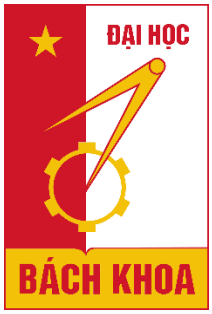
ĐẠI HỌC BÁCH KHOA HÀ NỘI



BÁO CÁO DEEP LEARNING

Transfer Learning sử dụng kiến trúc SqueezeNet để tối ưu hóa cho hệ thống nhận diện khuôn mặt

NGUYỄN ĐÌNH KHÁNH
khanh.nd241787e@sis.hust.edu.vn

Khoa Tự động hóa
Trường Điện- Điện tử

Giảng viên hướng dẫn: PGS. TS. Nguyễn Hoài Nam

Chữ ký của GVHD

HÀ NỘI, 1/2026

MỤC LỤC

CHƯƠNG 1. TỔNG QUAN

1.1 Giới thiệu chung

Trong kỷ nguyên Công nghiệp 4.0, nhận diện khuôn mặt không còn là một công nghệ xa lạ mà đã trở thành một phần thiết yếu trong đời sống hàng ngày, từ việc mở khóa điện thoại thông minh đến các hệ thống giám sát an ninh phức tạp. Tuy nhiên, thách thức lớn nhất hiện nay là làm sao để triển khai các mô hình Deep Learning mạnh mẽ lên các thiết bị đầu cuối (Edge Devices) có tài nguyên hạn chế về bộ nhớ và năng lượng.

Việc xây dựng một mô hình mạng nơ-ron tích chập (CNN) từ đầu đòi hỏi hàng triệu hình ảnh gán nhãn và hàng tuần huấn luyện trên các hệ thống GPU đắt tiền. Do đó, phương pháp Học chuyển đổi (Transfer Learning) kết hợp với các kiến trúc mạng siêu nhẹ như SqueezeNet nổi lên như một giải pháp tối ưu, cho phép đạt được độ chính xác cao với chi phí tính toán thấp nhất.

1.2 Mục tiêu và động lực nghiên cứu

Yêu cầu thiết kế cho hệ thống nhận diện khuôn mặt là xử lý trên tập dự liệu cực kỳ hạn chế, với 28 class và 20 image đã được gán nhãn cho mỗi class. Điều này trở nên rất khó khăn cho các bài toán deep learning nói chung và bài toán classification nói riêng. Việc bộ dữ liệu nhỏ yêu cầu xây dựng mạng phù hợp để tránh overfitting và đảm bảo được độ chính xác cho mạng. Từ đó đặt ra các bài toán nhỏ cần xử lý như làm giàu dữ liệu và các phương pháp tiền xử lý cho tập dữ liệu nhận dạng này.

Báo cáo này trình bày dự án nhận diện khuôn mặt sử dụng kỹ thuật transfer learning trên mạng SqueezeNet – một kiến trúc mạng nơ-ron tích chập nhẹ, có độ chính xác tương đương AlexNet nhưng với số lượng tham số ít hơn 50 lần. Dự án bao gồm các bước tiền xử lý dữ liệu (phát hiện khuôn mặt bằng thuật toán Viola-Jones, cắt ảnh, tăng cường dữ liệu), fine-tuning mạng SqueezeNet và đánh giá kết quả. Kết quả huấn luyện đạt độ chính xác cao nhất 81.82% trên tập kiểm tra. Báo cáo cũng phân tích các thí nghiệm thay đổi tham số và ma trận nhầm lẫn để đánh giá hiệu suất mô hình.

CHƯƠNG 2. CƠ SỞ LÝ THUYẾT

Phần này trình bày chi tiết các nền tảng lý thuyết liên quan đến dự án, bao gồm nhận diện khuôn mặt trong học sâu, kỹ thuật transfer learning, kiến trúc mạng SqueezeNet, và thuật toán Viola-Jones dùng cho phát hiện khuôn mặt.

2.1 Tổng quan về Deep Learning và CNN

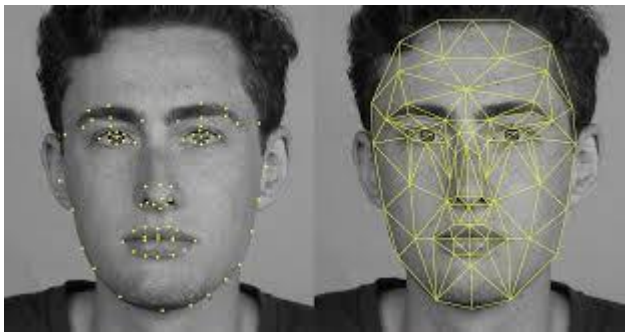
Mạng nơ-ron tích chập (CNN) là xương sống của các hệ thống thị giác máy tính hiện đại. CNN tự động học các đặc trưng từ mức độ thấp (cạnh, góc) đến mức độ cao (bộ phận khuôn mặt) thông qua các lớp tích chập, giúp loại bỏ việc trích xuất đặc trưng thủ công vốn tốn rất nhiều thời gian và công sức và kiến thức chuyên môn về đối tượng như machine learning.

2.2 Nhận diện khuôn mặt trong Deep Learning

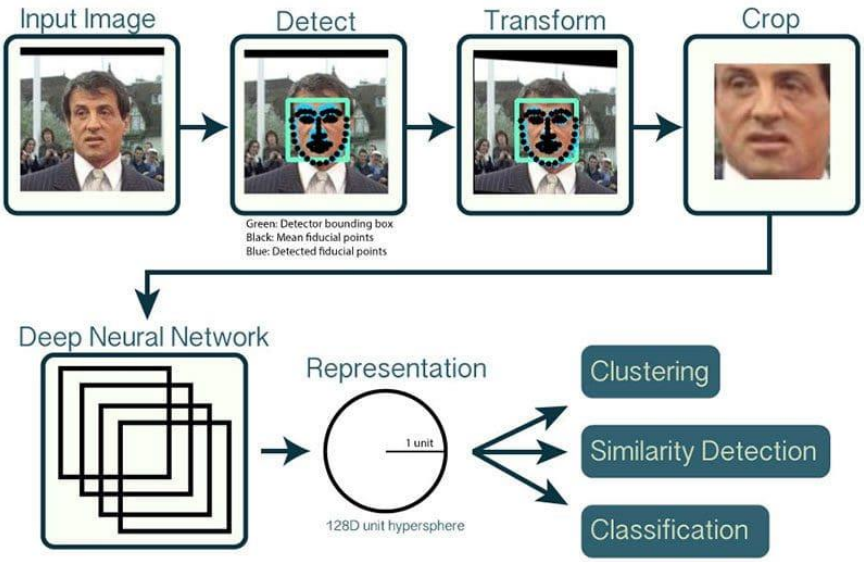
Nhận diện khuôn mặt (face recognition) là nhiệm vụ phân loại hoặc xác định danh tính cá nhân dựa trên đặc trưng khuôn mặt từ ảnh hoặc video. Trong học sâu, nhiệm vụ này thường được chia thành các giai đoạn chính:

- **Phát hiện khuôn mặt (Face Detection):** Xác định vị trí và bounding box của khuôn mặt trong ảnh.
- **Căn chỉnh khuôn mặt (Face Alignment):** Chuẩn hóa vị trí mắt, mũi, miệng để giảm biến đổi pose.
- **Trích xuất đặc trưng (Feature Extraction)**: Chuyển đổi khuôn mặt thành vector embedding (đặc trưng vector) trong không gian cao chiều.
- **Phân loại hoặc so sánh (Classification/Verification):** Phân loại đa lớp (multi-class) hoặc tính khoảng cách (e.g., Euclidean, Cosine) để xác định danh tính.

Quy trình điển hình của một hệ thống nhận diện khuôn mặt dựa trên deep learning được minh họa như sau:



Source: [medium.com]



Source: pyimagesearch.com

Các mô hình hiện đại như FaceNet (Google, 2015) sử dụng triplet loss để học embedding sao cho khoảng cách giữa các ảnh cùng người nhỏ hơn khác người. VGGFace hoặc ArcFace đạt độ chính xác >99% trên benchmark LFW (Labeled Faces in the Wild). Trong dự án này, chúng ta tập trung vào cách tiếp cận phân loại đa lớp sử dụng CNN với transfer learning.

2.3 Kiến trúc mạng SqueezeNet

SqueezeNet (Iandola et al., 2016) là kiến trúc CNN nhẹ được thiết kế để đạt độ chính xác tương đương AlexNet trên ImageNet nhưng với số lượng tham số ít hơn 50 lần (chỉ ~1.25 triệu tham số) và kích thước mô hình <0.5MB.

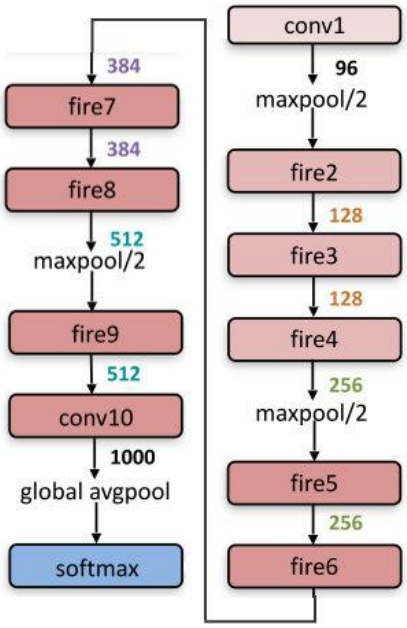
Các chiến lược chính để nén mô hình:

- Sử dụng chủ yếu convolution 1x1 để giảm chiều kênh.
- Giảm số lượng kênh input cho convolution 3x3.
- Đẩy các lớp downsampling về cuối mạng.

Đơn vị cốt lõi là **Fire Module**, bao gồm:

- **Squeeze layer**: Convolution 1x1 để nén số kênh (s1x1).
- **Expand layer**: Hỗn hợp convolution 1x1 (e1x1) và 3x3 (e3x3), sau đó concatenate.

Kiến trúc tổng thể của SqueezeNet và Fire Module:



SqueezeNet phù hợp cho các ứng dụng trên thiết bị nhúng (mobile, edge devices) nhờ kích thước nhỏ và tốc độ suy luận nhanh.

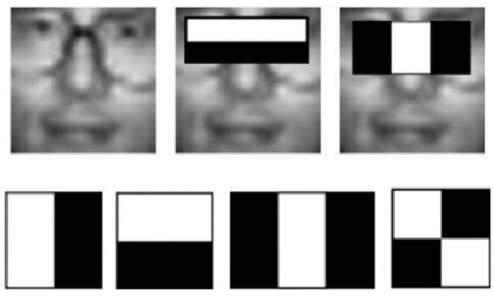
2.4 Thuật toán phát hiện khuôn mặt Viola-Jones

Trong dự án, phát hiện khuôn mặt được thực hiện bằng thuật toán Viola-Jones (Viola & Jones, 2001) – một phương pháp cổ điển nhưng hiệu quả và nhanh chóng dựa trên machine learning truyền thống.

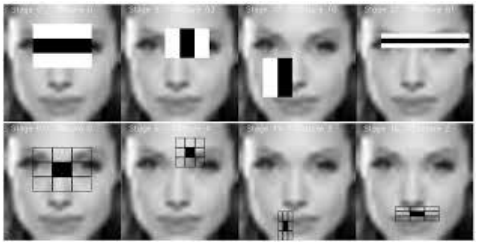
Các thành phần chính:

- **Haar-like features**: Các đặc trưng đơn giản tính toán sự chênh lệch cường độ giữa các vùng trắng-đen (edge, line, diagonal).
- **Integral Image**: Cho phép tính toán nhanh tổng pixel trong vùng chữ nhật bất kỳ.
- **AdaBoost**: Chọn lọc các đặc trưng tốt nhất và kết hợp thành strong classifier.
- **Cascade Classifiers**: Chuỗi các classifier đơn giản, loại bỏ nhanh vùng không phải mặt để tăng tốc độ.

Ví dụ các Haar features và quy trình cascade:



Source: researchgate.net

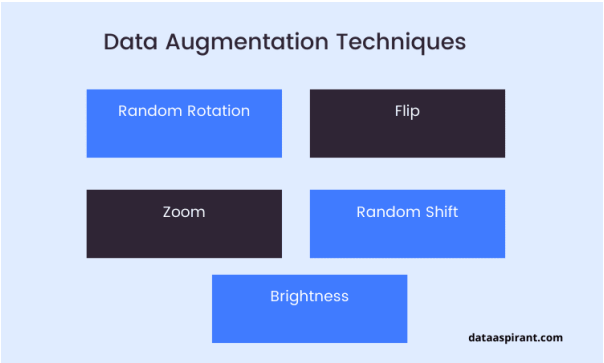


Source: medium.com

Mặc dù các phương pháp deep learning hiện đại (MTCNN, RetinaFace) chính xác hơn, Viola-Jones vẫn được sử dụng rộng rãi nhờ tốc độ thực thời và không cần GPU.

Để tăng độ đa dạng và giảm overfitting, đặc biệt với tập dữ liệu khuôn mặt nhỏ, chúng ta áp dụng data augmentation: xoay, scale, lật ngang, thay đổi độ sáng, thêm nhiễu Gaussian.

Các kỹ thuật phổ biến:



Source: dataaspirant.com

Phần cơ sở lý thuyết này cung cấp nền tảng vững chắc cho các lựa chọn phương pháp trong dự án. Bạn có thể thay thế phần này trực tiếp vào báo cáo để làm phần 2 chi tiết và chuyên nghiệp hơn (khoảng 5-7 trang khi định dạng với hình ảnh). Nếu cần mở rộng thêm phần khác hoặc chỉnh sửa, hãy cho tôi biết!

2.5 Phương pháp Transfer Learning

Transfer learning là kỹ thuật tái sử dụng kiến thức từ mô hình đã huấn luyện trên tập dữ liệu lớn (thường là ImageNet với >1 triệu ảnh và 1000 lớp) để giải quyết nhiệm vụ mới với dữ liệu hạn chế. Ưu điểm chính:

- Giảm thời gian huấn luyện và nhu cầu dữ liệu.
- Tránh overfitting khi tập dữ liệu mục tiêu nhỏ.
- Khai thác đặc trưng thấp cấp (edges, textures) học được từ các lớp đầu.

Có hai cách tiếp cận chính:

- **Feature Extractor**: Freeze toàn bộ mạng pre-trained, chỉ thay lớp phân loại cuối và huấn luyện lớp mới.

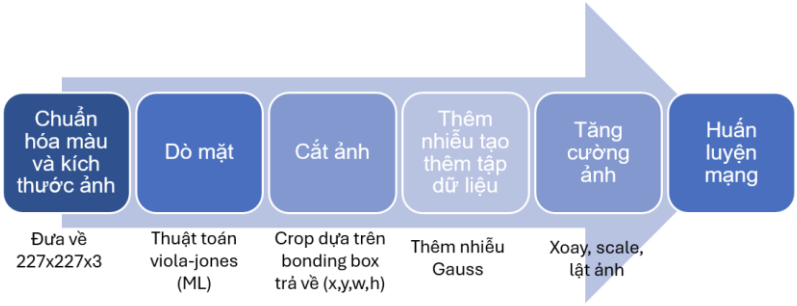
- **Fine-tuning**: Freeze các lớp đầu (học đặc trưng chung), mở khóa một số lớp cuối để huấn luyện lại với learning rate nhỏ.

Trong dự án, chúng ta áp dụng fine-tuning bằng cách thay lớp softmax thành 28 class và mở khóa lớp 8,9,10 của SqueezeNet.

CHƯƠNG 3. QUY TRÌNH TIỀN XỬ LÝ VÀ HUẤN LUYỆN

3.1 Xây dựng và Tiền xử lý dữ liệu

Quy trình được thực hiện qua 4 bước nghiêm ngặt:



1. **Dò tìm:** Sử dụng bộ dò Viola-Jones để xác định khuôn mặt trong ảnh gốc.
2. **Cắt ảnh (Crop):** Dựa trên tọa độ khung (Bounding Box) (x, y, w, h) trả về để tách lấy phần mặt.
3. **Chuẩn hóa:** Đưa ảnh về kích thước 227x227x3 (RGB) để khớp hoàn toàn với lớp đầu vào của SqueezeNet.
4. **Tăng cường (Augmentation):** Để tránh hiện tượng quá khớp (Overfitting), dữ liệu được nhân đôi bằng cách thêm nhiễu Gaussian và thực hiện các phép xoay, thay đổi tỉ lệ, lật ảnh.

3.2 Cấu hình thực nghiệm

Các tham số huấn luyện được khảo sát để tìm ra cấu hình tốt nhất. Một yếu tố quan trọng là Tần suất kiểm tra (Frequency) và số kỷ nguyên (Epoch):

- **Initial Learn Rate:** Được đặt thấp (thường là 0.0001) để không phá vỡ các trọng số đã học từ trước.
- **Mini-batch Size:** Được đặt cố định ở mức 8 để cân bằng giữa tốc độ và sự ổn định của gradient.

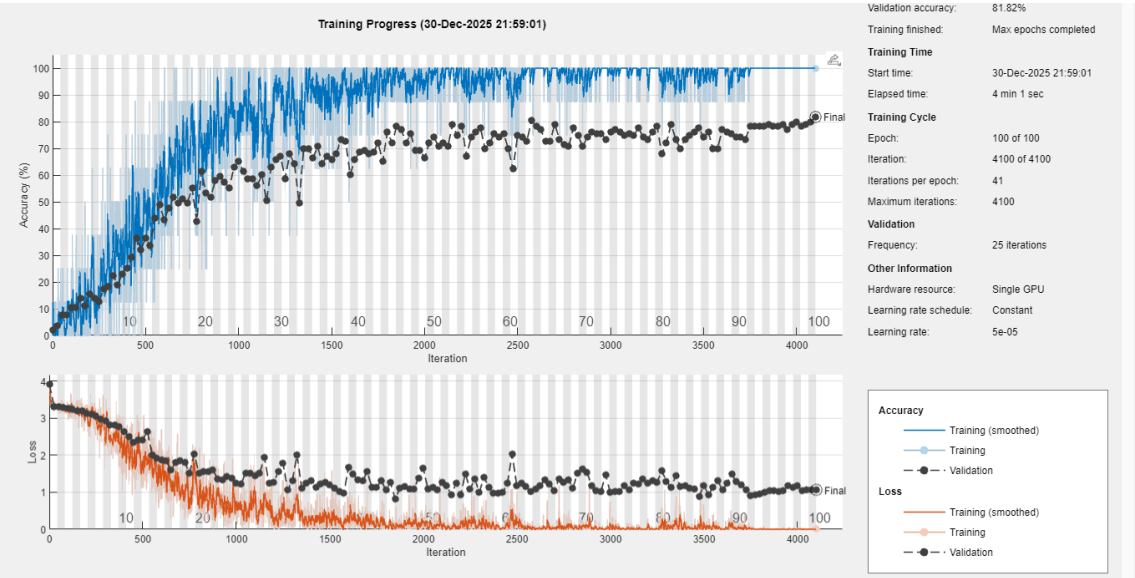
Frequency	Mini batch	epoch	Training time	Validated acurrency
5	8	30	2'35	67.66 %
	5		4'35	68.26 %
	3		10'02	66.47 %
10	8	30	1'56	61.08 %
15			1'35	61.08 %
25			1'24	71.86 %
25	8	15	'43	58.08 %
		50	2'13	61.08 %
		100	4'40	68.26 %

CHƯƠNG 4. KẾT QUẢ VÀ PHÂN TÍCH CHI TIẾT

4.1 Kết quả huấn luyện qua các kịch bản

Dữ liệu thực nghiệm cho thấy sự thay đổi rõ rệt khi điều chỉnh Frequency và Epoch:

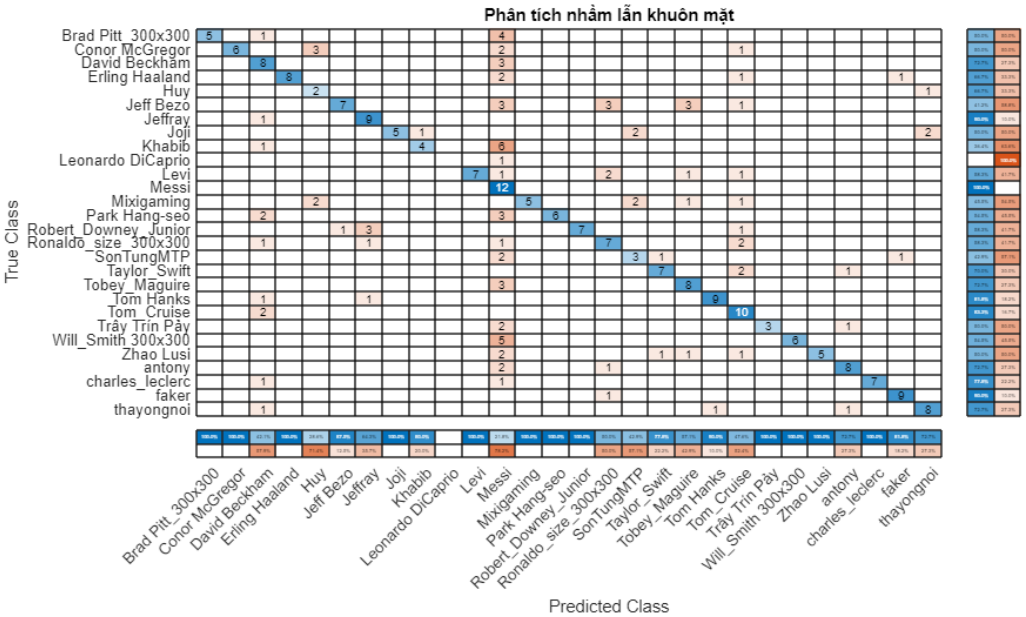
- Ở cấu hình Frequency 25 và 30 Epoch, độ chính xác đạt mức khá là 71.86% chỉ trong 1 phút 24 giây.



- Khi tăng cường độ huấn luyện lên 100 Epoch, mô hình đạt độ chính xác tối ưu là **81.82%** với tổng thời gian 4 phút 1 giây.

4.2 Phân tích Ma trận nhầm lẫn (Confusion Matrix)

Phân tích sâu vào ma trận nhầm lẫn giúp xác định các yếu điểm của mô hình:



- **Các đối tượng thành công:** Messi, Haaland, Tom Cruise có tỉ lệ nhận diện gần như tuyệt đối do các đặc điểm khuôn mặt đặc trưng và bộ dữ liệu sạch.
- **Các đối tượng gây nhầm lẫn:** Jeff Bezos bị sai lệch 10 lần, Khabib bị sai 7 lần. Nguyên nhân có thể do ảnh chụp của các đối tượng này trong bộ dữ liệu có góc nghiêng lớn hoặc điều kiện ánh sáng không đồng nhất, dẫn đến việc mạng nơ-ron khó trích xuất được các đặc trưng bất biến.

4.3 Kiểm tra trên thực tế

Sử dụng tập test bên ngoài, mô hình thể hiện sự tin cậy cao khi dự đoán các nhãn mới với độ tự tin (Confidence) cao:

- Erling Haaland: 99.97%.
- Mixigaming: 97.69%.
- Conor McGregor: 89.73%.

CHƯƠNG 5. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

5.1 Thành tựu đạt được

Đồ án đã xây dựng thành công một hệ thống nhận diện khuôn mặt hoàn chỉnh từ khâu thu thập ảnh thô đến khâu dự đoán thực tế. Việc sử dụng SqueezeNet giúp mô hình cực kỳ gọn nhẹ, phù hợp để tích hợp vào các ứng dụng nhúng mà không cần nâng cấp phần cứng quá mức.

5.2 Hạn chế và Hướng phát triển

- **Hạn chế:**

- Độ chính xác 81.82% vẫn còn dư địa để cải thiện, đặc biệt là với các nhãn có độ tương đồng cao.
- Tuy độ chính xác cao nhưng do dữ liệu hạn chế nên tỉ lệ phân loại tên tập test vẫn sai rất cao.

- **Hướng phát triển:** Trong tương lai, hệ thống có thể tích hợp thêm các mô hình dò mặt mạnh hơn như MTCNN thay cho Viola-Jones để xử lý các góc nghiêng khó. Ngoài ra, việc bổ sung dữ liệu ảnh chất lượng cao cho các đối tượng như Jeff Bezos sẽ giúp nâng cao độ chính xác tổng thể.

DANH MỤC TÀI LIỆU THAM KHẢO

[1] Iandola, F. N., et al. (2016). SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size. arXiv:1602.07360.

[2] Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. CVPR.