ĐẠI HỌC QUỐC GIA TP.HCM
TRƯỜNG ĐẠI HỌC BÁCH KHOA
KHOA KHOA HỌC & KỸ THUẬT MÁY TÍNH

———————— * ————————

ĐỀ CƯƠNGLUẬN VĂN TỐT NGHIỆP ĐẠI HỌC

XÂY DỰNG GIẢI PHÁP XÁC THỰC DỰA VÀO ẢNH CHỤP KHUÔN MẶT

Hội đồng	:	**Hệ Thống Mạng**
Giảng viên hướng dẫn	:	**PGS TS. Phạm Trần Vũ**
Giảng viên phản biện	:	**TS. Nguyễn Đức Thái**
Sinh viên thực hiện	:	**Nguyễn Khắc Quang Huy (1611288)**

TP. HỒ CHÍ MINH, THÁNG 12/2019

Lời cam đoan

Nhóm cam đoan mọi điều được trình bày trong báo cáo, cũng như mã nguồn là do nhóm tự thực hiện - trừ các kiến thức tham khảo có trích dẫn cũng như mã nguồn mẫu do chính nhà sản xuất cung cấp, hoàn toàn không sao chép từ bất cứ nguồn nào khác. Nếu lời cam đoan trái với sự thật, nhóm xin chịu mọi trách nhiệm trước Ban Chủ Nhiệm Khoa và Ban Giám Hiệu Nhà Trường.

Nhóm sinh viên thực hiện đề tài

Tóm tắt

Nhận diện khuôn mặt là một công nghệ được ứng dụng rộng rãi ngày nay trong các giải pháp xác thực sinh trắc học. Các hệ thống nhận diện khuôn mặt đã đạt đến mức độ chính xác rất cao trong việc nhận diện mặt người. Việc hiện thực công nghệ này trong xác thực không những giúp tăng trải nghiệm người dùng mà còn tăng độ bảo mật. Tuy nhiên, việc đó cũng làm nảy sinh một số vấn đề quan trọng: các hình thức mạo danh khác nhau được phát triển, trở thành mối đe dọa đối với việc xác thực. Face Anti-spoofing (chống mạo danh khuôn mặt) là một bước quan trọng trước khi đưa hình khuôn mặt vào xác thực. Đề cương này trình bày quá trình tìm hiểu nền tảng lý thuyết và đề xuất một giải pháp xác thực khuôn mặt với độ bảo mật và tin cậy cao.

Mục lục

Danh sách hình vẽ

1 Giới thiệu đề tài

1.1 Lý do và động lực thực hiện đề tài

Với sự phát triển liên tục của công nghệ thông tin, nhu cầu bảo mật và an toàn ngày càng cải thiện. Vì nhu cầu bảo mật, việc nhận diện khuôn mặt đã được nghiên cứu qua nhiều thập kỷ. Nhận diện khuôn mặt đã đươc sử dụng rộng rãi trong đời sống hằng ngày, đặc biệt là trong các hệ thống bảo mật, trong bảo mật thông tin và tương tác máy tính - con người. Các nghiên cứu nỗ lực vào việc cải thiện độ chính xác của việc nhận dạng và tốc độ phản hồi của những hệ thống nhận diện khuôn mặt. Những công nghệ mới nhất của nhận diện khuôn mặt đã được cải thiện đáng kể do sự xuất hiện của deep learning. Mặc dù những hệ thống này vận hành tốt trên lượng lớn các dữ liệu mặt từ web, hiệu năng và độ chính xác vẫn còn hạn chế khi chúng được áp dụng vào những trường hợp thực tiễn. [1]

Việc nhận diện khuôn mặt có thể gặp phải những trở ngại khác nhau từ trong ảnh chụp. Các vấn đề như độ sáng, tư thế, vật che chắn và độ tuổi người được chụp sẽ làm cho hệ thống nhận diện những bức hình của cùng một người là khác nhau.

Dữ liệu cũng là một vấn đề lớn cho hiệu năng của các hệ thống nhận diện khuôn mặt. Phân phối dữ liêu và kích thức dữ liệu sẽ làm ảnh hưởng rất nhiều đến kết quả. Càng nhiều dữ liệu sẽ làm cho hệ thống nhận diện tốt hơn, tuy nhiên vấn đề mất cân bằng dữ liệu có thể làm hại đến hiệu năng.

Hơn nữa, một vấn đề đáng lo ngại trong lĩnh vực nhận diện khuôn mặt là việc chống lại các kiểu tấn công mạo danh hòng đánh lừa những hệ thống xác thực. Những kiểu tấn công này ngày càng đa dạng trong hình thức để chống lại những hệ thống xác thực khác nhau. Các giải pháp chống lại các kiểu tấn công này còn

nhiều hạn chế. Việc rút trích ra những đặc trưng mà con người tự định nghĩa - những đặc trưng này dễ bị ảnh hưởng bởi điều kiện môi trường của ảnh chụp. Chỉ có một số ít giải pháp sử dụng Convolutional Neural Network đã được đề xuất nhằm chống lại các kiểu tấn công này. [7]

1.2 Mục tiêu và giới hạn đề tài

Mục tiêu đề tài là xây dựng một hệ thống xác thực bằng nhận diện khuôn mặt, với hiệu năng, độ chính xác và tính bảo mật cao. Hệ thống sẽ mô phỏng lại quá trình xác thực cho một máy bán hàng tự động, cho phép người dùng truy cập vào máy bán hàng chỉ qua xác thực khuôn mặt, dùng tài khoản liên kết để thực hiện các giao dịch trên máy bán hàng. Việc này đòi hỏi hệ thống phải cho phép nhận diện những khách hàng hoàn toàn mới chỉ bằng một lần chụp hình, đồng thời xác định được người đó trong những lần truy cập sau trong thời gian thực. Hệ thống cũng phải cung cấp những giải pháp hiệu quả giúp chống lại các kiểu tấn công mạo danh. Tuy nhiên, đề tài chỉ nghiên cứu giúp tìm ra một giải pháp nhận diện khuôn mặt hiệu quả chứ không chú trọng quá nhiều vào tổ chức, kiến trúc, quá trình phát triển, tính năng và trải nghiệm người dùng của ứng dụng đi kèm.

1.3 Cấu trúc báo cáo đề cương luận văn

Phần còn lại của báo cáo này được tổ chức như sau. Phần 2 trình bày các kiến thức nền tảng về Deep Learning, công nghệ nhận dạng khuôn mặt. Phần 3 trình bày giải pháp đề xuất của nhóm nghiên cứu và kiến trúc hệ thống được thiết kế cho giải pháp này. Phần 4 đề cập đến việc hiện thực và đánh giá ưu nhược điểm của giải pháp.

2 Quá trình tìm hiểu

2.1 Deep Learning

Trong những năm gần đây, Deep Learning và đặc biệt là CNN ngày càng trở nên hiệu quả trong các bài toán liên quan tới hình ảnh, như nhận dạng vật thể, phân loại hình ảnh, nhận dạng khuôn mặt...[1]

2.2 Nhận dạng khuôn mặt

Nhận diện khuôn mặt là việc tìm nhân dạng của một khuôn mặt trong một bức hình hoặc video trong một cơ sở dữ liệu khuôn mặt có sẵn. Nó bao gồm 2 giai đoạn: detection - phân biệt mặt người với các vật thể khác trong hình; identification - tìm ra nhân dạng của người có khuôn mặt này.

2.2.1 Face Detection

Face Detection là việc phân biệt một khuôn mặt đối với các hiện vật khác xung quanh trong một bức hình. Các hệ thống detection nhận vào một bức hình có thể chứa khuôn mặt trong đó và cho ra kết quả là vị trí của bức hình đó, thường là trong một hình chữ nhật (bounding box) bao gồm tọa độ của góc trên bên trái khuôn mặt và độ rộng, độ dài và hình chữ nhật. Việc nhận biết và căn chỉnh khuôn mặt trong những môi trường không bị giới hạn rất khó khăn do các yếu tố như nhiều tư thế, độ sáng và vật che chắn.

Mobile Net là một mạng CNN được sử dụng để phục vụ cho các bài toán liên quan đến phát hiện và nhận dạng vật thể. Nhờ áp dụng một kỹ thuật gọi

là depthwise separable convolution cùng với việc sử dụng các width multiplier và resolution multiplier để chịu mất một độ chính xác vừa phải nhằm giảm thiểu kích thước cũng như latency của mô hình và đạt hiệu năng nhanh hơn.[2]

Table 1. MobileNet Body Architecture

Type / Stride	Filter Shape	Input Size
Conv / s2	$3 \times 3 \times 3 \times 32$	$224 \times 224 \times 3$
Conv dw / s1	$3 \times 3 \times 32$ dw	$112 \times 112 \times 32$
Conv / s1	$1 \times 1 \times 32 \times 64$	$112 \times 112 \times 32$
Conv dw / s2	$3 \times 3 \times 64$ dw	$112 \times 112 \times 64$
Conv / s1	$1 \times 1 \times 64 \times 128$	$56 \times 56 \times 64$
Conv dw / s1	$3 \times 3 \times 128$ dw	$56 \times 56 \times 128$
Conv / s1	$1 \times 1 \times 128 \times 128$	$56 \times 56 \times 128$
Conv dw / s2	$3 \times 3 \times 128$ dw	$56 \times 56 \times 128$
Conv / s1	$1 \times 1 \times 128 \times 256$	$28 \times 28 \times 128$
Conv dw / s1	$3 \times 3 \times 256$ dw	$28 \times 28 \times 256$
Conv / s1	$1 \times 1 \times 256 \times 256$	$28 \times 28 \times 256$
Conv dw / s2	$3 \times 3 \times 256$ dw	$28 \times 28 \times 256$
Conv / s1	$1 \times 1 \times 256 \times 512$	$14 \times 14 \times 256$
$5\times$ Conv dw / s1	$3 \times 3 \times 512$ dw	$14 \times 14 \times 512$
$5\times$ Conv / s1	$1 \times 1 \times 512 \times 512$	$14 \times 14 \times 512$
Conv dw / s2	$3 \times 3 \times 512$ dw	$14 \times 14 \times 512$
Conv / s1	$1 \times 1 \times 512 \times 1024$	$7 \times 7 \times 512$
Conv dw / s2	$3 \times 3 \times 1024$ dw	$7 \times 7 \times 1024$
Conv / s1	$1 \times 1 \times 1024 \times 1024$	$7 \times 7 \times 1024$
Avg Pool / s1	Pool 7×7	$7 \times 7 \times 1024$
FC / s1	1024×1000	$1 \times 1 \times 1024$
Softmax / s1	Classifier	$1 \times 1 \times 1000$

Hình 2.1: *Kiến trúc mô hình MobileNet*

Kỹ thuật depthwise separable convolution là một giải pháp giúp rút ngắn thời gian tính toán một tác vụ convolution. Một lớp convolution tiêu chuẩn nhận vào một feature map **F** với chiều là $D_F \times D_F \times M$ và xuất ra một feature map **G** với số chiều là $D_G \times D_G \times M$, trong đó D_F là chiều rộng hoặc chiều dài trong không gian của feature map đầu vào, M là số lượng kênh đầu vào (input channels), D_G là chiều rộng hoặc chiều dài trong không gian của feature map đầu ra, N là số lượng kênh đầu ra.

Lớp convolution tiêu chuẩn có thông số là một kernel K với kích thước $D_K \times D_K \times M \times N$ trong đó D_K là chiều không gian của kernel được cho là hình vuông và M là số lượng kênh đầu vào, N là số lượng kênh đầu ra như đã định nghĩa trước.

Feature map đầu ra cho convolution tiêu chuẩn (với stride 0 và có padding) được tính như sau:

$$\mathbf{G}_{k,l,n} = \sum_{i,j,m} \mathbf{K}_{i,j,m,n} \cdot \mathbf{F}_{k+i-1,l+j-1,m}$$

Với chi phí tính toán:

$$D_K \cdot D_K \cdot M \cdot N \cdot D_F \cdot D_F$$

Depthwise separable convolution tách convolution tiêu chuẩn ra làm 2 bước. Đầu tiên, nó áp dụng một filter trên từng kênh input. Bước này gọi là depthwise convolution, được định nghĩa như sau:

$$\hat{\mathbf{G}}_{k,l,m} = \sum_{i,j} \hat{\mathbf{K}}_{i,j,m} \cdot \mathbf{F}_{k+i-1,l+j-1,m}$$

Trong đó $\hat{\mathbf{K}}$ là kernel depthwise convolution với kích thước $D_K \times D_K \times M$ trong đó filter thứ m_{th} trong $\hat{\mathbf{K}}$ được áp dụng cho kênh thứ m_{th} trong $\mathbf{F}$ để tạo ra kênh thứ m_{th} của feature map đầu ra $\hat{\mathbf{G}}$.

Chi phí tính toán cho depthwise convolution là:

$$D_K \cdot D_K \cdot M \cdot D_F \cdot D_F$$

Depthwise convolution chỉ mới filter được những kênh đầu vào và chưa kết hợp những feature đó lại với nhau để tạo ra những đặc trưng mới như trong convolution tiêu chuẩn. Do vậy, cần thêm một layer nữa để tính tổ hợp tuyến tính của kết quả depthwise convolution thông qua 1×1 convolution.

Sự kết hợp giữa depthwise convolution và 1×1 convolution được gọi là depthwise separable convolution. Chi phí tính toán là tổng chi phí tính toán của 2 thao tác:

$$D_K \cdot D_K \cdot M \cdot D_F \cdot D_F + M \cdot N \cdot D_F \cdot D_F$$

Lượng cải thiện chi phí tính toán đối với phép convolution tiêu chuẩn:

Layer/Modification	Million Mult-Adds	Million Parameters
Convolution	462	2.36
Depthwise Separable Conv	52.3	0.27
$\alpha = 0.75$	29.6	0.15
$\rho = 0.714$	15.1	0.15

Hình 2.2: *Tài nguyên sử dụng cho mỗi biến thể của convolution tiêu chuẩn. Mỗi hàng là thay đổi dựa trên hàng trước đó. Ví dụ này là cho một lớp trong MobileNet với* $D_K = 3$, $M = 512$, $N = 512$, $D_F = 14$*[2]*

Model	ImageNet Accuracy	Million Mult-Adds	Million Parameters
Conv MobileNet	71.7%	4866	29.3
MobileNet	70.6%	569	4.2

Hình 2.3: *So sánh Mobile Net depthwise separable covnlution với Mobile Net khi dùng full convolution[2]*

$$\begin{aligned} &\frac{D_K \cdot D_K \cdot M \cdot D_F \cdot D_F + M \cdot N \cdot D_F \cdot D_F}{D_K \cdot D_K \cdot M \cdot N \cdot D_F \cdot D_F} \\ =&\frac{1}{N} + \frac{1}{D_K^2} \end{aligned}$$

Việc làm giảm số lượng thông số so với convolution tiêu chuẩn là một điểm cần phải cân nhắc khi áp dụng depthwise separable convolution. Nếu số thông số trong mạng đã quá ít thì không nên sử dụng layer này thay thế cho convolution tiêu chuẩn. Tuy nhiên, nếu tận dụng hiệu quả, depthwise separable convolution sẽ tăng tốc độ tính toán và giảm kích thức của mạng đáng kể.

Một đặc điểm khác nữa của Mobile Net là nó cho phép ta tinh chỉnh độ dày của mỗi lớp depthwise separable convolution bằng việc giảm tải số lượng kênh đầu vào và kênh đầu ra qua thông số α (width multiplier). Điều này giúp giảm chi phí tính toán và số lượng thông số. Ta còn có thể làm cho mạng nhỏ gọn hơn bằng cách sử dụng thông số ρ (resolution multiplier) để làm giảm kích thước của hình ảnh đầu vào. Tổng chi phí tính toán với 2 thông số này như sau:

$$D_K \cdot D_K \cdot \alpha M \cdot \rho D_F \cdot \rho D_F + \alpha M \cdot \alpha N \cdot \rho D_F \cdot \rho D_F$$

Mobile Net đã cho thấy được hiêu quả của nó trong việc thu giảm kích

thước mô hình mà vẫn đem lại khả năng cạnh tranh với các đối thủ trong bài toán về nhận diện vật thể.

Framework Resolution	Model	mAP	Billion Mult-Adds	Million Parameters
SSD 300	deeplab-VGG	21.1%	34.9	33.1
	Inception V2	22.0%	3.8	13.7
	MobileNet	19.3%	1.2	6.8
Faster-RCNN 300	VGG	22.9%	64.3	138.5
	Inception V2	15.4%	118.2	13.3
	MobileNet	16.4%	25.2	6.1
Faster-RCNN 600	VGG	25.7%	149.6	138.5
	Inception V2	21.9%	129.6	13.3
	Mobilenet	19.8%	30.5	6.1

Hình 2.4: *Kết quả so sánh khi thực hiện phát hiện vật thể trên dataset COCO với framework và kiến trúc mạng khác nhau.[2]*

Hiệu quả của Mobile Net được thể hiện rõ nhất trong các thiết bị điện thoại. Đối với những thiết bị này, khả năng tính toán và bộ nhớ là có giới hạn, cho nên việc tối ưu hiệu quả việc nhận dạng cần phải tính đến lượng tài nguyên sử dụng của mô hình. Do đó, Mobile Net là một lựa chọn rất tốt cho các thiết bị này cũng như những thiết bị ít có khả năng tính toán cao khác.

Trong các bài toán nhận dạng vật thể, Mobile Net có thể hoạt động như một bộ trích xuất đặc trưng từ hình ảnh, qua đó thể hiện trong hình ảnh đó có những vật thể gì (xe cộ, chim chóc, chó mèo,...). Những đặc trưng đó có thể được đưa vào một bộ phát hiện vật thể như **SSD** (Single Shot Multibox Detector)[3] để xuất ra các hình chữ nhật (bounding box) biểu diễn cho vị trí từng object trong hình.

2.2.2 Face Identification

Sau khi bức hình khuôn mặt đã được phát hiện thành công, nó sẽ được đưa đến một bộ nhận dạng để tìm ra danh tính của người trong hình. Ngày nay, các hệ thống nhận dạng khuôn mặt tối tân nhất sử dụng Deep Learning mà cụ thể là mạng CNN do hiệu năng vượt trội của nó. Môt số kiến trúc nhận diện khuôn mặt tiêu biểu bao gồm DeepFace, DeepID hay FaceNet. Trong số các kiến trúc này, FaceNet đạt được độ chính xác cao nhất là 99.67% trên tập dữ liệu Labeled Face in the Wilds (LFW).[1]

Việc nhận dạng khuôn mặt đòi hỏi phải có những tập dataset rất lớn và

chất lượng để các kiến trúc CNN có thể học từ chúng một cách hiệu quả. Các tập dataset nổi tiếng như Labeled Face in the Wild, MS-Celeb-1M, VGGFace, IJB-A, IJB-B, IJB-C,... ra đời nhằm giúp cho việc training các model mới hiệu quả hơn cũng như giúp cho việc đánh giá. Một số giải pháp đã lựa chọn phương pháp sử dụng lượng hình ảnh rộng lớn từ trên các Search Engine hoặc mạng xã hội. Điển hình như Google có tới 500 triệu hình ảnh phục vụ cho việc training mô hình nhận diện khuôn mặt của họ; trong khi đó, Facebook sử dụng tới 100 triệu hình.

Việc thu thập hàng triệu hình ảnh, làm sạch và gán nhãn không hề đơn giản. Training và tuning một model trên những tập dữ liệu khổng lồ này sẽ cần rất nhiều thời gian. Những tập dữ liệu mã nguồn mở với hàng trăm nghìn hoặc hàng triệu ảnh đã ra đời, với chất lượng đảm bảo cho việc training các model. Dĩ nhiên, kèm theo đó là sự xuất hiện của các model đã đươc train sẵn (pretrained) trên các tập dữ liêu này.

(a) pose statistics (b) age statistics

(c) John Wesley Shipp (d) Leymah Gbowee

(e) Princess Haya Bint Al Hussein (f) Julio César Chávez Jr.

(g) Roy Jones Jr. (h) Ruby Lin

(i) Additi Gupta (j) Lee Joon-gi

Hình 2.5: *Một số hình ảnh từ tập dữ liệu VGGFace2*

VGGFace2 là một tập dữ liệu bao gồm 3.2 triệu ảnh của hơn 9000 người khác nhau. Những hình ảnh này có nguồn gốc từ Google Image Search, thông qua việc tìm kiếm dựa trên một name list những người nổi tiếng, sau đó thông qua các bước giảm nhiễu cả thủ công lẫn bằng máy để loại những hình ảnh không rõ ràng hoặc bị trùng. Tập dữ liệu này đảm bảo mỗi người có một số lượng hình ảnh đủ trong mức cho phép để không bị làm mất cân bằng dữ liệu, cũng như có thêm sự phong phú về độ tuổi và dân tộc (Ấn Độ, Trung Quốc,châu Âu...).[4]

Datasets	# of subjects	# of images	# of images per subject	manual identity labelling	pose	age	year
LFW [10]	5,749	13,233	1/2.3/530	-	-	-	2007
YTF [24]	1,595	3,425 videos	-	-	-	-	2011
CelebFaces+ [21]	10,177	202,599	19.9	-	-	-	2014
CASIA-WebFace [26]	10,575	494,414	2/46.8/804	-	-	-	2014
IJB-A [13]	500	5,712 images, 2,085 videos	11.4	-	-	-	2015
IJB-B [23]	1,845	11,754 images, 7,011 videos	36.2	-	-	-	2017
IJB-C [14]	3,531	31,334 images, 11,779 videos	36.3	-	-	-	2018
VGGFace [17]	2,622	2.6 M	1,000/1,000/1,000	-	-	Yes	2015
MegaFace [12]	690,572	4.7 M	3/7/2469	-	-	-	2016
MS-Celeb-1M [7]	100,000	10 M	100	-	-	-	2016
UMDFaces [5]	8,501	367,920	43.3	Yes	Yes	Yes	2016
UMDFaces-Videos [4]	3,107	22,075 videos	-	-	-	-	2017
VGGFace2 (**this paper**)	9,131	3.31 M	80/362.6/843	Yes	Yes	Yes	2018

Hình 2.6: *So sánh một số dataset phổ biến trong việc nhận diện khuôn mặt với VGGFace2*

Trong giai đoạn hiện tại, hệ thống đang sử dụng một mô hình **ResNet** được train sẵn trên tập dữ liệu VGGFace2. Mô hình ResNet sử dụng các khối residual network để vừa xây dựng những kiến trúc học sâu với khả năng mô hình hóa cao, vừa chống lại được hiện tượng vanishing gradient thường thấy trong các kiến trúc này. ResNet đạt giải nhất cuộc thi ILSVRC 2015 trong lĩnh vực phân loại hình ảnh. Thay vì sử dụng ResNet cho việc phân loại, lớp phân loại ở cuối kiến trúc này sẽ được gỡ ra, để lộ lớp CNN phía trước. Lớp CNN này sẽ output ra một ma trận được làm phẳng thành vector 2048 chiều, biểu diễn cho những đặc trưng mà kiến trúc ResNet đã học được từ hình ảnh đầu vào. Mỗi khuôn mặt khác nhau sẽ có những đặc trưng riêng giúp phân biệt chúng với những khuôn mặt khác. Như vậy, ta sử dụng ResNet để xuất ra một vector biểu diễn khuôn mặt (face embedding) đặc trưng nhất cho khuôn mặt đó. [4][5]

2.3 Vấn đề bảo mật trong xác thực sinh trắc học

Các hệ thống xác thực sinh trắc học tận dụng các đặc tính sinh lý (vân tay, khuôn mặt và con ngươi) hoặc hành vi (nhịp độ di chuyển hoặc gõ phím) để nhận dạng hoặc xác thực một đối tượng. Những hệ thống này được sử dụng rộng rãi trong đời sống, bao gồm xác thực trên điện thoại và quản lý truy cập. Do đó, các kiểu tấn công đánh lừa sinh trắc học (còn được gọi là Presentation Attack) đã trở thành một mối lo ngại lớn, trong đó một mẫu sinh trắc học giả mạo sẽ được thể hiện trước hệ thống hòng cố gắng xác thực.

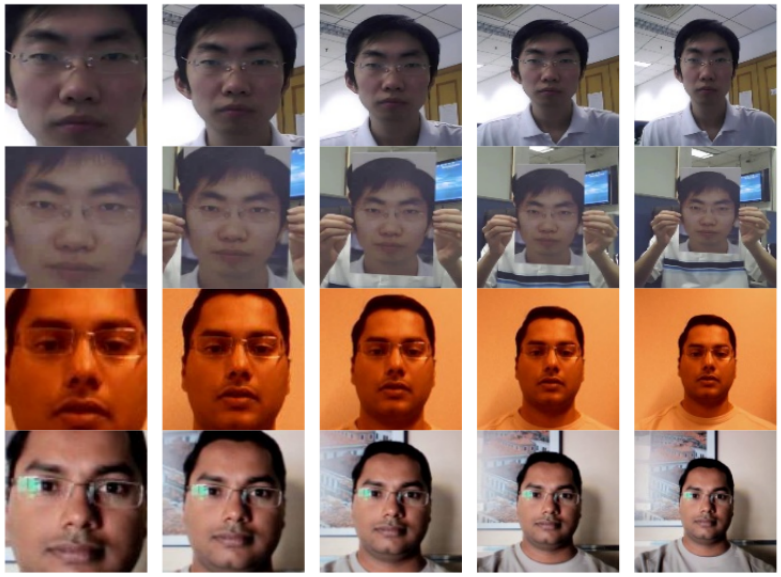
Bởi vì khuôn mặt là phương thức xác thực sinh trắc học dễ sử dụng nhất, nhiều kiểu tấn công sinh trắc học đã ra đời và phát triển, như print attack (tấn công bằng ảnh 2D được in ra), replay attack (tấn công bằng video), 3D masks (sử dụng mặt nạ 3D),...

Như vậy, một hệ thống nhận diên khuôn mặt có thể được sử dụng hiệu quả trong thực tế không những cần đáp ứng các yêu cầu về tốc độ, độ chính xác mà còn phải có khả năng chống loại các hình thức mạo danh khác nhau.

Những phương pháp chống mạo danh có thể được chia làm 3 loại lớn[7]:

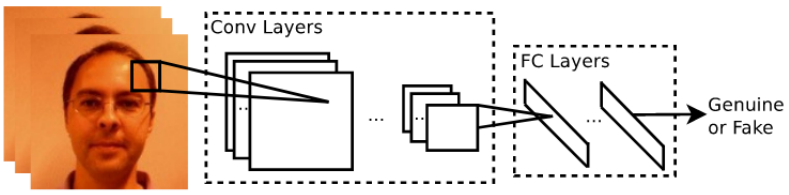
1. Dựa trên đặc điểm texture của các loại tấn công khác nhau. Do không có sự tương quan rõ ràng giữa cường độ của pixel và mỗi loại tấn công khác nhau, việc chắt lọc thuộc tính rất khó.
2. Yêu cầu người xác thực thực hiện một số hành vi phụ trợ, ví dụ như xoay sang trái hoặc phải hoặc cử động môi. Phương pháp này không chống lại được replay attack.
3. Dựa trên chất lượng hình ảnh và sự phản chiếu ánh sáng trên bức hình, đánh giá các thuộc tính liên quan đến đến độ sáng và độ nhiễu của hình.

Với sự phát triển của Deep Learning và sự hiệu quả của mạng CNN trong các bài toán về hình ảnh, ta hoàn toàn có thể sử xây dựng một neural network để học các đặc trưng của ảnh giả để từ đó bảo vệ hệ thống trước các kiểu tấn công mạo danh. Mạng neural network này sẽ bao gồm đầu vào là một bức hình có khuôn mặt của người dùng, những lớp mạng CNN để trích xuất đặc trưng từ bức hình đó và cuối cùng là những lớp mạng Fully Connected để quyết định ảnh đó là giả hay thật.



Hình 2.7: *Ảnh mẫu với scale khác nhau. Hai hàng đầu là từ tập dữ liệu CASIA và 2 hàng dưới là từ tập dữ liệu REPLAY ATTACK. Hàng lẻ là hình thật và hàng chẵn là hình giả.[10]*

Một số giải pháp thay vì sử dụng những đặc trưng tự định nghĩa bởi con người đã sử dụng mạng Convolutional Neural Network để học các đặc trưng từ một tập dữ liệu phân biệt ảnh thật, giả. Mô hình [10] sử dụng một mạng CNN đã được train và test trên các tập CASIA và REPLAY-ATTACK, đạt kết quả khá cao so với các giải pháp không sử dụng mạng CNN. Những kết quả thử nghiệm giữa 2 tập dữ liệu cho thấy CNN có thể học được những đặc tính với khả năng tổng quát hóa tốt hơn. Hơn nữa, khi mạng được train trên cả 2 tập dữ liệu thì bias thấp hơn cho từng tập dữ liệu. Tuy nhiên, khi mạng train chỉ trên một tập dữ liệu, nó có xu hướng overfit tập dữ liệu đó và hoạt động ít hiệu quả hơn trên tập dữ liệu khác.



Hình 2.8: *Kiến trúc của mô hình[10]*

3 Đề xuất giải pháp và thiết kế hệ thống

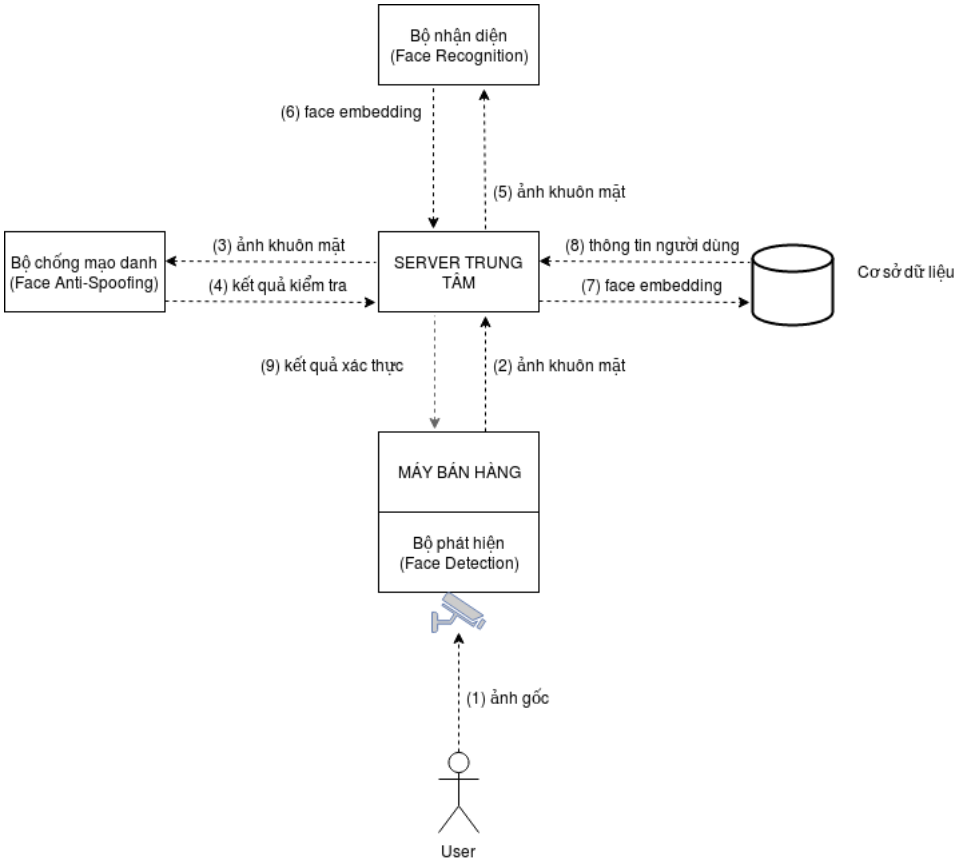
3.1 Mô tả giải pháp

Mô hình kiến trúc cơ bản của hệ thống xác thực bằng khuôn mặt cho máy bán hàng tự động sẽ bao gồm những phần sau: kiến trúc phát hiện khuôn mặt từ camera, cơ sở dữ liêu biểu diễn mỗi tài khoản và những khuôn mặt tương ứng, kiến trúc chống lại các kiểu tấn công giả mạo và kiến trúc định ra nhân dạng của khuôn mặt trong cơ sở dữ liệu. Tất cả những kiến trúc liên quan đến nhận diện khuôn mặt đều được sử dụng mạng CNN để đạt đến tối đa hiệu quả.

Người dùng sẽ đứng trước camera ở một khoảng cách nhất định, và kiến trúc phát hiện khuôn mặt sẽ phát hiện ngay khuôn mặt của người dùng trong thời gian thực, sau đó gửi về kiến trúc chống lại các kiểu tấn công giả mạo. Một khi đã xác định ảnh khuôn mặt đầu vào là ảnh thật, nó sẽ được gửi đến kiến trúc định ra nhân dạng của khuôn mặt trong cơ sở dữ liệu. Đối với giải pháp này, những metrics cơ bản cần phải đánh giá bao gồm:

1. Tốc độ phát hiện khuôn mặt: Thời gian một khuôn mặt được phát hiện kể từ khi người dùng đứng ở khoảng cách đủ gần
2. Độ chính xác của việc xác định hình giả
3. Recall và False Positive Rate của hệ thống nhận diện khuôn mặt
4. Latency: Thời gian từ khi khuôn mặt được phát hiện cho đến khi danh tính của khuôn mặt được xác định và máy bán hàng xác thực người dùng

Các kiến trúc được mô tả có thể được đặt tại các server riêng rẽ. Ví dụ, một server có thể chứa cơ sở dữ liệu của khuôn mặt, một server có thể chứa bộ



Hình 3.1: *Sơ đồ tổng quát của hệ thống, cho thấy luồng di chuyển dữ liệu cho phần xác thực và các các thực thể tham gia vào hệ thống*

nhận dạng và chống mạo danh,... vâng vâng. Việc tách các kiến trúc như vậy có thể giúp cho việc mở rộng khả năng tính toán tốt hơn, giảm tải và tránh những vấn đề như single point of value.

3.2 Mô hình kiến trúc hệ thống đề xuất

Dữ liệu hình ảnh khuôn mặt của người dùng sẽ được luân chuyển đi qua nhiều công đoạn khác nhau với mục đích cuối cùng là tìm ra nhân dạng của người trong hình một cách nhanh và chính xác nhất có thể. Những nhân tố xử lý chủ yếu trong các công đoạn bao gồm: máy bán hàng, bộ phát hiện khuôn mặt, bộ tiền xử lý hình ảnh, bộ chống mạo danh, bộ nhận diện và cơ sở dữ liệu. Quy trình di chuyển của dữ liệu như sau:

1. Ở máy bán hàng, một camera trắng đen hoạt động trong thời gian thực sẽ phát hiện khuôn mặt của người dùng và gửi khuôn mặt này về bộ phận xác

thực khi người đó đã đứng ở một khoảng cách nhất định - khoảng cách này đủ để người đó truy cập vào bàn phím của máy bán hàng nhưng không quá gần để làm khuất mất những phần quan trọng của khuôn mặt.

2. Sau khi khuôn mặt đã được phát hiện, hệ thống phát hiện sẽ khoanh vùng khuôn mặt đó bằng một hình chữ nhật (bounding box). Hệ thống sẽ gửi tọa độ góc trên bên trái, độ rộng, độ dài của hình chữ nhật cho một server trung tâm thông qua API của server đó.

3. Ở server trên, hình ảnh sẽ thông qua các bước biến đổi phù hợp, chuẩn hóa các pixel và chuyển sang dạng array có chiều phù phù hợp với các bộ nhận dạng và bộ chống mạo danh. Sau đó, hình ảnh đã được xử lý sẽ được gửi đến một server chứa bộ chống mạo danh thông qua API của server này.

4. Bộ chống mạo danh sẽ nhận vào hình trên và output kết quả 0 (không mạo danh) và 1 (có mạo danh) đối với hình này. Kết quả này được trả về cho server trung tâm.

5. Server trung tâm sẽ trả về cho máy bán hàng thông báo xác thực không thành công nếu như nó nhận được giá trị 1, ngược lại nó sẽ gửi hình khuôn mặt đã xử lý đến cho server chứa bộ nhận diện.

6. Bộ nhận diện sẽ trả về face embedding là một vector đặc trưng tương ứng với khuôn mặt đó cho server trung tâm.

7. Server trung tâm sẽ tìm trong cơ sở dữ liệu tài khoản nào có face embedding của khuôn mặt với khoảng cách gần nhất với vector đã nhận. Khoảng cách này có thể là khoảng cách cosine, euclid hoặc bất kỳ loại khoảng cách nào phục vụ tốt cho nhu cầu bài toán.

8. Nếu khoảng cách gần nhất nhỏ hơn một ngưỡng cho trước, ta coi như hai khuôn mặt là giống nhau và cho phép người dùng xác thực vào hệ thống với tài khoản tương ứng. Nếu không, không cho phép người dùng xác thực.

Đối với những người dùng mới, họ có thể đăng ký bằng cách nhập các thông tin cá nhân như tên tài khoản, mật khẩu, liên kết tài khoản ngân hàng tương ứng và chụp khuôn mặt của mình. Khuôn mặt này sẽ được đưa đến server trung tâm, được tiền xử lý và gửi đến bộ nhận dạng để ra được face embedding tương ứng, sau đó sẽ được lưu trong cơ sở dữ liệu cùng với thông tin cá nhân của người dùng mới. Hệ thống sẽ đảm bảo sao cho chỉ cần một ảnh chụp của người dùng là đủ để phân biệt với những người khác nhằm giảm số lượng ảnh phải lưu trên cơ sở dữ liệu.

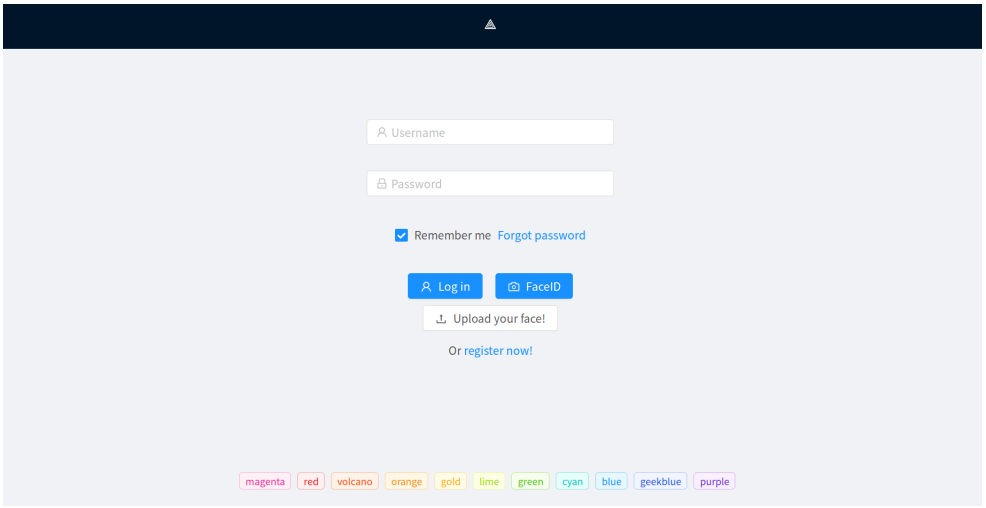
4 Hiện thực và đánh giá

Ở giai đoạn hiện tại, hệ thống sẽ hiện thực trên nền tảng web để đơn giản hóa việc đánh giá các mô hình, do chi phí hiện thực trên nền tảng web thấp hơn nhiều so với các nền tảng khác như di động hoặc các hệ thống nhúng.

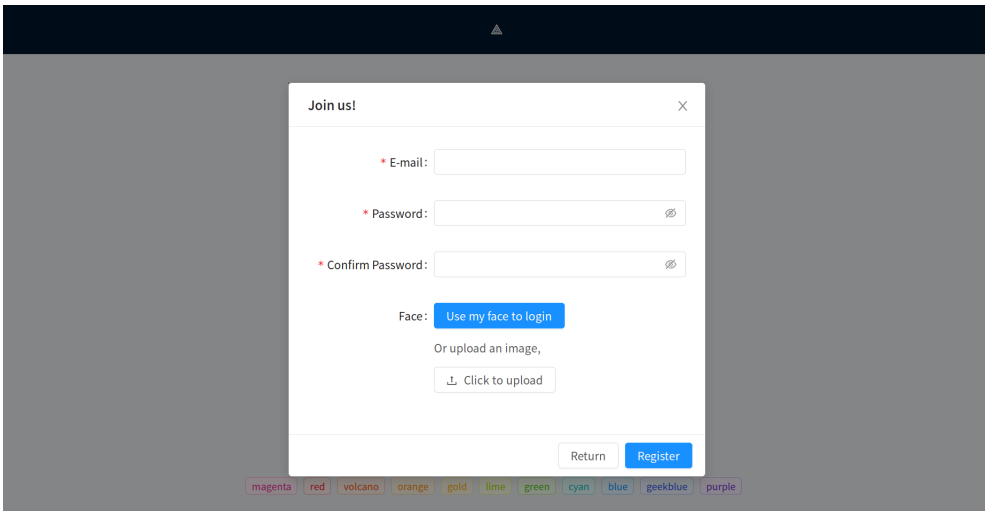
4.1 Ứng dụng mô phỏng

4.1.1 Giao diện web

Một giao diện web sẽ được hiển thị cho người dùng, hiện thực bằng ReactJS. Giao diện cho phép người dùng đăng ký tài khoản sử dụng một email hợp lệ bất kỳ, đồng thời phải chục ảnh có khuôn mặt của mình. Giao diện còn hiển thị một khung đăng nhập, cho phép người dùng đăng nhập bằng tên đăng nhập & mật khẩu hoặc bằng khuôn mặt.



Hình 4.1: *Màn hình login. Ấn vào "FaceID"sẽ hiển thị camera để phát hiện khuôn mặt*



Hình 4.2: *Màn hình đăng ký. Người dùng phải bật webcam để phát hiện khuôn mặt hoặc tải ảnh có khuôn mặt của mình*

4.1.2 Face Detection

Khi người sử dụng sử chụp các chức năng chụp hình khuôn mặt của ứng dụng, trình duyệt sẽ yêu cầu truy cập webcam của máy tính và một mô hình phát hiện khuôn mặt là SSD Mobilenet V1 - Mobile Net được pretrain trên một tập dữ liệu khuôn mặt gọi là WIDER FACE [6] và sau đó kết hợp với SSD, sẽ được sử dụng ngay trong trình duyệt để phát hiện và trích xuất khuôn mặt ra từ webcam trong thời gian thực. Khuôn mặt này sau đó sẽ được gửi đến một server trung tâm.

4.1.3 Server trung tâm

Server trung tâm được hiện thực bằng Flask, một framework ứng dụng Web trên ngôn ngữ Python. Server này chạy trên port 5000 của localhost. Flask server cung cấp các đầu API sau:

- *POST /signup*: đăng ký tài khoản. Người dùng (thông qua trình duyệt) sẽ gửi một form data bao gồm email, mật khẩu và ảnh khuôn mặt đã được phát hiện ở client-side. Nếu thông tin hợp lệ, Flask server sẽ tiền xử lý và gửi ảnh đến bộ nhận dạng để nhận về một face embedding, sau đó lưu tên tài khoản, mật khẩu cùng face embedding này vào trong cơ sỡ dữ liệu và báo về cho người dùng một HTTP response với status 200.
- *POST /faceid*: đăng nhập bằng khuôn mặt. Người dùng gửi ảnh được phát hiện ở client-side lên cho Flask server. Sau đó, Flask server sẽ tiền xử lý và gửi ảnh này đến bộ nhận dạng để lấy face embedding. Cuối cùng, face embedding này sẽ được so sánh với tất cả những face embedding khác trong cơ sở dữ liệu thông qua khoảng cách cosine. Tài khoản nào có face embedding gần nhất với hình đã được tải lên và nếu khoảng cách này nhỏ hơn 0.5 thì cho phép client xác thực với tài khoản đó. Ngược lại, nếu khoảng cách lớn hơn 0.5 thì không cho phép xác thực.

4.1.4 Bộ nhận dạng ảnh

Sử dụng công cụ Tensorflow Serve, chúng ta có thể deploy một mô hình học máy được hiện thực trên thư viện Tensorflow. Mô hình được sử dụng là ResNet đươc pretrain sẵn trên tập dataset VGGFace2. Nhiệm vụ chính của mô hình này là trả về một face embedding có thể phân biệt tốt với các khuôn mặt khác. Tensorflow Serve là một server chạy trong docker và xử lý request trên port 8501 của localhost.

Tensorflow Serve giao tiếp với Flask server bằng cách expose một API giúp trả về face embedding từ một ảnh khuôn mặt.

4.1.5 Cơ sở dữ liệu

Hệ thống sử dụng SQLite để lưu trữ các tài khoản cũng như face embedding tương ứng. Cơ sở dữ liệu gồm một table (*users*) với các cột sau đây:

- email (kiểu TEXT): chứa email của người dùng và là khóa chính

- password (kiểu BLOB): chứa mật khẩu của người dùng
- embedding (kiểu TEXT): là một chuỗi bao gồm các 2048 giá trị thập phân ngăn cách nhau bằng dấu cách. Đây là face embedding của khuôn mặt người dùng.

4.2 Đánh giá

Mô hình ResNet sau khi được train trên VGGFace2, đã được đánh giá trên tập dữ liệu IJB-A gồm có 5712 hình và 2085 videos từ 500 người, với trung bình 11.4 hình và 4.2 video mỗi người. Mọi hình và video được lấy từ những môi trường không có ràng buộc và thể hiện những thay đổi lớn trong biểu cảm và chất lượng hình ảnh. Kết quả đánh giá cho thấy VGGFace2 vượt trội hơn so với các dataset khác trong task verification (so sánh 1:1 giữa 2 hình) và identification (so sánh 1:N, lấy 1 hình ra và so sánh với N hình khác trong cơ sở dữ liệu). Đối với verification, hiệu năng được đánh giá sử dụng True Accept Rate (TAR) và false positive rates (FAR). Đối với identification, hiệu năng được đánh giá bằng true positive identificaiton rate (TPIR) và false positive identification rate (FPIR).

Training dataset	Arch.	1:1 Verification TAR			1:N Identification TPIR				
		FAR=0.001	FAR=0.01	FAR=0.1	FPIR=0.01	FPIR=0.1	Rank-1	Rank-5	Rank-10
VGGFace [17]	ResNet-50	0.620 ± 0.043	0.834 ± 0.021	0.954 ± 0.005	0.454 ± 0.058	0.748 ± 0.024	0.925 ± 0.008	0.972 ± 0.005	0.983 ± 0.003
MS1M [7]	ResNet-50	0.851 ± 0.030	0.939 ± 0.013	0.980 ± 0.003	0.807 ± 0.041	0.920 ± 0.012	0.961 ± 0.006	0.982 ± 0.004	0.990 ± 0.002
VGGFace2	ResNet-50	0.895 ± 0.019	0.950 ± 0.005	0.980 ± 0.003	0.844 ± 0.035	0.924 ± 0.006	0.976 ± 0.004	0.992 ± 0.002	0.995 ± 0.001
VGGFace2_ft	ResNet-50	0.908 ± 0.017	0.957 ± 0.007	0.986 ± 0.002	0.861 ± 0.027	0.936 ± 0.007	0.978 ± 0.005	0.992 ± 0.003	0.995 ± 0.001
VGGFace2	SENet	0.904 ± 0.020	0.958 ± 0.004	0.985 ± 0.002	0.847 ± 0.051	0.930 ± 0.007	0.981 ± 0.003	$\mathbf{0.994 \pm 0.002}$	$\mathbf{0.996 \pm 0.001}$
VGGFace2_ft	SENet	$\mathbf{0.921 \pm 0.014}$	$\mathbf{0.968 \pm 0.006}$	$\mathbf{0.990 \pm 0.002}$	$\mathbf{0.883 \pm 0.038}$	$\mathbf{0.946 \pm 0.004}$	$\mathbf{0.982 \pm 0.004}$	0.993 ± 0.002	0.994 ± 0.001
Crosswhite *et al.* [6]	-	0.836 ± 0.027	0.939 ± 0.013	0.979 ± 0.004	0.774 ± 0.049	0.882 ± 0.016	0.928 ± 0.010	0.977 ± 0.004	0.986 ± 0.003
Sohn *et al.* [20]	-	0.649 ± 0.022	0.864 ± 0.007	0.970 ± 0.001	-	-	0.895 ± 0.003	0.957 ± 0.002	0.968 ± 0.002
Bansalit et al. [4]	-	$0.730^{\dagger}$	0.874	$0.960^{\dagger}$	-	-	-	-	-
Yang *et al.* [25]	-	0.881 ± 0.011	0.941 ± 0.008	0.978 ± 0.003	0.817 ± 0.041	0.917 ± 0.009	0.958 ± 0.005	0.980 ± 0.005	0.986 ± 0.003

Hình 4.3: *Hiệu năng trên tập dữ liệu IJB-A. Giá trị lớn hơn là tốt hơn.*

Hiện tại, đối với mô hình nhận diện ảnh giả mạo, chưa có kết quả đánh giá do chưa hiện thực được mô hình cụ thể và hạn chế về thời gian.

4.3 Hạn chế

Nhìn chung, hệ thống xác thực cơ bản đã hoàn thành phần nhiều. Tuy nhiên, một số hạn chế trong giai đoạn đề cương này bao gồm: chưa xây dựng được mô hình chống ảnh giả mạo, chưa xây dựng được một tập dữ liệu thực tế đối với máy bán hàng và đánh giá hệ thống trên tập dữ liệu này.

5 Tổng kết

5.1 Những đóng góp chính của đề cương

Đề tài nhóm nghiên cứu tập trung giải quyết trong Deep Learning và Computer Vision hiện nay là xác thực bằng nhận diện khuôn mặt. Hai vấn đề lớn của đề tài này này là làm sao để đạt được độ chính xác cao nhất có thể, và làm sao để chống lại các kiểu tấn công mạo danh.

Hiện tại, nhóm đã phát triển được một ứng dụng tương đối hoàn thiện để kiểm tra khả năng xác thực của mô hình, đồng thời đi đến kết luận cho việc lựa chọn những mô hình nào là hợp lý cho từng công đoạn của quá trình nhận dạng khuôn mặt.

5.2 Kế hoạch sắp tới

Trong giai đoạn luận văn, nhóm sẽ tập trung chủ yếu vào việc tăng độ chính xác của mô hình nhận diện khuôn mặt hiện tại, đồng thời hiện thực và cải tiến giải pháp chống lại các kiểu tấn công mạo danh.

Việc tăng độ chính xác của mô hình nhận diện khuôn mặt có thể đạt được bằng việc sử dụng các mô hình tốt hơn, như SEResNet hoặc một biến thể khác phù hợp hơn với bài toán hiện tại. Đồng thời, nhóm cũng phải tìm cách đánh giá mô hình dựa trên một tập dữ liệu khuôn mặt được chụp từ camera thật.

Đối với việc chống lại các kiểu tấn công mạo danh, nhóm sẽ hiện thực một mạng CNN có thể chống lại overfit, tốc độ cao khi áp dụng trên thực tế và đồng thời với độ chính xác cao hơn trong [10]. Nhóm sẽ thử nghiệm việc trích xuất một

số đặc trưng nhằm phụ trợ cho mạng CNN này hoạt động hiệu quả hơn, ví dụ như đường viền, độ sâu của hình, các tín hiệu sinh học phát ra từ khung hình...

Tài liệu tham khảo

[1] Fang L., Fu M., Sun S., Ran Q. (2019) *Overview of Face Recognition Methods*. In: Sun S., Fu M., Xu L. (eds) Signal and Information Processing, Networking and Computers. ICSINC 2018. Lecture Notes in Electrical Engineering, vol 550. Springer, Singapore

[2] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, Hartwig Adam, *MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications*

[3] Wei Liu1, Dragomir Anguelov2, Dumitru Erhan3, Christian Szegedy3,Scott Reed4, Cheng-Yang Fu1, Alexander C. Berg1, *SSD: Single Shot MultiBox Detector*

[4] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, Andrew Zisserman, *VGGFace2: A dataset for recognising faces across pose and age*

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, *Deep Residual Learning for Image Recognition*

[6] Yang, Shuo and Luo, Ping and Loy, Chen Change and Tang, Xiaoou, *WIDER FACE: A Face Detection Benchmark*

[7] Xiaoguang Tu, Jian Zhao, Mei Xie, Guodong Du, Hengsheng Zhang, Jianshu Li, Zheng Ma, Jiashi Feng, *Face Anti-Spoofing Using Patch and Depth-Based CNNs*

[8] Zhiwei Zhang, Junjie Yan, Sifei Liu, Zhen Lei, Dong Yi, Stan Z. Li, *https://ieeexplore.ieee.org/abstract/document/6199754*

[9] Yaojie Liu, Amin Jourabloo, Xiaoming Liu, *Learning Deep Models for Face Anti-Spoofing: Binary or Auxiliary Supervision*

[10] Jianwei Yang, Zhen Lei and Stan Z. Li, *Learn Convolutional Neural Network for Face Anti-Spoofing*