TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP TP HỒ CHÍ MINH
KHOA CÔNG NGHỆ THÔNG TIN

TRẦN XUÂN THỦY
LÊ DUY TÂM

XÂY DỰNG HỆ THỐNG TRUY XUẤT HÌNH ẢNH BẰNG NỘI DUNG DỰA TRÊN CNN VÀ HASH

Chuyên ngành: Khoa học Dữ liệu

Giảng viên hướng dẫn: TS. BÙI THANH HÙNG

THÀNH PHỐ HỒ CHÍ MINH, THÁNG 5 NĂM 2023

INDUSTRIAL UNIVERSITY OF HO CHI MINH CITY

FACULTY OF INFORMATION TECHNOLOGY

TRAN XUAN THUY

LE DUY TAM

BUILDING A CONTENT-BASED IMAGE RETRIEVAL SYSTEM USING CNN AND HASH

Major: Data Science

Supervisor: Dr. BUI THANH HUNG

HO CHI MINH CITY, MAY 2023

LỜI CẢM ƠN

Chúng em muốn gửi lời cảm ơn sâu sắc nhất đến Thầy Bùi Thanh Hùng, người đã dành thời gian và kiến thức để trang bị cho chúng em những kiến thức và kỹ năng quan trọng trong quá trình hoàn thành đồ án tốt nghiệp.

Chúng em cũng muốn gửi lời cảm ơn đến các thầy cô giảng viên và nhà trường Đại học Công nghiệp thành phố Hồ Chí Minh đã tạo điều kiện thuận lợi cho chúng em trong quá trình học tập tại trường.

Chúng em vẫn còn nhiều thiếu sót trong quá trình học tập và nghiên cứu do chúng em chưa vững kiến thức chuyên ngành cũng như kinh nghiệm thực tế.

Chúng em rất biết ơn đến gia đình, bạn bè và người thân đã luôn ủng hộ, động viên và giúp đỡ chúng em trong suốt quá trình học tập và hoàn thành đồ án tốt nghiệp. Sự quan tâm và động viên của họ đã truyền động lực cho chúng em vượt qua những khó khăn và thử thách trong quá trình học tập.

Chúng em rất cảm ơn thầy vẫn luôn dành nhiều thời gian để hướng dẫn, chỉ bảo tận tình cho chúng em, giúp chúng em hoàn thiện đồ án cũng như bản thân mình.

LUẬN VĂN ĐƯỢC HOÀN THÀNH
TẠI TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP TP HỒ CHÍ MINH

Tôi xin cam đoan đây là sản phẩm đồ án của riêng chúng tôi và được sự hướng dẫn của TS. Bùi Thanh Hùng. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong luận văn còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

Nếu phát hiện có bất kỳ sự gian lận nào tôi xin hoàn toàn chịu trách nhiệm về nội dung đồ án của mình. Trường đại học Công nghiệp TP Hồ Chí Minh không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện (nếu có).

TP. Hồ Chí Minh, ngày tháng năm

Tác giả

(ký tên và ghi rõ họ tên)

Trần Xuân Thủy

Lê Duy Tâm

NHẬN XÉT VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN HƯỚNG DẪN

_____

_____

_____

_____

_____

_____

_____

____

Tp. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

NHẬN XÉT VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN PHẢN BIỆN

_____

_____

_____

_____

_____

_____

_____

____

Tp. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

TÓM TẮT

Trong những năm gần đây, truy xuất hình ảnh bằng nội dung đã và đang thu hút được sự quan tâm nghiên cứu của các chuyên gia bởi xu hướng gia tăng theo cấp số nhân của dữ liệu toàn cầu và tại Việt Nam. Thống kê của Seedscientic cho rằng, đến đầu năm 2020, vũ trụ dữ liệu số được ước tính bao gồm 44 zettabytes dữ liệu và đến năm 2025, ước tính sẽ có thêm 1% lượng dữ liệu này được tạo mới mỗi ngày. Đặc biệt với dạng dữ liệu hình ảnh với các ma trận điểm ảnh. Với sự phát triển của deep learning trong đó có mạng nơ-ron tích chập (CNN) thì các công việc liên quan đến thị giác máy tính đã có những lời giải tối ưu. Các công cụ tìm kiếm hình ảnh cũng nhờ đó đã ra đời, tuy nhiên tốc độ xử lý còn chậm bởi vì lí do kể trên. Do đó nếu giải quyết được vấn đề này thì tìm kiếm hình ảnh hoàn toàn có thể thay thế được cho các cách tìm kiếm thông thường. Nhược điểm của phương pháp trước đây là các vector đặc trưng có độ dài lớn dẫn đến việc tốn thời gian tìm kiếm và gây sự khó chịu cho cả người dùng và người quản lý. Để giải quyết những vấn đề đó, chúng tôi sử dụng phương pháp dựa trên mạng nơ-ron tích chập để trích xuất đặc trưng ảnh sau đó tiếp tục thực hiện việc sinh mã nhị phân (binary hashing). Kết quả thực nghiệm trên bộ dữ liệu fashion-product-images-dataset cho thấy việc sử dụng mạng nơ-ron tích chập kết hợp phương pháp sinh mã nhị phân trong tìm kiếm ảnh có MAP20 = 0.99 và cải thiện đáng kể thời gian truy vấn ảnh.

SUMMARY

In recent years, content-based image retrieval has attracted significant research attention from experts due to the exponential growth of global data and in Vietnam. According to Seedscientic's statistics, by early 2020, the digital universe was estimated to contain 44 zettabytes of data, and by 2025, it is projected that an additional 1% of this data will be generated every day. Especially, image data in the form of pixel matrices. With the development of deep learning, including convolutional neural networks (CNNs), computer vision tasks have been optimized. Image search tools have also been developed, but their processing speed is still slow due to the aforementioned reasons. Therefore, if this problem can be addressed, image search can completely replace traditional search methods. The drawback of previous methods is that the feature vectors are long, leading to time-consuming search and causing inconvenience for both users and managers. To address these issues, we utilize a convolutional neural network for feature extraction and then proceed with binary hashing. Experimental results on the fashion-product-images-dataset show that using a convolutional neural network combined with binary hashing in image search achieves a mean average precision (MAP20) of 0.99 and significantly improves the image retrieval time.

DANH MỤC CHỮ VIẾT TẮT

TBIR	Text-Based Image Retrieval	Truy xuất (tìm kiếm) ảnh dựa trên văn bản
CBIR	Content-based Image Retrieval	Truy xuất (tìm kiếm) ảnh theo nội dung
CNN	Convolutional Neural Network	Mạng nơ-ron tích chập
QBIC	Query By Image Content	Truy vấn dựa trên nội dung hình ảnh
TMĐT		Thương mại điện tử
ANN	Artificial Neural Network	Mạng nơ-ron nhân tạo
CSDL		Cơ sở dữ liệu
MAP	Mean Average Precision	Trung bình cộng giá trị AP của các lớp khác nhau

DANH MỤC BẢNG BIỂU

DANH MỤC HÌNH

MỤC LỤC

CHƯƠNG 1
GIỚI THIỆU CHUNG

Trong chương này, chúng tôi sẽ giới thiệu tổng quan về các vấn đề được nghiên cứu của đề tài, đồng thời xác định mục tiêu, phạm vi nghiên cứu cũng như những ý nghĩa mà kết quả của đề tài này mang lại cả về ý nghĩa khoa học và áp dụng thực tiễn

1.1. Lý do chọn đề tài

Trong thực tế hiện nay, với sự phát triển của công nghệ thông tin nhu cầu tìm kiếm thông tin đang tăng cao, do đó nhu cầu tìm kiếm ảnh cũng là một lĩnh vực nghiên cứu được quan tâm. Bài toán tìm kiếm hay truy xuất hình ảnh được chia thành hai phương pháp chính [1]. Thứ nhất là bài toán tìm kiếm ảnh dựa trên văn bản TBIR (Text-Based Image Retrieval). Phương pháp này có nhiều hạn chế như tốn thời gian để mô tả hình ảnh, kết quả không như ý muốn do sự mô tả của con người. Thứ hai là bài toán tìm kiếm ảnh dựa trên nội dung CBIR (Content-Based Image Retrieval), là tìm kiếm các hình ảnh có màu sắc, hình dạng, kết cấu hoặc bất kỳ thông tin nào khác có thể được lấy từ chính hình ảnh chứ không phải là từ khóa, thẻ hoặc mô tả được liên kết với hình ảnh, nghĩa là dựa trên các đặc trưng, thuộc tính và thông tin nội dung của hình ảnh để tìm kiếm những hình ảnh tương đồng. Phương pháp CBIR thực hiện tìm kiếm dựa trên đặc trưng thị giác của hình ảnh, vì thế nó khắc phục được các hạn chế của phương pháp TBIR. Đối với CBIR, vấn đề trích xuất đặc trưng giữ vai trò rất quan trọng để đánh giá hiệu quả tìm kiếm. Các hệ thống truy xuất cũ như QBIC, VisualSeek,…thường dựa vào các đặc trưng như màu sắc, kích thước, độ sáng, hình dạng, cấu trúc,... được trích xuất từ các ảnh[1]. Phương pháp này vẫn tồn tại nhiều hạn chế như khó tìm ra được đặc trưng để phù hợp cho việc tìm kiếm đạt kết quả tốt.

Xuất phát từ những vấn đề trên, trong đồ án này, chúng tôi sẽ giới thiệu một hệ thống truy xuất hình ảnh bằng nội dung dựa trên CNN và hash.

1.2. Mục tiêu nghiên cứu

Công cụ tìm kiếm hình ảnh sẽ được huấn luyện trên tập dữ liệu fashion-product-images-dataset để có thể đưa ra phản hồi hình ảnh tương tự nhanh chóng và chính xác nhất khi người dùng (đối tượng nghiên cứu trong bài báo) muốn tìm kiếm một món hàng nhất định.

1.3. Đối tượng, phạm vi nghiên cứu

Xây dựng công cụ tìm kiếm hình ảnh bằng phương pháp kết hợp thuật toán hash và các mô hình học sâu. Huấn luyện mô hình trên dữ liệu tập hình ảnh có sẵn.

Sử dụng các kiến thức đã học để tiến hành nghiên cứu liên quan đến lĩnh vực thị giác máy tính với các mô hình được đào tạo trước như VGG, Alexnet, Resnet, … kết hợp với giải thuật hash để thực nghiệm theo bộ dữ liệu, từ đó xây dựng công cụ có thể ứng dụng vào thực tế.

1.4. Phương pháp nghiên cứu

Các phương pháp nghiên cứu mà nhóm đã sử dụng để khám phá và đánh giá hiệu xuất của mô hình truy xuất hình ảnh. Cụ thể, tập trung vào hai phương pháp quan trọng: phương pháp nghiên cứu lý thuyết và phương pháp nghiên cứu thực nghiệm.

1.4.1. Phương pháp nghiên cứu lý thuyết

Thu thập và nghiên cứu mô hình, bài báo liên quan đến thị giác máy tính, trích xuất đặc trưng và các tài liệu liên quan đến đề tài truy xuất hình ảnh. Dựa trên nền tảng này, luận án đề xuất một phương pháp mới và xây dựng một mô hình truy xuất hình ảnh. Phương pháp và mô hình được thiết kế để tối ưu hóa quá trình trích xuất đặc trưng và cải thiện độ chính xác của hệ thống. Sau đó, phương pháp đề xuất được thực hiện và đánh giá hiệu suất dựa trên các tập dữ liệu thích hợp.

1.4.2 Phương pháp nghiên cứu thực nghiệm

Xây dựng mô hình truy xuất hình ảnh dựa trên nội dung và thực hiện thử nghiệm trên tập dữ liệu thu thập. Qua đó, tiến hành so sánh độ chính xác và thời gian thực hiện của mô hình với các phương pháp khác trên tập dữ liệu nhỏ hơn. Kết

quả từ thử nghiệm sẽ cung cấp thông tin về hiệu suất của mô hình và đóng góp vào sự phát triển của lĩnh vực truy xuất hình ảnh. Điều này giúp người dùng có trải nghiệm tốt hơn và tìm kiếm hình ảnh một cách hiệu quả trong thực tế.

1.5. Ý nghĩa khoa học và thực tiễn

Điều quan trọng của một luận án là ý nghĩa của nó, và nghiên cứu này không phải là ngoại lệ. Có hai vấn đề quan trọng khi nhận xét ý nghĩa của một luận án là: ý nghĩa khoa học và ý nghĩa thực tiễn.

1.5.1. Ý nghĩa khoa học

Luận án có ý nghĩa khoa học quan trọng bởi vì nó tìm hiểu và chứng minh sự ứng dụng hiệu quả của mạng nơ-ron tích chập (CNN) kết hợp với kỹ thuật hash trong việc cải thiện hiệu suất và độ chính xác của hệ thống truy xuất hình ảnh dựa trên nội dung. Bằng cách kết hợp khả năng trích xuất đặc trưng mạnh mẽ của CNN với tính năng nhanh chóng và khả năng xử lý song song của kỹ thuật hash, luận án đã đạt được kết quả tốt hơn trong việc truy vấn hình ảnh và tìm kiếm các hình ảnh tương tự. Kết quả này có thể đóng góp vào phát triển các ứng dụng thực tế như truy xuất hình ảnh, gợi ý sản phẩm và phân loại hình ảnh.

1.5.2 Ý nghĩa thực tiễn

Về ý nghĩa thực tiễn, hệ thống truy xuất hình ảnh có thể dựa trên nội dung mà nó cung cấp. Người dùng không cần phải mô tả chi tiết, chỉ cần sử dụng một hình ảnh có sẵn, hệ thống có khả năng tìm kiếm và đưa ra những sản phẩm tương tự ngay lập tức. Điều này giúp cải thiện trải nghiệm của người dùng, đồng thời giảm thời gian và công sức tìm kiếm thông qua quá trình tương tác đơn giản. Hệ thống truy xuất hình ảnh này có thể ứng dụng rộng rãi trong các lĩnh vực như thương mại điện tử, quảng cáo, thư viện ảnh và nhiều lĩnh vực khác, đáp ứng nhu cầu ngày càng tăng về tìm kiếm thông tin hình ảnh một cách nhanh chóng và hiệu quả.

1.6. Bố cục luận văn

Toàn bộ nội dung luận văn được trình bày trong 5 chương như sau:

Chương 1: Tổng quan về lĩnh vực nghiên cứu sơ lược tổng quan về vấn đề nghiên cứu trên phương diện tổng quan nhất, nêu ra mục tiêu, đối tượng nghiên cứu, phương pháp nghiên cứu và bố cục luận văn.

Chương 2: Cơ sở lý thuyết và các nghiên cứu liên quan. Giới thiệu tổng quan về thị giác máy tính, về CNN, giới thiệu về mạng neural nhân tạo, các mô hình mạng neural cải tiến là cơ sở của mạng học sâu. Trình bày về một số nghiên cứu về trích xuất, tìm kiếm hình ảnh, cùng với tình hình nghiên cứu trong nước và ngoài nước thời gian gần đây.

Chương 3: Mô hình đề xuất. Trình bày tổng quan về mô hình đề xuất và đi sâu phân tích các đặc trưng của mô hình đề xuất.

Chương 4: Thực nghiệm trình bày các kết quả huấn luyện mô hình đạt được và phân tích, đánh giá, so sánh kết quả đạt được với các mô hình trước.

Chương 5: Kết luận và hướng phát triển.

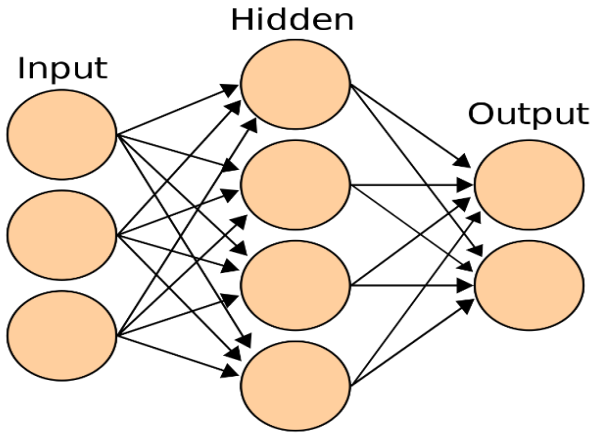
CHƯƠNG 2
CƠ SỞ LÝ THUYẾT VÀ CÁC NGHIÊN CỨU LIÊN QUAN

Trong chương này, chúng tôi sẽ giới thiệu về mạng nơ-ron tích chập, sơ lược về các mô hình CNN, các phương pháp học sâu cho tìm kiếm hình ảnh và đề xuất phương pháp nghiên cứu giải quyết bài toán tìm kiếm bằng hình ảnh.

2.1. Mạng nơ-ron tích chập

2.1.1. Mạng nơ-ron nhân tạo

Mạng nơ-ron nhân tạo (hay gọi là mạng nơ-ron) (Artificial Neural Network - ANN) là một mô hình tính toán được xây dựng dựa trên sự hoạt động của các nơ-ron trong hệ thần kinh sinh học. Nó bao gồm các nơ-ron nhân tạo được tổ chức thành các lớp được kết nối với nhau thông qua trọng số. Cấu trúc của mạng nơ-ron bao gồm 3 lớp chính:



Hình 2.1. Cấu trúc của mạng nơ-ron[1]

Lớp nhập (Input layer): Nhận đầu vào là các số hoặc vector từ bên ngoài. Số node của lớp này bằng với kích thước của vector đặc trưng đầu vào

Lớp ẩn (Hidden layer): Là lớp nằm giữa lớp nhập và lớp đầu ra, thực hiện tính toán và truyền thông tin từ lớp nhập đến lớp đầu ra thông qua trọng số.

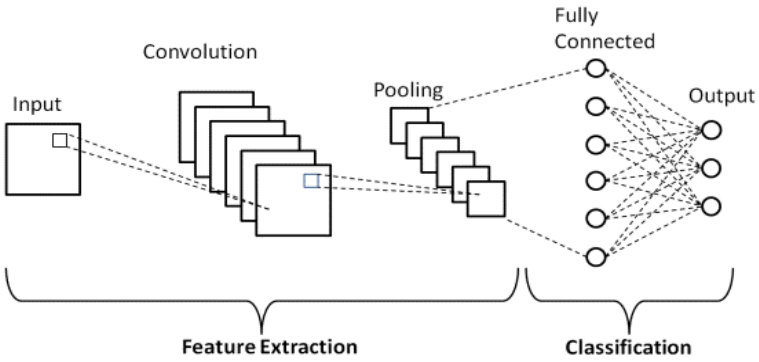
[1] https://vi.wikipedia.org/wiki/M%E1%BA%A1ng_th%E1%BA%A7n_kinh_nh%C3%A2n_t%E1%BA%A1o

Lớp đầu ra (Output layer): Tạo ra đầu ra dự đoán dựa trên thông tin từ lớp ẩn.

Mỗi nơ-ron trong mạng nơ-ron nhận đầu vào từ các nơ-ron trong lớp trước đó sau đó tổng hợp thông tin và áp dụng một hàm kích hoạt để tạo ra đầu ra. Các trọng số được gán cho các kết nối giữa các nơ-ron, và qua quá trình huấn luyện, các trọng số này được điều chỉnh để mạng nơ-ron có thể học và tạo ra các dự đoán chính xác. Mạng nơ-ron nhân tạo đã được ứng dụng rộng rãi trong nhiều lĩnh vực như xử lý ngôn ngữ tự nhiên, dự báo thị trường tài chính, hệ thống tự lái,… và đặc biệt là nhận dạng hình ảnh. Nó đóng vai trò quan trọng trong phát triển trí tuệ nhân tạo và là một công cụ mạnh mẽ để xử lý dữ liệu phức tạp và tìm ra các mẫu ẩn trong dữ liệu.

2.1.2. Mạng nơ-ron tích chập

Dữ liệu hình ảnh có kích thước rất lớn và cần một mạng sâu hơn để học được các đặc trưng cấp cao, điều này không dễ để thực hiện đối với mạng ANN thông thường bởi vì số lượng tham số quá lớn của nó . Mạng nơ-ron tích chập (Convolutional Neural Network), viết tắt là CNN ra đời để giải quyết vấn đề này. CNN đã đạt được sự thành công đáng kể trong các ứng dụng liên quan đến thị giác máy tính, bao gồm nhận dạng khuôn mặt, nhận dạng đối tượng, xử lý ngôn ngữ tự nhiên và đặc biệt là phân loại hình ảnh. Mô hình CNN bao gồm các lớp: Lớp Tích chập (convolution), Lớp Gộp (pooling), Đơn vị Tuyến tính Cải tiến (ReLU) và Lớp Fully Connected.



Hình 2.2. Cấu trúc mạng CNN[2]

Lớp Tích chập (Convolutional layer): Lớp này áp dụng các bộ lọc (filters) lên các vùng nhỏ của ảnh đầu vào để tạo ra các feature maps. Mỗi bộ lọc là một ma trận trọng số được trượt qua khắp các vùng ảnh. Kết quả của phép tích chập tạo thành các feature maps, trong đó mỗi giá trị đại diện cho một đặc trưng cụ thể trong ảnh.

Lớp Gộp (Pooling layer): Lớp này thực hiện việc giảm kích thước của feature maps bằng cách lấy giá trị tối đa (Max Pooling) hoặc giá trị trung bình (Average Pooling) trong các vùng không gian. Qua quá trình này, số lượng thông tin và tham số của mạng giảm đi, đồng thời giữ lại các đặc trưng quan trọng và giảm thiểu hiện tượng overfitting.

Đơn vị Tuyến tính Cải tiến (ReLU - Rectified Linear Unit): Đây là một hàm kích hoạt không tuyến tính được áp dụng sau lớp tích chập hoặc lớp gộp. ReLU chuyển đổi các giá trị âm thành 0 và giữ nguyên các giá trị không âm. Điều này giúp kích hoạt sự phi tuyến và cải thiện khả năng học của mạng.

Lớp Fully Connected (Hoàn toàn kết nối): Lớp này tạo ra các kết nối đầy đủ giữa các nút (neuron) trong lớp trước đó và lớp kết quả. Các nút trong lớp này nhận thông tin từ tất cả các nút trong lớp trước đó và tính toán kết quả dự đoán cuối cùng. Đây là lớp cuối cùng trong mạng CNN và thường được kết hợp với các lớp kích hoạt như softmax để đưa ra xác suất phân loại.

[2]https://www.researchgate.net/figure/CNN-Architecture-Convolutional-layer-This-layer-involves-a-mathematical-operation-that_fig2_365435384

Tóm lại, lớp Tích chập để trích xuất đặc trưng, lớp Gộp để giảm kích thước thông tin, lớp ReLU để kích hoạt phi tuyến và lớp Fully Connected để tính toán kết quả dự đoán cuối cùng. Sự kết hợp của các lớp này tạo nên kiến trúc mạng CNN mạnh mẽ để xử lý hình ảnh và các nhiệm vụ liên quan.

Mạng CNN đã chứng minh hiệu suất cao trong việc xử lý dữ liệu không gian và thể hiện khả năng học hiệu quả từ các mẫu dữ liệu lớn. Với những ưu điểm của mình, mạng CNN đang trở thành một trong những công cụ học máy quan trọng và phổ biến nhất hiện nay, đặc biệt là trong lĩnh vực nhận dạng hình ảnh.

2.2. Các mô hình học sâu CNN

2.2.1 Sự hình thành và phát triển

Mạng Neocognitron, mạng tương tự với mạng nơ-ron tích chập (CNN), được đề xuất bởi Kunihiko Fukushima, vào năm 1980 [2]. Tuy nhiên, mô hình này khác với CNN hiện đại trong cách thức hoạt động và cách xử lý dữ liệu. Năm 1998, Yann LeCun và các đồng nghiệp đã phát triển mô hình CNN đầu tiên cho bài toán nhận dạng đối tượng trong ảnh và đưa ra kết quả đáng kể trong các cuộc thi nhận dạng hình ảnh trong bài báo Object Recognition with Gradient-Based Learning [3].

2.2.2. Các mô hình tiêu biểu

Với sự phát triển vượt bậc của mạng nơron tích chập (CNN) trong lĩnh vực xử lí ảnh và thị giác máy tính như: Yangqing Jia và cộng sự đã đề xuất phương pháp trích xuất đặc trưng sử dụng mô hình CNN cho bài toán truy xuất ảnh, giúp cải thiện độ chính xác so với các phương pháp truy xuất ảnh truyền thống trong bài báo Learning Deep Features for Image Retrieval [4]. Alex Krizhevsky và cộng sự [5] đã giới thiệu mô hình AlexNet, với số lượng tham số và độ sâu lớn lớn hơn rất nhiều so với các mô hình trước đây. Mô hình này đã giành chiến thắng trong cuộc thi ImageNet Large Scale Visual Recognition Challenge vào năm 2012 và đánh dấu sự trỗi dậy của CNN. Karen Simonyan và Andrew Zisserman đã giới thiệu mô hình VGGNet trong bài báo Very Deep Convolutional Networks for Large-Scale Image Recognition [6], với cấu trúc mạng CNN có độ sâu lên tới 19 hoặc 16 lớp, giúp cải

thiện độ chính xác trong phân loại hình ảnh. Mô hình này đã được sử dụng rộng rãi trong các bài toán phân loại ảnh và truy xuất ảnh. Trong bài báo Deep Residual Learning for Image Recognition, Kaiming He và cộng sự đã giới thiệu mô hình ResNet [7], với cấu trúc mạng CNN sử dụng residual blocks để giải quyết vấn đề mất mát thông tin khi độ sâu mạng tăng lên. Mô hình này đã đạt được độ chính xác rất cao trong bài toán phân loại ảnh và truy xuất ảnh. Andrew G. Howard và cộng sự đã giới thiệu mô hình MobileNet [8] trong bài báo MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications, với cấu trúc mạng CNN nhẹ và hiệu quả về mặt tính toán để sử dụng trên các thiết bị di động. Mô hình này đã đạt được kết quả tốt trong các bài toán phân loại ảnh và truy xuất ảnh trên các thiết bị di động.

2.3. Tìm kiếm hình ảnh bằng phương pháp học sâu

2.3.1. Tổng quan

Tìm kiếm ảnh bằng phương pháp học sâu là một bài toán một trong những bài toán thuộc lĩnh vực truy xuất hình ảnh. Mục tiêu của bài toán là tìm kiếm và xếp hạng các hình ảnh dựa trên sự tương đồng dựa trên đặc trưng của chúng. Phương pháp học sâu được áp dụng để trích xuất đặc trưng của hình ảnh và biểu diễn chúng dưới dạng các vector số. Các mô hình học sâu như mạng nơ-ron tích chập (CNN) thường được sử dụng để trích xuất đặc trưng này. Mô hình mạng nơ-ron tích chập (CNN) được huấn luyện trên tập dữ liệu lớn để học cách phân biệt và nhận dạng các đặc trưng của hình ảnh. Sau khi đặc trưng của hình ảnh được trích xuất, chúng được biểu diễn dưới dạng vector. Các phương pháp đo độ tương đồng như cosine similarity, Euclidean distance hoặc khoảng cách Hamming có thể được sử dụng để so sánh các vector đặc trưng của hình ảnh và xác định độ tương đồng giữa chúng. Bài toán tìm kiếm hình ảnh bằng phương pháp học sâu có ứng dụng rộng rãi trong nhiều lĩnh vực như tìm kiếm ảnh trực tuyến, gợi ý hình ảnh, tìm kiếm sản phẩm dựa trên hình ảnh và nhiều ứng dụng khác liên quan đến truy xuất thông tin hình ảnh.

Tìm kiếm hình ảnh bằng phương pháp học sâu là một lĩnh vực nghiên cứu đang rất phát triển trong trí tuệ nhân tạo. Sử dụng các mô hình học sâu, như mạng CNN

đã cải thiện độ chính xác và tốc độ trong khả năng nhận dạng, phân loại và tìm kiếm hình ảnh.

2.3.2. Các nghiên cứu liên quan

Trong bài báo Image-based Product Recommendation System with Convolutional Neural Networks của Stanford University, 450 Serra Mall, Stanford, CA vào năm 2017. Chen, Luyang và cộng sự [9] sử dụng các mô hình học sâu Alexnet, VGG để huấn luyện trên tập dữ liệu sản phẩm của Amazon nhằm mục đích tìm kiếm hình ảnh. Hệ thống cũng chia ra hai phần. Đầu tiên là sử dụng mô hình đã được huấn luyện để trích xuất đặc trưng của ảnh (đây vốn dĩ là một mô hình phân lớp nhưng được cắt đi lớp cuối cùng). Sau đó sử dụng cosine similarity để đo độ tương đồng của đặc trưng của ảnh mới và các đặc trưng của ảnh trong tập dữ liệu. Ảnh đề xuất sẽ là ảnh có độ tương đồng cosine cao nhất so với ảnh cần tìm.

Do mô hình còn đơn giản nên độ chính xác trong việc phân lớp còn chưa cao (0.5) và chỉ sử dụng độ đo tương đồng cosine sau khi trích xuất đặc trưng khiến công việc tìm kiếm chưa tối ưu. Dữ liệu được sử dụng cũng chưa thật sự tốt khi chưa chi tiết hóa được các lớp.

Trong bài báo Image Retrieval Algorithm Based on Locality-Sensitive Hash Using Convolutional Neural Network and Attention Mechanism năm 2022, tác giả Youmeng Luo và cộng sự [10] sử dụng mạng Resnet-50 kết hợp với attention mechanism để tăng độ chính xác trong công việc trích xuất đặc trưng cho hình ảnh. Huấn luyện trên tập dữ liệu corel5K, với 20 lớp. Mô hình sau khi huấn luyện đạt được độ chính xác recall là 0.95 trên tập dữ liệu trên. Tuy nhiên mô hình chưa khai thác được các tính năng cấp thấp của ảnh cũng như tập dữ liệu chưa phong phú.

Tác giả Varga, D., và Szirányi, T. trong bài báo “Fast content-based image retrieval using Convolutional Neural Network and hash function” [11] được đăng trong hội nghị quốc tế IEEE 2016 sử dụng mạng CNN với 5 lớp convolution, hàm hash cũng được sử dụng để biến các đặc trưng thành mã băm. đã cho thấy sự hiệu quả đối với bộ dữ liệu nhỏ.

Trong bài báo Medical Image Retrieval Based on Convolutional Neural Network and Supervised Hashing vào năm 2019 Yiheng cai và cộng sự [12] đã sử dụng mạng CNN kết hợp mạng Siamese với phương pháp băm có kết quả vượt trội hơn so với các phương pháp băm và phương pháp CNN hiện có thể áp dụng được cho ảnh y tế.

2.3.3. Hướng nghiên cứu đề xuất

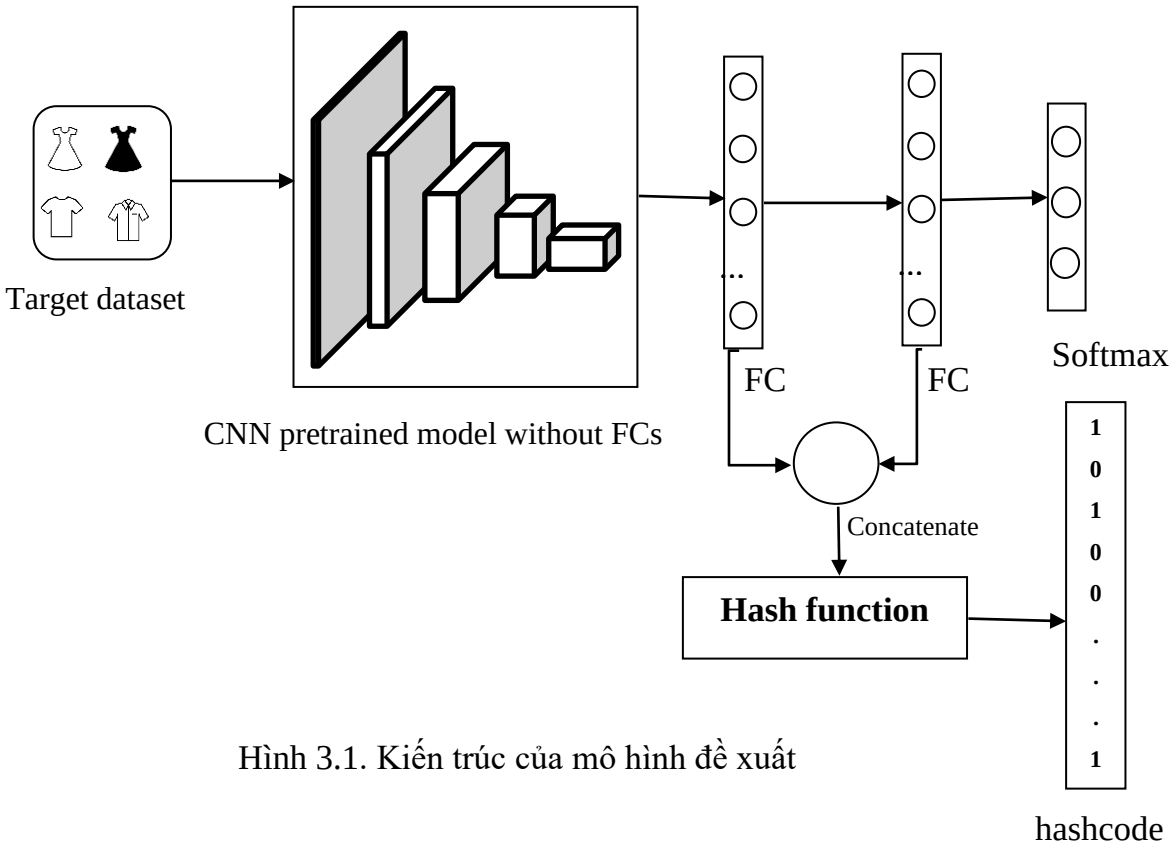
Chúng tôi sẽ giới thiệu phương pháp mới trong lĩnh vực truy xuất ảnh dựa trên mạng nơ-ron tích chập (CNN) kết hợp với việc áp dụng mã nhị phân (binary hashing). Phương pháp này nhằm chuyển đổi các đặc trưng trích xuất thành các mã băm (hash code), mỗi mã băm là một vector nhị phân. Qua việc biểu diễn dữ liệu dưới dạng nhị phân, phương pháp giúp giảm bộ nhớ lưu trữ cần thiết và tăng tốc độ tính toán và truy xuất dữ liệu. Đồng thời, việc sử dụng các độ đo như khoảng cách Hamming giữa các mã băm cũng giúp đánh giá sự tương đồng giữa các ảnh một cách nhanh chóng và hiệu quả. Phương pháp này hứa hẹn sẽ mang lại cải tiến đáng kể trong việc phân tích và tìm kiếm hình ảnh.

CHƯƠNG 3
MÔ HÌNH ĐỀ XUẤT

Trong phần này, chúng tôi sẽ trình bày từ tổng quan đến chi tiết của mô hình đề xuất, hàm băm được sử dụng. Để truy xuất hình ảnh chúng tôi cũng đưa ra phương pháp tính toán độ tương đồng giữa hai mã băm (hashcode) và cuối cùng là phương pháp đánh giá hiệu quả truy xuất của hệ thống.

3.1 Tổng quan mô hình đề xuất

Như được trình bày ở hình 3.1. Mô hình đề xuất của chúng tôi sẽ được chia làm 2 phần. Trong phần đầu tiên, với nhiệm vụ rút trích đặc trưng, mô hình gồm nhiều lớp CNN sẽ được sử dụng. Tiếp theo, các đặc trưng được trích xuất qua mô hình trên sẽ được đưa qua một hàm băm (hash function) để tạo mã băm (hashcode). Hashcode này sẽ được lưu trữ để biểu diễn cho hình ảnh trong bộ dữ liệu nhằm mục đích truy xuất.



Hình 3.1. Kiến trúc của mô hình đề xuất

3.2. Đặc trưng của mô hình đề xuất

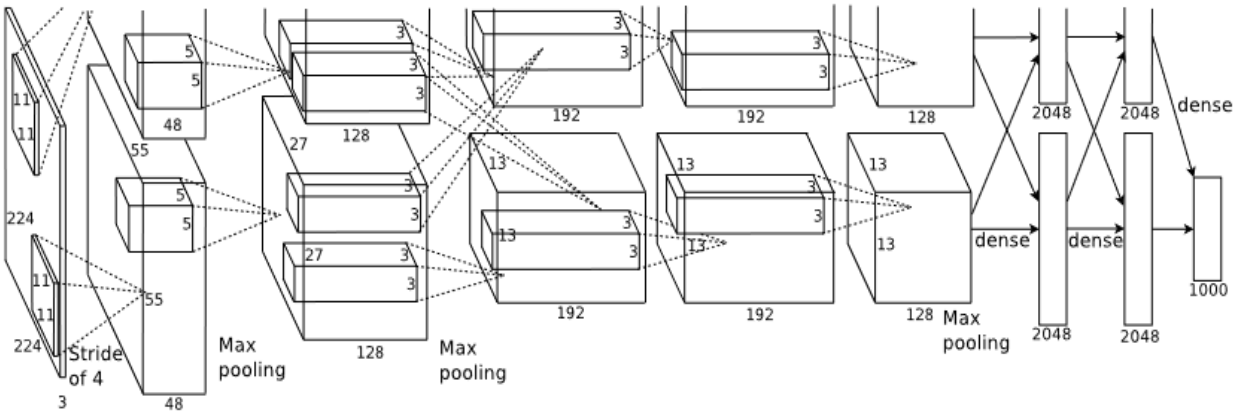
Ở phần này chúng tôi sẽ nói rõ hơn về từng phần của mô hình đề xuất. Cách thức ứng dụng nó trong bài toán tìm kiếm hình ảnh cũng được nêu ra trong phần cuối cùng của chương này.

3.2.1. Mô hình CNN

Mở đầu phần mô hình CNN, chúng tôi sẽ giới thiệu một số mô hình quan trọng trong lĩnh vực xử lý ảnh và truy xuất thông tin dựa trên nội dung(CBIR). Chúng tôi sẽ giới thiệu các mô hình CNN đáng chú ý như AlexNet, VGG19, Res-Net, MobileNet, và khả năng của chúng trong việc giải quyết bài toán tìm kiếm bằng hình ảnh trong các tiểu mục sau.

3.2.1.1. Mô hình Alexnet

Như đã nói, AlexNet là một mô hình CNN tiên phong trong việc xây dựng mạng học sâu được giới thiệu vào năm 2012 bởi Hinton và nghiên cứu sinh của ông là Alex Krizhevsky [5], và tên của mạng cũng được đặt theo tên nghiên cứu sinh đó. Nó đã chứng minh khả năng xuất sắc trong việc phân loại hình ảnh và khởi đầu cho sự phát triển của các mô hình CNN sau này. Với kiến trúc sâu hơn với 8 lớp tích chập và 3 lớp được kết nối đầy đủ (Fully Connected).



Hình 3.2. Kiến trúc mô hình Alexnet[3]

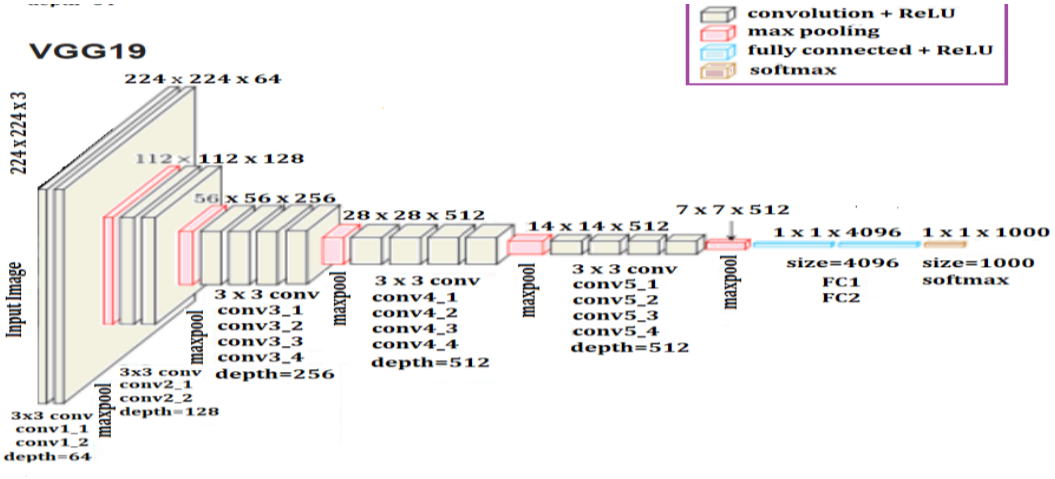
[3] https://tiendv.wordpress.com/2016/12/25/convolutional-neural-networks/

Với việc tăng một số lượng lớp tích chập lớn, sẽ xảy ra tình trạng mô hình quá phù hợp với dữ liệu huấn luyện (Overfitting). Để đối phó với vấn đề này, AlexNet đã sử dụng hàm kích hoạt ReLU trong kiến trúc mạng và đặc biệt là kĩ thuật Dropout, kỹ thuật này sẽ giúp cho mô hình không phụ thuộc quá nhiều vào một số đơn vị cụ thể trong mạng, do đó, nó phải học cách phân phối khái quát đặc trưng trên nhiều đơn vị khác nhau. Giúp cải thiện khả năng tổng quát hóa của mô hình, mô hình trích xuất và biểu diễn các đặc trưng của hình ảnh một cách hiệu quả hơn. Khả năng tìm kiếm hình ảnh tương tự từ đó được hiệu quả hơn.

Để áp dụng vào bài toán truy xuất hình ảnh chúng tôi sẽ loại bỏ đi ba lớp fully connected ở cuối cùng trong mạng. thay vào đó là ba lớp FC mới với 20 node ở lớp softmax, 514 node ở hai lớp áp chót. Đầu ra của hai lớp này sẽ dùng để trích xuất đặc trưng của ảnh.

3.2.1.2. Mô hình VGG19

Sự đột phá của mô hình AlexNet với việc sử dụng nhiều lớp CNN xếp chồng mang lại hiệu quả cao trong việc phân lớp, trích xuất đặc trưng đã thúc đẩy tạo ra các mô hình sâu hơn nữa. Mô hình VGG16 được giới thiệu ngay sau đó. Sâu hơn nữa chúng ta có VGG19 trong bài báo “Very Deep Convolutional Networks for Large-Scale Image Recognition” [13] được viết bởi Karen Simonyan và Andrew Zisserman với cơ sở là các khối VGG chứa các CNN, Pooling được xếp chồng. Với một kiến trúc mạng sâu như thế này thì mô hình sẽ học được các đặc trưng cấp cao hơn, từ đó các tính năng trích xuất cũng mô tả tốt hơn đặc điểm của ảnh, điều này làm tăng khả năng phân lớp cũng như truy xuất, tìm kiếm hình ảnh. VGG19 giành được top 1 độ chính xác trong cuộc thi ImageNet Large Scale Visual Recognition Challenge (ILSVRC) năm 2014(trước đó là của Alexnet).



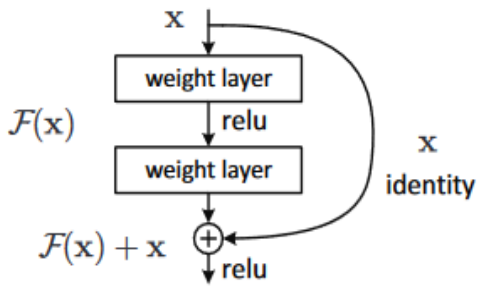
Hình 3.3. Kiến trúc mô hình VGG19[4]

Về kiến trúc thì mô hình VGG19 vẫn giữ các đặc điểm của AlexNet nhưng ngoài sự cải tiến là sâu hơn như đã nói thì mô hình còn dùng các lớp tích chập với kích thước nhỏ hơn là 3x3 (thay vì dùng nhiều kích thước tích chập như AlexNet). Điều này sẽ giảm số lượng tham số cho mô hình sâu, giúp trích xuất đặc trưng nhanh hơn, từ đó ứng dụng tốt trong bài toán truy xuất hình ảnh cần đến sự hiệu quả về thời gian này.

3.2.1.3. Mô hình Resnet

Khi các mạng càng ngày có xu hướng càng sâu, ngoài việc khó khăn về huấn luyện mô hình, độ phức tạp tính toán và đòi hỏi lượng dữ liệu lớn thì nguy cơ over-fitting cũng sẽ lớn hơn (như đã nói ở Alexnet). Giải pháp mà Kaiming He cùng cộng sự đưa ra trong bài báo Deep Residual Learning for Image Recognition [7] là sử dụng một kết nối "tắt" đồng nhất để xuyên qua một hay nhiều lớp. Một khối như vậy gọi là Residual Block, như trong hình sau:

[4] https://www.oreilly.com/library/view/hands-on-image-processing/9781789343731/a4cfa101-430d-4724-b0a5-0b577d0cea21.xhtml



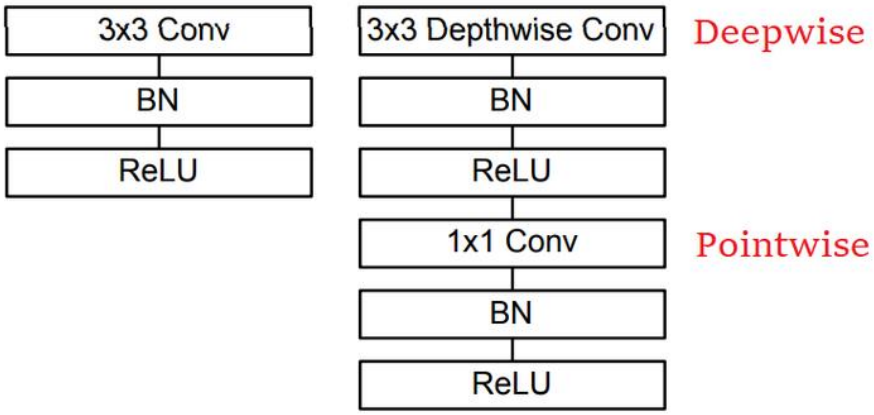
Hình 3.4.Residual block[5]

Resnet gần như tương tự với các mạng CNN thông thường như AlexNet, VGG, … với convolution, pooling, activation và fully connected layer. Khối Residual Block như trong hình 3.4 sẽ được sử dụng trong mạng điều này bổ sung thêm input X từ đầu ra của layer (hay chính là phép cộng). Việc này sẽ tránh cho đạo hàm bằng 0 trong quá trình huấn luyện mô hình bằng thuật toán lan truyền ngược (back propagation). Thực nghiệm cũng cho thấy những kiến trúc này có thể được huấn luyện mạng nơ-ron với độ sâu hàng nghìn lớp. Phù hợp cho bài toán phân lớp, rút trích đặc trưng khi dữ liệu lớn hơn cũng như phức tạp hơn.

3.2.1.4. Mô hình MobileNet

Các mô hình CNN càng ngày càng sâu, càng nhiều tham số, càng phức tạp. Nhờ có sự phát triển của sự hỗ trợ về phần cứng, GPU, TPU đã có những mô hình khổng lồ từ vài trăm triệu tham số (parameter) đến số tỷ. Tuy nhiên, những mô hình này không thể chạy được trên các thiết bị nhỏ gọn như điện thoại di động được. Vì vậy mạng MobileNet ra đời vào năm 2017 trong bài báo "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications" [8] của Andrew G. Howard cùng cộng sự để giải quyết điều này. Để rút gọn lại vài triệu tham số nhưng vẫn giữ được độ chính xác ổn, MobileNet đã sử dụng một cơ chế gọi là Depthwise Separable Convolutions.

[5] https://viblo.asia/p/paper-explain-vovnet-backbone-tiet-kiem-dien-nang-cho-object-detection-Eb85ovo4l2G



Hình 3.5. Trái: Convolution.

Phải: Depthwise separation convolution.[6]

Điều khác biệt của Depthwise separable convolution so với convolution thông thường là Depthwise separable sử dụng 2 kỹ thuật được gọi là deepwise và pointwise. Đầu ra của nó bằng với convolution thông thường tuy nhiên tham số được giảm đi nhiều lần. Ngoài ra MobileNet V2 [13] tăng độ chính xác bằng cách thêm kết nối "tắt" của Residual Block trong mạng ResNet ở trên. Mạng này sẽ thích hợp để ứng dụng hệ thống CBIR trên thiết bị nhỏ gọn, cấu hình và bộ nhớ không được tốt cũng như dữ liệu không đa dạng.

Nhắc tới kiến trúc CNN, không chỉ có những mô hình ở trên, còn GoogleNet, Efficient Net, … nhiều và rất, rất nhiều và sẽ còn có thêm những mạng CNN được thiết kế thêm nữa. Mỗi kiến trúc CNN đều có thế mạnh của riêng nó, được điều chỉnh để phù hợp với các mục đích khác nhau. Vì vậy chúng tôi chỉ đề cập đến những mạng có thế mạnh gần đây trong bài toán CBIR. Danh sách các kiến trúc CNN có tại đây[7].

[6] https://www.google.com/search?client=firefox-b-d&q=mobilenet+v1+paper
[7] https://paperswithcode.com/methods/category/convolutional-neural-networks

3.2.2 Hàm băm

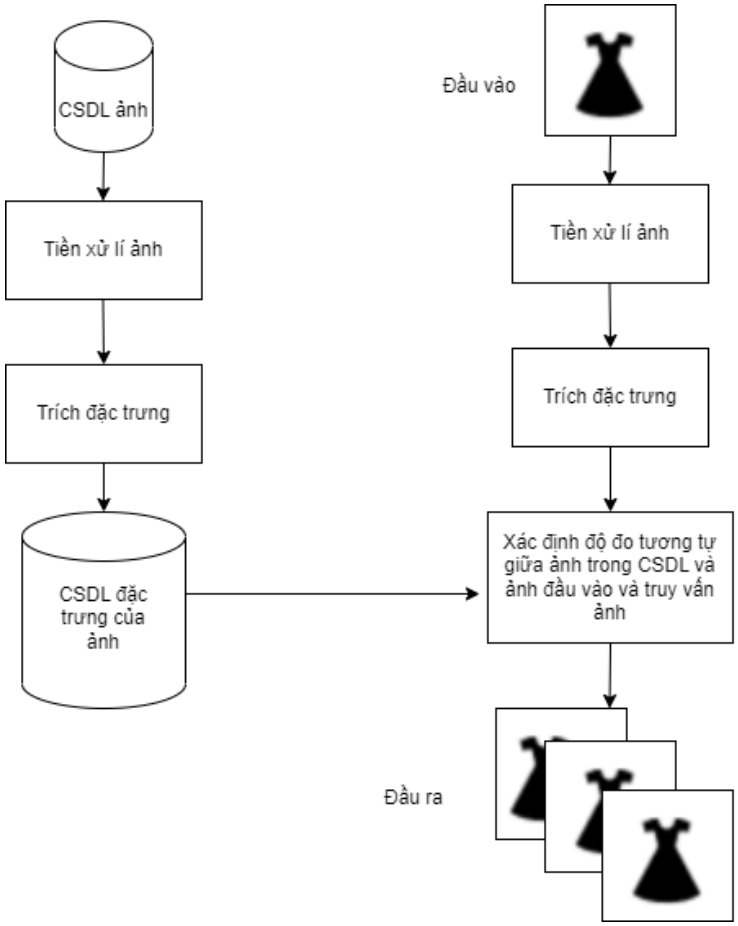
Sau khi thu được vector đặc trưng của hình ảnh, chúng ta có thể trực tiếp sử dụng chúng để đo độ tương đồng cosine, qua đó trả về hình ảnh có độ tương đồng cosine với hình ảnh truy xuất cao nhất trong tập hình ảnh mục tiêu. Tuy nhiên điều này là không khả thi trong bộ dữ liệu lớn. Khi đó để tính độ tương tự cosine của hình ảnh truy vấn với toàn bộ hình ảnh trong tập dữ liệu sẽ mất rất lớn thời gian. Giảm trải nghiệm người dùng. Phương pháp khắc phục được đưa ra ở đây là sử dụng một hash function (hàm băm) để tạo ra mã băm vừa giảm bộ nhớ cần để lưu trữ vừa tăng tốc độ tính toán, truy xuất.

Một cách tổng quát, một hàm băm $h: R^D \rightarrow \{-1,1\}^D$ ánh xạ một vector số thực D chiều thành một vector gồm các mã nhị phân có số chiều tương ứng. Cho một hình ảnh a, đầu tiên chúng tôi trích xuất đặc trưng thông qua phương pháp trên. Vector đặc trưng này sẽ có D chiều. Sau đó D-bit binary code sẽ được đưa vào một hash function $h(.)$. Cho mỗi bit $i = 1, 2, \ldots, D$ đầu ra của binary hashcode sẽ là:

$$\boldsymbol{H_i} = \begin{cases} \mathbf{1} & \boldsymbol{nếu}\ \boldsymbol{x_i} - \dfrac{\mathbf{1}}{\boldsymbol{D}} \displaystyle\sum_{\boldsymbol{i=1}}^{\boldsymbol{D}} \boldsymbol{x_i} \geq \mathbf{0} \\ \mathbf{-1} & \boldsymbol{nếu}\ \boldsymbol{x_i} - \dfrac{\mathbf{1}}{\boldsymbol{D}} \displaystyle\sum_{\boldsymbol{i=1}}^{\boldsymbol{D}} \boldsymbol{x_i} < \mathbf{0} \end{cases} \qquad (\mathbf{3.1})$$

$H_i = \{1, -1\}^q$ là viết tắt của mã nhị phân của mỗi ảnh a, xi là bit tương ứng thứ i của vector đặc trưng ban đầu . Vector mã nhị phân này cũng sẽ được lưu trữ trong cơ sở dữ liệu ứng với mỗi ảnh và sẽ được sử dụng để truy vấn, sẽ được nêu ra trong các tiểu mục sau.

3.2.3. Tìm kiếm bằng hình ảnh



Hình 3.6. Quá trình truy vấn hình ảnh

Quá trình truy vấn ảnh được chia thành 2 giai đoạn:

Giai đoạn 1: Đầu tiên, từ CSDL ảnh thực hiện tiền xử lý ảnh sau đó rút trích các đặc trưng từ mỗi hình ảnh trong tập dữ liệu ban đầu. Sau đó, chúng tôi trích xuất các vector đặc trưng cho từng đối tượng trong các hình ảnh. Cuối cùng, chúng tôi kết hợp các đặc trưng này để tạo thành vector đặc trưng cho mỗi hình ảnh. Vector này sau đó đi qua hàm băm (hash function) để thu được mã băm (hash code). Hash-code sẽ được lưu trữ để phục vụ cho vụ cho việc truy xuất ở bước sau.

Giai đoạn 2: Ảnh truy vấn cũng sẽ đi qua khung đề xuất ở trên và thu được hashcode của nó. Sau đó hashcode này sẽ so sánh độ tương đồng so với toàn bộ hashcode trong toàn bộ dữ liệu đã được lưu trữ ở trên. Việc so sánh độ tương đồng

ở đây sẽ được sử dụng bởi khoảng cách hamming. Khoảng cách hamming ở đây được định nghĩa là số bit khác biệt giữa 2 vector nhị phân.

So sánh độ tương đồng

Cho ví dụ: có hai vector mã nhị phân *bi* và *bj* với *D* chiều, chúng tôi định nghĩa một hàm khoảng cách để tính hamming distance như sau:

$$distH(b_i, b_j) = \frac{1}{2}\left(D - \langle b_i, b_j \rangle\right) \quad (3.2)$$

Ở đây *distH*(*,*) là khoảng cách hamming giữa hai mã nhị phân và <* , *> là inner product giữa hai vector. Hình ảnh đầu vào sau khi đã trích xuất ra vector mã nhị phân sẽ được tính toán khoảng cách hamming với tập vector mã nhị phân trong bộ dữ liệu theo công thức trên. Hệ thống sẽ đề xuất top k hình ảnh có hamming distance thấp nhất so với hình ảnh đầu vào.

Nói tóm lại, tìm kiếm hay truy xuất hình ảnh bằng nội dung là từ hình ảnh người dùng muốn tìm kiếm, hệ thống sẽ đưa ra hình ảnh tương đồng với nó. Một hệ thống CBIR cần đạt được hai kết quả: đặc trưng nhỏ gọn để giảm tài nguyên bộ nhớ và tăng tốc độ tính toán cũng như độ chính xác của truy vấn. Đặc trưng hình ảnh bây giờ đã được lưu trữ dưới dạng một vector mã nhị phân cho nên nó đã có thể gọi là tối ưu được vấn đề thứ nhất. Chúng tôi sẽ sử dụng khoảng cách hamming để tính toán nhanh và chính xác. Công thức sử dụng sẽ được nêu rõ hơn ở mục nhỏ kế tiếp.

3.2.4. Phương pháp đánh giá kết quả

Giống như các bài toán Deep Learning (học sâu) khác, để biết mô hình có tốt hay không và tốt như thế nào thì sẽ cần phải đánh giá hiệu quả của nó qua các độ đo. các phương pháp đánh giá được sử dụng phổ biến cho dạng bài toán CBIR là sử dụng Precision (độ chính giác), recall (độ nhạy), Mean Average Precision (MAP, độ chính xác trung bình). Điều khác biệt để đánh giá bài toán này so với các bài toán deep learning thông thường là phải xác định được hình ảnh trả về có đúng với hình ảnh truy xuất hay không (có tương đồng hay không). Có một cách đơn giản thường được sử dụng nhưng cũng không kém phần hiệu quả là đánh giá dựa trên sự tương

đồng của nhãn của hình ảnh trả về và nhãn của hình ảnh truy xuất. Bây giờ một truy vấn đúng sẽ có hình ảnh trả về có cũng nhãn với hình ảnh truy vấn.

Khi đó Precision sẽ được tính bằng số truy xuất trả về đúng chia cho số truy xuất được trả về. Tương tự cho Recall là số lượng hình ảnh trả về đúng chia cho số hình ảnh có liên quan với hình ảnh truy xuất có trong tập dữ liệu đích. Công thức như sau:

$$Precision = \frac{TP}{TP+FP} \tag{3.3}$$

$$Recall = \frac{TP}{TP+FN} \tag{3.4}$$

Tuy nhiên phương pháp thường được sử dụng để so sánh các mô hình CBIR là MAP. MAP được tính bằng công thức sau:

$$MAP = \frac{AP1 + AP2 + \ldots + APn}{n} \tag{3.5}$$

Nó tính toán trung bình của các giá trị Average Precision (AP) cho từng truy vấn trong tập dữ liệu. AP đo lường chính xác của danh sách kết quả truy vấn bằng cách tính tỉ lệ các hình ảnh truy xuất đúng tại mỗi điểm cắt, vị trí trong danh sách (công thức 3.6).

$$AP = \frac{P_1 + P_2 + \ldots + P_k}{k} \tag{3.6}$$

trong đó k là số lượng truy vấn, P được tính bằng công thức 3.3

Trần Xuân Thủy – 19446021

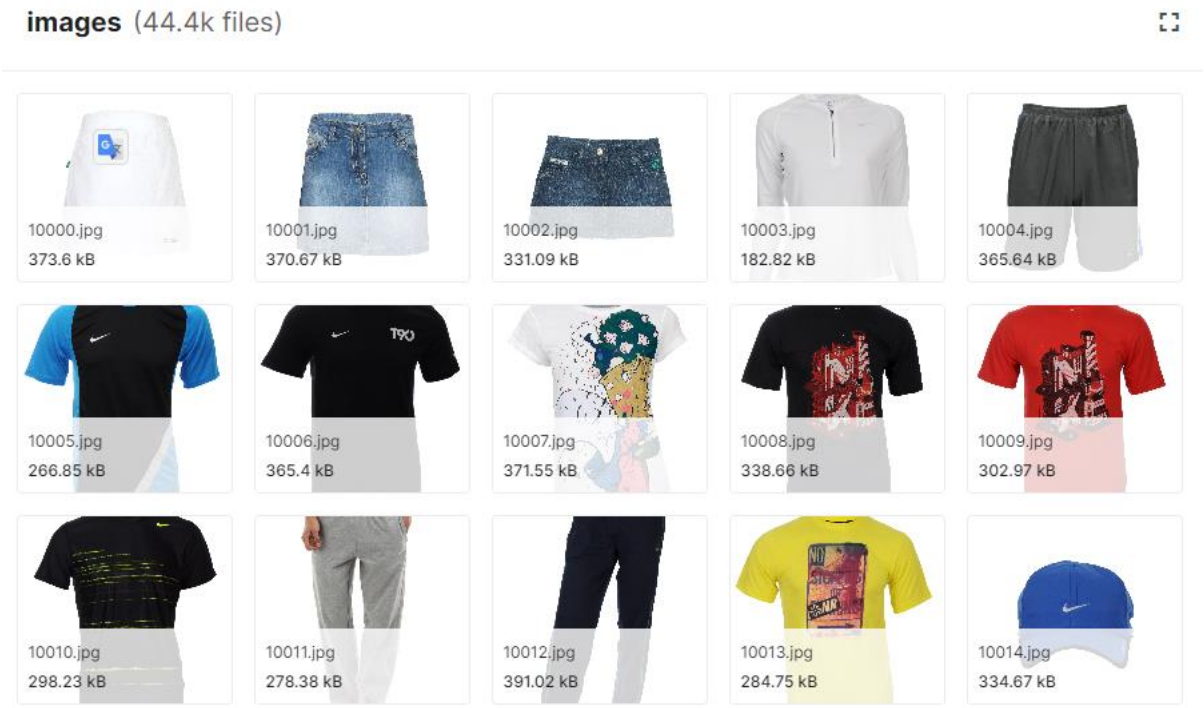
Lê Duy Tâm - 19445631

CHƯƠNG 4
THỰC NGHIỆM

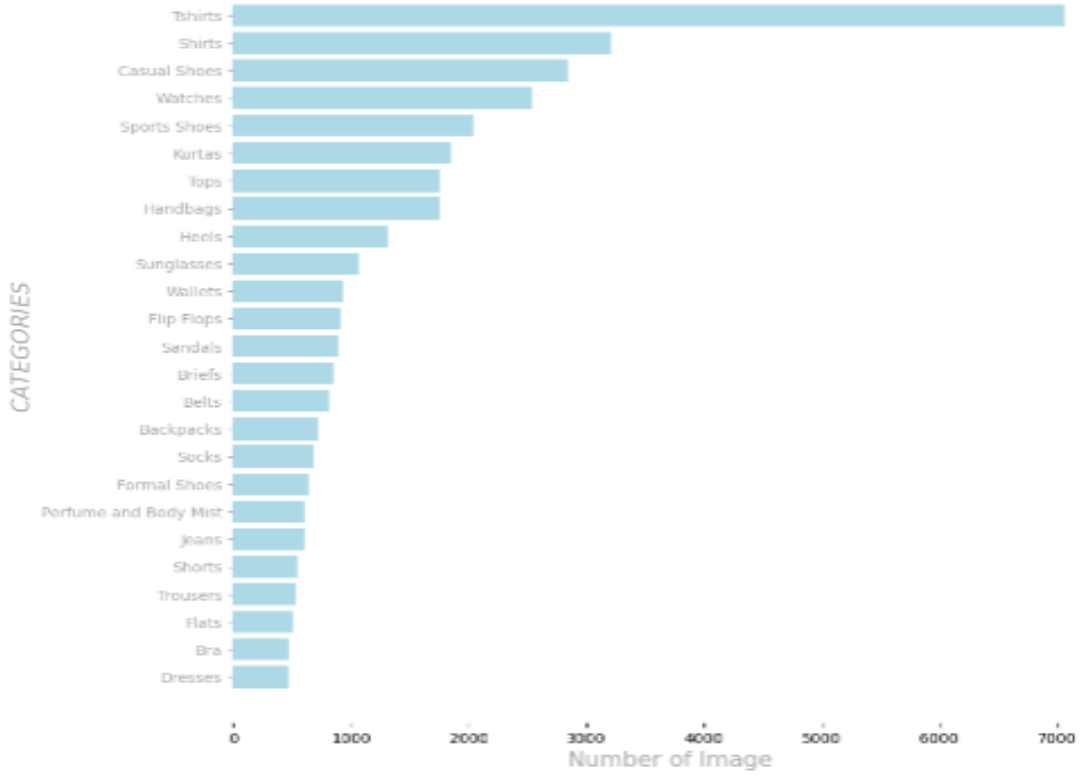
Trong chương này chúng tôi sẽ trình bày về thông tin bộ dữ liệu sử dụng để thực nghiệm mô hình đề xuất: quá trình thu thập, mô tả bộ dữ liệu, các bước xử lý dữ liệu. Trình bày và so sánh kết quả giữa các mô hinh, đưa ra nhận xét đánh giá về các kết quả thu được. Sau cùng giới thiệu về ứng dụng web thể hiện kết quả thu được từ mô hình đề xuất.

4.1. Dữ liệu

Bộ dữ liệu chính nhóm sử dụng là bộ dữ liệu fashion-product-images-dataset được lấy trên Kaggle chứa khoảng 44 nghìn hình ảnh sản phẩm thời trang (hình 4.1) phân phối không đồng đều vào 45 nhãn lớp khác nhau (hình 4.2).

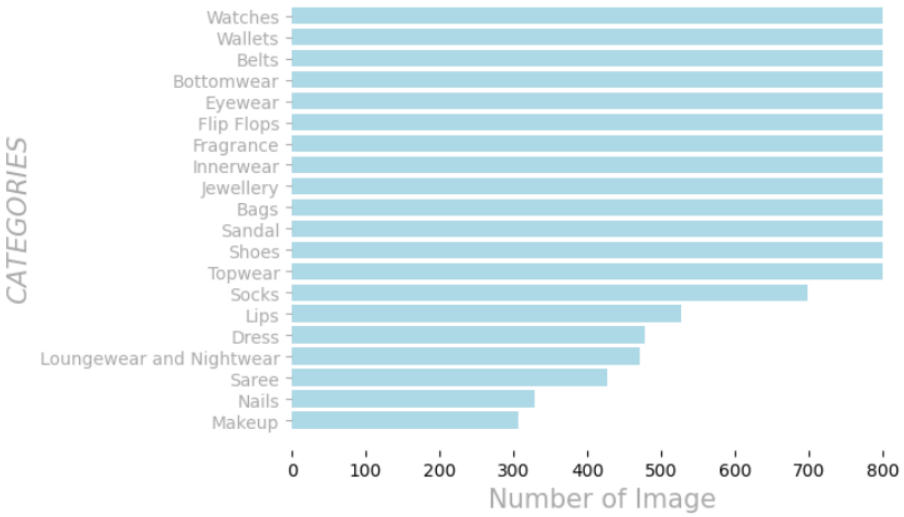


Hình 4.1. Bộ dữ liệu fashion-product-images-dataset



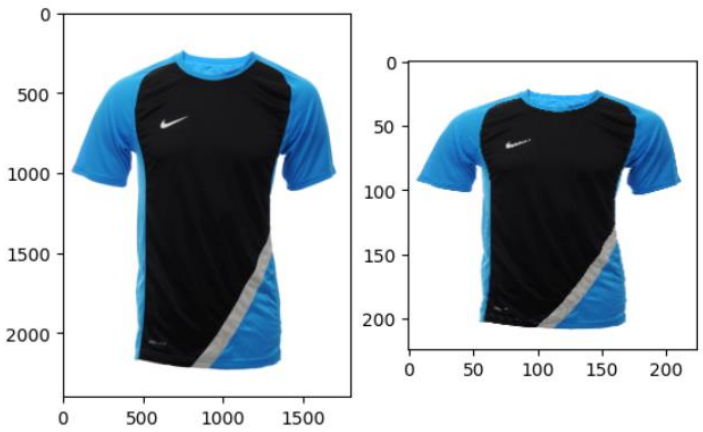
Hình 4.2. Số lượng ảnh giữa các lớp trong bộ dữ liệu gốc.

Chúng tôi sẽ sử dụng 20 nhãn có số lượng mẫu nhiều nhất và lấy tối đa 800 hình cho mỗi nhãn. Phân phối số lượng mẫu trên từng danh mục của bộ dữ liệu như sau:



Hình 4.3. Sự phân bố ban đầu của tập dữ liệu chính

Xử lý dữ liệu

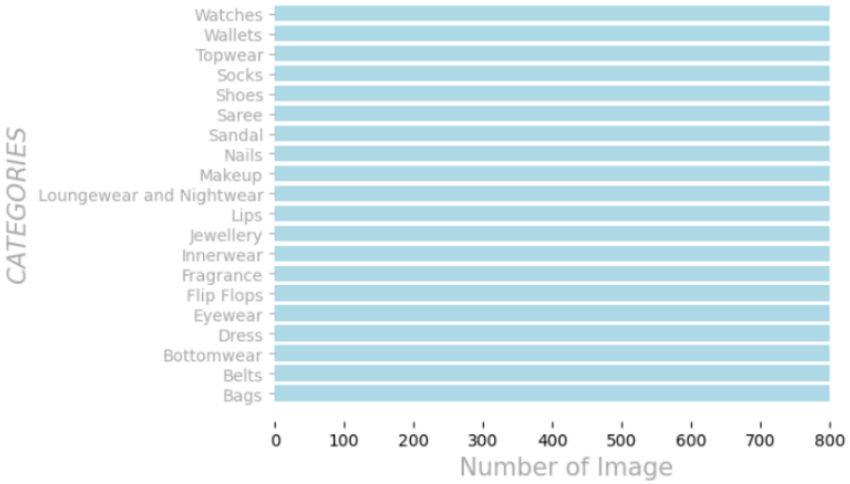


Hình 4.4. Bên phải là hình ảnh ban đầu
Bên trái là hình ảnh sau khi điều chỉnh lại kích thước

Sự đồng đều trong số lượng mẫu trên mỗi nhãn là rất quan trọng trong một bài toán Deep learning nói chung và CBIR nói riêng. Ví dụ như trong bài toán phân lớp nhị phân gồm 2 nhãn là chó và mèo. Nếu tập dữ liệu có sự không đồng đều khi số lượng hình ảnh chó chỉ chiếm 20%, khi đó nếu một dự đoán luôn là mèo thì độ chính xác đã lên đến 80%. Như ta đã biết bài toán CBIR dưa trên bài toán phân lớp, vì vậy để giải quyết vấn đề này, chúng tôi thu thập thêm dữ liệu của các nhãn bị thiếu quá nhiều hình ảnh như makeup, nail, saree, dress, lips, … từ các trang web thương mại điện tử cũng như sử dụng các chuyển đổi lật ngang, xoay và chuyển đổi độ sáng trong Data augmentation để dữ liệu thêm đa dạng và cân bằng hơn. Những hình ảnh này cũng được chuyển thành kích thước 224 x 224. Điều này là cần thiết vì đây là kích thước ảnh đầu vào cho các mô hình CNN trong phần trích xuất đặc trưng.

Hình 4.5. Các phép tăng dữ liệu

Bộ dữ liệu sau cùng sẽ có phân phối như hình 4.3 đảm bảo sự đồng đều cũng như đa dạng của dữ liệu khi tổng số lượng hình lên đến 16000 chia đều cho 20 lớp. Dữ liệu này sau đó sẽ được chia ra thành 3 tập: huấn luyện (train set), kiểm thử (validation set) và đánh giá (test set) với tỉ lệ 6:2:2.



Hình 4.6. Sự phân bố của tập dữ liệu sau khi tăng và cào thêm.

4.2. Kết quả thực nghiệm

Phần này chúng tôi sẽ trình bày về các công nghệ giúp xây dựng và đào tạo mô hình CNN cũng như kết quả thực nghiệm của chúng.

4.2.1 Công nghệ sử dụng

Chúng tôi sử dụng các thư viện:

- Tensorflow backend: xây dựng và huấn luyện mô hình Deep Learning.
- OpenCv: Tải, tiền xử lý ảnh

Ngôn ngữ lập trình được sử dụng trong đồ án này là Python với tính năng mạnh mẽ, dễ sử dụng của nó. Các ngôn ngữ HTML, CSS, Javascript, Bootstrap để xây dựng ứng dụng CBIR. Cơ sở dữ liệu sử dụng là SQLite

4.2.2. Huấn luyện mô hình phân lớp

Chúng tôi sẽ sử dụng các mô hình: AlexNet, VGG19, Resnet50 và MobileNetV2 để thực nghiệm trên bộ dữ liệu nhỏ hơn được lấy mẫu từ bộ dữ liệu đích ở trên. Mỗi nhãn lớp của bộ thử nghiệm sẽ chứa 500 ảnh với 5 nhãn lớp. Dữ liệu sau đó cũng được chia ra ba tập huấn luyện, kiểm thử và đánh giá để lựa chọn mô hình tối ưu. Mỗi mô hình cũng sẽ thử nghiệm với hai thuật toán tối ưu khác nhau đó là Stochastic Gradient Descent (SGD), Adam. Để nhanh chóng và hiệu quả hơn chúng tôi sẽ sử dụng mô hình được huấn luyện trước trên bộ dữ liệu ImageNet với 14 triệu hình ảnh và 1000 lớp, tinh chỉnh ở lớp softmax cuối cùng với số units là 5(ứng với 5 nhãn lớp). Chúng tôi sẽ so sánh độ chính xác và thời gian huấn luyện của các mô hình.

Bảng 4.1. Kết quả thực nghiệm trên thuật toán SGD thông thường

Model	Train acc	Val acc	Test acc
AlexNet	0.53	0.50	0.45
VGG19	0.86	0.83	0.82
ResNet50	0.82	0.81	0.79
MobileNetV2	0.75	0.71	0.73

Bảng 4.2. Kết quả thực nghiệm trên thuật toán Adam

Model	Train acc	Val acc	Test acc
AlexNet	0.56	0.54	0.51
VGG19	0.98	0.96	0.93
ResNet50	0.91	0.81	0.79
MobileNetV2	0.73	0.72	0.68

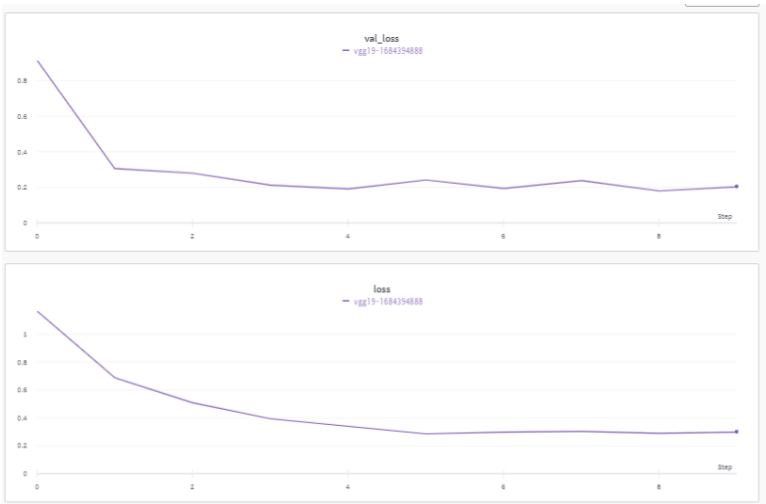
Kết quả thực nghiệm cho thấy độ chính xác trên tập train của mô hình ResNet và VGG19 cao nhất, tuy nhiên độ chính xác trên tập test của mô hình ResNet lại rất thấp. Tuy nhiên cả 2 đang có dấu hiệu của overfitting. Như đã nói đây là vấn đề khi mô hình quá phức tạp so với dữ liệu. Giải thuật Adam cũng cho thấy độ chính xác cao hơn và hội tụ nhanh hơn so với SGD thông thường. Về thời gian huấn luyện, mô hình VGG19 có nhiều tham số hơn so với các mô hình khác cho nên cũng mất thời gian huấn luyện hơn, nhanh nhất là MobileNet bởi vì mô hình này khá đơn giản và phù hợp để chạy trên các thiết bị nhỏ gọn.

Sau các thực nghiệm, chúng tôi đã lựa chọn VGG19 làm mô hình trích xuất đặc trưng vì độ chính xác nó mạng lại và phù hợp với điều kiện tài nguyên có thể đáp ứng. Bộ dữ liệu đích lên tới 20 nhãn lớp, có sự đa dạng trong dữ liệu, vì vậy mô hình sẽ không quá phức tạp đội với bộ dữ liệu này.

Để lựa chọn siêu tham số learning rate, chúng tôi cũng sử dụng một phương pháp gọi là GridSearchCV**.** Đây là một phương pháp tối ưu siêu tham số cho các mô hình học máy. Nó tìm kiếm tất cả các kết hợp có thể của các tham số đã chọn và đánh giá hiệu suất của mô hình với từng kết hợp đó. Điều này giúp tìm ra các tham số tối ưu cho mô hình và cải thiện hiệu suất của mô hình VGG19 của chúng tôi. Kết quả của việc thực nghiệm trên mô hình VGG19 + GridSearchCV của chúng tôi cho thấy các tham số tối ưu là learning rate = 0.01 và batch_size được chọn là 32 để sử dụng được khả năng tính toán vector.

Chúng tôi sẽ sử dụng mô hình VGG19 pretrained model với learning rate là 0.01, batchsize là 32 kết hợp với các kĩ thuật dropout và Batch normalization để

tránh overfitting và giúp mô hình hội tụ nhanh hơn. Kết quả của công việc huấn luyện mô hình được trình bày ở hình 4.4.



Hình 4.7. Train loss và valid loss khi huấn luyện trên bộ dữ liệu đích

4.2.3. Trích xuất đặc trưng

Sau khi huấn luyện được mô hình phân lớp khá tốt, chúng tôi sẽ thực hiện trích xuất đặc trưng như đã đề xuất ở phần trước. Đầu ra của 2 lớp FC áp chót sẽ được nối vào nhau và đi qua hàm hash function. Hashcode thu được sau khi được trích xuất qua framework sẽ được lưu trữ vào database. Bảng dữ liệu chứa hashcode sẽ có id để kết nối với bảng chứa các thông số, mô tả của ảnh.

	id	hashcode
0	56049	[-1, 1, -1, 1, 1, -1, -1, -1, -1, -1, -1, 1, -...
1	41681	[-1, 1, -1, 1, -1, 1, -1, 1, -1, 1, -1, 1, -1,...
2	43513	[1, -1, -1, 1, 1, -1, 1, 1, 1, -1, -1, 1, -1, ...
3	30576	[-1, -1, -1, 1, -1, -1, 1, 1, -1, 1, 1, -1, -1...
4	8444	[-1, -1, -1, 1, 1, -1, 1, 1, 1, -1, -1, 1, -1,...
...	...	...

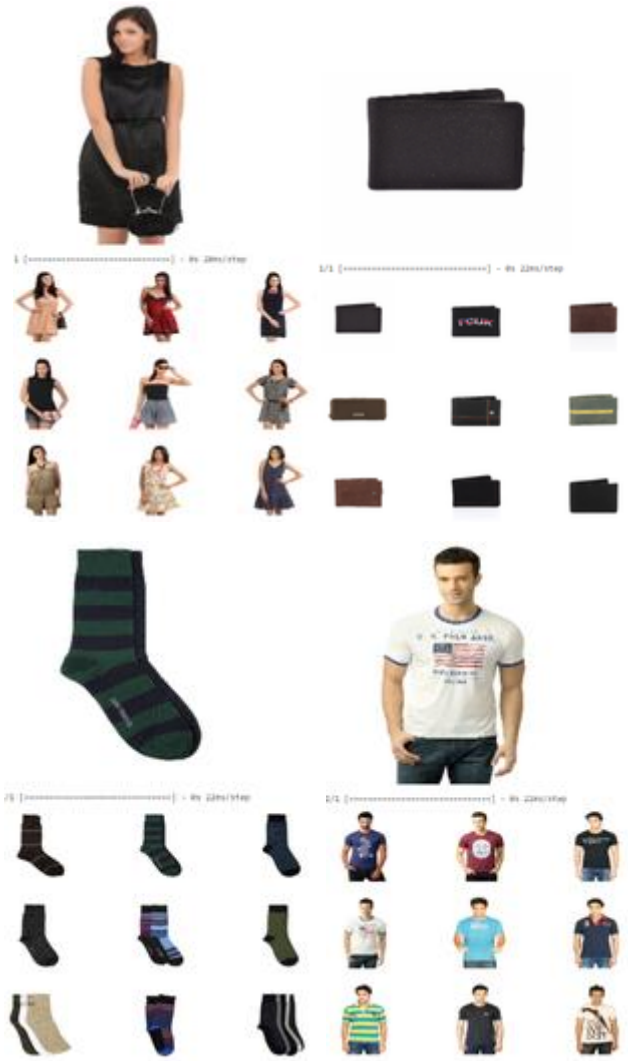
Hình 4.8. Bảng lưu trữ hashcode của tập dữ liệu đích

4.2.4 Kết quả tìm kiếm sử dụng CNN và Hash

Sau khi đã trích xuất được đặc trưng hashcode với độ dài 1024 và sử dụng hamming distance để đo lường sự tương đồng của hình ảnh truy xuất so với CSDL đặc trưng. Như đã đề cập phương pháp đánh giá độ hiệu quả truy xuất, chúng tôi tính toán Mean Average Precision (MAP) trên bộ dữ liệu kiểm tra 20 hình với top 9 hình ảnh kết quả trả về và đã đạt được 0.99 và thời gian truy xuất nhanh chưa đến 1s. Điều này chứng minh mô hình có thể ứng dụng thực tế trong công việc tìm kiếm sản phẩm thời trang. Hình 4.8 thể hiện một số ảnh kết quả thu được sau khi truy xuất ảnh, với 9 kết quả trả về cho mỗi hình. Có thể thấy ở các nhãn lớp có tính đặc thù như giày, đồng hồ,…thì mô hình có khả năng trích xuất với độ chính xác cao hơn các lớp dữ liệu khác.

	Kết quả
MAP1	1.0
MAP2	1.0
MAP20	0.99
MAP200	0.93

Bảng 4.3. Độ chính xác MAP



Hình 4.9. Kết quả khi truy xuất hình ảnh

4.3. Xây dựng ứng dụng

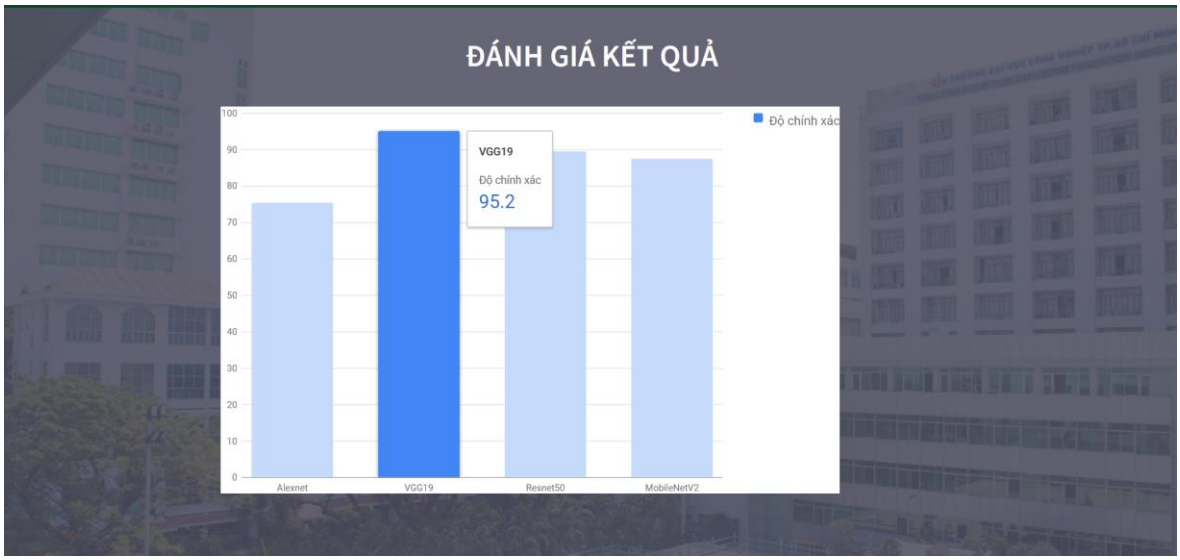
Trang chủ của ứng dụng chúng tôi mang đến một giao diện hấp dẫn và chuyên nghiệp, được chia thành ba tab chính: CBIR, Phân tích dữ liệu và Đánh giá kết quả. Trang chủ, như được thể hiện trong hình 4.10, giới thiệu về ứng dụng và cung cấp các tab để người dùng điều hướng đến các phần chức năng khác

Hình 4.10. Giao diện trang chủ

Ở phần Phân tích dữ liệu (hình 4.11), chúng tôi sử dụng một biểu đồ tròn để thể hiện tỉ lệ dữ liệu huấn luyện, kiểm thử và kiểm tra. Kèm theo đó là một bảng miêu tả nguồn của bộ dữ liệu đích và số lượng của từng mục. Và cuối cùng là một dữ liệu cột nằm ngang miêu tả số lượng mẫu trên mỗi nhãn dữ liệu, có thể thấy dữ liệu sau khi xử lý và tổng hợp đã phân bố rất đồng đều.

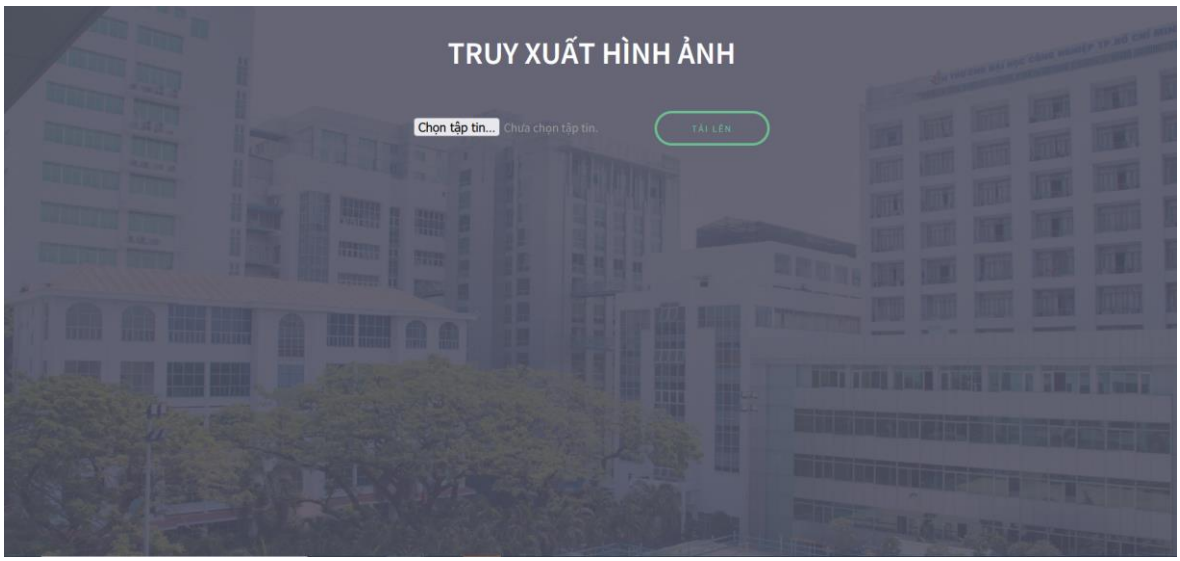
Hình 4.11. Giao diện phân tích dữ liệu

Tiếp theo, giao diện Đánh giá kết quả (hình 4.12) sử dụng biểu đồ cột để đánh giá độ chính xác của từng mô hình huấn luyện. Người dùng có thể dễ dàng nhận thấy và so sánh hiệu suất của các mô hình thông qua biểu đồ này.



Hình 4.12. Giao diện đánh giá kết quả

Ở giao diện CBIR (hình 4.13) cho phép người dùng tải lên một ảnh để thực hiện truy vấn.



Hình 4.13. Giao diện CBIR

Cuối cùng, trong giao diện kết quả (hình 4.14), chúng tôi thể hiện kết quả truy xuất ảnh sau khi người dùng đưa một ảnh vào. Giao diện này hiển thị ảnh truy xuất được, thường đi kèm với các thông tin như tên, mô tả hoặc thuộc tính của ảnh.

Hình 4.14. Giao diện kết quả

Đây là trang web được được thiết kế nhằm mục đích thử nghiệm và đánh giá mô hình. Có thể thấy mô hình hoàn toàn có thể sử dụng trong các trang web thương mại điện tử với khả năng tìm ra hình ảnh yêu cầu của người dùng với tốc độ khá nhanh, tạo trải nghiệm tốt cho người dùng. Các tab CBIR, Phân tích dữ liệu và Đánh giá kết quả cung cấp cái nhìn về dữ liệu, kết quả huấn luyện của mô hình.

CHƯƠNG 5
KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

5.1. Kết luận

Trong đồ án này, chúng tôi xây dựng một hệ thống tìm kiếm hình ảnh bằng nội dung. Chúng tôi đã sử dụng các mô hình mạng tích chập và nơ-ron khác nhau để đánh giá phân loại hình ảnh và sử dụng khoảng cách hamming. Chúng tôi đạt được độ chính xác phân loại là 0.95 trên bộ dữ liệu đích. Và với truy xuất, MAP200 lên tới 0.93 có thể thấy mô hình truy xuất có độ chính xác cao trên tất cả các danh mục, với sự nhanh chóng là 0.5s cho mỗi truy xuất.

5.2. Hướng phát triển

Trong bước tiếp theo, chúng tôi có thể sẽ tăng thêm dữ liệu cũng như sử dụng mô hình phân loại phức tạp hơn để rút trích đặc trưng, việc sử dụng nhiều lớp hơn để rút trích đặc trưng cũng là một ý tưởng mới, nó sẽ sử dụng được cả những đặc trưng thấp và cao trong mô hình deep learning. Thông tin danh mục của chúng tôi hiện tại cũng chưa chi tiết hóa khi sử dụng chỉ 20 danh mục. Các sản phẩm trong cùng danh mục có sự khác nhau. Chúng tôi sẽ cố gắng tìm thông tin danh mục cụ thể hơn.

Bên cạnh đó, chúng ta có thể úng dụng phương pháp tìm kiếm hình ảnh và mã nhị phân vào lĩnh vực y khoa, như phân loại hay tìm kiếm ảnh y tế. Điều này có thể là một bước đột phá lớn hỗ trợ bác sĩ có thể ra một quyết định chuẩn đoán và điều trị bệnh một cách nhanh chóng mà vẫn chính xác.

TÀI LIỆU THAM KHẢO

[1]. Văn Thế Thành (2017), Tìm kiếm ảnh dựa trên đồ thị chữ ký nhị phân.

[2]. Kunihiko Fukushima (1980), Neocognitron: A Self-Organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position.

[3]. LeCun, Bottou, Bengio & Haffner (1998), Object Recognition with Gradient-Based Learning.

[4]. Yangqing Jia, Eric T. Lin, Yulong Li, Jianchao Yang & Jiebo Luo (2014), Learning Deep Features for Image Retrieval.

[5]. Krizhevsky A., Sutskever I., & Hinton G (2012), ImageNet classification with deep convolutional neural networks.

[6]. Karen Simonyan & Andrew Zisserman (2014), Very Deep Convolutional Networks for Large-Scale Image Recognition

[7]. Kaiming He, Xiangyu Zhang, Shaoqing Ren & Jian Sun (2016), Deep Residual Learning for Image Recognition.

[8]. Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto & Hartwig Adam (2017), MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications.

[9]. Chen, Luyang, Fan Yang & Heqing Yang (2017), Image-based Product Recommendation System with Convolutional Neural Networks. Stanford University, 450 Serra Mall, Stanford, CA.

[10]. Youmeng Luo, Wei Li, Xiaoyu Ma and Kaiqiang Zhang (2022), Image Retrieval Algorithm Based on Locality-Sensitive Hash Using Convolutional Neural Network and Attention Mechanism.

[11]. Varga, D., & Szirányi, T. (2016), Fast content-based image retrieval using convolutional neural network and hash function.

[12]. Yiheng Cai, Yuanyuan Li, Changyan Qiu, Jie Ma & Xurong Gao (2019), Medical Image Retrieval Based on Convolutional Neural Network and Supervised Hashing.

[13]. Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, Liang-Chieh Chen (2017), MobileNetV2: Inverted Residuals and Linear Bottlenecks.

[14]. Navneet Dalal and Bill Triggs (2015), Histograms of Oriented Gradients for Human Detection.

[15]. Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, Hartwig Adam (2019), Searching for MobileNetV3.

[16]. Chang Zhou, Lai Man Po (2021), Angular Deep Supervised Hashing for Image Retrieval.

[17]. Domonkos Varga, Tamas Szir (2016), Fast content-based image retrieval using Convolutional Neural Network and hash function.