete discriminant evaluation and feature extraction in kernel space for face recognition*. Machine Vision and Applications, Springer 20(1), 35-46.

[2]. M. Korkmaz, N. Yilmaz, 2015. *Face Recognition by Using Back Propagation Artificial Neural Network and Windowing Method*. 2015 2nd International Conference on Artificial Intelligence (ICOAI 2015), vol. 4, no. 1, pp. 15-19, 2015.

[3]. O. M. Parkhi, A. Vedaldi, A. Zisserman, 2015. *Deep Face Recognition*. Proceedings of the British Machine Vision Conference 2015, no. Section 3, pp. 41.1-41.12.

[4]. F. Schroff, D. Kalenichenko, J. Philbin, 2015. *Facenet: A unified embedding for face recognition and clustering*. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 815-823.

[5]. Y. Sun, D. Liang, X. Wang, X. Tang, 2015. *Deepid3: Face recognition with very deep neural networks*. arXiv preprint arXiv:1502.00873, 2015.

[6]. Y. Taigman, M. Yang, M. A. Ranzato, L. Wolf, 2014. *Deepface: Closing the gap to human-level performance in face verification.* in Proceedings of the IEEE conference on computer Vision and Pattern Recognition.

[7]. K. Q. Weinberger, J. Blitzer, L. K. Saul, 2006. *Distance metric learning for large margin nearest neighbor classification.* In NIPS. MIT Press, 2, 3.

[8]. F. Rahman, I. J. Ritun, N. Farhin, JiaUddin, 2019. *An Assistive Model for Visually Impaired People using YOLO and MTCNN.* ICCSP '19 Proceedings of the 3rd International Conference on Cryptography, Security and Privacy, pp. 225-230.

[9]. Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, Yu Qiao, 2016. *Joint Face Detection and Alignment using Multi-task Cascaded Convulotional Networks.* in IEEE Signal Processing Letters (SPL), vol.23, no. 10, pp. 1499-1503.

[10]. H. Li, Z. Lin, X. Shen, J. Brandt, G. Hua, 2015. *A convolutional neural network cascade for face detection.* in IEEE Conference on Computer Vision and Pattern Recognition, pp. 5325-5334.

[11]. I. Goodfellow, Y. Bengio, A. Courville, 2016. *Deep Learning.* The MIT Press.

[12]. S. H. Bach, B. D. He, A. Ratner, C. Re, 2017. *Learning the structure of generative models without labeled data.* in ICML, pp. 273–282.

[13]. Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, Alex Alemi, 2016. *Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning*. in ICLR.

[14]. K. He, X. Zhang, S. Ren, J. Sun, 2016. *Deep residual learning for image recognition.* In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778.

[15]. Szegedy C., Liu W., Jia Y., Sermanet P., Reed S., Anguelov D., Erhan D., Vanhoucke V., Rabinovich A., 2015. *Going deeper with convolutions.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1–9.

[16]. Min Lin, Qiang Chen, Shuicheng Yan, 2013. *Network in network*. CoRR, abs/1312.4400.

[17]. Krizhevsky A., Sutskever I., Hinton G. E., 2012. I*mageNet classification with deep convolutional neural networks*. In: NIPS, vol. 1.

[18]. A. Usha Ruby, Prasannavenkatesan Theerthagiri, I. Jeena Jacob, Y. Vamsidhar, 2020. *Binary cross entropy with deep learning technique for Image classification.* In: International Journal of Advanced Trends in Computer Science and Engineering, vol. 9, No.4.

[19]. Florian Schroff, Dmitry Kalenichenko, James Philbin, 2015. *FaceNet: A Unified Embedding for Face Recognition and Clustering*. arxiv 1503.03832.

[20]. Chauvin Y., D. E. Rumelhart, 1995. *Backpropagation: Theory, Architectures and Applications.* Erlbaum, Mahwah, NJ., ISBN: 080581258X, pp: 561.

[21]. Raniah Zaheer, Humera Shaziya, 2019. *A Study of the Optimization Algorithms in Deep Learning.* International Conference on Inventive System and Control (ICISC 2019), IEEE Xplore Part Number: CFP19J06-ART; ISBN:978-1-5386-3950-4.

[22]. Boser B. E., Guyon I. M., Vapnik V. N., 1992. *A training algorithm for optimal margin classifiers*. In D. Haussler, editor, 5th Annual ACM Workshop on COLT, pp. 144–152, Pittsburgh, PA. ACM Press.

[23]. Xue Ying, 2019. *An Overview of Overfitting and it Solutions.* IOP Conf.Series: Journal of Physics: Conf.Series 1168, 022022.

AUTHORS INFORMATION

Pham Viet Anh, Le Xuan Hai, Vuong Trung Hieu

Hanoi University of Industry