BỘ GIÁO DỤC
VÀ ĐÀO TẠO

VIỆN HÀN LÂM KHOA HỌC
VÀ CÔNG NGHỆ VIỆT NAM

HỌC VIỆN KHOA HỌC VÀ CÔNG NGHỆ

AN HỒNG SƠN

TRA CỨU ẢNH DỰA VÀO NỘI DUNG VỚI HỌC BIỂU DIỄN VÀ GIẢM CHIỀU DỮ LIỆU

LUẬN ÁN TIẾN SĨ NGÀNH KHOA HỌC MÁY TÍNH

Hà Nội - Năm 2023

BỘ GIÁO DỤC
VÀ ĐÀO TẠO

VIỆN HÀN LÂM KHOA HỌC
VÀ CÔNG NGHỆ VIỆT NAM

HỌC VIỆN KHOA HỌC VÀ CÔNG NGHỆ

AN HỒNG SƠN

TRA CỨU ẢNH DỰA VÀO NỘI DUNG VỚI HỌC BIỂU DIỄN VÀ GIẢM CHIỀU DỮ LIỆU

LUẬN ÁN TIẾN SĨ NGÀNH KHOA HỌC MÁY TÍNH
Mã số: 9 48 01 01

Xác nhận của Học viện
Khoa học và Công nghệ

Người hướng dẫn
(Ký, ghi rõ họ tên)

PGS.TS. Nguyễn Hữu Quỳnh

Hà Nội - Năm 2023

LỜI CAM ĐOAN

Tôi xin cam đoan đề tài nghiên cứu trong luận án này là công trình nghiên cứu của tôi dựa trên những tài liệu, số liệu do chính tôi tự tìm hiểu và nghiên cứu. Chính vì vậy, các kết quả nghiên cứu đảm bảo trung thực và khách quan nhất. Đồng thời, kết quả này chưa từng xuất hiện trong bất cứ một nghiên cứu nào. Các số liệu, kết quả nêu trong luận án là trung thực, nếu sai tôi hoàn toàn chịu trách nhiệm trước pháp luật.

Tác giả luận án

NCS. An Hồng Sơn

LỜI CẢM ƠN

Luận án này được hoàn thiện nhờ vào sự nỗ lực của bản thân cùng với sự hướng dẫn tận tình của Thầy hướng dẫn khoa học, sự giúp đỡ quý báu từ các thầy, cô Viện Công nghệ thông tin, Ban lãnh đạo, phòng Đào tạo, các phòng chức năng của Học viện Khoa học và Công nghệ, Viện Hàn lâm Khoa học và Công nghệ Việt Nam, Ban lãnh đạo Trường Đại học Công nghiệp Việt - Hung, các chuyên gia, nhà khoa học cùng gia đình, bạn bè và đồng nghiệp.

Trước tiên, tôi xin được bày tỏ lòng biết ơn chân thành đến Thầy hướng dẫn khoa học PGS.TS. Nguyễn Hữu Quỳnh đã trực tiếp hướng dẫn, định hướng khoa học, truyền tải những kinh nghiệm nghiên cứu quý giá và tạo mọi điều kiện thuận lợi trong suốt quá trình nghiên cứu và phát triển luận án.

Tôi xin được gửi lời cảm ơn chân thành đến Ban lãnh đạo Viện Công nghệ thông tin, phòng Đào tạo, các phòng chức năng của Học viện Khoa học và Công nghệ, Viện Hàn lâm Khoa học và Công nghệ Việt Nam vì đã tạo mọi điều kiện thuận lợi và giúp đỡ tôi trong quá trình nghiên cứu và hoàn thành luận án của mình.

Tôi xin chân thành cảm ơn Ban lãnh đạo Trường Đại học Công nghiệp Việt - Hung, các thầy cô Khoa Công nghệ thông tin, phòng Quản lý khoa học đã quan tâm giúp đỡ và tạo điều kiện để tôi hoàn thành nhiệm vụ học tập và nghiên cưu của mình. Xin cảm ơn sự động viên, sự quan tâm giúp đỡ và những ý kiến đóng góp quý báu của quý đồng nghiệp.

Cuối cùng, xin bày tỏ lòng biết ơn vô hạn tới mọi thành viên trong gia đình, bạn bè đã thông cảm, khuyến khích động viên và giúp đỡ cho tôi có đủ nghị lực để hoàn thành luận án này.

NCS. An Hồng Sơn

MỤC LỤC

DANH MỤC CÁC KÝ HIỆU, CÁC CHỮ VIẾT TẮT

Ký hiệu	Diễn giải tiếng Anh	Diễn giải tiếng Việt
AIR	Autoencoders for Image Retrieval	Autoencoder cho tra cứu ảnh
ANN	Artificial Neural Network	Mạng nơ ron nhân tạo
AP	Average Precision	Độ chính xác trung bình
CBIR	Content-Based Image Retrieval	Tra cứu ảnh dựa vào nội dung
CNN	Convolutional Neural Network	Mạng nơ ron tích chập
DBN	Deep Belief Network	Mạng niềm tin sâu
DNN	Deep Neural Network,	Mạng nơ ron sâu
GBL	GBL	Gần bỏ lỡ
GT	GT	Gần trúng
HOG	Histogram of Oriented Gradient	Lược đồ gradient có hướng
LBP	Local Binary Pattern	Mẫu nhị phân cục bộ
LDA	Linear Discriminant Analysis	Phân tích phân biệt tuyến tính
LSR	Latent Space Representation	Biểu diễn không gian ẩn
mAP	Mean Average Precision	Độ đo tổng hợp kết quả của nhiều truy vấn
PCA	Principal Component Analysis	Phân tích thành phần chính
RBM	Restricted Boltzmann Machine	Máy boltzmann giới hạn
RF	Relevant Feedback	Phản hồi liên quan
RSLDA	Robust Sparse Linear Discriminant Analysis	Phân tích phân biệt tuyến tính thưa mạnh
SDAIR	Sparse Discriminant Analysis for Image Retrieval	Phân tích phân biệt thưa cho tra cứu ảnh
SGD	Stochastic Gradient Descent	Thuật toán giảm gradient
SIFT	Scale-Invariant Feature Transform	Biến đổi đặc trưng bất biến tỉ lệ
SURF	Speeded-Up Robust Feature	Đặc trưng mạnh và nhanh
SVM	Support Vector Machine	Máy véc tơ hỗ trợ
TBIR	Text-Based Image Retrieval	Tra cứu ảnh dựa vào văn bản

DANH MỤC BẢNG BIỂU

DANH MỤC CÁC HÌNH VẼ, ĐỒ THỊ

MỞ ĐẦU

1. Lý do chọn đề tài

Trong những năm gần đây, với sự xuất hiện của công nghiệp 4.0, các thiết bị di động thông minh và sự phát triển nhanh chóng của mạng xã hội, việc xử lý và lưu trữ ảnh số đã trở nên phổ biến hơn bao giờ hết. Ảnh số đã trở thành một thành phần không thể thiếu trong các lĩnh vực hoạt động của cuộc sống như y học, kiến trúc, thời trang, giáo dục và phòng chống tội phạm. Do đó, việc tra cứu nhanh chóng và chính xác một bức ảnh yêu thích trong một cơ sở dữ liệu (CSDL) ảnh số lớn và đa dạng là một nhiệm vụ hết sức khó khăn, đầy thách thức trong lĩnh vực thị giác máy tính hiện nay.

Trong tra cứu ảnh, có hai phương pháp thường được sử dụng như: Tra cứu ảnh dựa vào văn bản (TBIR - Text-Based Image Retrieval) và Tra cứu ảnh dựa vào nội dung (CBIR - Content-Based Image Retrieval) [1]. Phương pháp TBIR có ưu điểm là đơn giản, nhanh chóng và hiệu quả, tuy nhiên nó cũng có nhược điểm là yêu cầu độ nhân công lớn cho việc chú thích thủ công và độ chính xác của các ảnh được chú thích thủ công có thể bị ảnh hưởng bởi sự chủ quan trong nhận thức của người dùng [1]. Do đó, phương pháp CBIR đã ra đời và được giới thiệu vào đầu những năm 1990 để khắc phục những hạn chế này.

Trong lĩnh vực thị giác máy tính, CBIR đang là một trong những hướng được nghiên cứu rất tích cực hiện nay. Mục tiêu của CBIR là tìm kiếm các ảnh dựa trên việc phân tích các nội dung trực quan của chúng. Vì vậy, biểu diễn ảnh là mấu chốt quan trọng của CBIR [2].

CBIR là phương pháp tìm kiếm ảnh trong CSDL dựa trên nội dung trực quan của ảnh truy vấn [3]. Tuy nhiên, phương pháp này gặp phải vấn đề "khoảng trống ngữ nghĩa" giữa các đặc trưng mức thấp mô tả ảnh và các khái niệm mức cao được con người nhận biết [4], do đó có thể dẫn đến các ảnh không liên quan được trả về. Để khắc phục điều này, nhiều phương pháp đã được đề xuất để chuyển đổi các khái niệm mức cao trong ảnh sang các đặc trưng mức thấp. Các đặc trưng này được phân loại thành các đặc trưng toàn cục (bao gồm màu sắc, hình dạng, kết cấu và thông tin không gian) và các đặc trưng cục bộ tùy thuộc vào phương pháp trích rút đặc trưng [4]. Biểu diễn của các đặc trưng này là nền tảng cho CBIR. Chúng có ưu điểm là nhanh hơn trong việc tính toán độ tương tự và trích rút đặc trưng [5]. Mặt khác, chúng không phân biệt được giữa nền và đối tượng trong ảnh (các phần ảnh khác nhau). Điều này

làm cho chúng không phù hợp để tra cứu trong các cảnh phức tạp hoặc nhận dạng đối tượng [6], nhưng chúng phù hợp để phân loại và phát hiện đối tượng [7]. Khi so sánh đặc trưng cục bộ với đặc trưng toàn cục, thì đặc trưng cục bộ thích hợp cho việc tra cứu, đối sánh và nhận dạng [6]. Nhận dạng đối tượng là nhiệm vụ nhận dạng và gắn nhãn đối tượng trong một hình ảnh [8] trong khi phát hiện đối tượng liên quan đến sự tồn tại của một đối tượng thuộc một lớp được xác định trước trong ảnh và vị trí của nó [9]. Do đó, phân lớp là một nhiệm vụ con của phát hiện đối tượng [9]. Các đặc trưng cục bộ được định nghĩa là các điểm chính hoặc một số phần của ảnh, chẳng hạn như góc, đốm màu và cạnh. Chúng mạnh với tỉ lệ, xoay, dịch chuyển, các thay đổi nền, các che lấp [6].

Đặc trưng được trích rút là quá trình đầu tiên trong CBIR, nhằm chuyển nhận thức của người vào một mô tả số mà máy có thể thao tác được. Độ chính xác của các ảnh được tra cứu "bị ảnh hưởng rất nhiều bởi các đặc trưng được trích rút" [10]. Tuy nhiên, việc lựa chọn này dựa trên yêu cầu của người dùng. Việc cung cấp các đặc trưng được trích rút cho các thuật toán học máy (có giám sát hoặc không giám sát) có thể cải thiện được hiệu năng đối với phương pháp CBIR [11].

Số các đặc trưng mà biểu diễn các mẫu dữ liệu được xem như chiều của dữ liệu. Đặc trưng trong tra cứu ảnh có thể thuộc một trong ba loại sau: (1) đặc trưng liên quan, (2) đặc trưng không liên quan, và (3) đặc trưng dư thừa. Đặc trưng liên quan là những đặc trưng quan trọng để cải thiện độ chính xác của mô hình phân lớp và nâng cao hiệu suất của tra cứu ảnh. Các đặc trưng không liên quan không đóng góp vào quá trình cải thiện chất lượng phân lớp và do đó không cải thiện được hiệu năng của tra cứu ảnh. Các đặc trưng dư thừa là các đặc trưng có thể là liên quan, nhưng chúng không đóng góp vào việc cải tiến chất lượng mô hình, trái lại, các đặc trưng này có thể dẫn đến quá trình học không hiệu quả, tốn nhiều thời gian.

Các cách tiếp cận CBIR truyền thống thường chọn các hàm khoảng cách cứng trên một số đặc trưng mức thấp được trích rút, như Euclide hoặc độ tương tự cosine. Tuy nhiên, các hàm khoảng cách cứng có thể không luôn tối ưu đối với các nhiệm vụ tra cứu ảnh dựa vào nội dung phức tạp. Nguyên nhân của sự không tối ưu này là do khoảng trống giữa các đặc trưng trực quan mức thấp được trích rút bởi máy tính và các khái niệm mức cao được nhận thức bởi con người. Do đó, trong những năm gần đây, đã có rất nhiều nỗ lực nghiên cứu để thiết kế các độ đo khoảng cách trên các đặc trưng mức thấp thông qua khai thác các kỹ thuật học máy.

Học máy là một công cụ quan trọng để khai thác các cấu trúc dữ liệu, thu được biểu diễn dữ liệu tốt hơn và khám phá các mẫu dữ liệu ẩn để có thể trích rút được các thông tin liên quan. Trong học máy, có ba cách tiếp cận chính, bao gồm: học có giám sát, học không giám sát và học bán giám sát. Sự khác nhau của các cách tiếp cận này là ở chỗ sử dụng các mẫu có nhãn trong quá trình học. Trong học có giám sát, các nhãn dữ liệu được sử dụng để học. Tuy nhiên, điều này yêu cầu tất cả các mẫu dữ liệu đều phải có nhãn. Trong học không giám sát, các nhãn dữ liệu không được yêu cầu trong quá trình học. Thông tin nhãn không cần thiết cho tất cả các mẫu dữ liệu. Học bán giám sát là cách tiếp cận kết hợp giữa học có giám sát và học không giám sát. Nó sử dụng tất cả các mẫu huấn luyện có nhãn và không có nhãn để tạo ra cấu trúc hình học nội tại của toàn bộ dữ liệu huấn luyện.

Chiều của dữ liệu ảnh trong các ứng dụng thực tế thường rất cao. Dữ liệu chứa một số lượng lớn các đặc trưng hoặc là dư thừa hoặc là không liên quan. Vì vậy, nếu loại đi các đặc trưng này sẽ giúp giảm thời gian và tăng độ chính xác của các nhiệm vụ học và phân lớp. Trong các bài toán học phân lớp trên dữ liệu nhiều chiều, giảm chiều được xem là một trong những kỹ thuật hiệu quả nhất, nó được đề xuất để giải quyết vấn đề thuộc về "Vấn đề của chiều - Curse of dimensionality". Gần đây, nhiều mô hình học phân lớp đã được đề xuất như học đa thể hiện (Multiple-instance learning) và học không gian con (Subspace learning). Các phương pháp học không gian chiếu nổi tiếng nhất bao gồm phân tích thành phần chính (PCA - Principal Component Analysis) và phân tích phân biệt tuyến tính (LDA - Linear Discriminant Analysis).

Trong những năm gần đây, ở Việt Nam đã có nhiều Nghiên cứu sinh, Nhóm nghiên cứu tiếp cận và khai thác hiệu quả các kỹ thuật học máy cho bài toán CBIR với phản hồi liên quan (RF), giúp thu hẹp "khoảng trống ngữ nghĩa" và cải thiện độ chính xác tra cứu của hệ thống tra cứu ảnh. Tuy nhiên, các công trình này chưa tập trung giải quyết vấn đề cỡ lớp nhỏ. Ở đây, khái niệm ***cỡ lớp nhỏ*** được hiểu là lớp âm và lớp dương trong cơ chế RF (nó không phải là số chủ đề của tập ảnh). Bên cạnh đó, các công trình này vẫn chưa khai thác được thuộc tính thưa dòng của ma trận chiếu. Ở đây, khái niệm ***ma trận chiếu*** ma trận giúp biến đổi dữ liệu từ không gian gốc sang không gian chiếu (trong luận án này, ma trận chiếu thu được còn giúp xác định được đặc trưng gốc nào là quan trọng nhất). Khái niệm ***thuộc tính thưa dòng của ma trận chiếu*** được hiểu là dòng của ma trận chiếu mà giá trị của các

phần tử đều bằng không. Thuộc tính này sẽ giúp phương pháp xác định đặc trưng nào của dữ liệu gốc là dư thừa hoặc không liên quan. Ngoài ra, tính ưu việt của các kỹ thuật học sâu cho tra cứu ảnh trên tập dữ liệu cỡ lớn, không có nhãn và dữ liệu cao chiều vẫn chưa được khai thác. Đây là một định hướng nghiên cứu phù hợp với xu thế nghiên cứu chung của thế giới, mang tính cấp thiết cao và có khả năng ứng dụng hiệu quả trong thực tiễn và đây cũng chính là các hướng nghiên cứu mà nhiều Nhóm nghiên cứu và Nghiên cứu sinh đang theo đuổi.

Học sâu là một kỹ thuật đột phá, mà bao gồm một họ các thuật toán học máy để mô hình các khái niệm mức cao trong dữ liệu. Kỹ thuật học sâu này sử dụng các kiến trúc sâu bao gồm nhiều phép biến đổi phi tuyến. Học sâu mô phỏng bộ não người được tổ chức theo kiến trúc sâu và xử lý thông tin qua nhiều giai đoạn biến đổi và biểu diễn. Nó không giống như các phương pháp học máy truyền thống mà thường sử dụng kiến trúc nông. Bằng việc khai thác các kiến trúc sâu để học tự động các đặc trưng ở nhiều mức trừu tượng từ dữ liệu, các phương pháp học sâu cho phép hệ thống học các hàm phức tạp mà ánh xạ dữ liệu đầu vào sang đầu ra.

Từ sự thành công của các kỹ thuật học máy và học sâu, cùng những hướng tiếp cận khả thi của các Nhóm nghiên cứu ở Việt Nam trong những năm gần đây, đã thúc đẩy Nghiên cứu sinh khám phá các kỹ thuật học máy và học sâu vào bài toán CBIR để cải tiến độ chính xác và tốc độ tra cứu của hệ thống. Đây cũng chính là lý do mà Nghiên cứu sinh đã chọn đề tài **"Tra cứu ảnh dựa vào nội dung với học biểu diễn và giảm chiều dữ liệu"** để góp phần khám phá và giải quyết các vấn đề đã đặt ra ở trên.

2. Mục tiêu nghiên cứu của luận án

Mục tiêu chung:

Nghiên cứu, đề xuất một số phương pháp cải tiến độ chính xác và thời gian tra cứu đối với hệ thống tra cứu ảnh dựa vào nội dung với RF.

Mục tiêu cụ thể:

Đề xuất được một số cải tiến đối với hệ thống CBIR với RF, bao gồm:

- Kết hợp mô hình trích rút đặc trưng với mô hình phân lớp trong hệ thống CBIR, sử dụng thuộc tính thưa dòng của ma trận chiếu để cải tiến độ chính xác tra cứu và thời gian truy vấn khi cỡ mẫu và cỡ lớp nhỏ.

- Huấn luyện bán giám sát bằng mạng nơ ron tích chập autoencoder, trích rút đặc trưng ảnh và phân lớp SVM trong RF, giúp tăng cường khả năng học các đặc

trưng phân biệt dùng cho tra cứu ảnh.

3. Đối tượng và phạm vi nghiên cứu của luận án

Đối tượng nghiên cứu:

Luận án tiến hành tìm hiểu và nghiên cứu một số đối tượng liên quan đến bài toán tra cứu ảnh được đề xuất như:

- CBIR và các thành phần của một hệ thống CBIR; khoảng trống ngữ nghĩa trong CBIR và các kỹ thuật giảm khoảng trống ngữ nghĩa trong CBIR;
- Kỹ thuật học máy, học sâu và mạng Autoencoder;
- Một số độ đo tương tự giữa các ảnh và phương pháp đánh giá hiệu năng;
- Một số phương pháp phân tích phân biệt tuyến tính (LDA, RSLDA);
- Mạng phần dư (ResNet) và Shortcut Connections.

Phạm vi nghiên cứu:

Luận án tập trung nghiên cứu trên phạm vị một số nội dung chính sau:

- Học ma trận chiếu với việc khai thác thuộc tính thưa dòng của ma trận chiếu để giải quyết vấn đề cỡ lớp nhỏ.
- Học biểu diễn ảnh hiệu quả thông qua mạng nơ ron sâu trên tập dữ liệu không có nhãn.
- Phương pháp tra cứu ảnh tận dụng các mẫu huấn luyện thông qua cơ chế RF của người dùng.

4. Phương pháp nghiên cứu của luận án

Nghiên cứu lý thuyết:

Nghiên cứu các cơ sở lý thuyết liên quan đến CBIR, kỹ thuật RF và vấn đề giảm khoảng trống ngữ nghĩa thông qua tiếp cận kỹ thuật học máy, học sâu, các độ đo tương tự cho tra cứu ảnh.

Khảo sát, phân tích ưu điểm, nhược điểm và những vấn đề tồn tại của một số công trình nghiên cứu liên quan về CBIR theo cách tiếp cận sử dụng học mày vào quá trình tra cứu ảnh với RF ở trong nước và trên thế giới, từ đó đề xuất một số vấn đề cần nghiên cứu và giải quyết, làm tiền đề thực hiện đối với các chương nội dung của luận án.

Các tư liệu và thông tin liên quan sử dụng trong luận án được thu thập, tổng hợp và sưu tầm từ các nguồn như: (1) công trình khoa học trên các tạp chí khoa học có uy tín trong và ngoài nước, Internet,..; (2) cùng Thầy hướng dẫn khoa học và các đồng nghiệp nghiên cứu, trao đổi và thực nghiệm; (3) seminar khoa học hoặc báo cáo

tại các hội thảo khoa học giúp nâng cao kỹ năng cách trình bày và kiểm chứng, đánh giá các kết quả đã nghiên cứu của luận án.

Nghiên cứu thực nghiệm:

Đề xuất môi trường thực nghiệm (gồm nền tảng, ngôn ngữ lập trình và cấu hình máy tính), tập CSDL ảnh thực nghiệm (đã được sử dụng nhiều, chuyên nghiệp) và phương pháp đánh giá hiệu năng phù hợp cho bài toán CBIR với RF đã được xác định.

Cài đặt, chạy thử nghiệm và tiến hành đánh giá, so sánh kết quả giữa phương pháp đề xuất của luận án với các phương pháp tiêu biểu khác, nhằm chứng minh hiệu năng của phương pháp và mô hình đã đề xuất.

5. Những đóng góp mới của luận án

Các đóng góp mới của luận án là đề xuất được hai phương pháp CBIR sử dụng RF, gồm: phương pháp **SDAIR** (Sparse Discriminant Analysis for Image Retrieval) [CT4, CT2] và phương pháp **AIR** (Autoencoders for Image Retrieval) [CT1, CT3].

- Phương pháp SDAIR kết hợp mô hình trích rút đặc trưng quan trọng dựa trên phương pháp RSLDA với mô hình phân lớp trong hệ thống CBIR nhằm cải tiến độ chính xác và thời gian truy vấn. Phương pháp này giải quyết được ba vấn đề: *Thứ nhất,* số lượng phản hồi mà người dùng cung cấp nhỏ hơn so với chiều của không gian đặc trưng. *Thứ hai,* số lượng mẫu phản hồi dương thường thấp hơn rất nhiều so với số lượng mẫu phản hồi âm. *Thứ ba*, số lớp quá nhỏ, mà có nghĩa rằng số các hướng chiếu bị giới hạn bởi số các lớp.

- Phương pháp AIR dựa trên ba thành phần: Huấn luyện bán giám sát bằng mạng nơ ron tích chập autoencoder, trích rút đặc trưng ảnh và phân lớp SVM trong RF nhằm cải tiến độ chính xác và thời gian truy vấn. Phương pháp này giải quyết được hai hạn chế: *Thứ nhất*, khả năng phân biệt kém của các phương pháp đã có. *Thứ hai*, giảm nhẹ vấn đề vanishing/exploding gradients và quá trình hội tụ nhanh.

6. Bố cục của luận án

Luận án này được trình bày với bố cục bao gồm phần mở đầu, 3 chương nội dung, phần kết luận, danh mục công trình của tác giả và tài liệu tham khảo, cụ thể như sau:

Phần mở đầu, trình bày về ý nghĩa khoa học và tính cấp thiết của đề tài, cũng như giải thích lý do chọn đề tài. Sau đó, trình bày về nội dung, đối tượng, phạm vi, phương pháp và mục tiêu nghiên cứu của luận án.

Chương 1, giới thiệu tổng quan về Tra cứu ảnh. Chương này trình bày khái niệm và sơ đồ của một hệ thống CBIR; các đặc trưng mức thấp và cách thức lựa chọn, trích rút các đặc trưng hữu ích; cơ chế RF và vấn đề giảm khoảng trống ngữ nghĩa thông qua tiếp cận học máy. Bên cạnh đó, chương này sẽ trình bày một số độ đo khoảng cách cho tra cứu ảnh. Ngoài ra, tình hình nghiên cứu liên quan đến các giai đoạn trong tra cứu ảnh cũng được phân tích để từ đó làm động cơ nghiên cứu cho luận án.

Chương 2, trình bày "Phương pháp tra cứu ảnh với phân tích phân biệt thưa". Chương này tập trung vào việc cải tiến hiệu suất cho bài toán tra cứu ảnh với RF bằng cách sử dụng thuộc tính thưa dòng của ma trận chiếu phân biệt, gồm bốn phần chính: phần đầu tiên giới thiệu về giảm chiều dữ liệu và bài toán CBIR với RF, các nghiên cứu gần đây và những thách thức hiện tại cho bài toán. Phần thứ hai, trình bày phương pháp tra cứu ảnh được đề xuất với 2 thuật toán: (1) Chọn tập đặc trưng quan trọng và (2) Xây dựng mô hình phân lớp. Phần thứ ba, mô tả chi tiết thuật toán được đề xuất SDAIR. Phần thứ tư, đánh giá độ chính xác và thời gian truy vấn của phương pháp đề xuất trên tập ảnh CIFAR-100.

Chương 3, trình bày phương pháp tra cứu ảnh dựa trên mạng nơ ron tích chập sâu autoencoder. Phương pháp được đề xuất cho phép tự động học véc tơ đặc trưng trực tiếp từ ảnh thô theo cách không giám sát và có giám sát để nâng cao hiệu năng tra cứu. Nội dung chương này có 3 phần: Phần thứ nhất, giới thiệu các nghiên cứu có liên quan và đặt vấn đề cho bài toán. Phần thứ hai, trình bày phương pháp đề xuất với hai nội dung: (1) Học các biểu diễn ảnh với mạng nơ ron tích chập sâu autoencoder và (2) Tra cứu ảnh với RF dựa vào máy véc tơ hỗ trợ. Phần thứ ba, đánh giá hiệu năng của phương pháp đề xuất thông qua thực nghiệm so sánh phương pháp đề xuất với 3 phương pháp khác ở ba lần lặp phản hồi đầu tiên.

Kết luận và kiến nghị, luận án tổng hợp kết quả đạt được và đưa ra một số kết luận, đồng thời trình bày một số định hướng nghiên cứu của luận án trong tương lai.

Danh mục công trình của tác giả, luận án liệt kê 04 công trình là các bài báo của tác giả được đăng trên các tạp chí, kỷ yếu hội thảo trong nước và quốc tế.

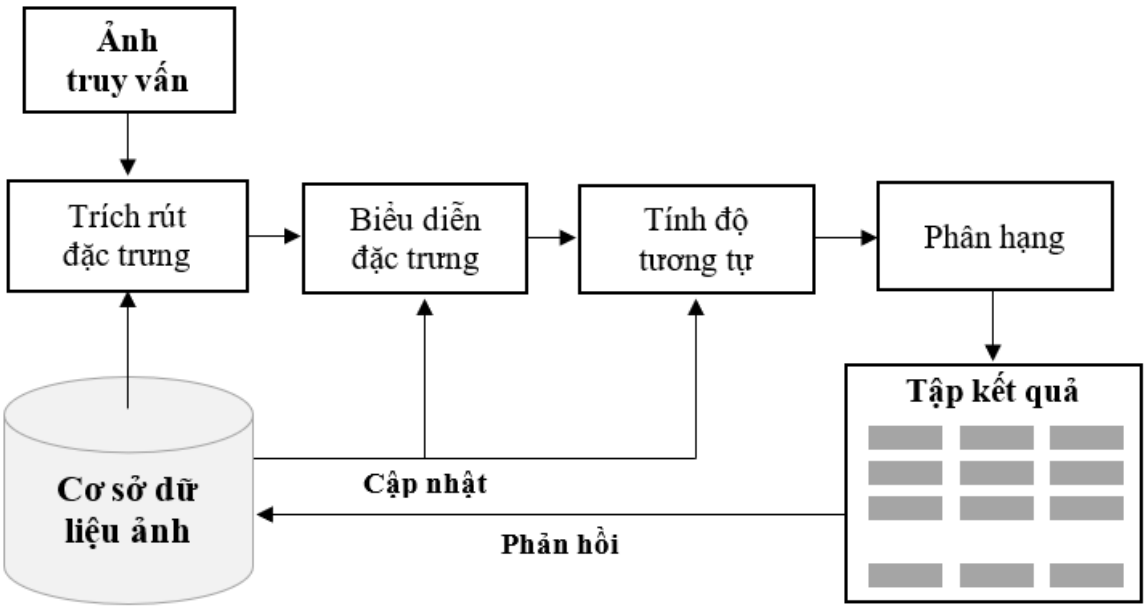
Cuối cùng, là danh mục các tài liệu tham khảo đươc sử dụng trong luận án.

Chương 1. TỔNG QUAN VỀ TRA CỨU ẢNH DỰA VÀO NỘI DUNG VỚI PHẢN HỒI LIÊN QUAN

Chương này trình bày các kiến thức lý thuyết cơ bản liên quan đến CBIR với phản hồi liên quan (RF), được sử dụng làm cơ sở lý luận trong luận án. Các kiến thức lý thuyết cơ bản về CBIR được mô tả bao gồm các đặc trưng mức thấp và cách thức lựa chọn, trích rút các đặc trưng hữu ích; cơ chế RF quan và vấn đề giảm khoảng trống ngữ nghĩa thông qua tiếp cận kỹ thuật học máy, học sâu; các độ đo tương tự cho tra cứu ảnh. Ngoài ra môi trường, tập dữ liệu ảnh thực nghiệm và phương pháp đánh giá hiệu năng cũng được trình bày trong chương này. Bên cạnh đó, một số công trình nghiên cứu liên quan về CBIR và các giai đoạn trong CBIR theo cách tiếp cận sử dụng học mày vào quá trình tra cứu với RF ở trong nước và trên thế giới được khảo sát và phân tích. Dựa trên những ưu điểm, hạn chế đối với các phương pháp được đề xuất trong các công trình nghiên cứu này để định hướng một số vấn đề cần giải quyết, làm tiền đề thực hiện đối với các chương tiếp theo của luận án.

1.1. Tra cứu ảnh dựa vào nội dung

Tra cứu ảnh dựa vào nội dung (CBIR) là một lĩnh vực nghiên cứu của thị giác máy tính [12]. Mục tiêu của CBIR là tìm kiếm các ảnh trong một CSDL ảnh lớn dựa trên các đặc trưng trực quan của chúng, bao gồm hình dạng, kết cấu, màu và các thông tin khác có thể trích rút được từ bản thân ảnh. Khung làm việc của CBIR được mô tả như trong Hình 1.1 dưới đây.



Hình 1.1. Sơ đồ hệ thống CBIR

Bước đầu tiên trong sơ đồ hệ thống CBIR là đưa ảnh truy vấn vào hệ thống bởi người dùng. Bước tiếp theo là trích rút đặc trưng, đây là bước quan trọng nhất, mà một khái niệm trực quan được chuyển sang dạng số. Các đặc trưng được trích rút trong tra cứu ảnh có thể là các đặc trưng mức thấp, chẳng hạn như màu, kết cấu, hình dạng và thông tin không gian hoặc các mô tả cục bộ của ảnh. Quá trình trích rút đặc trưng ảnh truy vấn cũng được thực hiện tương tự như trong trường hợp ảnh CSDL. Bước tiếp theo là tính độ tương tự giữa các đặc trưng được trích rút từ ảnh truy vấn và tất cả các ảnh trong CSDL để phục vụ phân hạng các ảnh. Bước cuối cùng là phân hạng các ảnh theo thứ tự về độ tương tự với ảnh truy vấn để được tập kết quả. RF là một bước mà được sử dụng để tăng cường các kết quả thông qua tương tác của người dùng bằng việc quyết định các ảnh được trả về là liên quan hay không liên quan. Nhiều kỹ thuật RF đã được đề xuất để áp dụng RF vào việc tăng cường hiệu năng của hệ thống CBIR [13].

1.2. Các đặc trưng mức thấp

Trong tra cứu ảnh, vấn đề chính là cách đo hiệu quả độ tương tự giữa các ảnh. Bởi vì các cảnh hoặc các đối tượng trực quan có thể có nhiều thay đổi hoặc biến đổi, nên việc so sánh trực tiếp các ảnh ở mức pixel (điểm ảnh) là không khả thi. Thông thường, các đặc trưng trực quan được trích rút từ các ảnh và sau đó được biến đổi thành một véc tơ có cỡ cố định cho biểu diễn ảnh.

Các đặc trưng có thể được chia thành các đặc trưng toàn cục và các đặc trưng cục bộ. Các đặc trưng toàn cục, bao gồm màu sắc, hình dạng, kết cấu, và thông tin không gian, mà mô tả toàn bộ ảnh. Trong khi đó, các đặc trưng cục bộ thường thu được thông qua việc chia các ảnh thành các đoạn hoặc thông qua việc tính một số điểm chính nào đó như các góc, các đóm màu và các cạnh. Các đặc trưng cục bộ là bất biến với tỉ lệ, xoay và dịch chuyển [14]. Hai loại đặc trưng này sẽ được mô tả ở phần dưới.

1.2.1. Các đặc trưng toàn cục

Các đặc trưng như màu, kết cấu, hình dạng và thông tin không gian được sử dụng rộng rãi trong các nhiệm vụ tra cứu ảnh.

1.2.1.1. Đặc trưng màu

Trong tra cứu ảnh, một trong những đặc trưng quan trọng nhất là màu sắc. Các đặc trưng màu được sử dụng để phân tích và nhận diện các đối tượng trong ảnh, và

được tính toán dựa trên các không gian màu khác nhau. Không gian màu được sử dụng phổ biến trong CBIR bao gồm RGB, HSV (LSV), YCbCr và LAB.

Các không gian màu này được mô tả sử dụng các mô men màu [15], tương quan màu, lược đồ màu, bộ mô tả màu trội, ma trận đồng xuất hiện màu [16] và nhiều bộ mô tả màu khác.

Các đặc trưng màu được coi là đặc trưng mạnh bởi vì chúng bất biến với tỉ lệ, xoay và dịch chuyển [19]. Tuy nhiên, đặc trưng màu bị hạn chế về thông tin không gian nên nó cần có sự hỗ trợ của các bộ mô tả khác [20].

1.2.1.2. Đặc trưng kết cấu

Kết cấu là các mẫu mà không thể đứng riêng lẻ như màu hoặc cường độ duy nhất. Kết cấu được coi là đặc trưng quan trọng trong thị giác máy tinh bởi vì các đặc trưng này tồn tại trong nhiều ảnh thế giới thực do đó nó thường được sử dụng trong nhận dạng mẫu và tra cứu ảnh. Hạn chế chính của tra cứu ảnh dựa vào kết cấu là độ phức tạp tinh toan và nhạy cảm với nhiễu [20].

Phân tích kết cấu đã được sử dụng cho nhiều thuật toán như lọc Gabor, trường ngẫu nhiên Markov, biến đổi wavelet, phân rã kim tự tháp, ma trận đồng xuất hiện mức xám, và bộ mô tả lược đồ cạnh [23].

1.2.1.3. Đặc trưng hình

Hình là một trong những đặc trưng mức thấp dùng cho nhận dạng đối tượng. Đặc trưng hình được trích rút trên cơ sở của một biên hoặc một vùng [25]. Trong cách tiếp cận dựa vào vùng, trích rút được thực hiện cho toàn bộ vùng trong khi cách tiếp cận trích rút dựa vào biên được thực hiện theo biên của vùng. Nhiều phương pháp như bộ mô tả Fourier và các bất biến mô men [27] được sử dụng cho quá trình trích rút các đặc trưng hình. Các bộ mô tả hình là bất biến với tỉ lệ và dịch chuyển. Do đó, chúng thường được kết hợp với các bộ mô tả khác để tăng độ chinh xác.

1.2.1.4. Thông tin không gian

Đặc trưng không gian đề cập đến vị trí của đối tượng trong một ảnh hai chiều. Đối sánh tháp không gian là một trong những phương pháp tốt nhất để thu các thuộc tính không gian của các ảnh [28].

Ở giai đoạn đầu của tra cứu ảnh, các hệ thống thường sử dụng một đặc trưng để tra cứu các ảnh. Tuy nhiên, kết quả thường cho độ chính xác thấp bởi vì các ảnh

thường bao gồm một số đặc trưng [29]. Để thu được độ chính xác cao hơn, các phương pháp tra cứu ảnh sau đó thường sử dụng việc kết hợp nhiều đặc trưng như trong [19].

1.2.2. Các đặc trưng cục bộ

Các đặc trưng toàn cục đã được sử dụng trong nhiều phương pháp CBIR và thu được độ chính xác tốt, tuy nhiên, các đặc trưng cục bộ đang phổ biến bởi vì chúng có ưu điểm hơn hẳn các đặc trưng toàn cục về tính bất biến với tỉ lệ và xoay. Bên cạnh đó, các đặc trưng cục bộ cũng cung cấp các đối sánh đáng tin cậy trong các điều kiện ảnh khác nhau [14].

1.2.2.1. Biến đổi đặc trưng bất biến tỉ lệ

Biến đổi đặc trưng bất biến tỉ lệ (SIFT - Scale-Invariant Feature Transform) do David Lowe [14] đề xuất. Nó là một trong những bộ đặc trưng cục bộ được sử dụng rộng rãi nhất, mà chứa một bộ mô tả và một bộ phát hiện cho các điểm chính (key point). SIFT là mạnh đối với xoay và tỉ lệ ảnh, nhưng nó thực hiện kém khi đối sánh với các chiều cao và cần một véc tơ cỡ cố định cho mã hóa để thực hiện kiểm tra độ tương tự ảnh. Trong tra cứu ảnh, SIFT có hạn hạn chế đó là nó sử dụng nhiều bộ nhớ và có chi phí tính toán cao [30].

1.2.2.2. Các đặc trưng mạnh và nhanh

Các đặc trưng mạnh và nhanh (SURF - Speeded-Up Robust Feature) là một bộ mô tả cục bộ mạnh [31], nó khắc phục hạn chế về chiều cao của SIFT. SURF thì nhanh hơn và mạnh hơn SIFT bởi vì nó đòi hỏi ít thời gian cho tính toán đặc trưng và đối sánh bằng việc sử dụng một lược đồ đánh chỉ số dựa vào dấu hiệu Laplacian. Tuy nhiên, SURF hoạt động kém trong trường hợp xoay.

1.2.2.3. Mẫu nhị phân cục bộ

Mẫu nhị phân cục bộ (LBP - Local Binary Pattern) so sánh điểm ảnh trung tâm và các lân cận của nó, ở đây điểm ảnh trung tâm được xem như là ngưỡng. LBP là mạnh bởi vì nó bất biến đối với các biến đổi về đa cấp xám. Hơn nữa, nó đơn giản về mặt tính toán. Hạn chế chính của LBP là nó làm mất thông tin không gian toàn cục.

1.3. Lựa chọn đặc trưng

Lựa chọn đặc trưng là một quá trình quan trọng trong phân tích dữ liệu, nó giúp chọn ra tập các đặc trưng có liên quan nhất đến đối tượng dữ liệu và biểu diễn chúng một cách hiệu quả nhất. Tập các đặc trưng này được chọn từ các đặc trưng dữ liệu ban đầu (gốc) và được xếp theo thứ tự giảm dần của độ quan trọng. Một số kỹ

thuật lựa chọn đặc trưng khác nhau đã được đề xuất trong lĩnh vực nhận dạng mẫu [32].

Trong những năm gần đây, đã có một số tiếp cận đề xuất như: trọng số Fisher (Fisher score) [33], nổi trội (Relief), nổi trội F (Relief-F) [35], thông tin tương hỗ (mutual information) [36], điều kiện độc lập của Hilbert Schmidt (HSIC-Hilbert Schmidt Independence Criterion) [37], điểm số Laplace (Laplacian score) [38].

Một số kỹ thuật lựa chọn đặc trưng phổ biến nhất được sử dụng trong lĩnh vực nhận dạng mẫu gồm kỹ thuật trọng số Fisher, Relief, Relief-F và một số biến thể khác.

1.3.1. Kỹ thuật trọng số Fisher

Phương pháp này tính toán trọng số cho mỗi đặc trưng và sau đó lựa chọn các đặc trưng dựa trên những trọng số đó. Thuật toán Fisher tính trọng số của đặc trưng thứ i ký hiệu là F_i như sau:

$$Score\_F_i = \frac{\sum_{j=1}^{C} n_j(\mu_{ij}-\mu_i)^2}{\sum_{j=1}^{C} n_j(\rho_{ij})^2} \qquad \textbf{(1.1)}$$

ở đây C là số các lớp n_j biểu thị số các mẫu trong lớp thứ j và μ_i là trung bình của đặc trưng thứ i. ρ_{ij} và μ_{ij} biểu diễn phương sai và trung bình của đặc trưng thứ i được kết hợp với lớp thứ j.

Hầu hết các kỹ thuật lựa chọn đặc trưng tính trọng số của các đặc trưng đơn lẻ trong khi bỏ qua việc kết hợp của các đặc trưng. Điều này dẫn đến kết quả thiếu chính xác, chẳng hạn: khi xét hai đặc trưng F_i và F_j và giả sử trọng số của cả hai đặc trưng này là thấp, các thuật toán lựa chọn đặc trưng đơn lẻ sẽ loại hai đặc trưng này đi. Tuy nhiên, trọng số kết hợp của của hai đặc trưng này có thể cao, đáng lẽ chúng phải được giữ lại.

1.3.2. Thuật toán Relief

Thuật toán Relief không dựa vào giả thiết có điều kiện về tính độc lập của các đặc trưng, do đó nó phù hợp đối với các bài toán thực tế, nơi mà sự tương tác đặc trưng (sự phụ thuộc giữa các đặc trưng) thường xuất hiện.

Thuật toán Relief được chỉ ra là tin cậy, bao gồm thông tin ngữ cảnh, ước lượng hiệu quả và sự liên quan của các đặc trưng trong các bài toán mà đặc trưng có sự phụ thuộc lẫn nhau cao. Thuật toán này dựa vào khái niệm về các lề cục bộ cho mỗi đặc trưng. Các lề này nên là đủ lớn cho các đặc trưng liên quan. Thuật toán này được sử dụng trong pha tiền xử lý dữ liệu để lựa chọn tập con các đặc trưng cho huấn luyện mô hình và nó vẫn là một trong những thuật toán phổ biến nhất về tiền xử lý

dữ liệu cho đến ngày nay [35]. Nó cũng là một bộ ước lượng đặc trưng chung mà đã được sử dụng phổ biến trong nhiều môi trường. Lấy ý tưởng từ học dựa trên mẫu (instance-based learning), các tác giả trong [35] đã đề xuất thuật toán Relief cổ điển. Nó được tối ưu cho các bài toán hai lớp. Nguyên lý cơ bản của thuật toán là ngoài việc xem xét sự chênh lệch giữa các giá trị đặc trưng và phương sai trong các lớp, nó còn xem xét cả khoảng cách giữa các mẫu (instance).

Xét véc tơ đặc trưng f và các véc tơ đặc trưng của mẫu gần nhất với f từ mỗi lớp. Mẫu gần nhất thuộc về cùng lớp được xem như là gần trúng (GT), và mẫu gần nhất với một lớp khác được xem như là gần bỏ lỡ (GBL).

Thuật toán Relief [35] tính trọng số của đặc trưng thứ i theo công thức lặp sau:

$$w_i = w_i - (v_i - GT_i)^2 + (v_i - GBL_i)^2 \qquad (\mathbf{1.2})$$

1.3.3. Thuật toán Relief-F

Thuật toán Relief-F [35] là một cải tiến của thuật toán Relief. Nó ước lượng các lề tin cậy hơn. Các đặc trưng dư thừa, không liên quan hoặc nhiễu, có thể ảnh hưởng đến việc lựa chọn các lân cận gần nhất. Đây là nguyên nhân làm cho việc ước lượng các lề thiếu tin cậy. Để giải quyết vấn đề này, Relief-F tìm kiếm k GT gần nhất và GBL gần nhất thay vì GT và GBL, và tính trung bình đóng góp của tất cả k GT gần nhất và GBL gần nhất. Lựa chọn các lân cận gần nhất là rất quan trọng trong Relief-F. Mục tiêu là tìm các lân cận gần nhất đối với các thuộc tính quan trọng. Nhiều nghiên cứu đã được thực hiện để khai thác khả năng lựa chọn đặc trưng của thuật toán Relief-F (xem [40]).

1.4. Trích rút đặc trưng

Việc trích rút đặc trưng là một phương pháp quan trọng để tạo ra các đặc trưng mới dựa trên sự kết hợp hoặc biến đổi nào đó của các đặc trưng gốc. Có rất nhiều phương pháp trích rút đặc trưng khác nhau được đề xuất để đáp ứng nhu cầu trong các ứng dụng khác nhau.

Kỹ thuật trích rút đặc trưng có nhiều lợi ích trong các ứng dụng thực tế. Chúng cho phép làm việc với dữ liệu chiều thấp hơn, dẫn đến chi phí tính toán giảm đi. Các phương pháp trích rút đặc trưng cũng giúp thu được các biểu diễn dữ liệu phân biệt hơn. Dữ liệu phân biệt làm tăng hiệu năng cho phân lớp trong khi cho phép sử dụng các bộ phân lớp đơn giản hơn. Nhiều nghiên cứu đã được thực hiện để thu được các biểu diễn dữ liệu phân biệt [41], [42], [43].

Trích rút đặc trưng được thực hiện thông qua việc chiếu dữ liệu gốc vào các không gian nhúng. Các phương pháp tiêu biểu có thể kể đến bao gồm Phân tích phân biệt tuyến tính (LDA - Linear Discriminant Analysis) [44], Phân tích phân biệt tuyến tính thưa mạnh (RSLDA - Robust Sparse Linear Discriminant Analysis) [41], và trích rút đặc trưng sử dụng giảm gradient (FE_GD - Feature Extraction using Gradient Descent) [43], Phân tích thành phần chính (PCA - Principal Component Analysis) [45].

1.4.1. Phân tích thành phần chính

Phân tích thành phần chính (PCA) là hữu ích nhất khi dữ liệu nằm trên hoặc gần với một không gian con tuyến tính của tập dữ liệu. Với loại dữ liệu này, PCA tìm một cơ sở cho không gian con tuyến tính và cho phép bỏ qua các đặc trưng không liên quan.

Với một tập dữ liệu được cho, nơi mỗi mẫu dữ liệu có D chiều (tức là D đặc trưng), PCA tính một tập các véc tơ đặc trưng D chiều được gióng với các hướng mà có phương sai cực đại của dữ liệu.

Các thành phần chính có một số cách sử dụng: (1) Chiếu dữ liệu gốc lên các thành phần chính này; (2) Sử dụng các thành phần chính này để tạo ra các điểm mới. (1) có thể được thực hiện bằng việc áp dụng tích vô hướng của một điểm dữ liệu đầu vào với thành phần chính để nhận về giá trị vô hướng mà là chiếu của điểm đó lên thành phần chính này. Về nguyên lý, dữ liệu đầu vào D chiều có thể được chiếu lên D thành phần chính của nó, tuy nhiên, chỉ lựa chọn các thành phần chính mà biểu diễn một phương sai dữ liệu cao để chiếu lên. Các thành phần chính này có thể được lựa chọn thủ công hoặc dựa vào một ngưỡng được thiết lập.

Ma trận biến đổi trực giao $\mathrm{P} \in \mathbb{R}^{D \times D}$ gồm các thành phần chính. Ma trận này được tính toán theo các ràng buộc sau:

$\mathrm{Y} = \mathrm{P}^T\mathrm{X}$, ở đây $\mathrm{X} \in \mathbb{R}^{N \times D}$ biểu thị ma trận dữ liệu gốc gồm N mẫu với D chiều, và các cột của Y chứa chiếu trên các thành phần chính $\mathrm{PP}^T = \mathrm{I}$. Hơn nữa, $\mathrm{YY}^T = \mathrm{U}$, ở đây U biểu thị ma trận hiệp phương sai (ma trận đường chéo) của các điểm được chiếu Y mà không tương quan.

Về mặt toán học, ma trận hiệp phương sai $\mathrm{U} = \mathrm{YY}^T$ có thể được biểu diễn như sau:

$$\mathrm{YY}^T = (\mathrm{P}^T\mathrm{X})(\mathrm{P}^T\mathrm{X})^T = \mathrm{P}^T(\mathrm{XX}^T)\mathrm{P} \qquad (\mathbf{1.3})$$

Đại lượng thu được trong phương trình (1.3) ở trên là một ma trận đường chéo U trong đó:

$$P^T(XX^T)P = U \qquad (\mathbf{1.4})$$

Nếu ta nhân vế trái của phương trình (1.4) với P và vế phải với P^T, ta thu được:

$$XX^T = PUP^T \qquad (\mathbf{1.5})$$

Biết rằng $PP^T = I$, và phân rã SVD (Singular Value Decomposition) của đại lượng XX^T là:

$$XX^T = VSW^T \qquad (\mathbf{1.6})$$

ở đây V và W chứa các véc tơ riêng trái và phải của đại lượng XX^T và S là một ma trận đường chéo chứa các giá trị riêng tương ứng. Bằng việc kết hợp hai phương trình (1.5 và 1.6) ở trên, ta có $PUP^T = VSW^T$ và bởi vì XX^T được xây dựng như một ma trận đối xứng, các véc tơ riêng trái và phải W và V sẽ bằng nhau, dẫn đến $PUP^T = VSW^T$, biết rằng P và V là trực giao. Điều này kết luận rằng $P = V$ và $U = S$. Vì thế, U là ma trận trực giao và dữ liệu chiếu Y là không tương quan. Điều này cũng cho thấy rằng các thành phần chính tương ứng với ma trận dữ liệu X được cho bởi các véc tơ riêng của ma trận hiệp phương sai XX^T của dữ liệu gốc.

Nói chung, Có thể sử dụng PCA như một phương pháp trích rút đặc trưng hoặc giảm chiều dữ liệu. PCA sử dụng các vector riêng của ma trận hiệp phương sai để thực hiện giảm chiều dữ liệu. PCA tập trung vào tìm các hàm cơ sở trực giao để thu các hướng phương sai cực đại trong dữ liệu. Nó bảo toàn các khoảng cách cặp dạng Euclide. Nhược điểm chính của PCA là sử dụng tiếp cận học không giám sát nên độ chính xác của các nhiệm vụ phân lớp bị hạn chế.

1.4.2. Phân tích phân biệt tuyến tính

Phân tích phân biệt tuyến tính (LDA) là một trong các phương pháp trích rút đặc trưng phổ biến nhất được sử dụng trong học có giám sát. Cho đến nay, LDA vẫn là một công cụ được yêu thích cho các nhiệm vụ phân lớp có giám sát. Do tính đơn giản và mạnh mẽ của nó [46]. Đặc trưng được trích rút bởi LDA có tính phân biệt tốt nên nó hiệu quả cho nhiệm vụ phân lớp. Tuy nhiên, LDA thất bại trong trường hợp số lượng biến dự đoán quá lớn so với số lượng quan sát. Trong trường hợp này, ma trận trong phạm vi lớp (within-class matrix) sẽ là suy biến (singular), do đó không

thể áp dụng LDA. Một viễn cảnh khác mà LDA cũng thất bại đó là khi các đường biên (boundary) không cung cấp sự phân tách tốt các lớp trong dữ liệu. Có nhiều phương pháp được đề xuất để giải quyết các hạn chế của LDA cổ điển và chúng được chứng minh là có tính hiệu quả tốt trong các nhiệm vụ phân lớp ảnh. Điều này đã giúp LDA trở thành một trong những cơ sở thành công nhất cho các thuật toán phân lớp ảnh mới sau này. Vì vậy, cách tiếp cận dựa trên LDA được cho là có hiệu quả vượt trội đối với phân lớp ảnh.

LDA đòi hỏi thông tin nhãn của dữ liệu huấn luyện để tính không gian con chiếu tốt nhất trong đó các mẫu kiểm tra sẽ được chiếu vào để được phân lớp. Cho C biểu thị số các lớp trong dữ liệu và n ký hiệu cho số các mẫu trong lớp c. LDA tìm một chiếu tuyến tính mà tăng khoảng cách giữa các mẫu thuộc về các lớp khác nhau và giảm khoảng cách giữa các mẫu thuộc về cùng một lớp. Nói cách khác, LDA ước lượng ma trận biến đổi nơi mà không gian mong muốn cực đại phương sai giữa các lớp và cực tiểu phương sai trong một lớp.

Giả sử μ là trung bình của tất cả các mẫu dữ liệu và μ_i trung bình của các mẫu thuộc về lớp thứ i. Các trung bình này có thể được tính như $\mu = \frac{1}{n}\sum_{i=1}^{C}\sum_{j=1}^{n_i} \mathrm{x}_j^i$ và $\mu_i = \frac{1}{n_i}\sum_{j=1}^{n_i} \mathrm{x}_j{}^i$

Đầu tiên, LDA tính ma trận phân tán giữa các lớp S_b sử dụng công thức:

$$\mathrm{S}_b = \frac{1}{n}\sum_{i=1}^{C} n_i(\mu_i - \mu)(\mu_i - \mu)^T \qquad \textbf{(1.7)}$$

Tiếp theo, ma trận phân tán trong vi lớp S_w được tính như sau:

$$\mathrm{S}_w = \frac{1}{n}\sum_{i=1}^{C}\sum_{j=1}^{n_i}(\mathrm{x}_i{}^j - \mu_i)(\mathrm{x}_i{}^j - \mu_i)^T \quad \textbf{(1.8)}$$

LDA nhằm ước lượng một không gian chiếu mà cực đại phương sai giữa các lớp và cực tiểu phương sai trong phạm vi lớp. Trong trường hợp chỉ có một trục chiếu, trục chiếu p có thể thu được thông qua giải tiêu chuẩn Fisher [33]:

$$\mathrm{p} = \underset{\mathrm{p}}{\operatorname{argmax}} \frac{\mathrm{p}^T \mathrm{S}_b \mathrm{p}}{\mathrm{p}^T \mathrm{S}_w \mathrm{p}} \qquad \textbf{(1.9)}$$

Bài toán (1.9) trên có thể biến đổi thành một dạng khác [47] như sau:

$$\mathrm{p} = \underset{\mathrm{p}^T\mathrm{p}=1}{\operatorname{argmax}} \frac{\mathrm{p}^T(\mathrm{S}_w - \lambda \mathrm{S}_b)\mathrm{p}}{\mathrm{p}^T \mathrm{S}_w \mathrm{p}} \qquad \textbf{(1.10)}$$

ở đây λ là một hằng số dương nhỏ. Thông qua giải phương trình (1.10) trên, có thể thấy rằng véc tơ chiếu tối ưu p là véc tơ riêng liên kết với giá trị riêng nhỏ nhất của

$S_w - \lambda S_b$. Đối với trường hợp có nhiều hơn một trục chiếu, ma trận chiếu $P \in \mathbb{R}^{d \times k}$ gồm k véc tơ riêng liên kết với k giá trị riêng nhỏ nhất của $S_w - \lambda S_b$.

1.5. Học máy cho tra cứu ảnh dựa vào nội dung

Gần đây, các hệ thống CBIR đã chuyển sang sử dụng các phương pháp học máy để cải tiến hiệu năng tra cứu. Phần này luận án sẽ trình bày về học không giám sát, học có giám sát và học sâu mà được sử dụng trong CBIR.

1.5.1. Học không giám sát cho CBIR

Quá trình phân cụm được thực hiện sau quá trình trích rút đặc trưng và xây dựng véc tơ đặc trưng. Phân cụm được xem là phương pháp học không giám sát bởi vì nó không biết trước các nhãn của các mẫu. Phương pháp phân cụm K-means và K-means++ [48] là hai phương pháp được sử dụng phổ biến trong CBIR, đặc biệt là khi sử dụng các đặc trưng cục bộ. Quá trình phân cụm thường được sử dụng trong các phương pháp này để quyết định một ảnh trong CSDL thuộc về nhóm ngữ nghĩa nào.

Trong [49], các tác giả đã áp dụng K-means lên từ vựng trực quan được xây dựng từ việc kết hợp các từ vựng trực quan của SIFT và mẫu thứ tự cường độ cục bộ (LIOP - local intensity order pattern) để tăng cường quá trình tra cứu. Hạn chế của K-mean khi được áp dụng vào CBIR là cần chỉ rõ số lượng các cụm khi bắt đầu. Hơn nữa, việc lựa chọn trọng tâm ban đầu sẽ ảnh hưởng đến hiệu năng của thuật toán phân cụm và làm cho nó dừng ở mức tối ưu cục bộ nếu không có trọng tâm ban đầu phù hợp nào được chọn. Mặc dù một số lượng lớn các cụm sẽ giảm sai số, tuy nhiên rủi ro của quá khớp vẫn là cao. K-means có nhược điểm là không xử lý được các dữ liệu ngoại lai và nhiễu. Các tác giả trong [50] đã sử dụng K-means++ lên từ điển trực quan được xây dựng từ việc kết hợp các từ điển trực quan lược đồ gradient có hướng (HOG - Histogram of Oriented Gradient) và SURF. K-means++ khắc phục hạn chế của K-mean thông qua việc gán các trọng số cho các trọng tâm khởi tạo. Mặc dù quá trình chọn trọng tâm ban đầu phức tạp và tốn thời gian hơn so với K-mean, nhưng việc phân cụm là chính xác hơn và có ít lần lặp hơn, giúp giảm chi phí tính toán.

1.5.2. Học có giám sát cho CBIR

Phương pháp học có giám sát khác với học không giám sát, bởi vì nó biết trước nhãn của ảnh trong tập huấn luyện. Do đó, nó được sử dụng để thực hiện các nhiệm vụ phân lớp. Phần này sẽ giới thiệu một số phương pháp học máy có giám sát thường được sử dụng trong CBIR.

1.5.2.1. Máy véc tơ hỗ trợ

Máy véc tơ hỗ trợ (SVM - Support Vector Machine) [51] là một trong các bộ phân lớp có giám sát phổ biến nhất được sử dụng trong nhận dạng mẫu và phân lớp ảnh. Khi dữ liệu mới được đưa vào, SVM sẽ quyết định lớp mà nó thuộc về. Có hai loại SVM [52] là tuyến tính và phi tuyến tính. Trong SVM tuyến tính, các đặc trưng có thể được tách thành hai lớp thông qua sử dụng siêu phẳng tách. Trong khi đó, SVM phi tuyến sử dụng cho tập dữ liệu không thể tách được bởi siêu phẳng tách, nó sử dụng các hàm nhân để cho phép tách được bằng việc bổ sung chiều mới. Hàm nhân được xem như một phần quan trọng mà ảnh hưởng đến hiệu năng của SVM. Nhiều nghiên cứu đã sử dụng bộ phân lớp SVM để dự đoán lớp của một ảnh đầu vào [49], [50]. Tất cả họ đều sử dụng SVM với nhân Hellinger [53], có nguồn gốc từ nhân Additive, có chi phí tính toán thấp và hiệu năng tốt hơn các nhân khác. Để phân hạng các ảnh, khoảng cách từ siêu phẳng tách đến một ảnh được sử dụng làm giá trị phân hạng, các ảnh thuộc lớp dương và càng xa siêu phẳng tách thì có điểm số càng cao và được xếp ở top của danh sách phân hạng.

1.5.2.2. Mạng nơ ron nhân tạo

Mạng nơ ron nhân tạo (ANN - Artificial Neural Network) được sử dụng rộng rãi để tìm các nghiệm tốt cho hầu hết các bài toán thế giới thực, trong đó có tra cứu ảnh. Sự phát triển của các mạng này giống như hành vi của hệ thống nơ ron con người. Các đặc điểm xử lý thông tin vượt trội của ANN, chẳng hạn như độ mạnh, tính toán song song, khả năng chịu lỗi, nhiễu và tính phi tuyến, làm cho nó trở thành một lựa chọn đầu tiên để giải quyết nhiều vấn đề [55].

Về cơ bản, ANN bao gồm các nơ ron và các liên kết kết nối các nơ ron với nhau. ANN gồm ba lớp: lớp đầu vào, lớp ẩn và lớp đầu ra. Lớp đầu vào gồm n nơ ron, với mỗi nơ ron cho một biến độc lập trong mạng. Người dùng chọn số nơ ron trong lớp ẩn dựa trên thực nghiệm. Lớp đầu ra có số nơ ron bằng với số lớp và được coi là biến phụ thuộc. Trọng số được gán cho mỗi kết nối giữa các nơ ron và được điều chỉnh trong mỗi lần lặp của quá trình huấn luyện mạng. Việc chọn mạng là theo từng bài toán cụ thể.

Để sử dụng ANN như một bộ phân lớp, nó cần thực hiện việc huấn luyện và kiểm tra. Các tác giả trong [56] đã đề xuất một hệ thống CBIR mà tra tự động các ảnh trên cơ sở đối tượng chính của chúng. Với trích rút đặc trưng, các tác giả đã áp dụng

biến đổi Bandelet trên đối tượng chính của ảnh. Các tác giả đã sử dụng mạng nơ ron lan truyền ngược để phân lớp kết cấu theo một trong bốn loại (không có khối đường viền, dọc, ngang, chéo phải/trái). ANN bao gồm 20 nơ ron trong một lớp ẩn và 4 nơ ron ở lớp đầu ra. Các tác giả đã sử dụng lọc Gabor cho các đặc trưng kết cấu dựa trên đầu ra ANN. Để tăng cường hiệu năng hệ thống, các đặc trưng màu được trích rút trong các không gian màu $\mathrm{YC}_b\mathrm{C}_r$ và RGB bằng việc sử dụng lược đồ màu và các wavelet màu. Một mạng ANN khác được sử sụng để phân lớp các ảnh trong cùng một lớp với ảnh truy vấn và so sánh chúng với toàn bộ các ảnh trong lớp đó. Hệ thống của họ được xây dựng trên khái niệm phân đoạn, cho ra kết quả chính xác hơn nhưng chậm.

Kiến trúc của ANN có tác động lớn đến hiệu năng của hệ thống. Kiến trúc này được học bằng cách thử và sai [57]. Lỗi huấn luyện và độ không đảm bảo tăng lên nếu dữ liệu đầu vào bị nhiễu và nhận thức sẽ không chính xác [20].

1.5.3. Học sâu cho CBIR

Trải qua nhiều thập kỷ, kỹ thuật học sâu đã thu hút được sự quan tâm đáng kể trong lĩnh vực học máy nhằm giải quyết các vấn đề của thực tiễn. Kiến trúc học sâu bao gồm một họ các thuật toán học máy, thiết kế của nó được lấy cảm hứng từ bộ não con người. Các thuật toán này tổ chức và thao tác thông tin bằng cách chuyển chúng qua các giai đoạn biểu diễn và biến đổi. Sự thành công của các thuật toán học sâu trong nhiều lĩnh vực (ví dụ: nhận dạng đối tượng) đã khiến nó trở thành một lựa chọn hàng đầu trong lĩnh vực tra cứu ảnh, đặc biệt là CBIR để khắc phục vấn đề khoảng trống ngữ nghĩa.

Kiến trúc của học sâu giúp nó có khả năng ánh xạ dữ liệu ở lớp đầu vào sang dữ liệu ở lớp đầu ra mà không phụ thuộc vào các đặc trưng do người cung cấp [58]. Thuật toán học sâu bao gồm mạng nơ ron tích chập (CNN - Convolutional Neural Network), mạng nơ ron sâu (DNN - Deep Neural Network), mạng niềm tin sâu (DBN - Deep Belief Network), và máy Boltzmann (Boltzmann machine). Trong đó, CNN thể hiện hiệu năng vượt trội trong các ứng dụng thị giác máy tính như nhận dạng khuôn mặt, phát hiện đối tượng và phân đoạn ngữ nghĩa [59] và đặc biệt trong lĩnh vực CBIR. CNN gồm ba loại lớp: lớp tích chập (convolutional layer), lớp gộp (pooling layer) và lớp kết nối đầy đủ (fully connected layer) [60]. Các bộ lọc được áp dụng đối với các ảnh thông qua lớp tích chập để học các đặc trưng, chức năng của các lớp trung gian (lớp pooling) là lấy mẫu giảm các đầu vào hiện tại (mà là đầu ra

của lớp ngay trước). Lớp cuối cùng (lớp kết nối đầy đủ) dự đoán lớp của ảnh đầu vào. Sự khác biệt giữa ANN và CNN là lớp cuối cùng của CNN chỉ là lớp được kết nối đầy đủ trong khi ở ANN, tất cả các nơ ron đều được kết nối với các nơ ron khác [61]. CNN cũng bất biến đối với dịch chuyển, tỉ lệ và xoay, do đó nó đặc biệt có ích cho các ứng dụng thị giác máy tính [59]. CNN không yêu cầu trích rút đặc trưng thủ công [62].

Các tác giả trong [58] đã kiểm tra hành vi của CNN trong các thiết lập khác nhau trong lĩnh vực CBIR nhằm cung cấp biểu diễn đặc trưng cho các ảnh, và nó cũng nhằm thực hiện đo độ tương tự. Các tác giả đã kết luận rằng CNN có thể được sử dụng để trích rút các đặc trưng cho cải tiến hiệu năng tra cứu ảnh. Tuy nhiên, do từ điển trực quan lớn được sử dụng, thời gian huấn luyện và chi phí bộ nhớ lớn đã làm giảm khả năng tra cứu của CBIR. Các tác giả trong [63] đã sử dụng CNN song tuyến tính cho đề xuất một hệ thống CBIR. Họ đã sử dụng CNN để trích rút các đặc trưng từ nội dung ảnh theo cách không giám sát mà không phụ thuộc vào chú thích hoặc nhãn lớp. Các đặc trưng được trích rút có chiều giảm đi bởi vì các tác giả phụ thuộc vào sử dụng lược đồ gộp trong quá trình trích rút, do đó giảm việc sử dụng bộ nhớ và giảm chi phí tính toán. Lược đồ đề xuất được chứng minh là cho hiệu năng tốt. Tuy nhiên, thời gian tra cứu cho CSDL lớn còn xa so với đòi hỏi của thực tế.

Để nâng cao hiệu năng tra cứu về mặt chi phí tính toán và sử dụng bộ nhớ, Gogul và Kumar trong [64] đã đề xuất phương pháp CBIR sử dụng CNN cho biểu diễn đặc trưng bằng việc sử dụng gộp cực đại (maximum pooling) sau các lớp tích chập thay vì sử dụng các lớp kết nối đầy đủ bởi vì các lớp kết nối đầy đủ loại bỏ thông tin không gian do kết nối đến toàn bộ các nơ ron đầu vào. Kiến trúc này giảm chiều của bộ mô tả đặc trưng trong khi giữ lại thông tin không gian, do đó nó thu được hiệu quả tra cứu cao. Đề xuất một cách tiếp cận gồm ba lược đồ, tùy thuộc vào thông tin có sẵn: huấn luyện không giám sát, được sử dụng khi không có nhãn dữ liệu; thông tin liên quan, được sử dụng khi có dữ liệu được gán nhãn; và huấn luyện dựa trên RF, được sử dụng khi có phản hồi từ người dùng.

G. Alain và Y. Bengio trong [65] đã đề xuất phương pháp CBIR end-to-end, mà dựa vào VGGNet [66]. Để huấn luyện CNN trong trích rút đặc trưng, họ đã sử dụng tập dữ liệu trường hấp dẫn và nhãn điểm tương tự thay vì nhãn thông thường. Hệ thống đã cho độ chính xác tốt, tuy nhiên hệ thống sử dụng nhiều thời gian để xây

dựng CSDL trường hấp dẫn và cần được tăng cường hơn nữa các giai đoạn huấn luyện và kiểm tra.

Các tác giả trong [67] đã đề xuất một cách tiếp cận CBIR mà sử dụng CNN để trích rút các đặc trưng mức cao. Lớp cuối cùng của Alexnet [2] được sử dụng để trích rút các đặc trưng bởi vì lớp cuối cùng có véc tơ đặc trưng nhỏ nhất. Biểu diễn thưa được sử dụng để giảm chi phí tính toán và nó là hiệu quả về mặt nén dữ liệu. Mặc dù biểu diễn thưa thu được tốc độ tra cứu nhanh hơn nhưng việc tính độ tương tự trên các véc tơ giá trị thực còn khá chậm đối với các CSDL lớn.

Như vậy, việc sử dụng các phương pháp học sâu có giám sát để tra cứu ảnh có thể cải thiện độ chính xác, tuy nhiên, để huấn luyện mô hình cần phải có một lượng dữ liệu lớn được gán nhãn và việc tra cứu trên các CSDL lớn lại tốn thời gian và không thực tế. Vì vậy, để khai thác các mẫu không có nhãn sẵn, cách tiếp cận học không giám sát cũng đang được quan tâm.

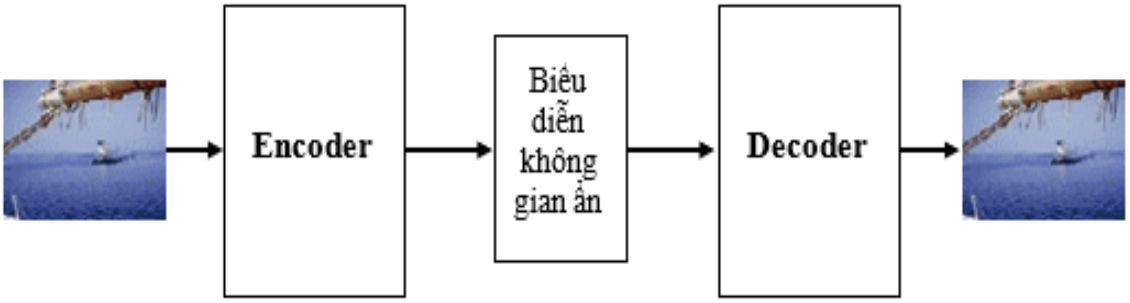
1.5.3.1. Mạng autoencoder

Một autoencoder là một mạng nơ ron nhân tạo, một phương pháp học không giám sát áp dụng thuật toán lan truyền ngược bằng việc thiết lập các giá trị mục tiêu xấp xỉ với các đầu vào. Autoencoder chuyển một ảnh đầu vào sang một biểu diễn không gian ẩn (LSR - Latent Space Representation) được nén (giống như giảm chiều). Do không có các nhãn lớp được gán, auencoder được xem là thuật toán học không giám sát. Tuy nhiên, autoencoder thường được xem như kỹ thuật học tự giám sát hơn là không giám sát bởi vì các giá trị mục tiêu được sinh ra từ dữ liệu đầu vào (học không giám sát được chuyển thành học có giám sát bằng cách tự động tạo nhãn). Autoencoder có thể tái cấu trúc dữ liệu đầu vào được cung cấp từ biểu diễn không gian ẩn của nó.

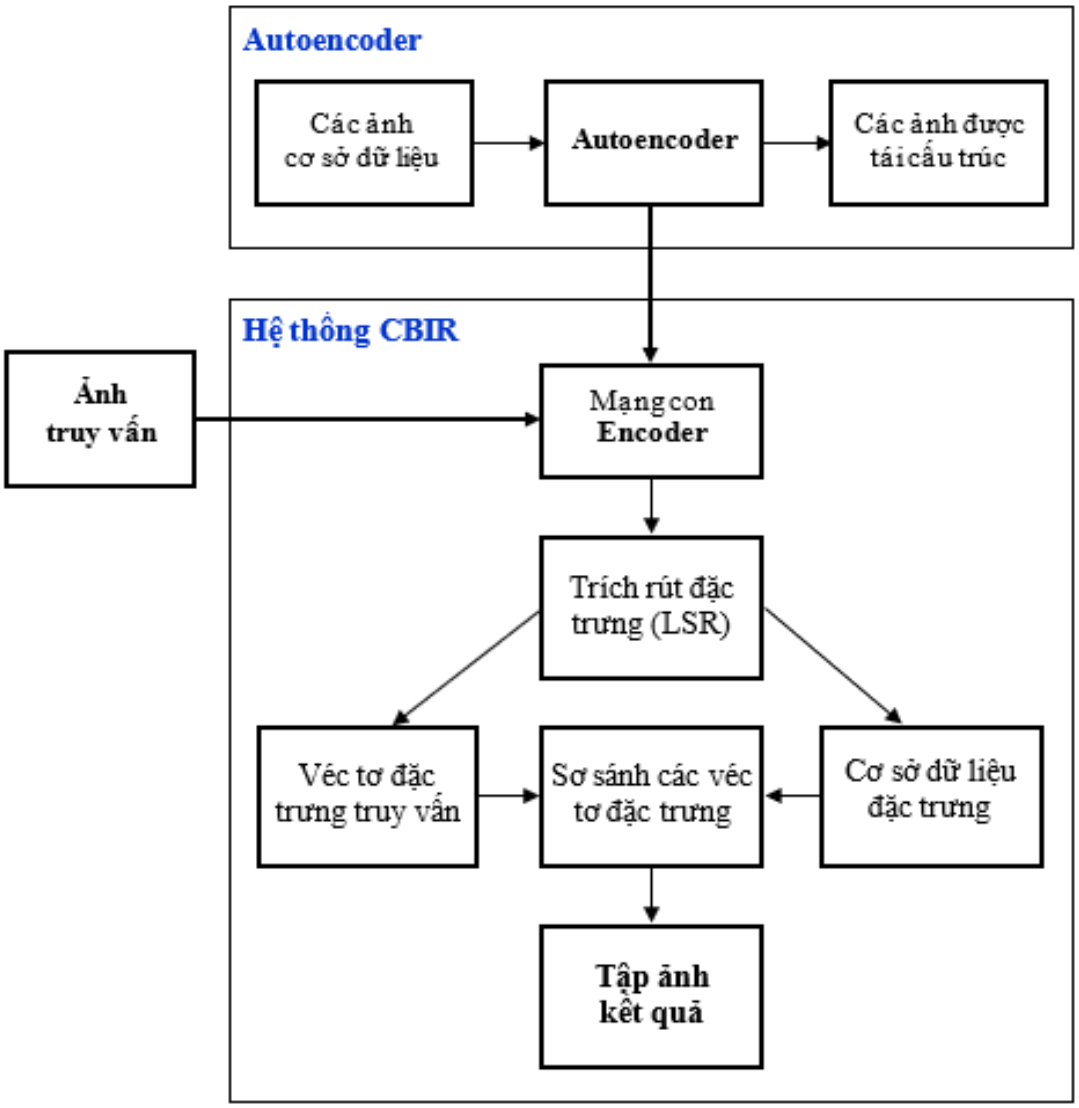
Điển hình, các autoencoder có hai mạng con (xem Hình 1.2): thứ nhất, một encoder mà chấp nhận dữ liệu đầu vào và chuyển dữ liệu đầu vào này sang biểu diễn không gian ẩn. Thứ hai, một decoder mà chấp nhận dữ liệu đầu vào là biểu diễn không gian ẩn để tái cấu trúc dữ liệu gốc. Autoencoder tập trung vào dữ liệu, có nghĩa là chúng chỉ có thể nén dữ liệu mà chúng đã được đào tạo. Bởi vì decoder không đảm bảo chỉ số chất lượng tương tự cho sinh ra dữ liệu gốc từ biểu diễn không giản ẩn, vì vậy có thể thấy chất lượng giá trị mục tiêu bị giảm một chút so với dữ liệu gốc.

Do autoencoder không yêu cầu nhãn lớp, nó có thể được sử dụng cho tính toán

biểu diễn không gian ẩn của toàn bộ ảnh trong CSDL. LSR biểu diễn một véc tơ chứa tất cả các đặc trưng của ảnh và có thể được xem như các véc tơ đặc trưng. Khi một ảnh được tra cứu, LSR của ảnh đầu vào được lấy ra và được tính toán với các véc tơ đặc trưng của các ảnh khác trong CSDL. Các ảnh trong CSDL mà có LSR gần với LSR của ảnh truy vấn sẽ được xem là các ảnh có liên quan với ảnh truy vấn nhất.



Hình 1.2. Mạng Autoencoder



Hình 1.3. Tích hợp autoencoder với mô hình CBIR [165]

Hình 1.3 ở trên là một mô hình CBIR với autoencoder. Đầu tiên, autoencoder được huấn luyện trên CSDL ảnh đầu vào theo cách không giám sát. Mạng con encoder được sử dụng để tính LSR của tất cả các ảnh trong CSDL ảnh. Các đặc trưng của ảnh truy vấn và các ảnh CSDL đều được trích rút và lưu trữ dưới dạng các véc tơ đặc trưng trong CSDL đặc trưng. Khi tra cứu, véc tơ đặc trưng của ảnh truy vấn sẽ được so sánh với các véc tơ đặc trưng của các ảnh trong CSDL để giúp hệ thống CBIR trả về các ảnh có khoảng cách nhỏ nhất so với ảnh truy vấn.

1.5.3.2. Mạng phần dư (ResNet)

Mạng phần dư [68] được đề xuất để giải quyết vấn đề suy giảm hiệu năng của các mạng nơ ron sâu. Ký hiệu ánh xạ cơ bản là $H(x)$, cho các lớp phi tuyến được xếp chồng khớp với ánh xạ còn lại $F(\mathrm{x}) = H(\mathrm{x}) - \mathrm{x}$. Như được chỉ ra trong Hình 1.4, ánh xạ gốc được chỉnh lại thành $F(\mathrm{x}) + \mathrm{x}$, tức là

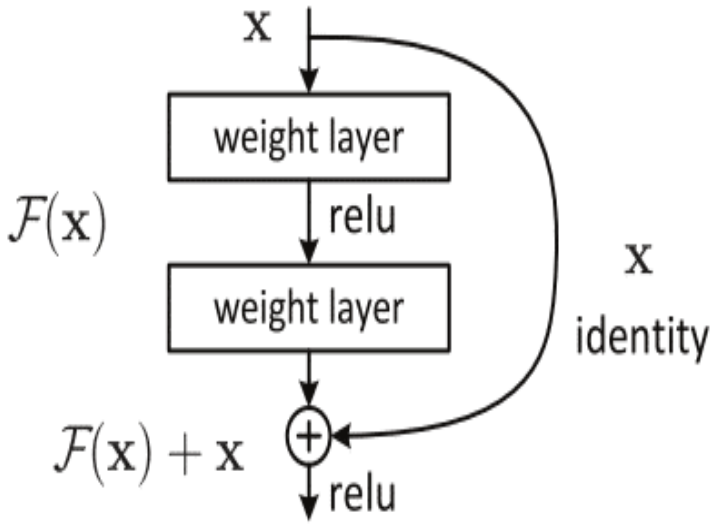
$$H(\mathrm{x}) = F(\mathrm{x}) + \mathrm{x} \qquad \textbf{(1.11)}$$

Lưu ý trong phương trình (1.11), chiều của x và $F(\mathrm{x})$ là như nhau. Bên cạnh đó, Hình 1.4 cũng xác định một khối xây dựng (building block) của ánh xạ đồng nhất bởi các đường tắt để học mỗi tầng được xếp chồng và được mô tả như sau:

$$\mathrm{y} = F(\mathrm{x}, \{\mathrm{W}_i\}) + \mathrm{x} \qquad \textbf{(1.12)}$$

ở đây x và y là các véc tơ đầu vào và đầu ra của các lớp được xét. Các chiều của x và F phải như nhau. Nếu không, một chiếu tuyến tính W_s phải được thực hiện bởi các kết nối đường tắt để khớp các chiều:

$$\mathrm{y} = F(\mathrm{x}, \{\mathrm{W}_i\}) + \mathrm{W}_s\mathrm{x} \qquad \textbf{(1.13)}$$



Hình 1.4. Một khối xây dựng của mạng phần dư

Như vậy, để nâng cao độ chính xác tra cứu của hệ thống thì cần phải áp dụng các kỹ thuật học máy, tuy nhiên các vấn đề sẽ gặp phải bao gồm chi phí thu thập mẫu huấn luyện cỡ lớn và thời gian tra cứu lâu. Các vấn đề này cũng là các vấn đề mà luận án sẽ tập trung giải quyết.

1.5.4. Học kết hợp

Trong lĩnh vực máy học, đặc biệt là khi nói về các phương pháp mà mục tiêu chính là cung cấp không gian nhúng phân biệt, ta thường tìm một mô hình duy nhất. Làm việc với các mô hình duy nhất được cung cấp bởi các thuật toán mạnh luôn là một cách tiếp cận hiệu quả trong các nhiệm vụ phân loại. Tuy nhiên, trong thực tế, học với một mô hình duy nhất không luôn dẫn đến hiệu năng tối ưu. Lưu ý ở đây là, "mô hình" có nghĩa ở trong ngữ cảnh của học kết hợp (Ensemble) chứ không phải ngữ cảnh học sâu. Trong ngữ cảnh này, mỗi mô hình khai thác một khía cạnh nào đó của đối tượng và để có thể khai thác được nhiều khía cạnh, chúng ta phải kết hợp nhiều mô hình.

Để giải quyết vấn đề này và nghiên cứu cách cải thiện hiệu năng của các phương pháp khác nhau, một số nghiên cứu đã đề xuất sử dụng tiếp cận học kết hợp (ensemble learning). Kỹ thuật học kết hợp giúp kết hợp các dự đoán từ nhiều mô hình học máy vào một mô hình duy nhất, từ đó giảm thiểu sai số tổng quát. Phương pháp này có tính linh hoạt cao và có thể được mở rộng cho quy mô lớn hơn với lượng dữ liệu huấn luyện sẵn có. Bagging [69] và boosting [70] là hai cách tiếp cận học kết hợp được sử dụng phổ biến.

Ý tưởng chính của học kết hợp là pha trộn và kết hợp các dự đoán từ nhiều mô hình. Những mô hình này thường là những mô hình tốt và mỗi mô hình trong số chúng cung cấp một thuộc tính phân biệt tốt của riêng nó. Bằng cách kết hợp các mô hình này, người ta thu được một mô hình duy nhất mà có khả năng phân biệt tốt hơn, dẫn đến phân lớp tốt hơn. Trong trường hợp các mô hình được kết hợp đúng, điều này có thể dẫn đến các mô hình sẽ mạng và chính xác hơn. Học kết hợp gồm một số phương pháp luận như stacking, boosting, bagging,...

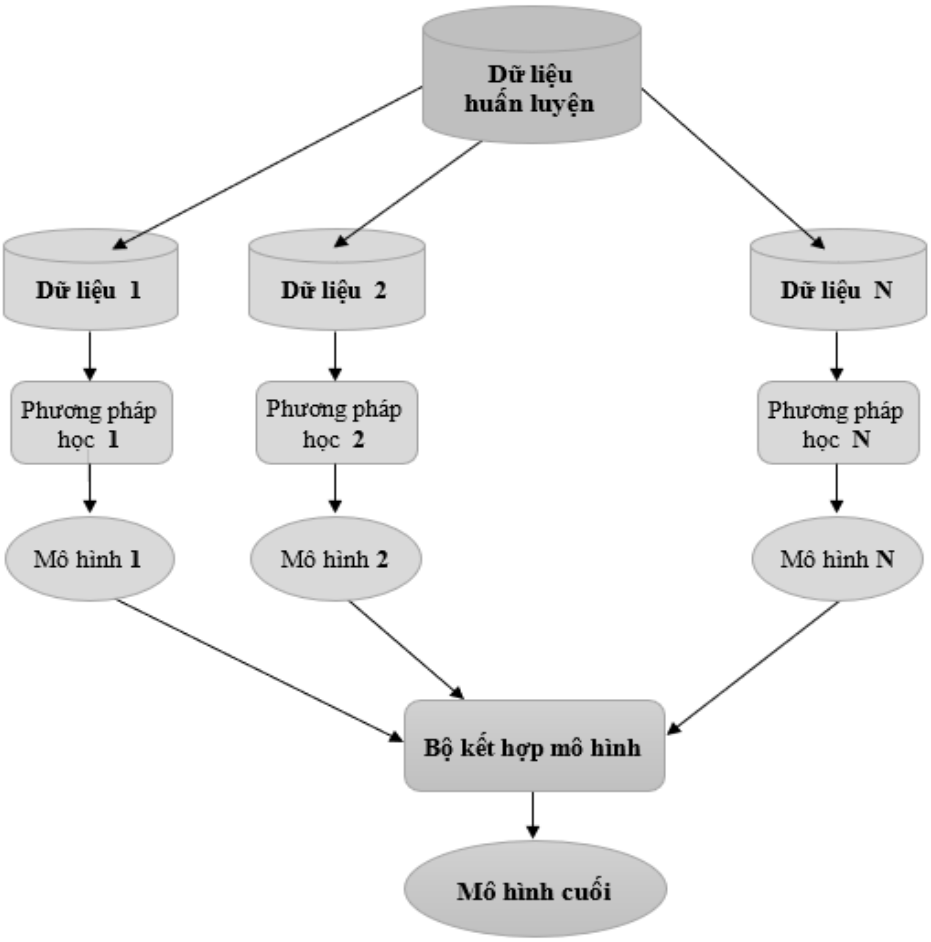
Nhiều phương pháp học kết hợp đã được áp dụng vào các nhiệm vụ phân lớp, sử dụng phổ biến nhất là phân lớp với mạng nơ ron. Lý do mà học kết hợp được chỉ ra là có đóng góp tốt cho cải tiến hiệu năng của mạng nơ ron [71].

Hiệu năng của một mô hình duy nhất thường được đo bởi khả năng của nó

trong xác định bộ dự đoán tốt nhất cho dữ liệu. Điều này chỉ có thể được suy ra sau khi quá trình phân lớp hoàn thành. Không có cách nào để nhận ra thông tin này trước bằng cách chỉ khai thác dữ liệu được xử lý và bài toán tối ưu hóa [72].

Brieman đề cập đến một số nghiên cứu liên quan về các thuộc tính lý thuyết của học kết hợp trong [69]. Một chiến lược nổi tiếng khác được sử dụng trong học kết hợp được gọi là "stacking", nó liên quan đến việc kết hợp các dự đoán của nhiều mô hình khác nhau trên cùng một tập dữ liệu. Nhiều nhà nghiên cứu đã đề xuất các phương pháp kết hợp tuyến tính đưa việc stacking vào tập hợp các mô hình [69].

Để biết sự kết hợp hiệu quả nhất của các mô hình, nghiên cứu được mô tả trong [69] đã kiểm tra hồi quy xếp chồng (stacked regression) bằng cách sử dụng xác thực chéo (cross-validation). Nghiên cứu dựa trên xác thực chéo được mở rộng với mục đích tìm ra sự kết hợp tốt nhất của các yếu tố dự đoán bằng cách đề xuất phương pháp "Super Learner" [72]. Phương pháp này cho thấy tính ưu việt và đóng góp tốt trong một số lĩnh vực: học trực tuyến (online learning) [74], các ứng dụng dự đoán không gian (spatial prediction applications) [75].

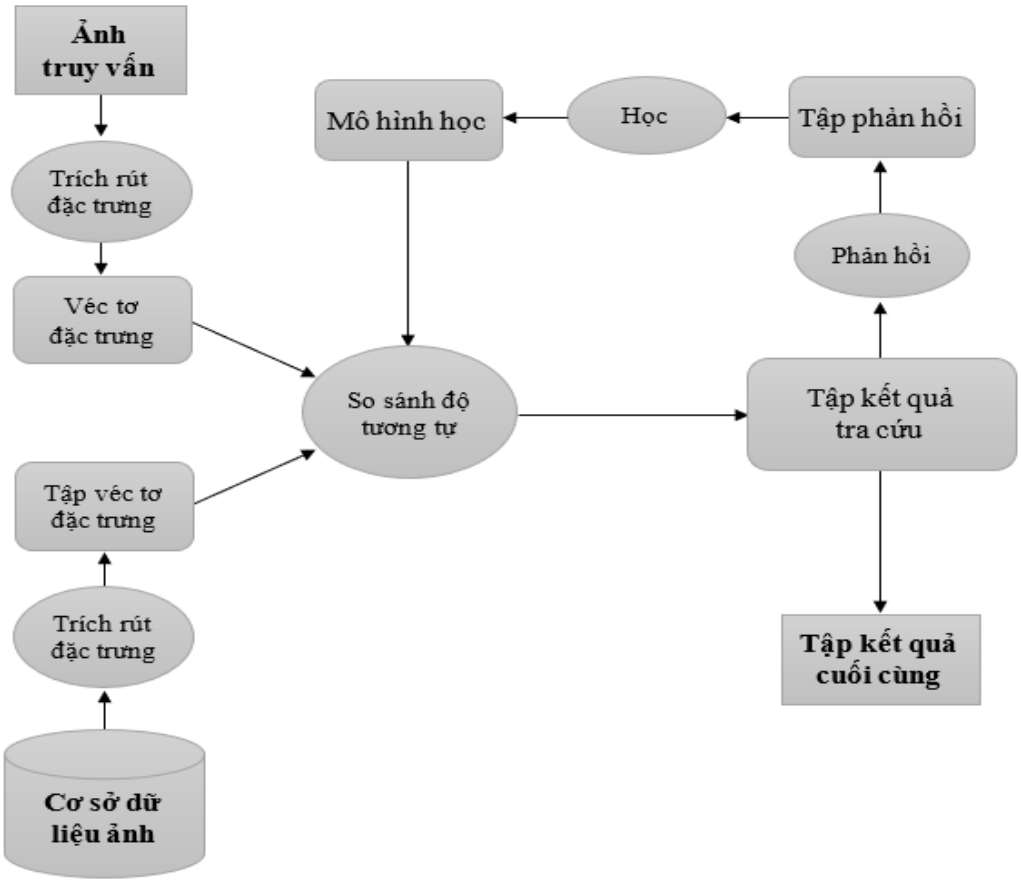


Hình 1.5. Học kết hợp

Hình 1.5 biểu diễn cấu trúc của phương pháp luận học kết hợp. Trong mô hình này, nhiều tập con dữ liệu huấn luyện được sử dụng để tạo ra nhiều mô hình. Các mô hình thu được đươc đưa vào bộ kết hợp mô hình để sinh ra một mô hình cuối cùng. Mô hình cuối cùng sẽ được sử dụng cho nhiệm vụ mong muốn.

1.6. Cơ chế phản hồi liên quan

Phản hồi liên quan (RF) là một công cụ phổ biến và mạnh mẽ trong các hệ thống CBIR. Nó được giới thiệu vào đầu những năm 1990 với mục đích giảm khoảng trống ngữ nghĩa giữa các truy vấn (đặc trưng mức thấp) và những gì người dùng nghĩ, bằng cách đưa người dùng tham gia vào quá trình tra cứu. Sự tương tác liên tục với người dùng đã giúp RF cải thiện hiệu suất của các hệ thống CBIR một cách đáng kể [77]. Hình 1.6 minh họa quá trình hoạt động của RF trong CBIR.



Hình 1.6. Sơ đồ mô tả hoạt động của RF trong CBIR

Quy trình chung của hệ thống RF trong CBIR được thể hiện như sau:

Bước 1: Người dùng chọn một ảnh truy vấn và hệ thống sẽ trích xuất đặc trưng mức thấp của ảnh.

Bước 2: Hệ thống trả về kết quả ảnh, với hai trường hợp xảy ra: (1) Trong giai đoạn ban đầu, ảnh kết quả được xếp hạng dựa trên mức độ tương tự của đặc trưng

mức thấp giữa ảnh truy vấn và các ảnh trong CSDL. (2) Trong các vòng lặp phản hồi, một hàm phân lớp được sử dụng để xếp hạng ảnh kết quả.

Bước 3: Người dùng đánh giá tính liên quan của kết quả trả về và chọn những ảnh có mức độ tương tự cao nhất.

Bước 4: Dữ liệu được gắn nhãn từ kết quả lựa chọn của người dùng được sử dụng cho thuật toán học máy, sau đó quá trình tra cứu được lặp lại từ bước 2.

Từ bước 2 đến bước 4 sẽ được lặp lại nhiều lần để tăng độ chính xác cho đến khi người dùng hài lòng với kết quả tra cứu của mình.

Tóm lại, RF là một bài toán phân loại nhị phân, trong đó ảnh mẫu được cung cấp bởi người sử dụng để huấn luyện một lớp phân loại. Lớp này được áp dụng để phân loại các ảnh trong CSDL thành hai nhóm: những ảnh liên quan đến truy vấn và những ảnh không liên quan đến truy vấn.

Những thách thức trong phản hồi liên quan:

Từ khi được giới thiệu vào năm 2007 bởi Liu và các cộng sự, kỹ thuật RF đã đạt được nhiều thành tựu đáng kể. Tuy nhiên, vẫn còn tồn tại những nhược điểm mà các phương pháp mới cần được cải tiến liên tục để khắc phục chúng. Cho đến nay, các nhà nghiên cứu vẫn đang tiếp tục nghiên cứu những nhược điểm nguyên thủy của kỹ thuật này để cải thiện chất lượng của nó.

Một số hạn chế trong RF của hệ thống CBIR như sau:

- Không thể trích chọn ngữ nghĩa mức cao: các kỹ thuật RF gặp khó khăn trong việc trích chọn ngữ nghĩa mức cao của ảnh khi chỉ có các đặc trưng mức thấp được sử dụng trong RF, nhưng vẫn tồn tại cách tiếp cận này trong việc tra cứu thông tin văn bản. Điều này là bởi vì việc tra cứu dựa trên từ khoá, chứ không phải trên các đặc trưng mức thấp.

- Sự khan hiếm và mất cân bằng các mẫu phản hồi: để đạt được kết quả tốt nhất, người dùng không muốn thực hiện nhiều lần lặp phản hồi. Tuy nhiên, trong một phiên phản hồi thì số lượng mẫu gắn nhãn thu được luôn là nhỏ so với số chiều của không gian đặc trưng. Vì vậy, đối với các tập dữ liệu huấn luyện có kích thước nhỏ, thì phần lớn các thuật toán học máy đều không thể cho ra kết quả có độ chính xác cao được. Ngoài ra, số lượng mẫu có nhãn âm thường lớn hơn số lượng mẫu có nhãn dương, điều này khiến cho độ tin cậy đối với việc học phân lớp trở nên kém hơn. Những vấn đề này đặc biệt nghiêm trọng đối với các mẫu dương, và khiến cho độ chính xác của RF giảm.

- Xử lý thời gian thực: để xử lý quá trình học trong RF một cách hiệu quả, tất cả các vòng lặp phản hồi bao gồm cả huấn luyện và kiểm tra đều phải được thực hiện trực tuyến, điều này tốn rất nhiều thời gian. Tuy nhiên, để giải quyết vấn đề này, có thể sử dụng phương pháp biểu diễn ảnh và cấu trúc lưu trữ dưới dạng một cây phân cấp. Điều này sẽ giúp nâng cao hiệu năng của hệ thống.

1.7. Đo độ tương tự giữa các ảnh

Quá trình trích rút đặc trưng và đo độ tương tự đóng vai trò quan trọng trong hiệu suất của hệ thống tra cứu ảnh. Để xác định ảnh liên quan nhất đến ảnh truy vấn, đo độ tương tự giữa các ảnh được sử dụng. Lựa chọn độ đo tương tự phù hợp sẽ ảnh hưởng trực tiếp đến độ chính xác và tính toán của hệ thống CBIR. Việc chọn độ đo tương tự phù hợp phụ thuộc vào cấu trúc của véc tơ đặc trưng được trích rút. Các yếu tố này cần được xem xét cẩn thận để tối ưu hóa hiệu suất của hệ thống CBIR.

Độ đo khoảng cách là một phương pháp thường được sử dụng để đo sự không tương tự giữa hai véc tơ đặc trưng. Khi sử dụng độ đo khoảng cách, giá trị nhỏ nhất được tính toán để xác định các ảnh tương tự nhất đối với ảnh truy vấn. Độ đo khoảng cách được chia thành hai loại chính: từng cặp thành phần của véc tơ đặc trưng (bin-by-bin) và chéo các thành phần của véc tơ đặc trưng (cross-bin). Trong độ đo khoảng cách bin-by-bin, các thành phần (bin) từ hai véc tơ đặc trưng được so sánh. Nếu u = $(u_1, u_2, \dots . u_n)$ và v = $(v_1, v_2, \dots . v_n)$ và $u_i \in$ u và $v_i \in$ v, thì hàm khoảng cách bin-by-bin so sánh u_i với chỉ v_i. Loại này được sử dụng rộng rãi bởi vì tính đơn giản và dễ thực hiện nhưng nó có nhược điểm là bị ảnh hưởng bởi tỷ lệ, lượng hóa, nhiễu, biến dạng và thay đổi ánh sáng [78]. Mặt khác, hàm khoảng cách cross-bin xem xét tương quan chéo giữa các thành phần trong véc tơ đặc trưng. Nó mạnh và mô tả tốt hơn khoảng cách bin-by-bin nhưng nó có độ phức tạp tính toán cao.

Khoảng cách họ Minkowski thuộc loại hàm khoảng cách bin-by-bin, được CBIR sử dụng rộng rãi bởi vì nó đơn giản trong cài đặt và tính toán. Công thức toán học của khoảng cách này như sau:

$$L_p(\mathrm{u}, \mathrm{v}) = \left(\sum_{i=1}^{N} |u_i - v_i|^p \right)^{1/p} \qquad \textbf{(1.14)}$$

Ở đây u và v là hai véc tơ trong R^N và N là tổng số chiều của không gian Euclide. Khoảng cách Minkowski cũng có tên là chuẩn L_p. Các khoảng cách L_1 (trường hợp $p = 1$) và L_2 (trường hợp $p = 2$) là các độ đo khoảng cách phổ biến được sử dụng

trong CBIR và các lĩnh vực xử lý ảnh khác. L_2 cũng được biết đến là khoảng cách Euclide. Nó có một tính chất đặc biệt là bất biến đối với phép biến đổi trực giao. Bên cạnh đó, L_1 được biết đến như là khoảng cách Manhattan hoặc City block. Nó biến đổi với xoay hệ tọa độ nhưng mạnh với phản xạ và dịch chuyển. L_∞có tên là khoảng cách Chessboard hay Chebyshev, cũng thuộc họ Minkowski.

$$L_\infty(\mathrm{u},\mathrm{v}) = \max_{i=1}|u_i - v_i| = \lim_{p\to\infty}\left(\sum_{i=1}^{N}|u_i - v_i|^p\right)^{1/p} \qquad \textbf{(1.15)}$$

Trên bàn cờ mà biểu diễn cho không gian 2 chiều, L_∞ là số lần di chuyển tối thiểu mà một quân vua cần để di chuyển giữa hai ô vuông. Thành viên cuối cùng của họ Minkowski là khoảng cách phân số (fractional distance), nó xuất hiện khi $p \in (0,1)$. Khoảng cách này không được coi là một độ đo bởi vì nó không tuân theo điều kiện bất đẳng thức tam giác [79]. Nó được ưu tiên nếu chiều của dữ liệu là cao [78]. Nó là mạnh với nhiễu so với các thành viên khác của họ Minkowski.

Thống kê Chi-square là một độ đo khoảng cách khác mà được sử dụng rộng rãi cho tính toán sự khác nhau giữa các hàm lược đồ [80]. Nó được biểu diễn bởi công thức như sau:

$$Chi - square(\mathrm{u},\mathrm{v}) = \sum_{i=1}^{N}\left(\frac{(\mathrm{u}-\mathrm{v})^2}{(\mathrm{u}+\mathrm{v})}\right) \qquad \textbf{(1.16)}$$

Chi-square [80] đã chỉ ra sự thành công khi nó được áp dụng cho phân lớp hình, đối sánh các mô tả cục bộ [81], phát hiện đường bao [82], và phân lớp các loại đối tượng [83].

Khoảng cách lược đồ giao là một độ đo khác mà được sử dụng trong CBIR và các thuật toán thị giác máy tính khác như phân đoạn, phân cụm, phân lớp. Nó là mạnh với sự thay đổi độ phân giải, che lấp và thay đổi góc nhìn [84]. Nó được biểu diễn bởi công thức toán học như sau:

$$HistogramIntersectionDis(\mathrm{M},\mathrm{S}) = \sum_{i=1}^{N}\min(\mathrm{m_i},\mathrm{s_i}) \qquad \textbf{(1.17)}$$

ở đây M và S là hai lược đồ với N thành phần.

Khoảng cách Mahalanobis được sử dụng để đo khoảng cách giữa một phân bố và một véc tơ đặc trưng. Nó được biểu diễn bằng toán học như sau:

$$\mathrm{MahalanobisDis}(\mathrm{v},\mathrm{m}) = \sqrt{(\mathrm{v}-\mathrm{m})^T\mathrm{C}^{-1}(\mathrm{v}-\mathrm{m})} \qquad \textbf{(1.18)}$$

Ở đây T biểu diễn chuyển vị của ma trận, v là véc tơ đặc trưng, m là véc tơ dòng trung bình và C là ma trận hiệp phương sai. Việc tính toán khoảng cách này sẽ có chi phí cao đối với dữ liệu cao chiều do tính ma trận hiệp phương sai.

Khoảng cách Canberra được sử dụng cho các số không dấu. Phiên bản điều chỉnh cho các số có dấu được giới thiệu trong [85] và được biểu diễn bởi công thức sau:

$$CanberraDis(\mathrm{u},\mathrm{v}) = \sum_{i=1}^{N} \frac{|u_i - v_i|}{|u_i + v_i|} \qquad \textbf{(1.19)}$$

ở đây u và v là các véc tơ với các giá trị thực. Độ đo này nhạy cảm với các giá trị gần 0 nhưng nó phù hợp khi sự khác biệt về dấu chỉ ra sự khác biệt về các lớp [85].

Squared Chord là một độ đo khoảng cách khác mà được sử dụng trong tra cứu ảnh nhưng nó không phù hợp cho các véc tơ đặc trưng có các giá trị âm [86]. Nó được xác định như sau:

$$SquaredChordDis(\mathrm{u},\mathrm{v}) = \sum_{i=1}^{N} \left(\sqrt{\mathrm{u}} - \sqrt{\mathrm{v}}\right)^2 \qquad \textbf{(1.20)}$$

Không giống như các khoảng cách ở trên, mà đo độ không tương tự, khoảng cách cosin được sử dụng để đo độ tương tự, tức là giá trị lớn của độ đo chỉ ra các ảnh tương tự nhất với ảnh truy vấn. Khoảng cách cosin [86] đo góc giữa hai véc tơ như sau:

$$\mathrm{CosinDis}(\mathrm{u},\mathrm{v}) = \frac{(\mathrm{u}.\mathrm{v})}{|\mathrm{u}|.|\mathrm{v}|} \qquad \textbf{(1.21)}$$

Việc chọn đúng và phù hợp độ đo tương tự là một thách thức lớn, và nhiều nghiên cứu đã tiến hành việc này dựa trên các thực nghiệm. Chẳng hạn các tác giả trong [87] đã so sánh giữa các khoảng cách Canberra, Chi-square, Manhattan và Oclit (Euclidean). Họ nhận thấy rằng sử dụng khoảng cách Euclide làm độ đo tương tự sẽ thu được độ chính xác cao hơn. Trong khi các tác giả trong [88] thực nghiệm với L_1, L_2, L_1 có trọng số, Chi Square, và Square Chord. Họ nhận thấy rằng sử dụng L_1 có trọng số là lựa chọn tốt nhất đối với bộ mô tả được đề xuất, bởi vì L_1 có trọng số cho các kết quả tốt hơn về độ chính xác và độ phức tạp trong tính toán. Mặc dù L_1 và L_2 có độ phức tạp thấp hơn Square Chord, Canberra và Chi Square, độ chính xác tra cứu cũng thấp hơn bởi vì cả L_1 và L_2 là nhạy cảm với nhiễu và không xem xét các thành phần lân cận. Các nghiên cứu đã chứng minh rằng không nên xem nhẹ việc sử dụng ngang bằng giữa các đặc trưng khác nhau trong quá trình đo độ tương tự giữa ảnh truy vấn và CSDL ảnh.

Trong không gian mã nhị phân [89], khoảng cách giữa u và v là khoảng cách Hamming. Nó được xác định là số các bít nơi mà các giá trị là khác nhau và được phát biểu về mặt toán học như sau:

$$\text{HammingDis}(\text{u},\text{v}) = \sum_{i=1}^{M} \delta[u_i \neq v_i] \qquad \textbf{(1.22)}$$

Công thức (1.22) tương đương với $\text{HammingDis}(\text{u},\text{v}) = \|u_i - v_i\|_1$ nếu mã có giá trị là 1 và 0. Khoảng cách cho các mã có giá trị bằng 1 và −1 được xác định tương tự. Độ tương tự dựa vào khoảng cách Hamming được xác định như $\text{HammingSim}(\text{u},\text{v}) = M - \text{HammingDis}(\text{u},\text{v})$ cho các mã có giá trị 1 và 0, tính số các bít nơi các giá trị là giống nhau. Tích vô hướng $\text{HammingSim}(\text{u},\text{v}) = \text{u}^T\text{v}$ được sử dụng làm độ tương tự cho các mã có giá trị 1 và -1. Các độ đo này cũng được mở rộng sang trường hợp có trọng số $\text{HammingDis}(\text{u},\text{v}) = \sum_{i=1}^{M} \lambda_i \delta[u_i \neq v_i]$ và $\text{HammingSim}(\text{u},\text{v}) = \text{u}^T\Lambda\text{v}$, ở đây $\Lambda = Diag(\lambda_1, \lambda_2, \ldots . \lambda_M)$ là ma trận đường chéo và mỗi thành phần trên đường chéo là trọng số của mã băm tương ứng.

Mặc dù đo độ tương tự đã được nghiên cứu nhiều trong thời gian qua nhưng việc chọn một độ đo tương tự phù hợp với mỗi bài toán cụ thể vẫn là một vấn đề cần được tiếp tục nghiên cứu.

1.8. Một số nghiên cứu về CBIR

1.8.1. Nghiên cứu quốc tế

Trong [90], các tác giả đã giới thiệu một hệ thống CBIR dựa trên sự tích hợp của màu sắc, hình dạng và kết cấu. Biểu đồ sự tương quan màu, biến đổi Gabor và biến đổi wavelet được sử dụng để trích xuất màu sắc, hình dạng và kết cấu tương ứng. Các tác giả đã sử dụng khoảng cách Manhattan làm độ đo tương tự giữa ảnh truy vấn và tập ảnh dữ liệu. Hạn chế chính của hệ thống là độ phức tạp tính toán tăng lên do tích hợp nhiều đặc trưng. Phân tích ảnh ở một mức độ phân giải duy nhất có thể làm mất một số chi tiết có giá trị. Do đó, các tác giả trong [91] đã phát triển một thuật toán phân tích đa độ phân giải mới giúp phân tích ảnh ở nhiều cấp độ, với các cấp độ khác nắm bắt thông tin mà một cấp độ đã bỏ qua. Cách tiếp cận này dựa trên việc trích rút các đặc trưng kết cấu và hình dạng bằng cách sử dụng bộ mô tả mẫu nhị phân cục bộ (LBP) để trích rút các đặc trưng kết cấu và mô men để trích rút các đặc trưng hình dạng từ các đặc trưng kết cấu ở các mức đa độ phân giải. Mặc dù LBP được sử dụng để trích rút các đặc trưng cục bộ, nhưng nó cũng tạo ra một véc tơ đặc trưng có

ảnh hưởng khi các đặc trưng cục bộ được kết hợp với các đặc trưng toàn cục.

Trong [92], các tác giả đã đề xuất một hệ thống CBIR bất biến đối với xoay và thay đổi màu. Hệ thống được đề xuất dựa trên việc ghép các đặc trưng màu và kết cấu để tạo thành một véc tơ đặc trưng chung. Để trích rút các đặc trưng màu, các ảnh được chuyển đổi sang không gian màu HSV và được lượng hóa thông qua biểu đồ màu. Để bất biến với sự thay đổi ánh sáng, chỉ có các kênh màu (Hue) và độ bão hòa (Saturation) được sử dụng. Mẫu nhị phân cục bộ xoay được sử dụng để trích rút các đặc trưng kết cấu bất biến xoay.

Các tác giả trong [93] đã đề xuất một phương pháp mới để lựa chọn các điểm hạt giống cho kỹ thuật tra cứu ảnh dựa trên màu trội. Tuy nhiên, phương pháp được đề xuất cần được hợp nhất với các phương pháp trích rút đặc trưng khác (hình dạng, kết cấu và thông tin không gian) để giảm khoảng trống ngữ nghĩa, do cùng một thông tin màu có thể được gán cho các ảnh trong các lớp ngữ nghĩa khác nhau. Để giảm khoảng trống ngữ nghĩa, các tác giả trong [94] đã đề xuất một hệ thống CBIR kết hợp các đặc trưng màu và cạnh để tạo thành một bộ mô tả đặc trưng. Để trích rút màu, các tác giả đã sử dụng biểu đồ màu. Để trích rút cạnh, biểu đồ cạnh canny đã được sử dụng trong không gian màu YCbCr. Để thu được sự tăng cường tốt hơn, wavelet rời rạc đã được tính toán, và để đẩy nhanh quá trình tính toán wavelet rời rạc, các tác giả đã sử dụng Haar wavelet, mà nhanh hơn về mặt tính toán [95]. Lược đồ đề xuất đã sử dụng ANN để hiểu lớp ngữ nghĩa của ảnh, mà cần thêm thời gian cho mục đích huấn luyện và kiểm tra. Hệ thống đã sử dụng khoảng cách Manhattan để đo độ tương tự của nó và các kết quả được báo cáo về độ chính xác trung bình đã chứng minh tính hiệu quả. Tuy nhiên, nó cũng bị thiếu thông tin không gian và không có thông tin về hiệu quả chi phí tính toán.

Trong [96], Song và cộng sự đã giới thiệu một phương pháp luận cho hệ thống CBIR trên cơ sở kết hợp các đặc trưng mức thấp (kết cấu và màu). Các mô men màu trong không gian màu HSV được sử dụng để trích rút các đặc trưng màu, và DWT và Gabor wavelet được sử dụng để trích rút các đặc điểm kết cấu. Để nâng cao hơn nữa, bộ mô tả hướng màu và cạnh đã được tính toán và đưa vào véc tơ đặc trưng, với kích thước 1×250. Kích thước véc tơ đặc trưng lớn hơn cho kết quả tra cứu chính xác hơn nhưng tốn nhiều thời gian hơn để tìm kiếm và so sánh. Tuy nhiên, lược đồ được đề xuất thiếu thông tin về kết cấu và không gian, như nhiều nghiên cứu khác.

Các tác giả trong [97] đã giới thiệu một kỹ thuật CBIR mới có lợi thế từ việc kết hợp màu sắc, hình dạng và kết cấu. Biểu đồ cạnh Canny và biến đổi DWT trong không gian màu YCbCr được sử dụng để trích rút các đặc trưng màu, trong khi GLCM được sử dụng để trích rút các đặc trưng kết cấu. Phương pháp cạnh canny trong không gian màu RGB được sử dụng để trích rút các đặc trưng hình dạng. Kỹ thuật được đề xuất đã áp dụng thuật toán di truyền, do đó nâng cao chất lượng giải pháp. Tuy nhiên, nó chịu mức độ quan trọng của quá trình và cần lặp lại nhiều lần, làm chậm thời gian tính toán.

Một hệ thống CBIR mới được trình bày bởi các tác giả trong [98], dựa trên việc trích rút các đặc trưng kết cấu toàn cục và cục bộ trong cả miền tần số và không gian cũng như các đặc trưng màu trong miền không gian. Để giảm hiệu ứng nhiễu, trước tiên ảnh được lọc bằng bộ lọc Gauss (Gaussian), sau đó các đặc trưng kết cấu toàn cục được trích rút trong miền không gian. Biểu đồ màu được lượng hóa trong không gian màu RGB được sử dụng để trích rút các đặc trưng màu. Để nâng cao hiệu năng tra cứu, các đặc trưng kết cấu cục bộ được trích rút thông qua bộ lọc Gabor. Hệ thống được đề xuất cho thấy các giá trị có độ chính xác cao và được so sánh với các phương pháp hiện đại khác. Ngoài ra, nó được báo cáo là bất biến với quay và ít nhạy cảm với nhiễu, nhưng nó có thời gian chạy cao do sử dụng các đặc trưng khác nhau.

Trong [87], T. Jolliffe và cộng sự đã trình bày phương pháp cho CBIR trên cơ sở tích hợp các đặc trưng phi tham số (kết cấu) và các đặc trưng tham số (màu sắc và hình dạng). Để trích rút các đặc trưng tham số, các mô men màu và các bất biến mô men đã được sử dụng và biến đổi xếp hạng được sử dụng để trích rút các đặc trưng phi tham số. Véc tơ đặc trưng được xây dựng có độ dài 247, làm tăng thời gian chạy và được coi là hạn chế chính của thuật toán này.

Một cách tiếp cận CBIR mới được trình bày bằng cách kết hợp các đặc trưng màu, hình dạng và kết cấu trong [99]. Entropy phân phối màu được sử dụng để trích rút các đặc trưng màu trong khi các mô men màu (Hue Moments) được sử dụng để trích rút các đặc trưng hình dạng. Để trích rút các đặc trưng kết cấu, ma trận xuất hiện mức độ màu đã được sử dụng. Đối với các độ đo tương tự giữa ảnh truy vấn và ảnh CSDL, các tác giả đã sử dụng phép đo độ tương tự chuẩn hóa có trọng số và các trọng số được quyết định dựa trên kinh nghiệm của người dùng. Mặc dù hệ thống được đề xuất thu được độ chính xác cao, nhưng hiệu năng của hệ thống bị ảnh hưởng khi ảnh

truy vấn chứa nhiều đối tượng phức tạp. Điều này có thể là do việc sử dụng các mô men màu (Hue Moments) để trích rút các đặc trưng hình dạng đôi khi không có khả năng nhận dạng ảnh mà chứa nhiều đối tượng hơn hoặc coi các cạnh khác nhau là một cạnh.

Một hệ thống CBIR trong miền cosin rời rạc (Discrete Cosine Domain) cũng được Latif và các cộng sự trong [100] đề xuất sử dụng. Các mô men màu, biểu đồ màu và biểu đồ cạnh đã được trích rút trực tiếp từ miền cosin rời rạc và thuật toán di truyền được sử dụng để gán độ quan trọng khác nhau cho các đặc trưng được trích rút nhằm cải thiện khả năng tra cứu ảnh. Mặc dù việc sử dụng thuật toán di truyền có tác động tích cực lớn đến độ chính xác của hệ thống, nhưng nó lại làm tăng thời gian sử dụng.

Như vậy, các phương pháp trích rút đặc trưng được sử dụng trong [91] được ưu tiên sử dụng khi yêu cầu độ chính xác; tuy nhiên, nó có chi phí tính toán cao do tính chất của các đặc trưng được trích rút. Mặt khác, khi chi phí tính toán đóng vai trò chính trong hiệu năng hệ thống, thì các phương pháp trong [99] có thể được xem xét.

1.8.2. Nghiên cứu trong nước

Tại Việt Nam, đã có nhiều công trình nghiên cứu, luận án tiến sĩ liên quan đến bài toán CBIR được công bố, cụ thể như:

Trong [166], Lư Minh Phúc và Trần Công Án đã đề xuất phương pháp để xây dựng một hệ thống tìm kiếm ảnh theo nội dung dựa trên mô hình học sâu (mạng nơ ron tích chập CNNs) nhằm tận dụng tối đa sức mạnh tính toán của máy tính trong việc tìm kiếm hình ảnh theo nội dung. Đồng thời, hệ thống cũng tích hợp ngữ nghĩa vào việc tìm kiếm dựa trên một domain-ontology để mô tả các mối quan hệ giữa các chủ đề ảnh cần phân lớp. Phương pháp tìm kiếm này không những khắc phục được các hạn chế của phương pháp tìm kiếm dựa trên metadata mà còn cho phép mở rộng và đa dạng hóa kết quả tìm kiếm thông qua việc kết hợp ngữ nghĩa vào việc tìm kiếm.

Trong [167], Nguyễn Thị Uyên Nhi và Văn Thế Thành đã đề xuất kỹ thuật trích xuất đặc trưng màu trội MPEG-7, kỹ thuật phát hiện biên với LoG, phép lọc Sobel, nâng cao cường độ ảnh với Gauss... Mỗi hình ảnh trong tập dữ liệu được trích xuất thành một vec-tơ đặc trưng, tạo thành CSDL đặc trưng, và lưu trữ trên cây phân cụm C-Tree cho bài toán tìm kiếm ảnh. Phương pháp đã cho độ chính xác vượt trội so với các phương pháp khác trên cùng tập ảnh, thời gian tìm kiếm nhanh (bộ COREL).

Các tác giả trong [168] đã giới thiệu một kỹ thuật dựa trên mạng nơron tích chập để trích chọn đặc trưng ảnh, sau đó tiếp tục thực hiện việc sinh mã nhị phân (binary hashing) để biến các đặc trưng này thành một véctơ nhị phân có độ dài nhỏ, véctơ này được gọi là mã nhị phân (hash code). Sau khi có được mã nhị phân cho từng bức ảnh, việc tính toán sự tương đồng giữa các bức ảnh sẽ trở nên đơn giản hơn vì số chiều thấp hơn và chỉ phải làm việc với các toán tử nhị phân đơn giản, từ đó cải thiện được tốc độ tìm kiếm. Kết quả thực nghiệm cho thấy việc sử dụng mạng CNN vào bài toán tìm kiếm ảnh theo nội dung cho kết quả tìm kiếm với độ chính xác cao, tuy nhiên thời gian truy vấn khá lâu.

Trong những năm gần đây, đã có nhiều công trình nghiên cứu liên quan đến bài toán CBIR được công bố, đặc biệt là các công trình nghiên cứu do nhóm nghiên cứu của PGS.TS. Nguyễn Hữu Quỳnh, PGS.TS. Ngô Quốc Tạo, cùng Nghiên cứu sinh và các cộng sự. Một số công trình tiêu biểu được công bố trong các luận án tiến sĩ đã bảo vệ thành công trong thời gian gần đây như:

- Năm 2017, Vũ Văn Hiệu đã bảo vệ thành công luận án tiến sĩ "Nghiên cứu một số kỹ thuật phân hạng trong tra cứu ảnh dựa vào nội dung" [101]. Công trình này đã đề xuất được hai giải pháp sử dụng RF: (1) cải tiến kỹ thuật hiệu chỉnh trọng số và dịch chuyển truy vấn để thu hẹp "khoảng trống ngữ nghĩa" trong CBIR và (2) nâng cao chất lượng CBIR thông qua cách tiếp cận tối ưu Pareto để xây dựng tập ứng viên có kích cỡ nhỏ và hỗ trợ nâng cao độ chính xác của máy phân lớp. Với hai giải pháp trên, đã giúp giảm khoảng trống ngữ nghĩa, rút gọn không gian tìm kiếm, có thể xem như sơ lọc trên CSDL lớn và giảm được số mẫu dữ liệu, cải thiện độ chính xác phân lớp, áp dụng cho bất kỳ kỹ thuật học máy nào trong việc phân lớp ảnh theo truy vấn. Tuy nhiên, độ chính xác của tập kết quả trong luận án còn thấp do cách tiếp cận của luận án là xét đến một vùng duy nhất chứa các điểm liên quan mà bỏ qua thực tế các ảnh được phân tán trong toàn bộ không gian đặc trưng. CSDL ảnh thực nghiệm là tập OXFORD Building và tập con của tập CALTECH 101. Điểm lưu ý ở đây là mặc dù luận án thu được các mẫu huấn luyện qua cơ chế RF nhưng cách tiếp cận của luận án không theo hướng học ma trận chiếu.

- Năm 2019, Đào Thị Thuý Quỳnh đã bảo vệ thành công luận án tiến sĩ "Nâng cao độ chính xác tra cứu ảnh dựa vào nội dung sử dụng kỹ thuật điều chỉnh trọng số hàm khoảng cách" [102]. Công trình này đã đề xuất được hai phương pháp tra cứu

ảnh liên quan ngữ nghĩa, giải quyết được các vấn đề như: (1) Danh sách kết quả gồm các ảnh thuộc về các vùng khác nhau trong khi chỉ sử dụng một truy vấn; (2) các cụm trong tập phản hồi không cần phải phân cụm lại; (3) độ quan trọng ngữ nghĩa của mỗi truy vấn được xác định; (4) trọng số quan trọng của mỗi đặc trưng được tính toán; (5) tận dụng được thông tin địa phương của mỗi vùng điểm trong không gian đặc trưng để xây dựng hàm khoảng cách. Với các giải pháp này, độ chính xác của phương pháp đã được cải tiến đáng kể, tuy nhiên vẫn còn giới hạn do phương pháp không xét đến sự không đồng nhất của không gian đặc trưng và không giải quyết vấn đề truy cập xấp xỉ trên không gian non-metric. CSDL ảnh thực nghiệm là tập COREL. Điểm lưu ý ở đây là mặc dù luận án thu thập các mẫu huấn luyện qua cơ chế RF nhưng cách tiếp cận của luận án là ma trận chiếu trên cơ sở tận dụng tính địa phương của mỗi vùng điểm đặc trưng.

- Gần đây nhất, năm 2022 NCS. Cù Việt Dũng đã thực hiện luận án tiến sĩ "Nghiên cứu phát triển một số thuật toán tra cứu ảnh dựa vào khái niệm mức cao sử dụng kỹ thuật học sâu" [103]. Công trình này đã đề xuất được hai phương pháp: (1) phương pháp học phép chiếu tối ưu cho dữ liệu đa tạp (xem xét cấu trúc cục bộ của các mẫu dương và âm thuộc hai lân cận khác nhau để học một phép chiếu mà dữ liệu có thể phân biệt trên không gian chiếu); (2) phương pháp học bán giám sát dựa trên đồ thị (tự động bổ sung các mẫu dương vào tập huấn luyện để giải quyết vấn đề mất cân bằng tập huấn luyện và tận dụng các khía cạnh khác nhau của đối tượng để tạo ra một bộ phân lớp mạnh). Hai phương pháp này đã sử dụng kỹ thuật giảm chiều với thông tin phản hồi từ người dùng, giúp cải thiện đáng kể độ chính xác tra cứu của hệ thống tra cứu ảnh khi số chiều của các đặc trưng là rất lớn (lớn hơn số lượng mẫu huấn luyện). Tuy nhiên, mặc dù cách tiếp cận của luận án là học ma trận chiếu với các mẫu huấn luyện được thu từ cơ chế RF nhưng việc tra cứu ảnh được thực hiện trên không gian chiếu.

Nhìn chung, các công trình nghiên cứu trong nước được công bố ở trên đã tập trung nghiên cứu về bài toán CBIR với RF sử dụng các kỹ thuật học máy, giúp thu hẹp "khoảng trống ngữ nghĩa" và cải thiện đáng kể độ chính xác tra cứu của hệ thống tra cứu ảnh. Các công trình này đã tiếp cận và khai thác hiệu quả các kỹ thuật học máy cho CBIR và thực nghiệm trên các tập dữ liệu ảnh chuyên nghiệp, phổ biến như tập OXFORD Building (5.062 ảnh), tập con của tập CALTECH 101 (8.000 ảnh), tập

COREL (10.800 ảnh), tập SIMPLIcity (1.000 ảnh). Tuy nhiên, các công trình này chưa khai thác được thuộc tính thưa dòng của ma trận chiếu và học biểu diễn ảnh theo tiếp cận học sâu. Đây vừa là một hướng nghiên cứu thiết thực, có tính khả thi cao mà nhóm nghiên cứu do PGS.TS. Nguyễn Hữu Quỳnh cùng các cộng sự đang theo đuổi và cũng chính là tiền đề để Nghiên cứu sinh hướng đến trong các nội dung nghiên cứu tiếp theo của mình tại luận án này.

1.9. Tổ chức thực nghiệm và đánh giá hiệu năng

1.9.1. Môi trường thực nghiệm

Để xác định hiệu quả của các mô hình và phương pháp đề xuất, thực nghiệm được xây dựng trên nền tảng dotNET, ngôn ngữ lập trình C#, Python và Matlab. Cấu hình máy tính sử dụng để thực nghiệm: Intel(R) Core(TM) i7-8550U CPU @ 1.80GHz, DDRam - 16GB và hệ điều hành Windows 11 Professional.

Thực nghiệm được mô tả dưới hai dạng gồm: đồ thị và bảng biểu; trong đó, hiệu suất tra cứu về độ chính xác và phạm vi được mô tả bằng đồ thị, các bảng biểu mô tả chỉ số đánh giá trung bình và so sánh giữa các phương pháp với nhau.

1.9.2. Cơ sở dữ liệu ảnh thực nghiệm

Trong luận án này, các tập ảnh chuyên nghiệp, phổ biến đã được sử dụng để thực hiện các thực nghiệm và đánh giá hiệu năng của hệ thống CBIR [104], bao gồm tập dữ liệu ảnh COREL[1] và CIFAR-100[2].

1.9.2.1. Tập dữ liệu ảnh COREL

Tập ảnh COREL chứa 10.800 ảnh được phân thành 80 chủ đề (khái niệm ngữ nghĩa) khác nhau, bao gồm mùa thu, cây cảnh, đám mây, hàng không, lâu đài, hổ, chó, voi, tàu thủy, tảng băng trôi, thạch nhũ và thác nước,.... Khoảng 100 ảnh hoặc nhiều hơn cho mỗi chủ đề, và kích thước của mỗi ảnh là 80×120 hoặc 120×80. Hình 1.7 là một số ảnh đại diện được thể hiện trong tập dữ liệu ảnh COREL.

Mỗi ảnh trong tập ảnh COREL được biểu thị bởi một véc tơ đặc trưng 190 chiều, biểu diễn cho các đặc trưng mức thấp. Trong số này, có 102 thành phần là đặc trưng màu và 88 thành phần là đặc trưng kết cấu. Trong đó, đặc trưng màu được thể hiện bằng 6 thành phần khoảnh khắc màu (color moment), 32 thành phần lược đồ màu (color histogram), và 64 thành phần tương quan màu (color correlation).

[1] https://sites.google.com/site/dctresearch/Home/content-based-image-retrieval
[2] https://www.cs.toronto.edu/~kriz/cifar.html

Hình 1.7. Một số ảnh đại diện trong tập dữ liệu ảnh COREL

1.9.2.2. Tập dữ liệu ảnh CIFAR-100

Tập dữ liệu ảnh CIFAR-100 là một tập con của bộ dữ liệu Tiny Images (80 triệu ảnh). Nó chứa 60.000 ảnh màu, được chia thành 100 lớp với 600 ảnh cho mỗi lớp như: máy bay, ô tô, chim, mèo, hươu, chó, ếch, ngựa, tàu và xe tải,…. Kích cỡ của mỗi ảnh trong tập này là 32x32 . Hình 1.8 chỉ ra một số ảnh đại diện trong tập dữ liệu ảnh CIFAR-100.

Hình 1.8. Một số ảnh đại diện trong tập dữ liệu ảnh CIFAR-100

1.9.3. Phương pháp đánh giá hiệu năng

Bài toán tra cứu ảnh thường thực hiện trên một CSDL ảnh và kết quả là một danh sách các ảnh tương tự. Số các ảnh trong tập kết quả thuộc về cùng chủ đề với ảnh truy vấn mà càng nhiều thì độ chính xác của hệ thống tra cứu ảnh càng cao. Các thước đo thường được sử dụng trong CBIR gồm độ chính xác P (Precision), độ triệu hồi R (Recall), độ chính xác trung bình AP (Average Precision) và độ đo tổng hợp kết quả của nhiều truy vấn mAP (Mean Average Precision) [105].

(1) Độ chính xác là tỷ lệ giữa số lượng ảnh có liên quan được tra cứu với tổng số ảnh tra cứu được trong một lần lặp và được tính toán theo công thức (1.23):

$$\text{Precision} = \frac{\text{Sum (relevant)}}{\text{Sum (retrieval)}} \qquad \textbf{(1.23)}$$

(2) Độ triệu hồi là tỷ lệ của các ảnh có liên quan được tra cứu trong một lần tra cứu trên số lượng tất cả ảnh liên quan trong CSDL ảnh, được tính toán theo công thức (1.24) như sau:

$$\text{Recall} = \frac{\text{Sum (relevant\_session)}}{\text{Sum (relevant\_in\_database)}} \qquad \textbf{(1.24)}$$

(3) Độ chính xác trung bình AP: đề cập đến vùng phủ phía dưới đường cong triệu hồi chính xác (precision-recall curve). AP lớn hơn hàm ý rằng đường cong triệu hồi chính xác cao hơn và độ chính xác tra cứu tốt hơn. AP được tính theo công thức (1.25) như sau:

$$\text{AP} = \frac{\sum_{k=1}^{N} P(k).rel(k)}{R} \qquad \textbf{(1.25)}$$

ở đây R biểu thị số các kết quả liên quan cho ảnh truy vấn từ tổng số N ảnh. $P(k)$ là độ chính xác của k ảnh được tra cứu, và $rel(k)$ là một hàm chỉ số mà có giá trị bằng 1 nếu ảnh thứ k trong danh sách phân hạng là liên quan và bằng 0 nếu ngược lại. Độ đo tổng hợp kết quả của nhiều truy vấn mAP được chấp nhận cho đánh giá trên tất cả các ảnh truy vấn. Công thức tính mAP như sau:

$$mAP = \frac{1}{Q}\sum_{q=1}^{Q} AP(q) \qquad \textbf{(1.26)}$$

ở đây Q là số các ảnh truy vấn.

Trong luận án này, AP và mAP được sử dụng cho thực nghiệm tại Chương 2 và 3 để đánh giá hiệu quả của các phương pháp được đề xuất.

1.10. Kết luận Chương 1

Với sự tăng nhanh của CSDL ảnh như hiện nay, việc nghiên cứu các phương pháp CBIR hiệu quả là rất cần thiết. Với hệ thống thống CBIR, hai đòi hỏi cần thiết là tăng độ chính xác tra cứu và tăng tốc độ tra cứu. Để giải quyết được hai đòi hỏi này, hệ thống CBIR phải tập trung vào hai giai đoạn quan trọng nhất đó là trích rút đặc trưng và tính độ tương tự.

Trong chương này, luận án đã hệ thống lại những kiến thức lý thuyết cơ bản và nghiên cứu liên quan đến CBIR như kiến trúc tổng quan; các đặc trưng mức thấp gồm màu sắc, hình dạng, kết cấu và thông tin không gian; lựa chọn và trích rút đặc trưng; các kỹ thuật học máy, học sâu cho tra cứu ảnh; các độ đo tương tự cho tra cứu ảnh; tổ chức thực nghiệm và đánh giá hiệu năng. Đặc biệt, Chương này đã phân tích nghiên cứu liên quan đến các giai đoạn trong CBIR để thấy được ưu điểm và hạn chế của các nghiên cứu hiện nay. Trên cơ sở các phân tích này, định hướng nghiên cứu của luận án sẽ tập trung vào nghiên cứu cơ chế RF, kỹ thuật học máy, và học sâu nhằm giải quyết vấn đề giảm khoảng trống ngữ nghĩa giữa các đặc trưng mức thấp và các khái niệm ngữ nghĩa mức cao để cải thiện tốc độ và độ chính xác tra cứu ảnh.

Nội dung cụ thể cần nghiên cứu ở các chương tiếp theo bao gồm:

(1) Cải thiện chất lượng của biểu diễn đặc trưng được trích rút;

(2) Khai thác các mẫu không có nhãn;

(3) Giảm chiều của biểu diễn đặc trưng;

(4) Chuyển biểu diễn đặc trưng sang dạng hiệu quả;

(5) Chọn độ đo tương tự hiệu quả.

Chương 2. PHƯƠNG PHÁP TRA CỨU ẢNH VỚI PHÂN TÍCH PHÂN BIỆT THƯA

Trong chương này, luận án trình bày phương pháp tra cứu ảnh mới có tên là SDAIR, sử dụng thuộc tính thưa dòng của ma trận chiếu phân biệt để cải tiến hiệu năng tra cứu. SDAIR khác biệt so với các phương pháp tra cứu ảnh dựa trên tiếp cận phân lớp đã có, mà nó phân lớp trên không gian chiếu nhưng không bị ảnh hưởng bởi vấn đề cỡ lớp nhỏ. Điều này đã làm cho SDAIR trở lên phù hợp hơn đối với tra cứu ảnh với RF, nơi mà cỡ lớp thường là rất nhỏ (chỉ có hai lớp). Các kết quả thực nghiệm trên tập dữ liệu CIFAR-100 minh chứng rằng phương pháp đề xuất thu được độ chính xác cạnh tranh so với một số phương pháp tra cứu ảnh tiêu biểu khác.

2.1. Giới thiệu

CBIR với RF đã thu hút sự quan tâm của cộng đồng nghiên cứu trong mấy thập kỷ qua [5], [104], [106]. Tuy nhiên, đo độ tương tự giữa hai ảnh bằng khoảng cách Euclide trong không gian cao chiều thường không hiệu quả do sự khác biệt ngữ nghĩa giữa các đặc trưng trực quan mức thấp và các khái niệm ngữ nghĩa mức cao của ảnh (tồn tại khoảng trống ngữ nghĩa). Một cách tiếp cận để giảm khoảng trống ngữ nghĩa giữa các đặc trưng trực quan mức thấp và các khái niệm ngữ nghĩa mức cao của ảnh là bao gồm học máy trong quá trình tra cứu ảnh với RF. Cơ chế này cho phép người dùng gán nhãn dương cho các ảnh có cùng chủ đề (tương tự ngữ nghĩa) với ảnh truy vấn (được gọi là các mẫu dương) và gán nhãn âm cho các ảnh khác chủ đề với ảnh truy vấn (không tương tự) (được gọi là các mẫu âm) trong danh sách kết quả trả về. Các mẫu này được sử dụng làm tập huấn luyện cho kỹ thuật học máy. Tuy nhiên, số lượng mẫu thu được từ quá trình RF thường rất nhỏ so với chiều của không gian đặc trưng, làm cho việc huấn luyện mô hình học máy trở nên rất khó khăn (khó thu được mô hình tốt). Trong trường hợp này, cách giải quyết là giảm chiều dữ liệu ảnh thông qua học một không gian chiếu thấp hơn [107], [108], nơi mà các kỹ thuật học máy có thể được áp dụng để học các khái niệm ngữ nghĩa mức cao.

Trong các bài toán học phân lớp trên dữ liệu nhiều chiều, giảm chiều được xem là một trong những kỹ thuật hiệu quả nhất [109]. Nó được đề xuất để giải quyết vấn đề thuộc về "Vấn đề của chiều - Curse of dimensionality" [110], mà điều này có nghĩa rằng các mô hình học máy không thể xử lý dữ liệu cao chiều một cách hiệu

quả. Gần đây, nhiều mô hình học phân lớp đã được đề xuất như học đa thể hiện (Multiple-instance learning) [89] và học không gian con (Subspace learning). Các phương pháp học không gian chiếu nổi tiếng nhất bao gồm phân tích thành phần chính (PCA - Principal Component Analysis] [111] và phân tích phân biệt tuyến tính (LDA - Linear Discriminant Analysis) [112]. PCA học một phép chiếu mà có thể bảo toàn được thông tin của dữ liệu, trong khi đó, LDA tìm một không gian chiếu phân biệt tối ưu sao cho tỉ số phân tán giữa các lớp (Between-class scatter) và phân tán trong phạm vi lớp (Within-class scatter) là cực đại.

Dưới góc nhìn về tái cấu trúc và bảo toàn thông tin, cách tiếp cận PCA được sử dụng rộng rãi như một công cụ tiền xử lý dữ liệu cho phân tích dữ liệu [113]. Các phương pháp theo tiếp cận PCA đã thu được một số thành công trong trích rút đặc trưng, tuy nhiên các véc tơ đặc trưng được trích rút từ các phương pháp này không có thông tin phân biệt do đó chúng không phù hợp cho nhiệm vụ phân lớp [114]. LDA học một phép chiếu phân biệt để cải tiến độ chính xác của phân lớp. Nhiều phương pháp cải tiến của LDA đã được đề xuất để nâng cao độ chính xác của phân lớp. Tuy nhiên, LDA và các phương pháp mở rộng của LDA đề cập ở trên không thể thực hiện lựa chọn đặc trưng của dữ liệu gốc trong khi tính phép chiếu. Các tác giả trong [41] đã đề xuất một phương pháp trích rút và lựa chọn các đặc trưng phân biệt từ dữ liệu gốc. Phương pháp của họ sử dụng ràng buộc thưa dòng chuẩn $\ell_{2,1}$ trên ma trận chiếu kết hợp với phân tích phân biệt tuyến tính.

Phương pháp tra cứu ảnh sử dụng RF thông qua kỹ thuật học phân lớp chỉ gồm có hai lớp là dương và âm, do đó nó đối diện với một số vấn đề như: (1) Số lượng các mẫu thu thập được "thường quá nhỏ so với chiều của không gian đặc trưng" [115], (2) Số lượng "các mẫu âm thường lớn hơn rất nhiều so với số lượng các mẫu dương" [115], và (3) Số các lớp là quá nhỏ, dẫn đến số các hướng chiếu phải nhỏ bởi vì số các hướng chiếu có liên quan chặt chẽ đến số các lớp. Như vậy, trong bài toán này, không thể áp dụng các phương pháp giảm chiều và trích rút đặc trưng ở trên vào quá trình chiếu dữ liệu ảnh gốc sang không gian chiếu và phân lớp. Trong chương này, luận án đề xuất một phương pháp tra cứu ảnh có giám sát mới, gọi là Phân tích phân biệt thưa cho tra cứu ảnh **SDAIR** (Sparse Discriminant Analysis for Image Retrieval). SDAIR giải quyết ba vấn đề gặp phải bên trên bằng cách tận dụng thuộc tính thưa dòng của ma trận chiếu phân biệt. Khác biệt so với các phương pháp truyền thống về

tra cứu ảnh theo tiếp cận phân lớp đã có, mà nó phân lớp trên không gian chiếu nhưng không bị ảnh hưởng bởi vấn đề cỡ lớp nhỏ. Điều này đã làm cho SDAIR trở lên phù hợp hơn đối với tra cứu ảnh với RF, nơi mà cỡ lớp thường là rất nhỏ.

SDAIR có những đặc điểm tiên tiến sau: (1) Khác với các mô hình tra cứu ảnh truyền thống, SDAIR có tính linh hoạt cao vì không phụ thuộc vào một độ đo tương tự hoặc mô hình học cụ thể nào (xem Hình 2.1), nó có thể áp dụng với bất kỳ độ đo tương tự ảnh nào, mô hình học lựa chọn đặc trưng nào, và mô hình học phân lớp nào. (2) Nó không bị ảnh hưởng bởi kích thước lớp nhỏ, trong khi nó vẫn có khả năng loại bỏ các đặc trưng dư thừa và không liên quan, cũng như tận dụng được thông tin phân biệt. (3) SDAIR không yêu cầu một số lượng lớn các mẫu dương trong tập huấn luyện, bởi vì nó có khả năng tự động bổ sung thêm mẫu dương vào tập huấn luyện bằng cách áp dụng mô hình học chiếu đã học trước đó. (4) SDAIR sử dụng cơ chế học ma trận chiếu hiệu quả, nó không chỉ tăng tính phân lớp trên ma trận chiếu thu được, mà nó còn đồng thời hỗ trợ đối với hai nhiệm vụ quan trọng đó là bổ sung mẫu huấn luyện dương và lựa chọn tập đặc trưng quan trọng.

2.2. Nghiên cứu liên quan

Phần này, luận án sẽ giới thiệu một số khái niệm và mô tả một số phương pháp liên quan như: phương pháp LDA, RSLDA là cơ sở cho nghiên cứu của luận án.

Dưới góc nhìn về tái cấu trúc và bảo toàn thông tin, cách tiếp cận PCA được sử dụng rộng rãi như một công cụ tiền xử lý dữ liệu cho phân tích dữ liệu [113]. Phép chiếu locality preserving, phép chiếu sparsity preserving bảo toàn và phép nhúng neighborhood preserving [116] là các phương pháp trích rút đặc trưng phổ biến nhất, mà chúng học các chiếu từ các cấu trúc hình học khác nhau của dữ liệu gốc. Các phương pháp này thu được một số thành công trong trích rút đặc trưng, tuy nhiên các véc tơ đặc trưng được trích rút từ các phương pháp này không có thông tin phân biệt do đó chúng không phù hợp cho nhiệm vụ phân lớp [114].

LDA học một phép chiếu phân biệt để cải tiến độ chính xác của phân lớp. Nhiều phương pháp cải tiến của LDA đã được đề xuất để nâng cao độ chính xác của phân lớp. Một số phương pháp cải tiến bao gồm Orthogonal LDA [117], Uncorrelated LDA và phân tích phân biệt tuyến tính hai chiều [118], mà chúng giải quyết vấn đề cỡ mẫu nhỏ của naive LDA. Đối với dữ liệu phân phối non-Gauss, LDA không xử lý tốt, do đó một số phương pháp cải tiến của LDA được đề xuất bao gồm: Marginal

Fisher analysis [119], Discriminative locality alignment [120], và Manifold partition discriminant analysis [121]. Tuy nhiên, LDA và các phương pháp mở rộng của LDA đề cập ở trên tính các ma trận phân tán bởi chuẩn ℓ_2, mà phóng đại sai số và nhạy cảm với phần tử ngoại lai. Li và cộng sự đã đề xuất một phương pháp để giải quyết vấn đề này thông qua sử dụng một bất biên quay chuẩn ℓ_1 để tính hai ma trận phân tán [122]. Tuy nhiên, phương pháp của Li và cộng sự là khó tìm giá trị trọng số tối ưu cho các nhiệm vụ học khác nhau. Wang và cộng sự cũng đề xuất một phương pháp cải tiến của LDA, mà sử dụng chuẩn ℓ_1 trong Fisher criterion function [123]. Tuy nhiên, phương pháp của Wang và cộng sự cần được giải lặp để tìm véc tơ chiếu, do đó nó là không hiệu quả.

Dữ liệu ảnh trong bài toán tra cứu ảnh với RF thường có nhiều đặc trưng không liên quan và dưa thừa. Các đặc trưng này làm giảm hiệu năng của mô hình phân lớp [47]. Các phương pháp học không gian chiếu thường sử dụng ràng buộc thưa để loại bỏ các đặc trưng không liên quan và dư thừa và chỉ giữ lại các đặc trưng quan trọng. Một số phương pháp tiêu biểu về ràng buộc thưa bao gồm: "Sparse discriminant analysis" [124], "Sparse linear discriminant analysis" [114], và "Sparse uncorrelated linear discriminant analysis" [125]. Các phương pháp này trích rút đặc trưng thông qua học một không gian chiếu phân biệt thưa. Kỹ thuật lựa chọn đặc trưng theo thuộc tính thưa dòng thông qua chuẩn $\ell_{2,1}$ là hiệu quả. Li và cộng sự đã đề xuất phương pháp học không gian chiếu, mà sử dụng chuẩn $\ell_{2,1}$ trong các công thức của hàm mất mát để chọn các đặc trưng phân biệt cho dự đoán nhãn [126]. Trong phương pháp của Tao và cộng sự, họ áp đặt một ràng buộc thưa dòng trên ma trận biến đổi của LDA thông qua chuẩn $\ell_{2,1}$ [127]. Tuy nhiên, các phương pháp này vẫn còn nhiều hạn chế. Với các phương pháp sử dụng chuẩn ℓ_1, chúng không thể biết được những đặc trưng nào là quan trọng nhất cho nhiệm vụ phân lớp. Với các phương pháp sử dụng chuẩn $\ell_{2,1}$, chúng nhạy cảm với việc lựa chọn số các chiều, điều này làm giảm nghiêm trọng hiệu năng của bộ phân lớp cho những bài toán có số lớp nhỏ. Để giải quyết vấn đề này, Wen và cộng sự đã đề xuất phương pháp trong [41] mà áp đặt ràng buộc chuẩn $\ell_{2,1}$ lên ma trận chiếu của LDA. Phương pháp của họ có thể đồng thời thực hiện trích rút và lựa chọn đặc trưng. Một mở rộng của phương pháp này được đề xuất bởi Dornaika và cộng sự, trong đó họ cực tiểu chiếu theo mỗi lớp để đảm bảo cùng cấu trúc thưa được biến đổi thuộc về cùng các lớp [42]. Mặc dù phương pháp của Wen

và của Dornaika đã đưa thêm vào một ràng buộc ma trận trực giao, tuy nhiên hai phương pháp này vẫn không giải quyết được vấn đề giảm hiệu năng của bài toán có cỡ lớp quá nhỏ. Gần đây, một số phương pháp trích rút đặc trưng dựa vào học sâu cũng đã được đề xuất [128]. Trong đó, phải kể đến phương pháp đại diện là DeepLDA [128]. DeepLDA là mạng nơ ron sâu, mà được mở rộng của LDA. Mô hình được học sẽ tập trung năng lượng phân biệt nhiều nhất lên C - 1 hướng (C là số lớp) và thu được hiệu năng tốt trên các tập ảnh lớn. Tuy nhiên, DeepLDA bị hạn chế đó là nó cần lượng lớn mẫu cho huấn luyện và mô hình không có tính giải thích (không có tính giải thích ở đây có nghĩa rằng nó không giúp tìm được đặc trưng gốc nào là quan trọng nhất). DeepLDA cũng bị vấn đề cỡ lớp nhỏ.

2.2.1. Giới thiệu chuẩn $\ell_{2,1}$

Chiều của dữ liệu ảnh trong các ứng dụng thực tế thường rất cao. Dữ liệu chứa một số lượng lớn các đặc trưng hoặc là dư thừa hoặc là không liên quan. Vì vậy, nếu loại đi các đặc trưng này sẽ giúp giảm thời gian và tăng độ chính xác của các nhiệm vụ học và phân lớp. Do đó lựa chọn đặc trưng là rất quan trọng.

Cho $\mathrm{Q} \in \mathbb{R}^{d \times d}$ biểu thị ma trận chiếu, mà thao tác trên các mẫu dữ liệu gồm d chiều. Phép chiếu của mẫu x được cho bởi $\mathrm{Q}^T\mathrm{x}$. $\ell_{2,1}$ norm của Q là như sau.

$$\|\mathrm{Q}\|_{2,1} = \sum_{i=1}^{d} \sqrt{\sum_{j=1}^{d} q_{ij}^2} = \sum_{i=1}^{d} \|\mathrm{q}_i\|_2 \qquad \textbf{(2.0)}$$

Bài toán (2.0) chính là tổng của chuẩn ℓ_2 của tất cả các dòng của ma trận Q. Ở đây, q_i là dòng thứ i của ma trận Q, q_{ij} là phần tử thứ j của dòng q_i.

2.2.2. Một số phương pháp liên quan

2.2.2.1. Phương pháp LDA (phân tích phân biệt tuyến tính)

LDA [44] là một thuật toán trích rút đặc trưng có giám sát nổi tiếng. Nó đòi hỏi tập dữ liệu huấn luyện có nhãn để ước lượng không gian chiếu. Trong không gian chiếu này, các mẫu test có thể được phân lớp một cách dễ dàng. LDA tìm một ma trận chiếu tuyến tính sao cho có thể tăng khoảng cách giữa các mẫu thuộc về các lớp khác nhau và giảm khoảng cách giữa các mẫu thuộc về cùng một lớp.

Cho C biểu thị số các lớp trong tập dữ liệu và n_i biểu thị số các mẫu trong lớp thứ i. Gọi μ, $\mu^{(i)}$ lần lượt là giá trị trung bình của tất cả các mẫu dữ liệu (tổng số mẫu dữ liệu trong tập dữ liệu được ký hiệu là N) và trung bình của các mẫu (x_j là ký hiệu của mẫu dữ liệu thứ j) thuộc lớp thứ i tương ứng. Các giá trị trung bình này có thể

được tính như sau:

$$\mu = \frac{1}{N}\sum_{i=1}^{C}\sum_{j=1}^{n_i} x_j^{(i)} \quad \text{và} \quad \mu^{(i)} = \frac{1}{n_i}\sum_{j=1}^{n_i} x_j^{(i)} \quad \textbf{(2.1)}$$

LDA tính ma trận phân tán giữa các lớp S_b và ma trận phân tán bên trong lớp S_w theo các công thức sau:

$$S_b = \frac{1}{n}\sum_{i=1}^{c} n_i(\mu^{(i)} - \mu)(\mu^{(i)} - \mu)^T \quad \textbf{(2.2)}$$

$$S_w = \frac{1}{n}\sum_{i=1}^{c}\left(\sum_{j=1}^{n_i}(x_j^{(i)} - \mu^{(i)})(x_j^{(i)} - \mu^{(i)})^T\right) \quad \textbf{(2.3)}$$

Trong trường hợp chỉ cần một trục chiếu, trục chiếu **a** có thể thu được thông qua giải tiêu chuẩn Fisher như sau: [33]

$$a = \underset{a}{\operatorname{argmax}} \frac{a^T S_b a}{a^T S_w a} \quad \textbf{(2.4)}$$

Một dạng khác của bài toán (2.4) được cho bởi [47] như sau:

$$a = \underset{a^T a=1}{\operatorname{argmin}}\, a^T (S_w - \lambda S_b) a \quad \textbf{(2.5)}$$

Bằng việc giải (2.5), ta có a là véc tơ riêng tương ứng với giá trị riêng nhỏ nhất của $S_w - \lambda S_b$. Trong bài toán (2.5), λ là một hằng số dương nhỏ.

Trong trường hợp nhiều hơn một trục chiếu, ma trận chiếu $A = [a_1, a_2, \dots, a_d] \in \mathbb{R}^{m\times d}$ sẽ gồm d véc tơ riêng tương ứng với giá trị riêng d nhỏ nhất của $S_w - \lambda S_b$.

2.2.2.2. Phương pháp RSLDA (phân tích phân biệt tuyến tính thưa)

Áp đặt ràng buộc chuẩn thưa (như các chuẩn l_1 và $l_{2,1}$) lên ma trận chiếu có thể tạo ra mô hình mà có thể lựa chọn được đặc trưng. Khác biệt với chuẩn l_1, chuẩn $l_{2,1}$ có thuộc tính thưa dòng tốt, mà có thể tạo ra một ma trận chiếu có khả năng giải thích tốt hơn các đặc trưng (đặc trưng gốc nào quan trọng nhất). Lấy cản hứng từ ý tưởng này, phương pháp RSLDA được đề xuất trong [41].

RSLDA [41] là một phương pháp trích rút đặc trưng dựa vào LDA. Nó cực tiểu chuẩn $\ell_{2,1}$ của ma trận chiếu tuyến tính Q. RSLDA có thể khôi phục dữ liệu ban đầu từ dữ liệu được chiếu chiều thấp.

Nhằm trích rút các đặc trưng mà vẫn bảo toàn được năng lượng chính của dữ liệu, RSLDA giải bài toán tối ưu sau:

$$\min_{P,Q,E} Tr(Q^T(S_w - \lambda S_b)Q) + \lambda_1\|Q\|_{2,1} + \lambda_2\|E\|_1 \quad \textbf{(2.6)}$$

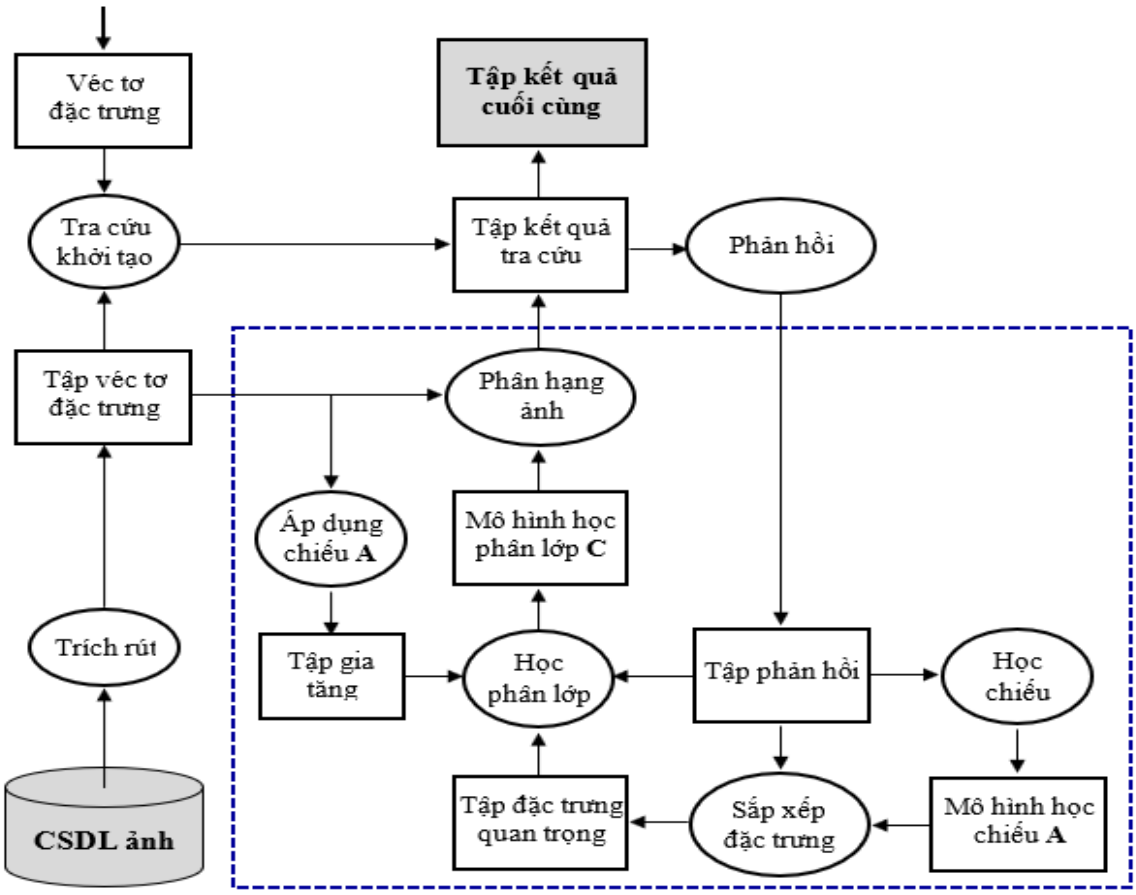
Thoả mãn $X = PQ^TX + E,\ P^TP = I$

Ở đây Q ∈ $\mathbb{R}^{d\times m}$ là ma trận chiếu phân biệt với $m < d$. S_w và S_b là các ma trận phân tán trong lớp và giữa các lớp tương ứng. λ_1 and λ_2 là tham số được dùng để xác định tính chất quan trọng của các số hạng khác nhau. λ là một hằng số được sử dụng để cân bằng hai ma trận S_w và S_b. X là ma trận dữ liệu đầu vào. $X = PQ^TX + E$ và $P^TP = I$ đảm bảo dữ liệu gốc có thể được khôi phục tốt [131]. $P \in \mathbb{R}^{m\times d}$ là một ma trận tái cấu trúc trực giao.

2.3. Phương pháp tra cứu ảnh được đề xuất

Trong phần này, luận án sẽ trình bày chi tiết mô hình tra cứu ảnh được đề xuất, kỹ thuật lựa chọn tập đặc trưng quan trọng, mô hình học phân lớp và thuật toán tổng thể để giải quyết bài toán tra cứu ảnh với RF.

2.3.1. Mô hình của phương pháp



Hình 2.1. Mô hình CBIR đề xuất.

Phương pháp được mô tả từ hướng nhìn của luồng dữ liệu tại Hình 2.1 như sau: Để thu được tập kết quả tra cứu khởi tạo, hệ thống trích rút đặc trưng của ảnh truy vấn và đối sánh véc tơ đặc trưng của ảnh truy vấn với từng véc tơ đặc trưng của ảnh CSDL. Các ảnh có khoảng cách đến ảnh truy vấn gần nhất sẽ được đưa lên trên đầu của danh sách.

Từ danh sách kết quả tra cứu khởi tạo ở trên, người dùng đánh dấu các ảnh có cùng chủ đề (gọi là ảnh liên quan) và các ảnh khác chủ đề (gọi là không liên quan) với ảnh truy vấn để nhận được tập phản hồi, cũng là tập mẫu huấn luyện cho mô hình học máy. Thuật toán học chiếu sử dụng tập mẫu huấn luyện này để thu được mô hình học chiếu A và sắp xếp các đặc trưng theo thứ tự giảm dần của độ quan trọng để thu được tập đặc trưng quan trọng. Tuy nhiên, giải pháp cho vấn đề cỡ mẫu nhỏ và số lượng mẫu âm nhiều hơn số lượng mẫu dương trong bài toán CBIR với RF, đó là tự động bổ sung mẫu dương vào tập huấn luyện thông qua áp dụng chiếu A. Quá trình này tạo ra tập gia tăng, kết hợp với tập đặc trưng quan trọng trên cả hai tập phản hồi và gia tăng để tạo ra tập huấn luyện cho học phân lớp. Dùng tập huấn luyện này để thu được mô hình học phân lớp C và phân hạng các ảnh sẽ được thực hiện theo mô hình này để được tập kết quả tra cứu. Quá trình này sẽ được lặp lại nếu người dùng chưa thỏa mãn với kết quả tra cứu, ngược lại thu được tập kết quả cuối cùng.

Sự khác biệt chính giữa mô hình của phương pháp đề xuất với các mô hình CBIR mà sử dụng RF đã có nằm ở các khối trong hình chữ nhật nét đứt. Hình 2.1 gợi ý rằng, nhóm các khối, mà nằm trong hình chữ nhật nét đứt, là độc lập với các thành phần còn lại bởi vì thông tin cần chỉ là một tập đặc trưng CSDL ảnh, một tập các RF của người dùng. Điều này hàm ý rằng mô hình đề xuất có thể nhúng trong bất cứ một hệ thống tra cứu ảnh với RF nào bất kể các đặc trưng ảnh và độ đo tương tự ảnh được sử dụng là gì. Ở đây cũng cần lưu ý rằng mô hình đề xuất rất là mềm dẻo, cụ thể: *Thứ nhất*, mô hình đề xuất có thể áp dụng bất cứ phương pháp học nào để lựa chọn đặc trưng, tức là, có thể thay đổi phương pháp học trong khối "Học chiếu" của Hình 2.1. *Thứ hai*, mô hình đề xuất cũng cho phép thay đổi phương pháp học phân lớp một cách phù hợp, tức là, phương pháp học máy phân lớp trong khối "Học phân lớp" của Hình 2.1 có thể được thay đổi tùy theo từng ngữ cảnh.

Phần tiếp theo của luận án, sẽ tập trung vào giới thiệu hai thành phần chính của mô hình, bao gồm mô hình học chiếu để lựa chọn tập đặc trưng quan trọng và mô hình học phân lớp.

2.3.2. Lựa chọn tập đặc trưng quan trọng qua mô hình học chiếu

Trong phần này, đầu tiên ràng buộc chuẩn $\ell_{2,1}$ được mô tả. Tiếp theo là lựa chọn tập đặc trưng quan trọng qua mô hình học chiếu. Phần này mô tả một số khối chính trên Hình 2.1 bao gồm: "Mô hình học chiếu A", "Học chiếu", "Sắp xếp đặc trưng" và "Tập đặc trưng quan trọng".

Dữ liệu ảnh thường có số chiều lớn và bao gồm nhiều đặc trưng không liên quan và dư thừa [129]. Vì vậy, việc loại bỏ những đặc trưng này không chỉ giúp giảm thời gian tính toán mà còn cải thiện hiệu năng của hệ thống phân lớp. Điều này sẽ giúp giảm thời gian và tăng độ chính xác tra cứu của hệ thống, vì phân lớp là một quá trình chính của hệ thống tra cứu ảnh.

Cho $\mathrm{A} = [\mathrm{a}_1, \mathrm{a}_2, \dots, \mathrm{a}_d] \in \mathbb{R}^{m \times d}$ biểu thị ma trận chiếu. Chiếu mẫu x sang không gian chiếu thấp chiều được cho bởi $\mathrm{A}^T\mathrm{x}$.

Phương pháp lựa chọn đặc trưng có thể được thực hiện bằng cách sử dụng cực tiểu chuẩn $\ell_{2,1}$ của ma trận chiếu, như đã được trình bày trong [130]. Trong phương pháp này, ràng buộc chuẩn $\ell_{2,1}$ được sử dụng như một công cụ để lựa chọn đặc trưng cho quá trình phân lớp. Bất cứ khi nào các dòng của ma trận A là bằng không (hoặc chuẩn ℓ_2 là rất nhỏ), các đặc trưng tương ứng với các dòng này là dư thừa và có thể loại bỏ. Để hiểu tại sao ma trận chiếu A có thể lựa chọn được đặc trưng quan trọng, luận án phân tích cấu trúc của ma trận A. Cho a_{ij} là các thành phần của ma trận biến đổi A. Nếu đặc trưng x_j là dư thừa, tất cả các thành phần của dòng thứ j của A phải bằng không, $\forall i, a_{ij} = 0$. Điều này được thực hiện thông qua cực tiểu $\|A\|_{2,1}$. Do đó, ép ma trận A có nhiều dòng không chính là lựa chọn đặc trưng.

Áp đặt ràng buộc "$\ell_{2,1}$-norm" lên ma trận chiếu theo tiếp cận LDA là một cách tiếp cận hiệu quả để trích rút đặc trưng [42], [41]. Phương pháp này giúp lựa chọn và trích rút các đặc trưng phân biệt nhất. Ma trận chiếu học được cho biết đặc trưng nào là quan trọng nhất.

Lấy động lực để khắc phục hạn chế của LDA, và kế thừa các ưu điểm của phương pháp RSLDA, luận án đề xuất một mô hình học bằng việc bổ sung một số hạng để khớp các nhãn lớp (các mẫu có cùng nhãn trong không gian chiếu sẽ gần nhau hơn trong khi các mẫu có nhãn khác nhau sẽ cách xa nhau hơn, tức tăng tính phân biệt). Mô hình đề xuất có thể tăng tính phân lớp của ma trận chiếu thu được.

Mô hình học được đề xuất là để học hai ma trận bằng việc cực tiểu hàm mục tiêu ở (2.7) dưới đây.

$$\min_{P,A,E} Tr(A^T(S_w - \lambda S_b)A) + \lambda_1\|A\|_{2,1} + \lambda_2\|E\|_1 + \frac{1}{2}\|Y - AX\|_F^2 \qquad \textbf{(2.7)}$$

Thoả mãn $X = PA^TX + E,\ P^TP = I$

Ở đây $\mathrm{A} \in \mathbb{R}^{m \times d}$ $(d < m)$ là ma trận chiếu phân biệt. S_w và S_b là các ma trận

phân tán trong lớp và giữa các lớp tương ứng. E là ma trận sai số. λ là hằng số được sử dụng để cân bằng hai ma trận phân tán. λ_1 và λ_2 là các tham số thỏa hiệp được sử dụng để xác định độ quan trọng của các thuật ngữ liên quan. $X = PA^TX + E$ và $P^TP = I$ đảm bảo dữ liệu gốc có thể được khôi phục tốt [131]. $P \in \mathbb{R}^{m\times d}$ là một ma trận tái cấu trúc trực giao. X là ma trận dữ liệu đầu vào với ma trận nhãn Y tương ứng. Cực tiểu số hạng đầu tiên để cung cấp ma trận chiếu mà liên kết với LDA. Cực tiểu số hạng thứ hai để thu được ma trận thưa và hỗ trợ việc xác định một tập các đặc trưng gốc quan trọng [41]. Cực tiểu số hạng thứ 3 là biểu thị các sai số và được sử dụng để mô hình nhiễu ngẫu nhiên. $\|.\|_1$ là chuẩn ℓ_1. Số hạng thứ tư được sử dụng để khớp với các nhãn lớp, tức là tăng cường khả năng phân lớp của ma trận chiếu A. $\|\quad\|_F$ là chuẩn của ma trận. Khi giải bài toán (2.7) để tìm số hạng P, A, E, số hạng $\|A\|_{2,1}$ trong công thức này sẽ ép $\|\mathrm{a}_i\|_2$ của dòng a_i của ma trận A bằng 0 nếu dòng đó là không quan trọng và dòng a_i nào quan trọng sẽ $\|\mathrm{a}_i\|_2$ lớn nhất.

Với phân lớp dữ liệu, các lề giữa các lớp khác nhau được kỳ vọng là lớn nhất có thể sau khi dữ liệu gốc được chiếu sang không gian nhãn của chúng (tức là Y trong (2.7)), do đó số hạng $\frac{1}{2}\|Y - AX\|_F^2$ trong (2.7) giúp cực đại lề được gia tăng, số hạng $\frac{1}{2}$ và mũ 2 được thêm vào để thu được hệ số 1 khi đạo hàm.

Trong (2.7), $\mathrm{Y} = [\mathrm{y}_1, \mathrm{y}_2, \ldots \mathrm{y}_n] \in \mathbb{R}^{C\times n}$ là ma trận nhãn nhị phân tương ứng. Y được xác định như sau: với mỗi mẫu huấn luyện x_i ($i = 1, 2, \ldots n$), $\mathrm{y}_i \in \mathbb{R}^C$ là véc tơ nhãn của nó. Nếu x_i là từ lớp thứ c (c=1, 2, ...C), thì chỉ mục thứ k của y_i là 1 và tất cả các mục còn lại là 0.

Để giải bài toán tối ưu (2.7), ta sử dụng Phương pháp Alternating direction of multipliers [132].

Để giải quyết vấn đề cỡ lớp rất nhỏ được đề cập ở trên, luận án đề xuất mô hình học chiếu cho việc lựa chọn tập đặc trưng quan trọng. Mô hình này bao gồm các bước như sau: trên tập phản hồi của người dùng, ta sẽ học một phép chiếu A theo cách mô tả trong bài toán (2.4). Sau đó tính $\|\mathrm{a}_i\|_2$ (khoảng cách Eucldie của véc tơ a_i), với a_i là dòng thứ i của ma trận A. Trong mô hình này, độ quan trọng của đặc trưng gốc thứ i chính là giá trị của $\|\mathrm{a}_i\|_2$. Sau đó, tiến hành sắp xếp các đặc trưng gốc theo thứ tự giảm dần của $\|\mathrm{a}_i\|_2$ tương ứng. Tập đặc trưng nhận được bao gồm các đặc trưng gốc được sắp xếp theo thứ tự giảm dần của độ quan trọng. Ví dụ ở dưới minh họa cho việc lựa chọn tập đặc trưng quan trọng.

Giả sử ta có ma trận dữ liệu X và A như sau:

$$\mathrm{X} = \begin{bmatrix} x_{11} & x_{12} & x_{13} & x_{14} \\ x_{21} & x_{22} & x_{23} & x_{24} \\ x_{31} & x_{32} & x_{33} & x_{34} \end{bmatrix} \text{ và } \mathrm{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} \qquad (\mathbf{2.8})$$

Giả sử sau khi tính, ta có $\|\mathrm{a}_3\|_2 > \|\mathrm{a}_2\|_2 > \|\mathrm{a}_1\|_2$ và muốn lấy k=2 đặc trưng quan trọng nhất. Kết quả sẽ được như sau:

$$\mathrm{A}_k = \begin{bmatrix} a_{31} & a_{32} \\ a_{21} & a_{22} \end{bmatrix} \text{ và } \mathrm{X}_k = \begin{bmatrix} x_{31} & x_{32} & x_{33} & x_{34} \\ x_{21} & x_{22} & x_{23} & x_{24} \end{bmatrix} \qquad (\mathbf{2.9})$$

Tóm tắt thuật toán 2.1 cho lựa chọn tập đặc trưng quan trọng với các bước cụ thể như sau:

Thuật toán 2.1: Chọn tập đặc trưng quan trọng

Input:

- X: ma trận mẫu huấn luyện; Y : ma trận nhãn.
- λ_1, λ_2: các tham số; k: số đặc trưng quan trọng.

Output:

- A (ma trận chiếu).
- X_k (Ma trận đặc trưng quan trọng)

Step 1: Tính S_b theo công thức (2.2); Tính và S_w theo công thức (2.3)

Step 2: Giải bài toán tối ưu (2.7) theo [132] để có ma trận chiếu A

Step 3: Tính $\|\mathrm{a}_i\|_2, i = 1,2,\dots,m$ của A

Step 4: Sắp xếp m dòng của X theo thứ tự giảm dần của $\|\mathrm{a}_i\|_2$. Xây dựng X_k gồm k dòng trên đỉnh của X.

Step 5: Return A và X_k

2.3.3. Mô hình học cho phân lớp

Để cải thiện độ chính xác của hệ thống tra cứu ảnh, trong phần này, luận án đề xuất mô hình học phân lớp. Ở đây cần hiểu rằng, nội dung này kế thừa giải pháp xử lý của vấn đề cỡ mẫu nhỏ trong Thuật toán 2.1 và tập trung vào giải quyết pha phân lớp của bài toán tra cứu ảnh với RF. Bên cạnh đó, một lưu ý nữa cũng cần được đề cập ở đây, đó là, nội dung phần này bao hàm các khối: "Học phân lớp", "Áp dụng

chiếu A", "Tập gia tăng" và "Mô hình học phân lớp C" trong mô hình trên Hình 2.1.

Việc giải quyết bài toán có cỡ lớp rất nhỏ, luận án đề xuất mô hình học phân lớp nhưng nó được thực hiện trên không gian đặc trưng gốc. Khi thực hiện phân lớp trên không gian đặc trưng gốc, gặp phải vấn đề về số chiều của không gian đặc trưng cao, do đó phải loại đi các đặc trưng dư thừa (xem Thuật toán 2.1). Tuy nhiên vẫn còn một vấn đề chưa được giải quyết đó là tập mẫu huấn luyện nhỏ và bị lệch.

Để giải quyết vấn đề tập mẫu huấn luyện nhỏ và bị lệch, luận án đề xuất cách giải quyết như sau. Ngay khi có được mô hình học chiếu A (đầu ra của Thuật toán 2.1), ta áp dụng mô hình chiếu A lên tập các véc tơ đặc trưng của CSDL ảnh và véc tơ đặc trưng của ảnh truy vấn để thu được các véc tơ phân biệt, cụ thể tính $Y = \mathrm{A}^T\mathrm{X}$. Cũng cần lưu ý ở đây rằng, ma trận chiếu A không phải học lại, do đó nó không tiêu tốn thời gian cho việc học lại. Trong không gian chiếu, ta nhận thấy rằng các véc tơ y_i, mà có vị trí gần với véc tơ ảnh truy vấn $\mathrm{y}_i^{(q)}$ hơn, sẽ có khả năng cao là cùng lớp với ảnh truy vấn, trái lại, những véc tơ y_i, mà có vị trí cách xa $\mathrm{y}_i^{(q)}$, sẽ có khả năng cao là khác lớp với ảnh truy vấn. Tận dụng thông tin này, trong đề xuất của luận án, các x_i, mà y_i gần với $\mathrm{y}_i^{(Q)}$, được chọn tự động để làm mẫu dương. Sau một số lần chọn như thế này, số mẫu dương sẽ tăng lên, do đó giải quyết được vấn đề tập mẫu bị lệch. Thuật toán phân lớp được tóm tắt trong Thuật toán 2.2 như sau:

Thuật toán 2.2: Xây dựng mô hình phân lớp

Input:

- Ma trận mẫu huấn luyện X, ma trận nhãn L
- Mô hình học chiếu A
- Ma trận đặc trưng quan trọng X_k
- Tập véc tơ đặc trưng F

Output:

Mô hình học phân lớp R

Step 1: Áp dụng mô hình học chiếu A lên tập véc tơ đặc trưng F.

Step 2: Xây dựng ma trận gia tăng $\mathrm{X}^{(e)}$ bao gồm e điểm x_i tương ứng với e điểm y_i mà là lân cận của $\mathrm{y}_i^{(q)}$. Xây dựng ma trận nhãn $\mathrm{L}^{(e)}$ bao gồm e nhãn dương của $\mathrm{x}_i \in \mathrm{X}^{(e)}$.

Step 3: Gộp ma trận $X^{(e)}$ vào ma trận X theo nguyên tắc cột đầu tiên của $X^{(e)}$ xếp ở bên phải cột cuối cùng của X. Tương tự trong việc gộp ma trận $L^{(e)}$ vào L.

Step 4: Huấn luyện phương pháp học phân lớp trên X và L để được mô hình phân lớp R.

Step 5: Return mô hình học phân lớp R.

2.3.4. Thuật toán tra cứu ảnh đề xuất

Trong tra cứu ảnh với RF, việc chọn tập con đặc trưng hợp lý để biểu thị thuộc tính ngữ nghĩa của các mẫu âm và dương là rất quan trọng trong việc xây dựng đối với mô hình RF hiệu quả (bao gồm cả phân lớp và phân hạng). Trong nội dung này, luận án đề xuất một thuật toán tra cứu ảnh tổng quát, cho phép lựa chọn một tập con đặc trưng hữu ích, tăng cường các mẫu dương và không bị ảnh hưởng bởi vấn đề về cỡ lớp nhỏ. Phần này bao gồm toàn bộ các khối và luồng được mô tả trên Hình 2.1.

Trong bước 2 (Step 2.2), Thuật toán tra cứu ảnh được đề xuất sử dụng Thuật toán 2.1 để giảm chiều và thu được tập đặc trưng quan trọng. Việc này giúp giải quyết vấn đề dữ liệu chiều cao và vấn đề cỡ lớp nhỏ (trong Thuật toán 2.2) của bài toán tra cứu ảnh với RF, mà sử dụng phân lớp. Bước 3 (Step 2.3) sử dụng Thuật toán 2.2 để giải quyết vấn đề cỡ lớp nhỏ, cỡ mẫu nhỏ và bị lệch. Thuật toán 2.3 được đề xuất như sau:

Thuật toán 2.3: SDAIR

Input:

F: tập đặc trưng của các ảnh CSDL, **q**: véc tơ đặc trưng ảnh truy vấn;

N: số các ảnh tại mỗi vòng lặp.

Output:

S: tập kết quả.

Step 1: Tra cứu ảnh với q để được tập kết quả khởi tạo và lấy N véc tơ ảnh ở top để được tập kết quả I

Step 2:

Repeat

Step 2.1: Người dùng phản hồi trên tập I để có tập phản hồi RF

Step 2.2: Thực hiện Thuật toán 2.1 để có ma trận đặc trưng quan trọng X_k

Step 2.3: Thực hiện Thuật toán 2.2 để có mô hình học phân lớp C

Step 2.4: Phân hạng tập đặc trưng F theo mô hình học phân lớp C để được danh sách phân hạng

Step 2.5: Lấy N ảnh ở trên TOP của danh sách phân hạng trong Step 2.4 làm tập ảnh kết quả S

Until (User stops responding)

Step 3: Return S.

2.4. Độ phức tạp tính toán

Trong Thuật toán 2.1, bước 2 (Step 2) có chi phí tính toán cao nhất. Trong bước này, chi phí tính toán là giải bài toán (2.7). Chi phí tính toán để tìm được nghiệm của bài toán (2.7) là $O(Inter(m^2n + m^3 + 2m^2d + d^3))$, trong đó m là số chiều của không gian gốc, n là số mẫu huấn luyện, d là số chiều của không gian chiếu, và $Inter$ là số vòng lặp. Bởi vì thuật toán tìm nghiệm của bài toán (2.7) thường hội tụ trong khoảng 10 vòng lặp do đó có thể xem $Inter$ là một hằng số và có thể bỏ qua [41]. Bài toán tra cứu ảnh với RF, số mẫu phản hồi n và số chiều d của không gian chiếu thường rất nhỏ, do đó có thể coi là hằng số. Như vậy, độ phức tạp tính toán của Thuật toán 2.1 là $\mathbf{O}(m^3)$.

Bởi vì Thuật toán 2.2 sử dụng chiếu đã được học ở Thuật toán 2.1 cho nên chi phí thời gian lớn nhất của Thuật toán 2.2 là chi phí thời gian cho bước 4 (Step 4). Độ phức tạp tính toán của Step 4 là chi phí thời gian để thực hiện một thuật toán phân lớp nào đó trên tập huấn luyện (X, L). Trong trường hợp dùng thuật toán Linear SVM [133], chi phí tính toán của Step 4 là $\mathbf{O}((n + e)k)$. Bởi vì, trong bài toán tra cứu ảnh với RF, số mẫu phản hồi $(n + e)$ thường rất nhỏ cho nên có thể coi $(n + e)$ là hằng số. Do đó, độ phức tạp tính toán của Thuật toán 2.2 là $\mathbf{O}(k)$.

Chi phí thời gian để thực hiện Thuật toán 2.3 là chi phí thời gian thực hiện bước 2 (Step 2). Trong Step 2, các bước có chi phí tính toán cao nhất là Step 2.2, Step 2.3 và Step 2.4. Chi phí tính toán của Step 2.2 là chi phí tính toán của Thuật toán 2.1, tức là $\mathbf{O}(m^3)$. Chi phí tính toán của Thuật toán 2.2 cũng chính là chi phí tính toán của Step 2.3, do đó nó là $\mathbf{O}(k)$. Step 2.4 thực hiện phân hạng l véc tơ đặc trưng trong tập đặc trưng F, vậy chi phí thời gian là $\mathbf{O}(l)$. Step 2 được lặp lại ir lần, tuy nhiên số vòng lặp trong quá trình phản hồi thường nhỏ (thấp hơn 5 lần), do đó có thể coi là hằng số. Như vậy, độ phức tạp tính toán của Thuật toán 2.3 là $\mathbf{O}(m^3 + l)$ với m và l là số chiều của dữ liệu gốc và số ảnh trong CSDL ảnh tương ứng.

2.5. Kết quả thực nghiệm

Trong phần này, luận án trình bày kết quả đánh giá thực nghiệm của phương pháp tra cứu ảnh được đề xuất. Kịch bản thực nghiệm thứ nhất là so sánh phương pháp đề xuất với các phương pháp tra cứu ảnh điển hình, nhằm chứng minh độ chính xác tổng thể của phương pháp được đề xuất cao hơn đối với các phương pháp còn lại. Thực nghiệm thứ hai nhằm kiểm tra hiệu quả của việc loại bỏ các đặc trưng dư thừa và không liên quan, đồng thời giải quyết vấn đề cỡ lớp nhỏ trên tập CIFAR-100. Độ đo mAP (trong 1.9.3) cũng được sử dụng để đánh giá độ chính xác của phương pháp tra cứu ảnh được đề xuất.

Phương pháp được đề xuất trong thực nghiệm sử dụng SVM cho khối phân lớp, tổng số truy vấn được thực hiện là 10.000 trên 100 chủ đề, trong đó, mỗi chủ đề sẽ được lấy ra 100 ảnh. 100 ảnh trên cùng được trả về cho người dùng là ảnh quan trọng bởi vì với mỗi khái niệm ngữ nghĩa trong tập ảnh bao gồm 100 ảnh cho kiểm tra. CSDL ảnh được sử dụng trong thực nghiệm này là CIFAR-100. Luận án thực hiện trên thư viện KERAS[3]. Các thực nghiệm được cài đặt trên môi trường Python 3 (cấu hình như trong 1.9.1), Bộ tối ưu (Optimizer) là Adam và hàm mất mát (Loss function) là sai số bình phương trung bình (MSE - Mean Squared Error).

2.5.1. Tập dữ liệu ảnh CIFAR-100

Tập dữ liệu ảnh CIFAR-100 có 60.000 ảnh màu (được mô tả trong 1.9.2.2 và Hình 1.8). Trong thực nghiệm, 10.000 ảnh được lấy làm tập các ảnh truy vấn. Tập ảnh truy vấn này được tạo ra thông qua việc lấy 100 ảnh từ mỗi trong 100 chủ đề. 50.000 ảnh còn lại được sử dụng làm tập huấn luyện.

2.5.2. Trích rút đặc trưng

Các đặc trưng trích rút từ tập CIFAR-100 được liệt kê trong Bảng 2.1 bao gồm:

Bảng 2.1. Các đặc trưng được trích rút từ tập CIFAR-100

TT	Tên đặc trưng	Số chiều
1	Color histogram	512
2	Color auto-correlogram	192

[3] https://github.com/fchollet/keras

3	Color moments	189
4	Gabor filters	64
5	Gray-level Co-occurrence matrix	24
6	HOG	324
Tổng số chiều		**1.305**

2.5.2.1. Lược đồ màu (Color histogram)

Véc tơ có độ dài 512 chiều được thể hiện như Hình 2.2 ở dưới.

	Topic	Example	27	28	29	35	...	474	475	476	482	483	484
88	tiger	panthera_tigris_s_000227.png	0.000977	0.028320	0.0	0.000000	...	0.0	0.000000	0.000000	0.0	0.000000	0.000000
34	fox	red_fox_s_000877.png	0.000000	0.000000	0.0	0.006836	...	0.0	0.000000	0.000000	0.0	0.000000	0.000000
84	table	table_s_000137.png	0.000000	0.000000	0.0	0.015625	...	0.0	0.003906	0.000977	0.0	0.044922	0.001953
25	couch	sofa_s_000902.png	0.000000	0.000000	0.0	0.069336	...	0.0	0.003906	0.004883	0.0	0.051758	0.006836
4	beaver	beaver_s_001511.png	0.000000	0.000000	0.0	0.031250	...	0.0	0.000000	0.000000	0.0	0.000000	0.000000
57	pear	pear_s_000775.png	0.000000	0.000000	0.0	0.000000	...	0.0	0.000000	0.000000	0.0	0.000000	0.000000
61	plate	plate_iron_s_001329.png	0.000000	0.000000	0.0	0.000000	...	0.0	0.010742	0.247070	0.0	0.045898	0.126953
16	can	oilcan_s_000747.png	0.000000	0.000977	0.0	0.000000	...	0.0	0.005859	0.030273	0.0	0.029297	0.233398
92	tulip	tulipa_clusiana_s_000224.png	0.006836	0.000000	0.0	0.049805	...	0.0	0.008789	0.003906	0.0	0.017578	0.012695
75	skunk	striped_skunk_s_000002.png	0.006836	0.000000	0.0	0.106445	...	0.0	0.018555	0.000000	0.0	0.000977	0.000000

Hình 2.2. Một số véc tơ đặc trưng theo Color histogram được trích rút

2.5.2.2. Tự tương quan màu (Color auto-correlogram)

Véc tơ có độ dài 192 chiều được thể hiện như Hình 2.3 ở dưới.

	Topic	Example	15	16	17	...	172	173	174	175	176
52	sp_ski	225018.jpg	0.222222	0.041667	0.000000	...	0.274272	0.133900	0.239583	0.091146	0.047743
18	obj_car	273063.jpg	0.395161	0.213710	0.030242	...	0.275093	0.132357	0.306569	0.154197	0.035280
58	texture_6	817001.jpg	0.117647	0.033088	0.012255	...	0.121693	0.077934	0.000000	0.000000	0.002193
63	wl_eagle	135033.jpg	0.000000	0.010417	0.003472	...	0.104167	0.000000	0.384615	0.185897	0.019231
54	texture_2	546027.jpg	0.346792	0.213081	0.104996	...	0.396373	0.165587	0.068966	0.025862	0.007184
66	wl_fox	109060.jpg	0.612342	0.503758	0.305314	...	0.000000	0.000000	0.000000	0.000000	0.000000
21	obj_dish	433097.jpg	0.000000	0.000000	0.000000	...	0.058333	0.000000	0.753058	0.652523	0.454829
51	sc_waves	312033.jpg	0.200000	0.100000	0.004167	...	0.000000	0.000000	0.815000	0.716875	0.360625
49	sc_sunset	345094.jpg	0.421053	0.187500	0.005482	...	0.000000	0.000000	0.000000	0.000000	0.000000
3	art_dino	644085.jpg	0.000000	0.000000	0.000000	...	0.583096	0.392163	0.760000	0.639052	0.407141

Hình 2.3. Một số véc tơ đặc trưng theo Color auto-correlogram được trích rút

2.5.2.3. Color moments

Véc tơ có độ dài 189 chiều được thể hiện như Hình 2.4 ở dưới.

	Topic	Example	0	1	2	...	184	185	186	187	188
6	bee	honeybee_s_001637.png	164.284180	133.436523	118.076172	...	0.481217	0.197864	1.745333	-0.665524	-0.330053
25	couch	couch_s_002287.png	119.235352	137.682617	117.459961	...	3.325591	4.449260	-16.339507	-0.769892	-3.504269
40	lamp	discharge_lamp_s_001196.png	160.379883	126.215820	129.949219	...	1.637493	1.560924	-66.770052	-1.380003	1.340385
74	shrew	sorex_cinereus_s_000170.png	68.692383	133.532227	112.991211	...	6.935113	6.590225	23.109494	6.085053	-6.004742
41	lawn_mower	power_mower_s_000577.png	221.184570	129.259766	124.196289	...	10.179112	23.402531	-78.053903	12.250742	-27.677006
46	man	man_s_000630.png	48.947266	128.000000	128.000000	...	0.000000	0.000000	4.292934	-0.000000	-0.000000
7	beetle	beetle_s_000380.png	153.963867	135.437500	120.772461	...	4.659859	4.311294	-41.415978	3.240352	-3.763453
44	lizard	gecko_s_002024.png	106.211914	134.007812	122.385742	...	0.806058	1.189990	-21.749318	-0.864734	1.163121
62	poppy	poppy_s_002182.png	144.912109	144.556641	99.530273	...	9.768755	13.441251	14.559486	9.338042	-15.498424
93	turtle	sea_turtle_s_001750.png	110.984375	91.949219	156.912109	...	13.973885	6.556200	46.072253	-12.133046	5.642789

Hình 2.4. Một số véc tơ đặc trưng theo Color moments được trích rút

2.5.2.4. Gabor filters

Véc tơ có độ dài 64 chiều được thể hiện như Hình 2.5 ở dưới.

	Topic	Example	0	1	2	...	59	60	61	62	63
95	whale	fin_whale_s_000110.png	254.996745	254.737630	191.373698	...	0.771955	0.000000	0.024490	0.678150	0.753221
53	orange	citrus_aurantium_s_001086.png	253.075521	250.978516	149.939779	...	1.606312	0.050779	0.218122	1.350261	1.551557
66	raccoon	raccoon_s_002505.png	253.161133	251.646810	169.723958	...	1.256774	0.335569	0.469198	1.116146	1.274577
88	tiger	panthera_tigris_s_000866.png	250.379232	248.671224	172.719401	...	1.049184	0.237501	0.417478	0.919868	1.086319
29	dinosaur	ornithischian_s_000319.png	208.895182	203.352539	152.831055	...	1.749578	1.691393	1.761789	1.222102	1.243386
39	keyboard	clavier_s_000216.png	255.000000	255.000000	204.311849	...	0.631697	0.000000	0.000000	0.557219	0.731992
2	baby	infant_s_000007.png	252.278646	250.336589	184.190755	...	1.183005	0.292943	0.380716	1.025588	1.168037
62	poppy	poppy_s_000474.png	254.760091	254.165690	155.232747	...	1.444014	0.018807	0.159171	1.140916	1.300357
76	skyscraper	skyscraper_s_001222.png	254.922201	254.601562	179.673503	...	0.834386	0.000000	0.111863	0.344892	0.581368
59	pine_tree	pine_s_000481.png	254.617513	253.076172	176.159505	...	1.225294	0.150515	0.265089	1.197328	1.275761

Hình 2.5. Một số véc tơ đặc trưng theo Gabor filters được trích rút

2.5.2.5. Gray-level Co-occurrence matrix

Véc tơ có độ dài 6 chiều được thể hiện như Hình 2.6 ở dưới.

	Topic	Example	0	1	2	...	19	20	21	22	23
96	willow_tree	willow_tree_s_002028.png	8.195565	12.532778	9.566532	...	0.001074	0.036646	0.032758	0.035882	0.032774
52	oak_tree	quercus_garryana_s_000447.png	13.379032	24.954214	23.316532	...	0.000709	0.030750	0.028412	0.027367	0.026621
89	tractor	dozer_s_000655.png	14.296371	21.942768	18.255040	...	0.000810	0.034979	0.029201	0.028972	0.028469
13	bus	bus_s_000008.png	15.996976	33.018730	25.591734	...	0.000674	0.029682	0.025648	0.026306	0.025962
8	bicycle	bike_s_000237.png	27.636089	30.949011	20.585685	...	0.000572	0.023951	0.023877	0.024517	0.023922
66	raccoon	raccoon_s_000786.png	13.142137	21.083247	15.970766	...	0.000661	0.026889	0.025658	0.026556	0.025700
4	beaver	beaver_s_001485.png	15.856855	19.543184	12.928427	...	0.000614	0.025212	0.024603	0.025181	0.024789
84	table	table_s_000489.png	5.535282	14.283039	12.177419	...	0.001160	0.041734	0.034457	0.036936	0.034054
88	tiger	panthera_tigris_s_000211.png	19.642137	27.328824	16.569556	...	0.001580	0.049013	0.038628	0.055034	0.039754
45	lobster	lobster_s_000040.png	22.992944	29.909469	23.826613	...	0.032794	0.222737	0.179527	0.200776	0.181090

Hình 2.6. Một số véc tơ đặc trưng theo Gray-level co-occurrence matrix được trích rút

2.5.2.6. Histogram of oriented gradients (HOG)

Véc tơ có độ dài 324 chiều được thể hiện như Hình 2.7 ở dưới.

	Topic	Example	0	1	2	...	319	320	321	322	323
22	clock	clock_s_002382.png	0.239091	0.267066	0.267066	...	0.159574	0.009587	0.047996	0.000000	0.176803
73	shark	mako_shark_s_001357.png	0.195074	0.063072	0.019464	...	0.241373	0.157168	0.083766	0.248552	0.292937
83	sweet_pepper	sweet_pepper_s_001779.png	0.267186	0.267186	0.267186	...	0.313111	0.108057	0.040891	0.018852	0.029822
24	cockroach	oriental_cockroach_s_000825.png	0.172477	0.159558	0.272679	...	0.043759	0.000000	0.000000	0.000000	0.042943
52	oak_tree	swamp_white_oak_s_000516.png	0.264969	0.264969	0.264969	...	0.283707	0.166598	0.068650	0.000000	0.042692
66	raccoon	raccoon_s_000193.png	0.016856	0.000862	0.003128	...	0.215635	0.081737	0.030327	0.012831	0.002930
87	television	television_s_002831.png	0.180942	0.087244	0.163470	...	0.057222	0.108552	0.034812	0.002569	0.084065
35	girl	female_child_s_000946.png	0.240817	0.118397	0.188784	...	0.227512	0.241976	0.128162	0.107824	0.241976
10	bowl	bowl_s_001292.png	0.291734	0.044494	0.004554	...	0.264626	0.027485	0.011192	0.004649	0.018280
19	cattle	cattle_s_001518.png	0.015313	0.000000	0.063160	...	0.203095	0.109021	0.178352	0.075533	0.072704

Hình 2.7. Một số véc tơ đặc trưng theo HOG được trích rút

2.5.3. Thực nghiệm về hiệu năng của phương pháp đề xuất

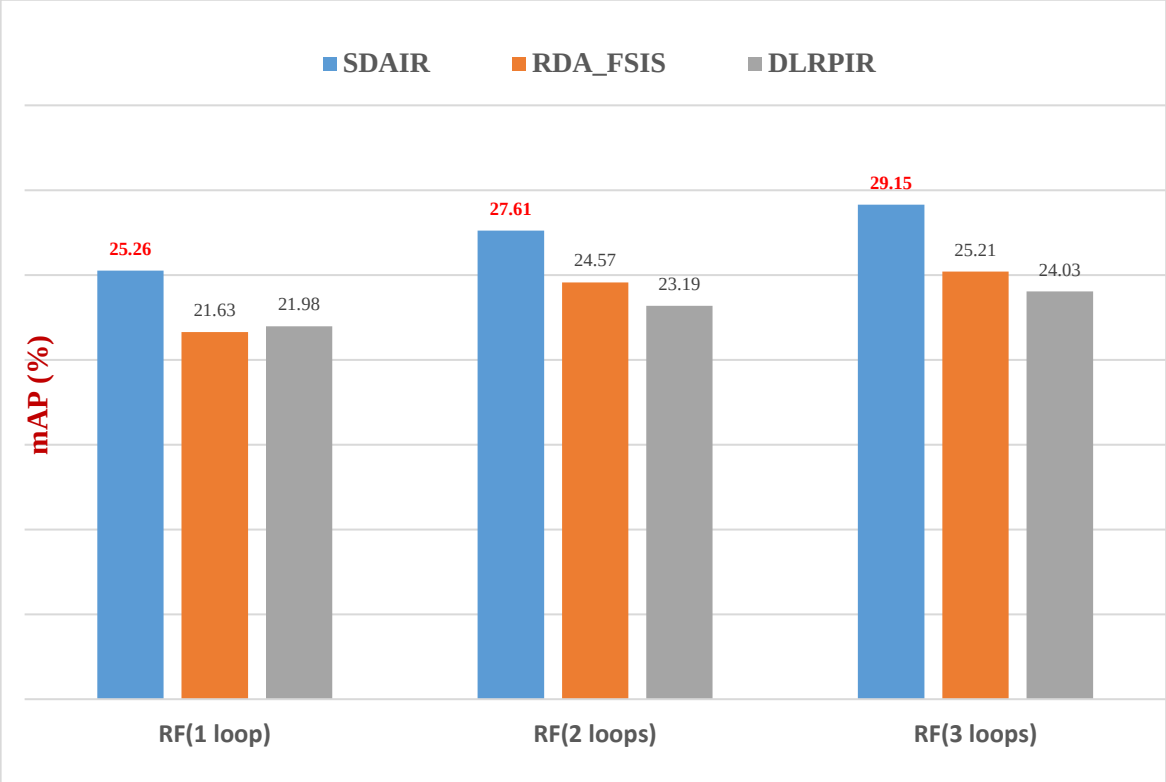
Phần này so sánh độ chính xác của phương pháp đề xuất với hai phương pháp khác là DLRPIR và RDA_FSIS [42] để chỉ ra độ chính xác của phương pháp được đề xuất. DLRPIR là phương pháp tra cứu ảnh sử dụng độ đo tương tự và cơ chế phản hồi tương tự như phương pháp đề xuất. Điều khác biệt so với phương pháp đề xuất trong luận án là DLRPIR sử dụng phương pháp chiếu hạng thấp phân biệt (DLRP - Discriminative Low-Rank Projection) [107] để chiếu dữ liệu gốc sang một không gian chiếu, sau đó thực hiện phân lớp trên không gian chiếu này để phân hạng các ảnh. Lý do mà luận án sử dụng phương pháp DLRPIR cho so sánh là bởi vì nó có thể giảm nhẹ vấn đề cỡ lớp nhỏ. Phương pháp RDA_FSIS được sử dụng cho so sánh là bởi vì nó không những là một phương pháp tốt, nó còn học một ma trận chiếu phân biệt thưa để trích rút đặc trưng như phương pháp đề xuất trong luận án.

Trong quá trình thực hiện thực nghiệm, luận án đã thiết lập giá trị của các tham số cho phương pháp như sau: Trong DLRP, luận án thiết lập giá trị của cả hai tham số λ và α đều được chọn là 10^{-4} bởi vì giá trị này nằm trong khoảng tối ưu [107], và luận án cũng chọn số chiều của không gian con là 128 bởi vì 128 nằm trong khoảng giá trị tối ưu của DLRP và cùng số chiều tối ưu với phương pháp đề xuất trong luận án. Trong phương pháp đề xuất, giá trị của các tham số λ_1 và λ_2 được chọn là 0.001 bởi vì đây là giá trị tối ưu được khuyên dùng [41]. Số chiều của phương pháp RDA_FSIS được chọn là 128 bởi vì giá trị này là xấp xỉ số chiều tối ưu. Trong phương pháp đề xuất, số chiều được chọn là 128 bởi vì giá trị này là số chiều mà nhiều bài báo khuyên dùng do vấn đề tốc độ truy vấn [42].

2.5.3.1. Kiểm tra hiệu năng toàn bộ của phương pháp đề xuất

Bắt đầu, hệ thống trả về tập kết quả tra cứu khởi tạo gồm 100 ảnh. Trên tập 100 ảnh này, người dùng sau đó đánh dấu các ảnh cùng chủ đề hay khác chủ đề với ảnh truy vấn để thu được tập phản hồi (bao gồm cả mẫu dương và mẫu âm). Quá trình phân hạng lại các ảnh trong tập ảnh CSDL được thực hiện qua một trong ba phương pháp là DLRPIR, RDA_FSIS và SDAIR. Sau khi phân hạng các ảnh, ta thu được ba tập kết quả tra cứu của ba phương pháp tương ứng. Độ chính xác trung bình của ba phương pháp trong top 100 ảnh sau ba lần lặp đầu tiên được thể hiện trên Hình 2.8. Trong tra cứu ảnh với RF, các lần lặp đầu tiên là rất quan trọng, vì vậy luận án đánh giá độ chính xác cho ba lần lặp đầu tiên. Ngoài ra, ở đây cần hiểu rằng, độ chính xác TB được tính trên tất cả các truy vấn, mỗi truy vấn tương ứng với một ảnh trong tập ảnh kiểm tra CIFAR-100.

Các kết quả trên Hình 2.8 cho thấy rằng, độ chính xác của phương pháp RDA_FSIS cao hơn DLRPIR là bởi vì nó học được một ma trận chiếu phân biệt thưa theo cấu trúc của từng lớp và giảm vấn đề cỡ lớp nhỏ. Phương pháp đề xuất có độ chính xác cao nhất trong ba phương pháp được so sánh, bởi vì nó đã loại bỏ được các đặc trưng dư thừa và không liên quan. Bên cạnh đó, nó cũng giải quyết hiệu quả vấn đề cỡ lớp rất nhỏ.



Hình 2.8. mAP của ba phương pháp trên top 100

2.5.3.2. Thực nghiệm về hiệu quả tra cứu ảnh khi loại bỏ các đặc trưng dư thừa và giải quyết vấn đề cỡ lớp nhỏ

Bởi vì hiệu quả của phương pháp đề xuất dựa vào việc loại đi các đặc trưng dưa thừa và không liên quan, do đó luận án chỉ ra điều này bằng thực nghiệm trên tập CIFAR-100 ở phần này.

Với mỗi chủ đề, ta huấn luyện một ma trận Q, tức là: với một chủ đề, ta lấy 500 mẫu dương (bởi vì trong tập ảnh CIFAR-100, mỗi chủ đề có 500 ảnh cho huấn luyện) và lấy 99 mẫu âm (bởi vì ta lấy một ảnh của mỗi trong 99 chủ đề còn lại). Do đó, đối với tập ảnh CIFAR-100, sẽ có 100 Q được huấn luyện.

Luận án thiết kế các kịch bản thực nghiệm như sau:

Kịch bản (**1**): So sánh hiệu quả tra cứu ảnh mà không sử dụng phản hồi (chỉ sử dụng khoảng cách Euclide) trên không gian gồm 1,305 chiều và trên không gian gốc nhưng loại đi các chiều dư thừa và không quan trọng.

Kịch bản (**2**): So sánh hiệu quả tra cứu ảnh mà không sử dụng phản hồi (chỉ sử dụng khoảng cách Euclide) trên không gian gốc (nhưng loại đi các chiều dư thừa và không quan trọng) và trên không gian chiếu.

Kịch bản (**3**): So sánh hiệu quả tra cứu ảnh mà sử dụng phản hồi trên các không gian bao gồm: (1) không gian gốc ban đầu (có 1,305 chiều); (2) không gian gốc (nhưng loại đi các chiều dư thừa và không quan trọng); và (3) không gian chiếu. Trong kịch bản này, SVM được sử dụng trong phương pháp tra cứu ảnh với RF, sau đó sử dụng mô hình SVM này để phân hạng các ảnh và thu về tập kết quả tra cứu.

Số chiều mà luận án thực nghiệm trong cả ba kịch bản ở trên bao gồm: 30 chiều gốc (loại đi 1,275 chiều gốc), 20 chiều gốc (loại đi 1,285 chiều gốc), và 10 chiều gốc (loại đi 1,295 chiều gốc). Luận án chọn số chiều là 128 là bởi vì nó là số chiều được nhiều bài báo khuyên nên dùng và việc tính toán khoảng cách giữa hai véc tơ có số chiều lớn hơn 128 sẽ làm quá trình tra cứu tốn nhiều thời gian hơn.

Khi thực nghiệm tra cứu ảnh với RF trên không gian gốc, luận án ký hiệu là OIRRF (trên không gian chiếu là PIRRF), trong trường hợp không có RF trên không gian gốc, luận án ký hiệu là OIR (trên không gian chiếu là PIR).

Khi áp dụng OIRRF trên không gian đặc trưng gốc, mà gồm 1305 chiều, luận án gọi là OIRRF_1305. Cũng OIRRF nhưng áp dụng trên tập đặc trưng quan trọng (kết quả của Thuật toán 2.1), mà gồm i chiều (với i=128, 30, 20, và 10), luận án gọi là OIRRF_i (trên không gian chiếu là PIRRF_i) với giá trị i là số chiều tương ứng.

Các Bảng 2.2, 2.3 và 2.4 là kết quả tương ứng với các kịch bản (1), (2), và (3).

Nhìn vào Bảng 2.2 cho thấy rằng, độ chính xác khi lựa chọn 128 chiều là cao nhất trong số các chiều gồm 128, 30, 20, và 10. Lý do của điều này là do, khi ta chọn số chiều quá ít, thông tin sẽ bị mất mát nhiều và dẫn đến độ chính xác giảm. Đặc biệt từ bảng này cho thấy rằng, độ chính xác tại số chiều 128 cao hơn số chiều 1350. Nguyên nhân của kết quả này là do trong số 1350 chiều ban đầu có nhiều chiều là dư thừa và không liên quan, chúng làm cho độ chính xác thấp hơn. Điều này là minh chứng để khẳng định hiệu quả khi loại bỏ các đặc trưng không liên quan và dư thừa của phương pháp đề xuất.

Bảng 2.2. Kết quả tra cứu ảnh theo kịch bản (1)

Phương pháp	**OIR_i**				
Số chiều	1305	128	30	20	10
mAP(%)	16,07	**18,27**	16,63	16,15	15,6

Quan sát Bảng 2.3, cho thấy rằng độ chính xác của phương pháp đề xuất trên không gian gốc là cao hơn độ chính xác trên không gian chiếu ở tất cả các chiều bao gồm 128, 30, 20, và 10. Lý do của việc này là bởi vì trên không gian gốc, có thể xác định được đặc trưng nào là quan trọng nhất để giữ lại trong khi trên không gian chiếu, ta không thể biết được đặc trưng nào là đặc trưng quan trọng để giữ lại, dẫn đến giữ lại những đặc trưng ít quan trọng nhưng có thể loại đi những đặc trưng quan trọng.

Bảng 2.3. Kết quả tra cứu ảnh theo kịch bản (2)

Phương pháp	**OIR_i**				**PIR_i**			
Số chiều	128	30	20	10	128	30	20	10
mAP(%)	**18,27**	**16,63**	**16,15**	**15,6**	17,21	15,68	15,3	15,05

Số liệu trên Bảng 2.4 cho thấy rằng, ở các chiều 128, 30, 20, và 10, độ chính xác của phương pháp đề xuất trên không gian gốc luôn cao hơn trên không gian chiếu. Lý do của điều này là ngoài việc loại đi được các đặc trưng dưa thừa và không liên quan, nó còn giảm được sự ảnh hưởng của vấn đề cỡ lớp nhỏ. Cũng trong bảng này, cho thấy rằng, độ chính xác của phương pháp đề xuất trên không gian gốc với số chiều ban đầu thấp hơn phương pháp đề xuất với số chiều 128. Nguyên nhân cho việc này là do phương pháp mới đề xuất đã loại bỏ được các đặc trưng dư thừa và không liên quan có thể gây ảnh hưởng đến kết quả tra cứu.

Bảng 2.4. Kết quả tra cứu ảnh theo kịch bản (3)

Phương pháp	**OIRRF_i**					**PIRRF_i**			
Số chiều	1305	128	30	20	10	128	30	20	10
mAP(%)	20,3	**25,26**	20,56	19,16	18,63	20,9	19,76	18,96	18,63

Bảng 2.5 ở dưới chỉ ra thời gian truy vấn của phương pháp tra cứu ảnh trên không gian gốc và không gian chiếu. Từ bảng này, cho thấy rằng thời gian truy vấn với số chiều 1305 ban đầu là lâu nhất. Lý do của việc này là bởi vì nó có số chiều lớn, dẫn đến tính toán khoảng cách giữa hai véc tơ sẽ cần nhiều thời gian. Cũng từ bảng này, để thấy rằng, nếu cùng số chiều (128, 30, 20, hoặc 10) thì thời gian truy vấn trên không gian gốc là nhanh hơn trên không gian chiếu. Lý do của việc này là do khi truy vấn trên không gian chiếu, thuật toán phải tính tích của ma trận Q (ma trận chiếu) với ma trận dữ liệu X (ma trận biểu diễn ảnh) trong khi truy vấn trên không gian gốc, thì không cần phải tính tích này.

Bảng 2.5. Thời gian truy vấn ảnh theo số chiều trên không gian gốc và không gian chiếu

Phương pháp	**Thời gian chạy của OIR_i**					**Thời gian chạy của PIR_i**			
Số chiều	1305	128	30	20	10	128	30	20	10
Thời gian (s)	0.5531	**0.35**	0.20	0.19	0.18	0.44	0.49	0.42	0.34

2.6. Kết luận Chương 2

Chương này, luận án đã tiến hành phân tích hạn chế và các ưu điểm của một số phương pháp đã có để từ đó đề xuất ra một phương pháp tra cứu ảnh mới, có tên là SDAIR. Phương pháp này kế thừa các ưu điểm của phương pháp RSLDA, đề xuất một mô hình học bằng việc bổ sung một số hạng để phù hợp với các nhãn lớp và có thể tăng thuộc tính phân lớp của ma trận chiếu thu được. Phương pháp mới được đề xuất trong luận án có khả năng cải thiện độ chính xác tra cứu ảnh ngay cả khi cỡ lớp của tập huấn luyện có thể là rất nhỏ nhưng vẫn loại bỏ được các đặc trưng dư thừa. Phương pháp này khác với các hệ thống tra cứu ảnh hiện có, thường được liên kết với một độ đo tương tự hoặc mô hình học cụ thể. Nó đưa ra một mô hình linh hoạt hơn, và sử dụng cơ chế bổ sung mẫu dương vào tập huấn luyện một cách tự động, mà không yêu cầu số các mẫu dương phải đủ lớn. Ngoài ra, phương pháp này có thể phục vụ đồng thời hai nhiệm vụ: bổ sung mẫu huấn luyện dương và lựa chọn tập đặc trưng quan trọng, và chỉ cần được huấn luyện một lần. Kết quả thực nghiệm trên tập CIFAR-100 cho thấy phương pháp đề xuất cải thiện hiệu suất cho bài toán tra cứu ảnh với RF, ngay cả khi có cỡ mẫu nhỏ, cỡ lớp nhỏ, và dữ liệu có chiều cao. Các kết quả đóng góp của chương này đã được Nghiên cứu sinh công bố trong các công trình [**CT4, CT2**].

Một trong những vấn đề mà tại chương này vẫn chưa giải quyết được đó là hiệu năng tra cứu bị giới hạn do các đặc trưng được thiết kế thủ công, không thể biểu diễn các đặc tính ảnh theo một cách chính xác, cùng với sự thiếu hụt đối với các mẫu có nhãn do số mẫu phản hồi của người dùng khá hạn chế (số lượng mẫu có nhãn nhỏ). Trong chương tiếp theo, luận án sẽ trình bày một phương pháp hiệu quả cho tra cứu ảnh với RF. Phương pháp này sử dụng mô hình mạng nơ ron tích chập autoencoder kết hợp cả học không giám sát và có giám sát để tối ưu hóa và giải quyết các vấn đề đã đề cập ở trên.

Chương 3. HỌC CÁC BIỂU DIỄN ẢNH VỚI MẠNG NƠ RON TÍCH CHẬP SÂU AUTOENCODER CHO TRA CỨU ẢNH VỚI PHẢN HỒI LIÊN QUAN

Các vấn đề thường gặp trong tra cứu ảnh với RF truyền thống bao gồm: (1) khả năng biểu diễn hạn chế của các đặc trưng được thiết kế thủ công, và (2) chiều của dữ liệu ảnh là rất cao. Trong chương này, luận án đề xuất một phương pháp dựa trên mạng nơ ron tích chập sâu autoencoder cho tra cứu ảnh, có tên là AIR (Autoencoders for Image Retrieval). Phương pháp được đề xuất cho phép tự động học biểu diễn ảnh trực tiếp từ ảnh thô theo cách không giám sát. Bên cạnh đó, Phương pháp tận dụng cách tiếp cận không giám sát và có giám sát để nâng cao hiệu năng tra cứu. Các kết quả thực nghiệm chỉ ra rằng phương pháp được đề xuất cho kết quả tốt hơn một số phương pháp đã có trên tập ảnh CIFAR-100 gồm 60.000 ảnh.

3.1. Giới thiệu

Hệ thống CBIR có mục tiêu tìm kiếm các ảnh giống nhất với ảnh truy vấn thông qua phân tích nội dung ảnh. Do đó các biểu diễn ảnh và đo độ tương tự trở nên quan trọng trong CBIR. Để tạo ra tra cứu mạnh đối với các thay đổi hình học và trực quan, sự tương tự giữa các ảnh được tính toán dựa vào nội dung của các ảnh. Nội dung của các ảnh được thể hiện qua màu sắc, kết cấu, hình dạng,... được biểu diễn ở dạng của một bộ mô tả đặc trưng [134]. Sự tương tự giữa các véc tơ đặc trưng của các ảnh tương ứng được xem như sự tương tự giữa các ảnh. Do đó, hiệu năng của bất cứ phương pháp CBIR nào cũng phụ thuộc chính vào biểu diễn mô tả đặc trưng của ảnh. Bất cứ một phương pháp biểu diễn mô tả đặc trưng nào cũng đều được kỳ vọng là có khả năng phân biệt, mạnh và chiều thấp. Nhiều phương pháp biểu diễn mô tả đặc trưng đã được nghiên cứu để tính độ tương tự giữa hai ảnh cho CBIR. Biểu diễn mô tả đặc trưng sử dụng các dấu hiệu trực quan của các ảnh được lựa chọn thủ công dựa trên nhu cầu [14]. Các cách tiếp cận này cũng được gọi là mô tả đặc trưng được thiết kế thủ công. Hơn nữa, nhìn chung, các phương pháp này là học không giám sát do chúng không cần dữ liệu để thiết kế phương pháp biểu diễn đặc trưng. Đặc trưng được thiết kế thủ công cho tra cứu ảnh là lĩnh vực nghiên cứu rất tích cực. Tuy nhiên, hiệu năng của nó bị giới hạn do các đặc trưng được thiết kế thủ công khó có thể biểu diễn các đặc tính của ảnh theo một cách chính xác [135].

Từ thập kỷ qua, chúng ta đã thấy sự dịch chuyển của biểu diễn đặc trưng từ thiết kế thủ công sang dựa vào học, đặc biệt là sự xuất hiện của học sâu [136]. Trong sự dịch chuyển này, học dựa vào các mạng nơ ron tích chập đã thay thế cho biểu diễn đặc trưng thiết kế thủ công truyền thống. Học sâu là một kỹ thuật để học các đặc trưng trừu tượng từ dữ liệu mà quan trọng cho ứng dụng và tập dữ liệu [137]. Dựa vào loại của dữ liệu được xử lý, các kiến trúc khác nhau đã ra đời như mạng nơ ron nhân tạo, perceptron đa lớp [138]. Các mạng nơ ron tích chập (CNN- Convolutional Neural Network) cho dữ liệu ảnh [2], [68] và các mạng nơ ron hồi quy (RNN - Reurrent Neural Network) cho dữ liệu chuỗi thời gian [139]. Sự tiến triển đã được tạo ra cho tra cứu ảnh sử dụng năng lực của học sâu [140]. Trong Chương 2, luận án đã trình bày phương pháp tra cứu ảnh đề xuất, phương pháp cải tiến độ chính xác và tốc độ tra cứu ảnh trên các đặc trưng thủ công. Trong chương này, luận án tập trung vào khai thác năng lực của học sâu vào việc đề xuất phương pháp dựa trên mạng nơ ron tích chập sâu autoencoder cho tra cứu ảnh.

Các phát triển của học sâu là theo cách có giám sát và sẽ khó khăn khi áp dụng vào bài toán với số lượng mẫu có nhãn nhỏ như bài toán tra cứu ảnh với RF, trong đó số mẫu phản hồi của của người dùng là khá hạn chế. Từ khía cạnh của học sâu không giám sát, Hinton và Krizhevsky [141] đã đề xuất thuật toán autoencoder với ứng dụng cho tra cứu ảnh, mà sau đó được sử dụng cho một số nhiệm vụ khác như gióng hàng khuôn mặt [142]. Huấn luyện autoencoder không yêu cầu các mẫu có nhãn. Autoencoder có thể được xem như một mạng mã hóa thưa đa tầng. Mỗi nút trong mạng autoencoder có thể được xem như một nguyên mẫu của ảnh đối tượng. Từ tầng dưới cùng đến tầng trên cùng, nguyên mẫu chứa thông tin ngữ nghĩa phong phú và trở thành một biểu diễn tốt hơn. Sau khi mạng autoencoder được học, các hệ số thu được bởi tái cấu trúc ảnh dựa vào các nguyên mẫu được sử dụng như đặc trưng cho tra cứu và đối sánh ảnh. Bởi vì autoencoder có thể học đặc trưng một cách thích nghi để huấn luyện, nó có thể nhận được hiệu năng tốt cho tra cứu ảnh.

Tuy nhiên, các phương pháp tra cứu ảnh sử dụng autoencoder ở trên phải đối mặt với vấn đề khả năng phân biệt của các đặc trưng kém hơn bởi vì các mô hình thường được huấn luyện cho phân lớp trong khi tra cứu ảnh cần học các đặc trưng cho đối sánh. Bên cạnh đó, các phương pháp này cũng bị mất mát thông tin do lượng hóa đặc trưng [143]. Hơn nữa, các mạng nơ ron sâu thường gặp phải vấn đề

vanishing/exploding gradients (biến mất/bùng nổ đạo hàm) và quá trình hội tụ nhanh. Bởi vì các autoencoder có nhiều lớp tích chập và giải chập (convolutional and deconvolutional) nên bị mất mát thông tin và làm giảm hiệu năng khi tái cấu trúc các ảnh.

Để giải quyết các hạn chế được nêu ở trên, chương này của luận án đề xuất một phương pháp bán giám sát dựa trên mạng nơ ron tích chập autoencoder cho tra cứu ảnh có tên là AIR. Phương pháp AIR khắc phục được hai vấn đề: (1) khả năng phân biệt các đặc trưng kém của các phương pháp trước do được tích hợp cơ chế RF và phân hạng qua máy véc tơ hỗ trợ SVM và (2) giảm nhẹ vấn đề vanishing/exploding gradients và quá trình hội tụ nhanh thông qua việc sử dụng các kết nối tắt (shortcut connections) trong kiến trúc autoencoder và dẫn đến có thể sử dụng các autoencoder sâu.

3.2. Nghiên cứu liên quan

Thông qua học có giám sát, dữ liệu mẫu được chuyển từ đầu vào đến tầng trên cùng cho dự đoán. Bằng việc cực tiểu giá trị của hàm chi phí giữa giá trị mục tiêu và giá trị dự đoán, thuật toán lan truyền ngược được sử dụng để tối ưu các tham số kết nối giữa mỗi cặp tầng. Cụ thể, CNN [2] là một biến đổi dựa vào mạng nơ ron, mà được sử dụng để biểu diễn các đặc trưng qua học có giám sát. CNN thường được thực hiện trong phân tích ảnh, nhận dạng tiếng nói [144] và phân tích văn bản,.... Đặc biệt trong phân tích ảnh, CNN đã thu được thành công rất lớn như nhận dạng khuôn mặt [145], phân tích cảnh [146], phân đoạn ô [147], và phân đoạn tổn thương não [148].

Trong các cách tiếp cận học không giám sát, dữ liệu không có nhãn được sử dụng để học các đặc trưng, trong khi một lượng nhỏ dữ liệu có nhãn được sử dụng để điều chỉnh các tham số, như máy boltzmann giới hạn (RBM - Restricted Boltzmann Machine) [149], mạng niềm tin sâu (DBN - Deep Belief Network) [150], autoencoders và các autoencoder được xếp chồng [151]. Kumar và cộng sự đã đề xuất một cách tiếp cận autoencoder cho học đặc trưng không giám sát [152]. Kalleberg và cộng sự đề xuất một cách tiếp cận autoencoder tích chập để phân tích ảnh [153]. Li và cộng sự đã thiết kế một cách tiếp cận dựa vào RBM cho phân lớp [154].

Các autoencoder được phát triển để học các đặc trưng hiệu quả cho biểu diễn nội dung ảnh [155]. Nó khai thác một mạng nơ ron để học các biểu diễn của một mẫu

được cho để cực tiểu sai số tái cấu trúc. Học đặc trưng với các thuật toán học không giám sát nhằm tái cấu trúc các mẫu đầu vào dựa trên các luật được xác định trước. Autoencoder [156] có thể học các đặc trưng đại diện để tái cấu trúc các mẫu đầu vào với sai số tái trúc cực tiểu. Các autoencoder được tận dụng để kết hợp âm thanh (audio) và lời (lyrics) cho phân lớp tâm trạng âm nhạc [157]. Các autoencoder và biến thể của nó cũng đã được áp dụng vào học biểu diễn đa phương thức [158]. Các tác giả trong [155] đã đề xuất một mạng autoencoder để học biểu diễn ẩn giữa nội dung văn bản và trực quan, cực tiểu sai số học tương quan giữa các biểu diễn ẩn của hai phương thức. Các tác giả trong [159] đã tận dụng autoencoder khử nhiễu để học các đặc trưng đại diện theo cách không giám sát và đã áp dụng vào huấn luyện các mô hình phát hiện nổi trội từ dữ liệu ảnh thô. Tuy nhiên, các phương pháp này phải đối mặt với vấn đề khả năng phân biệt của các đặc trưng kém hơn bởi vì các mô hình thường được huấn luyện cho phân lớp trong khi tra cứu ảnh cần học các đặc trưng cho đối sánh. Bên cạnh đó, các phương pháp này cũng bị mất mát thông tin do lượng hóa đặc trưng [143]. Ngoài ra, các phương pháp này không tận dụng được kiến trúc sâu của các mạng nơ ron và quá trình huấn luyện của chúng hội tụ chậm.

3.3. Phương pháp đề xuất

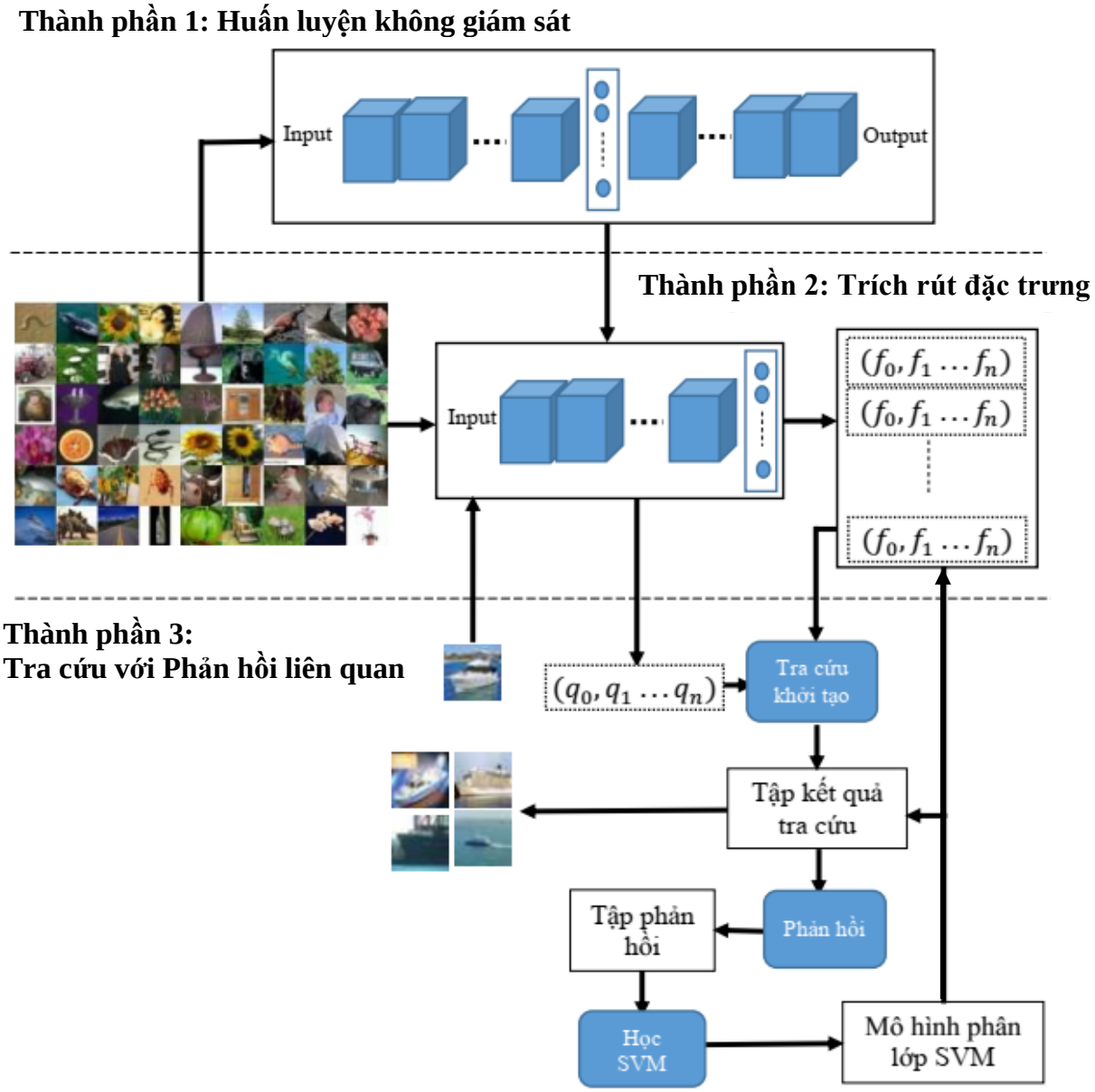
Phương pháp đề xuất gồm ba thành phần. Thành phần thứ nhất là huấn luyện không giám sát mạng nơ ron autoencoder sâu trên một tập con của tập ảnh. Thành phần thứ hai là áp dụng mô hình học từ thành phần thứ nhất encoder để trích rút các đặc trưng thấp chiều từ tập ảnh CSDL (ở đây cả thành phần thứ nhất và thành phần thứ hai đều được thực hiện ngoại tuyến (offline)). Thành phần thứ ba là tra cứu các ảnh tương tự với ảnh truy vấn dựa vào RF (Hình 3.1 chỉ ra mô hình của phương pháp được đề xuất). Mô hình mạng nơ ron tích chập sâu autoencoder được huấn luyện trên một tập con của tập ảnh CSDL. Trong trường hợp này, luận án sử dụng tập ảnh CIFAR-100.

3.3.1. Học các biểu diễn ảnh với mạng nơ ron tích chập sâu autoencoder

Phần này mô tả cấu trúc mạng nơ ron tích chập sâu autoencoder và huấn luyện các tham số.

Cách tiếp cận có giám sát là sẵn có cho học đặc trưng hướng dữ liệu, trong đó các trọng số kết nối được cập nhật thông qua thuật toán lan truyền ngược. So với cách tiếp cận học có giám sát, cách tiếp cận học không giám sát có thể nhận trực tiếp dữ

liệu đầu vào không có nhãn, làm giảm nhân lực cho việc gán nhãn. Autoencoder trích rút dữ liệu đầu ra để tái cấu trúc dữ liệu đầu vào, nó so sánh dữ liệu đầu vào với dữ liệu đầu vào gốc. Sau một số lần lặp, giá trị của hàm chi phí đạt đến mức tối ưu, mà có nghĩa là dữ liệu đầu vào tái cấu trúc có thể xấp xỉ dữ liệu đầu vào gốc.



Hình 3.1. Mô hình của phương pháp tra cứu ảnh đề xuất

3.3.1.1. Mạng nơ ron tích chập autoencoder

Mạng nơ ron tích chập autoencoder kết hợp tích chập kết nối cục bộ với Standard autoencoder, mà là một toán tử đơn giản để bổ sung đầu vào tái cấu trúc cho toán tử tích chập. Thủ tục chuyển đổi tích chập từ đầu vào bản đồ đặc trưng sang đầu ra được gọi là bộ giải mã tích chập. Sau đó, các giá trị đầu ra được tái cấu trúc thông qua toán tử tích chập ngược, mà được gọi là bộ mã hóa tích chập. Hơn nữa, thông qua

huấn luyện tham lam không giám sát autoencoder, các tham số của toán tử mã hóa và giải mã có thể được tính toán. Trong toán tử của tích chập autoencoder, $f(.)$ biểu diễn toán tử mã hóa tích chập và $f'(.)$ biểu diễn toán tử giải mã tích chập. Các bản đồ đặc trưng đầu vào p $\in \mathbb{R}^{n\times l\times l}$, mà thu được từ lớp đầu vào hoặc lớp trước đó. Nó chứa n bản đồ đặc trưng, và cỡ của mỗi đặc trưng là $l \times l$ pixel. Toán tử tích chập autoencoder bao gồm m nhân tích chập, và lớp đầu ra xuất ra m bản đồ đặc trưng. Khi các bản đồ đặc trưng đầu vào được sinh ra từ lớp đầu vào, n biểu diễn số các bản đồ đặc trưng đầu ra từ lớp trước. Cỡ của nhân tích chập là $d \times d$ với $d \leq l$.

Cho $\theta = \{W, \widehat{W}, b, \hat{b}\}$ biểu diễn các tham số của lớp tích chập autoencoder, mà cần được học. Trong đó, $W = \{w_j, j = 1,2, \dots, m\}$ và $b \in R^m$ biểu diễn các tham số của bộ mã hóa tích chập, ở đây $w_j \in R^{n\times l\times l}$ được xác định như một véc tơ $w_j \in R^{nl^2}$. Bên cạnh đó, $\widehat{W} = \{\widehat{w}_j, j = 1,2, \dots, m\}$ và $\hat{b}$ biểu diễn các tham số của bộ giải mã tích chập, ở đây $\hat{b} \in R^{nl^2}$, $w_j \in R^{1\times nl^2}$.

Đầu tiên, ảnh đầu vào được mã hóa mà mỗi thời điểm một mảng vá $d \times d$ pixel $p_i, i = 1,2, \dots k$, được lựa chọn ra từ ảnh đầu vào, và sau đó trọng số w_j của nhân chập j được sử dụng cho tính toán tích chập. Cuối cùng giá trị nơ ron $a_{ij}, j = 1,2, \dots, m$ được tính toán từ lớp đầu ra.

$$a_{ij} = f(p_i) = \sigma(w_j . p_i + b) \qquad \textbf{(3.1)}$$

Trong phương trình (3.1), σ là một hàm kích hoạt phi tuyến, trong phần này luận án sử dụng hàm RElu (Rectified Linear Function).

$$RElu(p) = \begin{cases} p \text{ nếu } p \geq 0 \\ 0 \text{ nếu } p < 0 \end{cases} \qquad \textbf{(3.2)}$$

Sau đó, đầu ra o_{ij} từ bộ giải mã tích chập được mã hóa mà p_i được tái cấu trúc qua a_{ij} để tạo ra $\hat{p}_i$.

$$\hat{p}_i = f'(a_{ij}) = \emptyset(w_i . a_{ij} + \hat{b}) \qquad \textbf{(3.3)}$$

$\hat{p}_i$ được tạo ra sau mỗi mã hóa và giải mã tích chập. Ta nhận được mảng vá P mà thu được từ toán tử tái cấu trúc. Sử dụng sai số bình phương trung bình giữa mảng vá gốc của ảnh đầu vào $p_i, i = 1,2, \dots k$ và mảng vá tái cấu trúc của ảnh $\hat{p}_i, i = 1,2, .., k$ như hàm chi phí. Ngoài ra, hàm chi phí được mô tả trong phương trình (3.4), và sai số tái cấu trúc được thể hiện trong phương trình (3.5).

$$L(\theta) = \frac{1}{k}\sum_{i=1}^{k} E(p_i, \hat{p}_i) \quad \textbf{(3.4)}$$

$$E(p_i, \hat{p}_i) = \|p_i - \hat{p}_i\|^2 = \|p_i - \emptyset(\sigma(p_i))\|^2 \quad \textbf{(3.5)}$$

3.3.1.2. Lớp pooling

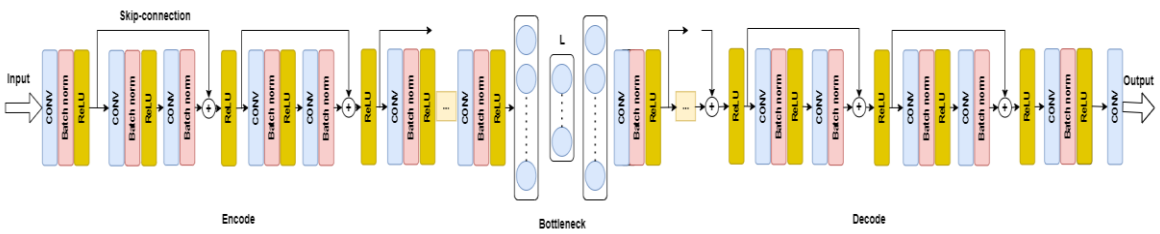
Giống như trong CNN, lớp tích chập được kết nối với lớp pooling [2]. Trong kiến trúc mạng nơ ron tích chập, lớp gộp cực đại (max pooling) được đặt sau lớp tích chập autoencoder, cụ thể:

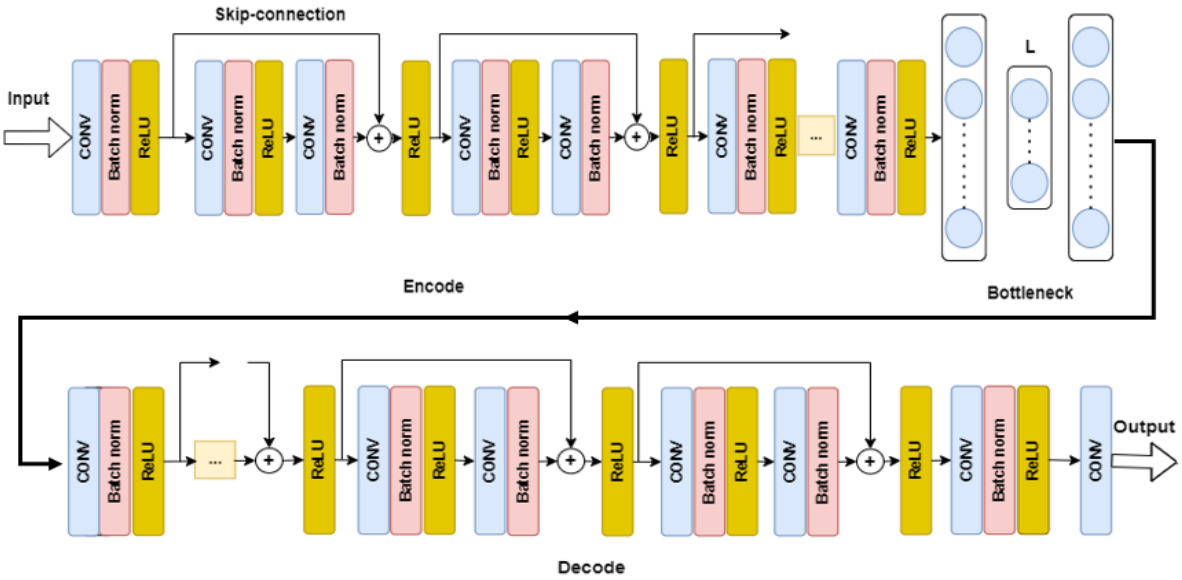
$$a_j^i = \max\,(p_j^i) \quad \textbf{(3.6)}$$

Bản đồ đặc trưng đầu vào được chia thành n vùng không giao nhau theo cỡ của vùng gộp. Trong phương trình (3.6), p_j^i biểu diễn vùng thứ i của bản đồ đặc trưng thứ j, và a_j^i biểu diễn nơ ron thứ i của bản đồ đặc trưng thứ j. Số các bản đồ đặc trưng đầu vào bằng với số các bản đồ đặc trưng đầu ra trong lớp gộp. Các nơ ron trong bản đồ đặc trưng có thể được giảm sau khi thực hiện toán tử gộp, do đó độ phức tạp tính toán cũng giảm theo. Trong phần thực nghiệm, luận án sẽ đánh giá thực nghiệm trên kiến trúc autoencoder bao gồm cả thiết lập có lớp gộp và không có lớp gộp.

3.3.1.3. Kiến trúc mạng tích chập autoencoder

Các mạng nơ ron sâu thường gặp phải vấn đề vanishing/exploding gradients và quá trình hội tụ nhanh. Bởi vì các autoencoder có nhiều lớp tích chập và giải chập nên bị mất mát thông tin và làm giảm hiệu năng khi tái cấu trúc các ảnh. Lấy cảm hứng từ mạng phần dư bao gồm các kết nối tắt [68], chúng ta bổ sung các kết nối tắt vào mạng autoencoder như trên Hình 3.2. Các kết nối này giúp gửi trực tiếp các bản đồ đặc trưng từ lớp đầu tiên của encoder đến một số lớp sau. Lý do của việc sử dụng kết nối tắt là bởi vì: thứ nhất, khi mạng quá sâu, các chi tiết ảnh có thể bị mất mát, trong khi các bản đồ đặc trưng được chuyển qua các kết nối tắt mang nhiều chi tiết ảnh. Thứ hai, các kết nối tắt giúp quá trình huấn luyện các mạng nơ ron sâu nhanh hội tụ hơn [68]. Ngoài ra, sử dụng kết nối tắt thu được lợi ích về trích rút đặc trưng cho tra cứu ảnh thông qua việc tăng độ sâu của mạng.





Hình 3.2. Kiến trúc mạng autoencoder đề xuất cho trích rút đặc trưng

3.3.1.4. Huấn luyện các tham số

Thông qua thuật toán giảm gradient (SGD - Stochastic Gradient Descent), các sai số đạt được cực tiểu, lớp tích chập autoencoder được tối ưu. Cuối cùng, các tham số đã huấn luyện được sử dụng để xuất ra các bản đồ đặc trưng mà được chuyển cho lớp tiếp theo.

Phần này, luận án sử dụng 50.000 mẫu không có nhãn để huấn luyện mạng nơ ron tích chập autoencoder thông qua học không giám sát ở lớp tích chập, gradient được tính toán thông qua hàm chi phí trong (3.4), và các tham số được tối ưu thông qua SGD. Mỗi một lô tối thiểu (mini batch) gồm 150 mẫu, và số các vòng lặp cho mỗi lô là 75. Số các kênh được thiết lập trong phương trình (3.2) cho bộ mã hóa tích chập và phương trình (3.3) cho bộ giải mã tích chập tương ứng.

3.3.2. Tra cứu ảnh với phản hồi liên quan dựa vào máy véc tơ hỗ trợ

3.3.2.1. Máy véc tơ hỗ trợ (SVM)

Trong phần này, luận án chọn máy véc tơ hỗ trợ SVM [160] cho việc phân lớp và phân hạng các ảnh. Lý do của việc chọn SVM là bởi vì: *Thứ nhất*, nó là một bộ phân lớp mạnh, đặc biệt cho phân lớp nhị phân, mà bài toán tra cứu ảnh với RF là bài toán có hai lớp. *Thứ hai*, thông qua siêu phẳng tối ưu tìm được, có thể sử dụng khoảng cách từ mỗi mẫu đến siêu phẳng tối ưu làm giá trị để phân hạng các ảnh.

Khoảng cách từ một điểm tới một siêu mặt phẳng. Trong không gian nhiều chiều: Khoảng cách từ một điểm tới siêu mặt phẳng (hyperplane) có phương trình $\mathrm{w}^T\mathrm{x} + b = 0$ được xác định bởi:

$$\frac{|\mathrm{w}^T\mathrm{x}+b|}{\|w\|_2} \quad \textbf{(3.7)}$$

Giả sử có tập mẫu huấn luyện như sau: $\{(x_i, y_i)\}_{i=1}^{N}$ và $y_i \in \{+1, -1\}$.
Ở đây x_i là một véc tơ n chiều và y_i là nhãn của lớp mà véc tơ thuộc về.
SVM là một phương pháp học phân lớp nhị phân rất hiệu quả. SVM tách hai lớp bởi một siêu phẳng.

$$\mathrm{w}^T\mathrm{x} + \mathrm{b} = 0 \quad \textbf{(3.8)}$$

Trong phương trình (3.8), x là một véc tơ đầu vào, w là một véc tơ trọng số, và b là độ lệch. SVM tìm các tham số w và b cho siêu phẳng tối ưu để cực đại lề $\frac{2}{\|\mathrm{w}\|}$, thỏa mãn:

$$y_i(\mathrm{w}^T\mathrm{x}_i + \mathrm{b}) \geq +1 \quad \textbf{(3.9)}$$

Nghiệm có thể tìm được thông qua bài toán đối ngẫu Lagrangian:

$$Q(\alpha) = \sum_{i=1}^{m} \alpha_i - \sum_{i,j=1}^{m} \alpha_i\, \alpha_j y_i y_j\, (\mathrm{x}_i.\mathrm{x}_j)/2 \quad \textbf{(3.10)}$$

Thoả mãn $\quad \alpha_i \geq 0, \sum_{i,j=1}^{m} \alpha_i\, y_j = 0$

Trong dạng đối ngẫu, các điểm dữ liệu chỉ xuất hiện dưới dạng tích vô hướng. Để nhận được biểu diễn dữ liệu tốt hơn, các điểm dữ liệu được ánh xạ sang một không gian tích vô hướng Hilbert thông qua một phép thế:

$$\mathrm{x}_i.\mathrm{x}_j \rightarrow \emptyset(\mathrm{x}_i).\emptyset(\mathrm{x}_j) = K(\mathrm{x}_i.\mathrm{x}_j) \quad \textbf{(3.11)}$$

ở đây K(.) là hàm nhân, ta nhận được công thức nhân của bài toán đối ngẫu Wolfe:

$$Q(\alpha) = \sum_{i=1}^{m} \alpha_i - \sum_{i,j=1}^{m} \alpha_i\, \alpha_j y_i y_j\, K(\mathrm{x}_i.\mathrm{x}_j)/2 \quad \textbf{(3.12)}$$

Do đó, với một hàm nhân được cho, bộ phân lớp SVM được cho bởi

$$F(x) = sgn(f(x)) \quad \textbf{(3.13)}$$

Ở đây $f(\mathrm{x}) = \sum_{i=1}^{l} \alpha_i y_i\, K(\mathrm{x}_i.\mathrm{x}) + b$ là hàm quyết định siêu phẳng đầu ra của SVM.

3.3.2.2. Tra cứu ảnh

Như mô hình phương pháp trên Hình 3.1, sau khi huấn luyện được mô hình mạng nơ ron tích chập autoencoder ở Thành phần 1, ta tiến hành bỏ phần decoder đi và giữ lại phần encoder để có mô hình học như trong Thành phần 2. Sử dụng mô hình

học trong Thành phần 2 của mô hình cho trích rút các véc tơ đặc trưng thấp chiều để thu được tập gồm n véc tơ đặc trưng $(f_0, f_1 \ ... \ f_n)$.

Trong quá trình tra cứu như trong Thành phần 3 của mô hình, người dùng cung cấp một ảnh truy vấn q, véc tơ của ảnh truy vấn sẽ được đưa qua mô hình học encoder để có véc tơ đặc trưng của ảnh truy vấn $(q_0, q_1 \ \ldots . q_n)$. Quá trình tra cứu khởi tạo sẽ so sánh (dùng khoảng cách Euclide) véc tơ của ảnh truy vấn với véc tơ của mỗi ảnh CSDL để thu được tập kết quả tra cứu. Trên tập kết quả này, người dùng phản hồi để thu được tập phản hồi (tập phản hồi này bao gồm các mẫu có nhãn âm và dương, nó cũng là tập huấn luyện). Học SVM được áp dụng trên tập huấn luyện để thu được mô hình phân lớp SVM. Áp dụng mô hình phân lớp trên tập véc tơ đặc trưng ảnh CSDL: những ảnh được dự đoán có nhãn dương mà có khoảng cách xa nhất từ siêu phẳng tối ưu) sẽ được xếp ở vị trí số một của danh sách kết quả, những ảnh được dự đoán có nhãn dương mà có khoảng cách xa thứ nhì từ siêu phẳng tối ưu sẽ được xếp ở vị trí số hai của danh sách kết quả,... Quá trình này lặp đi lặp lại cho đến khi người dùng dừng phản hồi.

3.4. Đánh giá thực nghiệm

Trong phần này, luận án sẽ trình bày thực nghiệm để đánh giá hiệu năng của phương pháp đề xuất và trong thực nghiệm này, CSDL ảnh được sử dụng là Corel *(trong 1.9.2.1)* và CIFAR-100 *(trong 1.9.2.2)*. Trong CIFAR-100, có 60.000 ảnh, 10.000 ảnh được lấy làm tập các ảnh truy vấn (tập ảnh truy vấn này được tạo ra thông qua việc lựa chọn ngẫu nhiên 100 ảnh từ mỗi trong 100 lớp), 50.000 ảnh còn lại được sử dụng làm tập huấn luyện.

Luận án thực hiện trên thư viện KERAS[4]. Các thực nghiệm được cài đặt trên môi trường Python 3 và Windows 11 (*cấu hình máy trong 1.9.1*). Bộ tối ưu (Optimizer) là Adam và hàm mất mát (Loss function) là sai số bình phương trung bình (MSE - Mean Squared Error).

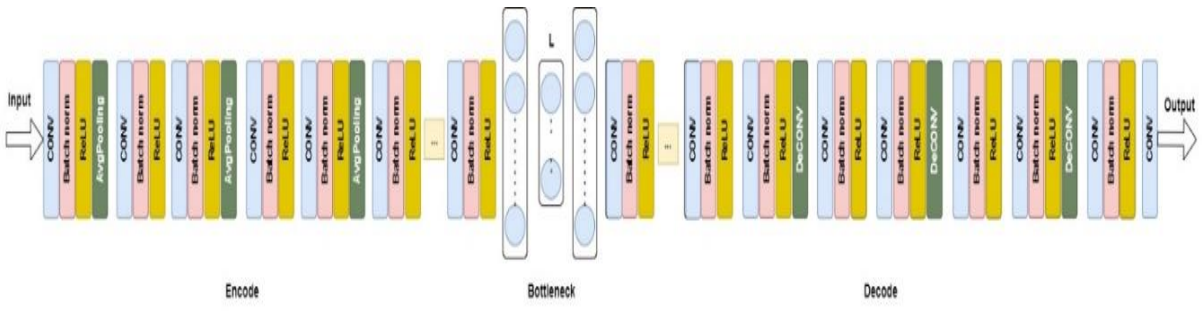
Độ đo được sử dụng trong phần đánh giá thực nghiệm này là độ chính xác trung bình (AP-Average Precision) và độ đo tổng hợp kết quả của nhiều truy vấn (mAP-Mean Average Precision). AP lớn hơn hàm ý rằng đường cong triệu hồi chính xác cao hơn và độ chính xác tra cứu tốt hơn. AP và mAP được tính theo công thức (1.25) và (1.26) (trong 1.9.3).

[4] https://github.com/fchollet/keras

3.4.1. Các kết quả trên tập dữ liệu ảnh CIFAR-100

Để chứng tỏ tính hiệu quả của kiến trúc autoencoder đề xuất cho tra cứu ảnh, luận án thực nghiệm tra cứu ảnh theo khoảng cách Euclide (sau đây gọi là Baseline) như sau:

Tại lúc bắt đầu tra cứu, khoảng cách Euclide trong không gian 128 chiều được sử dụng để phân hạng các ảnh trong CSDL. Lý do mà luận án sử dụng véc tơ đặc trưng 128 chiều là bởi vì đây là số chiều phù hợp với nhiều phương pháp tra cứu ảnh [161], [162]. Kết quả tra cứu tại lúc ban đầu (khi chưa có phản hồi) theo các độ sâu khác nhau trên tập CIFAR-100 được thể hiện trên Hình 3.3 ở dưới. Hình 3.3 thể hiện mAP của 5 cấu hình khác nhau cho 10, 20, 40, và 60 layer. Cấu hình thứ nhất, gọi là Classic (xem Hình 3.3), là một autoencoder sử dụng lớp pooling và không có shortcut connections. Cấu hình thứ hai, gọi là Shortcut (con-decon), là một autoencoder sử dụng lớp pooling, và có shortcut connections, mà là kết nối đối xứng [163]. Cấu hình thứ ba, gọi là NoP_ shortcut (con-decon), là một autoencoder không sử dụng lớp pooling và có shortcut connections, và nó cũng là kết nối đối xứng (xem Hình 3.4). Cấu hình thứ tư, gọi là Shortcut, là một autoencoder sử dụng lớp pooling và có shortcut connections nhưng nó không kết nối đối xứng (xem Hình 3.2). Cấu hình cuối cùng, gọi là NoP_ shortcut, là một autoencoder không sử dụng lớp pooling và có shortcut connections và nó cũng không kết nối đối xứng (xem Hình 3.2). Lý do mà luận án thực nghiệm với kiến trúc mạng theo cấu hình có cả lớp pooling và không có pooling là bởi vì muốn kiểm tra hiệu năng trên các ảnh có kích thước nhỏ như các ảnh trong tập ảnh CIFAR-100 (32×32). Các cấu hình mạng ở mức 40 layer được thể hiện ở dưới như sau:

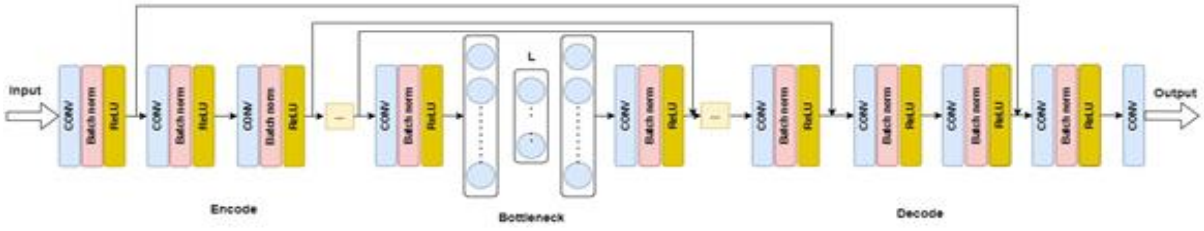


Hình 3.3. Kiến trúc mạng autoencoder chuẩn với lớp pooling

Bảng 3.1. Các tham số của kiến trúc mạng autoencoder chuẩn với lớp pooling (trên Hình 3.3)

Part	Layer Block	Layer Type	Window size	Stride	Padding	Input	Output
Encode	1	Conv2D BatchNorm ReLU AvgPooling2D	3	1	1	(32, 32, 3) (32, 32, 64)	(32, 32, 64) (16, 16, 64)
	2	Conv2D BatchNorm ReLU	3	1	1	(16, 16, 64)	(16, 16, 72)
	3	Conv2D BatchNorm2D ReLU AvgPooling2D	3	1	1	(16, 16, 72) (16, 16, 72)	(16, 16, 72) (8, 8, 72)
	4	Conv2D BatchNorm ReLU	3	1	1	(8, 8, 72)	(8, 8, 80)
	…						
	20	Conv2D BatchNorm ReLU	3	1	1	(4, 4, 136)	(4, 4, 144)
Bottleneck		Flatten				(4, 4, 144)	(2304,)
		Dense				(2304,)	(128,)
		Dense				(128,)	(2304,)
		Reshape				(2304,)	(4, 4, 144)
Decode	21	Conv2D BatchNorm ReLU	3	1	1	(4, 4, 144)	(4, 4, 144)

	22	Conv2D BatchNorm ReLU	3	1	1	(4, 4, 144)	(4, 4, 136)
	…						
	36	ConvTrans2D Conv2D BatchNorm ReLU	3 3	2 1	1 1	(4, 4, 88) (8, 8, 88)	(8, 8, 88) (8, 8, 80)
	37	Conv2D BatchNorm ReLU	3	1	1	(8, 8, 80)	(8, 8, 80)
	38	ConvTrans2D Conv2D BatchNorm ReLU	3 3	2 1	1 1	(8, 8, 80) (16, 16, 80)	(16, 16, 80) (16, 16, 72)
	39	Conv2D BatchNorm ReLU	3	1	1	(16, 16, 72)	(16, 16, 72)
	40	ConvTrans2D Conv2D BatchNorm ReLU Conv2D	3 3 3	2 1 1	1 1 1	(16, 16, 72) (32, 32, 72) (32, 32, 64)	(32, 32, 72) (32, 32, 64) (32, 32, 3)



Hình 3.4. Kiến trúc mạng autoencoder với kết nối tắt đối xứng (Symmetry Shortcut Connections) [163]

Bảng 3.2. Các tham số của kiến trúc mạng autoencoder với kết tối tắt đối xứng (trên Hình 3.4)

Part	Layer Block	Layer Type	Size	Stride	Padding	Input	Output	Connected
Encode	1	Conv2D BatchNorm ReLU	3	1	1	(32, 32, 3)	(32, 32, 32)	
	2	Conv2D BatchNorm ReLU	3	1	1	(32, 32, 32)	(32, 32, 32)	
	3	Conv2D BatchNorm2D ReLU	3	1	1	(32, 32, 32)	(32, 32, 32)	
	4	Conv2D BatchNorm ReLU	3	1	1	(32, 32, 32)	(32, 32, 32)	
	…							
	20	Conv2D BatchNorm ReLU	3	1	1	(32, 32, 32)	(32, 32, 32)	

Bottleneck		Flatten				(32, 32, 32)	(32768,)	
		Dense				(32768,)	(128,)	
		Dense				(128,)	(32768,)	
		Reshape				(32768,)	(32, 32, 32)	
Decode	21	Conv2D BatchNorm ReLU	3	1	1	(32, 32, 32)	(32, 32, 32)	
	22	Concatenate Conv2D BatchNorm ReLU	3	1	1	(32, 32, 32)	(32, 32, 32)	ReLU (Block 19)
	23	Conv2D BatchNorm ReLU	3	1	1	(32, 32, 32)	(32, 32, 32)	
	24	Concatenate Conv2D BatchNorm ReLU	3	1	1	(32, 32, 32)	(32, 32, 32)	ReLU (Block 17)

	…							
	39	Conv2D BatchNorm ReLU	3	1	1	(32, 32, 32)	(32, 32, 32)	
	40	Concatenate Conv2D BatchNorm ReLU Conv2D	3 3	1 1	1 1	(32, 32, 32) (32, 32, 32)	(32, 32, 32) (32, 32, 3)	ReLU (Block 1)

Bảng 3.3. Các tham số của kiến trúc mạng autoencoder với kết nối tắt đề xuất (trên Hình 3.2)

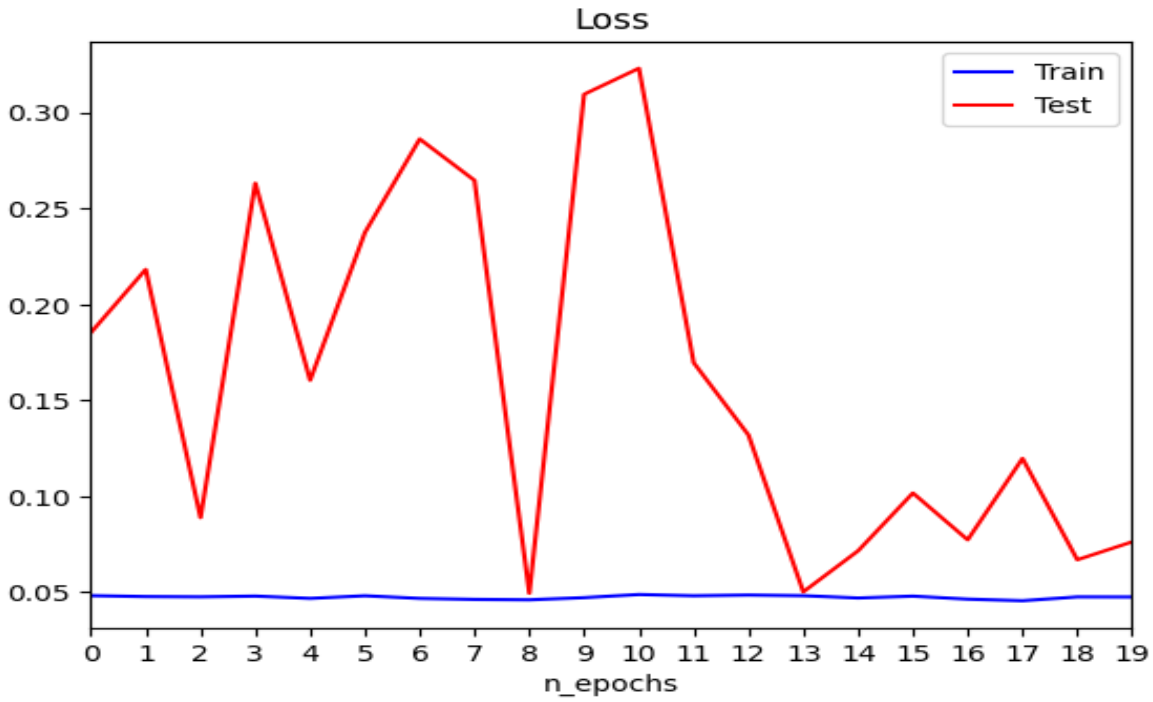
Part	Layer Block	Layer Type	Size	Stride	Padding	Input	Output	Connected
Encode	1	Conv2D BatchNorm ReLU	3	1	1	(32, 32, 3)	(32, 32, 32)	
	2	Conv2D BatchNorm ReLU	3	1	1	(32, 32, 32)	(32, 32, 32)	

	3	Conv2D BatchNorm2D Add ReLU	3	1	1	(32, 32, 32)	(32, 32, 32)	ReLU (Block 1)
	4	Conv2D BatchNorm ReLU	3	1	1	(32, 32, 32)	(32, 32, 40)	
	…							
	20	Conv2D BatchNorm ReLU	3	1	1	(32,32,104)	(32,32,104)	
Bottleneck		Flatten				(32,32,104)	(106496,)	
		Dense				(106496,)	(128,)	
		Dense				(128,)	(106496,)	
		Reshape				(106496,)	(32,32,104)	
Decode	21	Conv2D BatchNorm ReLU	3	1	1	(32,32,104)	(32,32,104)	

	22	Conv2D BatchNorm ReLU	3	1	1	(32,32,104)	(32,32,104)	
	23	Conv2D BatchNorm Add ReLU	3	1	1	(32,32,104)	(32,32,104)	ReLU (Block 21)
	24	Conv2D	3	1	1	(32,32,104)	(32, 32, 96)	
		BatchNorm						
		ReLU						
	…							
	39	Conv2D BatchNorm Add ReLU	3	1	1	(32, 32, 40)	(32, 32, 40)	ReLU (Block 37)
	40	Conv2D BatchNorm ReLU Conv2D	3 3	1 1	1 1	(32, 32, 40) (32, 32, 32)	(32, 32, 32) (32, 32, 3)	

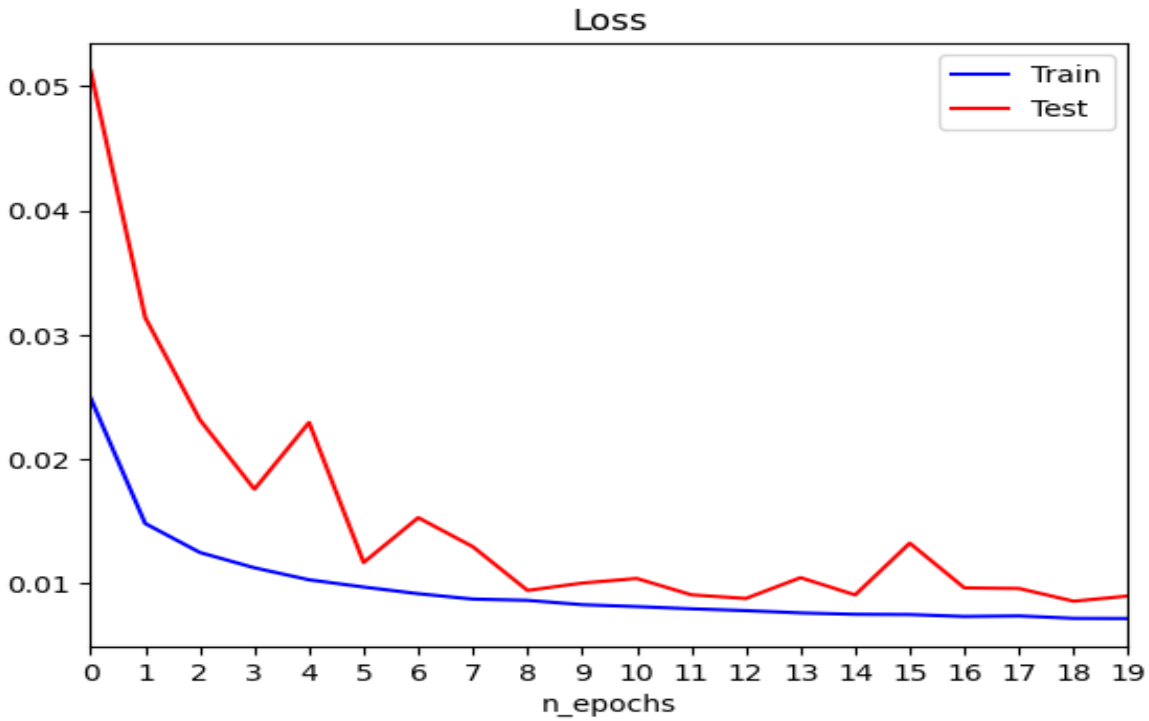
Việc huấn luyện mạng trong thực nghiệm ở chương này, luận án sử dụng bộ trọng số khởi tạo (không sử dụng học truyền). Thông tin về loss và val_loss cho 20 epoch trong quá trình huấn luyện các cấu hình mạng được thể hiện như sau:

- Autoencoder Classic:



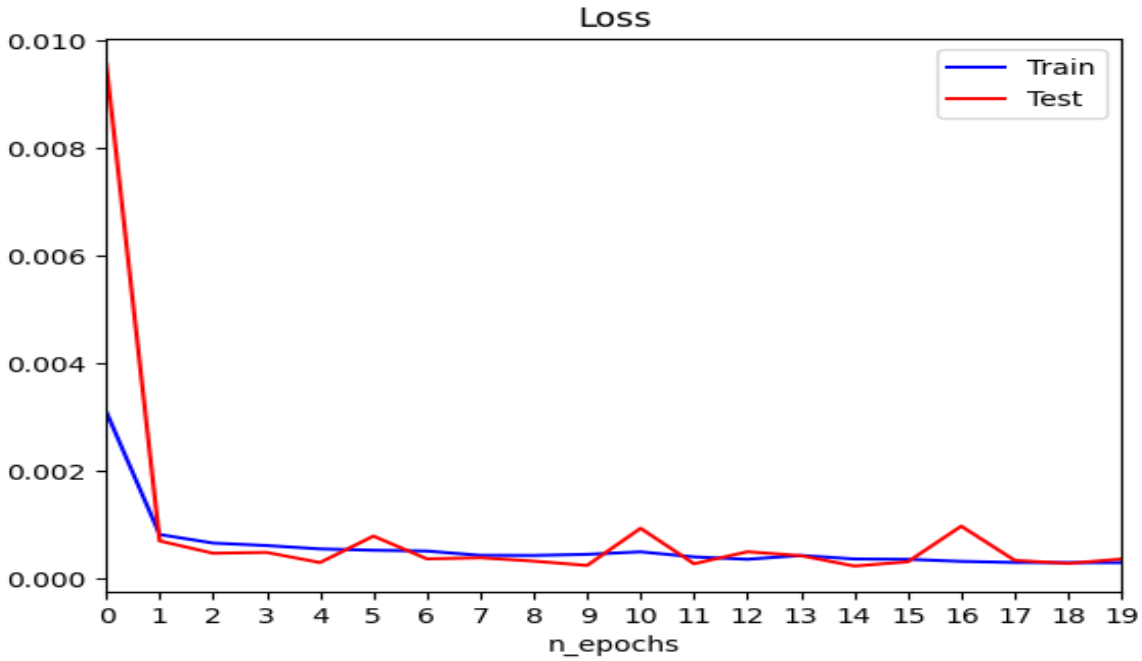
Hình 3.5. Huấn luyện Autoencoder Classic với 20 epoch

- AutoEncoder Shortcut(con-decon):



Hình 3.6. Huấn luyện Autoencoder Shortcut(con-decon) với 20 epoch

- AutoEncoder Shortcut:



Hình 3.7. Huấn luyện Autoencoder Shortcut với 20 epoch

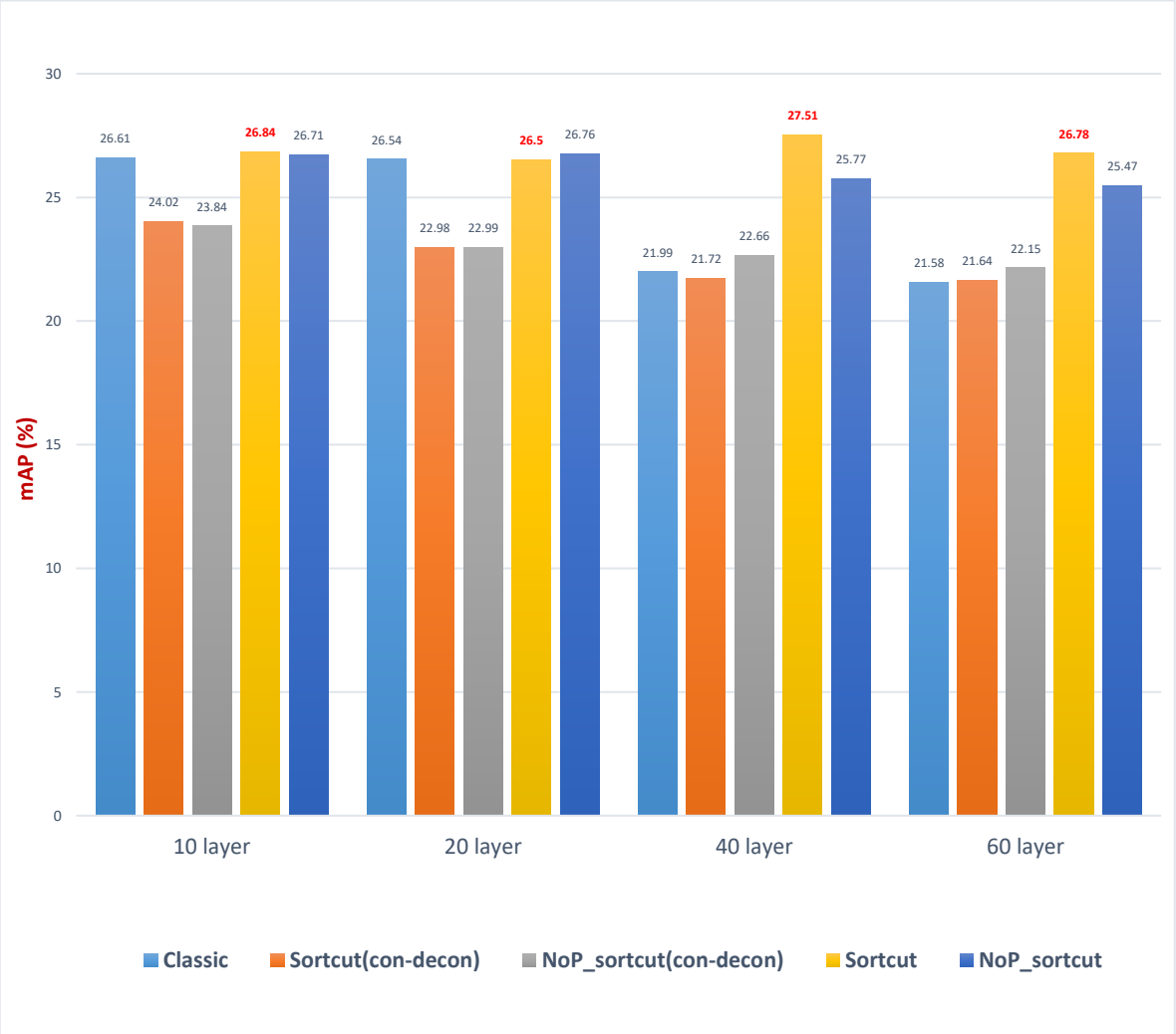
Sau khi huấn luyện xong các kiến trúc mạng ở trên, ta sử dụng phần Encoder để trích rút ra các véc tơ đặc trưng của các ảnh trong CSDL. Như đã đề cập ở phần trước, tập dữ liệu CIFAR-100 bao gồm 100 lớp/chủ đề và khoảng 600 ảnh cho mỗi chủ đề. Tập ảnh này được chia làm hai tập con: tập thứ nhất là tập kiểm tra gồm 10.000 ảnh và tập thứ hai là tập huấn luyện gồm 50.000 ảnh. Ta lấy 50.000 ảnh của tập con thứ hai cho việc đánh giá thực nghiệm.

Trong trường hợp lấy số chiều của véc tơ đặc trưng là 128 chiều thì tập đặc trưng gồm 50.000 dòng và 128 cột được mô tả như Hình 3.8 dưới đây.

Topic	Example	0	1	2	...	123	124	125	126	127
bear	bear_cub_s_000298.png	1.907470	4.116173	2.879014	...	3.140136	4.690528	3.639258	4.274721	2.355495
apple	macoun_s_001403.png	1.929587	3.324580	1.799607	...	3.028248	3.404996	2.346473	2.939421	1.796161
road	access_road_s_000002.png	1.146844	2.353577	0.803934	...	4.114978	5.313982	3.082025	4.625179	2.999759
tractor	tractor_s_001337.png	2.216322	4.857269	2.728326	...	3.862462	4.759552	2.178881	4.562580	1.924544
dolphin	common_dolphin_s_001524.png	1.628266	3.491193	1.803224	...	5.547873	5.252761	3.205729	6.232933	2.779444
bicycle	bike_s_001945.png	2.095326	3.670760	1.938436	...	1.893755	1.730091	0.787565	2.469499	1.285457
mountain	alp_s_003115.png	2.000207	5.525077	2.260195	...	3.901638	3.832243	2.576819	2.842464	2.314400
fox	red_fox_s_000589.png	0.942789	2.136873	1.158727	...	3.444645	4.882494	2.645034	4.478121	2.944619
worm	helminth_s_000630.png	2.915287	5.660411	2.854549	...	6.478815	8.472855	4.956523	6.669069	4.738469
skunk	striped_skunk_s_001711.png	2.188948	2.710665	1.885107	...	3.068597	3.759523	2.703760	3.387288	2.325939

Hình 3.8. Một số véc tơ đặc trưng được trích rút từ cơ sở dữ liệu CIFAR-100

Quan sát trên Hình 3.9 thấy rằng, số layer tối ưu của kiến trúc mạng autoencoder cho tra cứu ảnh trên tập CIFAR-100 mà trên tất cả các kiến trúc là 40 layers. Cũng từ Hình 3.9 này, cho thấy, cấu hình mạng sử dụng lớp pooling có hiệu quả cho kiến trúc mạng sâu.



Hình 3.9. Kết quả tra cứu ảnh theo các độ sâu khác nhau của mạng autoencoder trên tập CIFAR-100

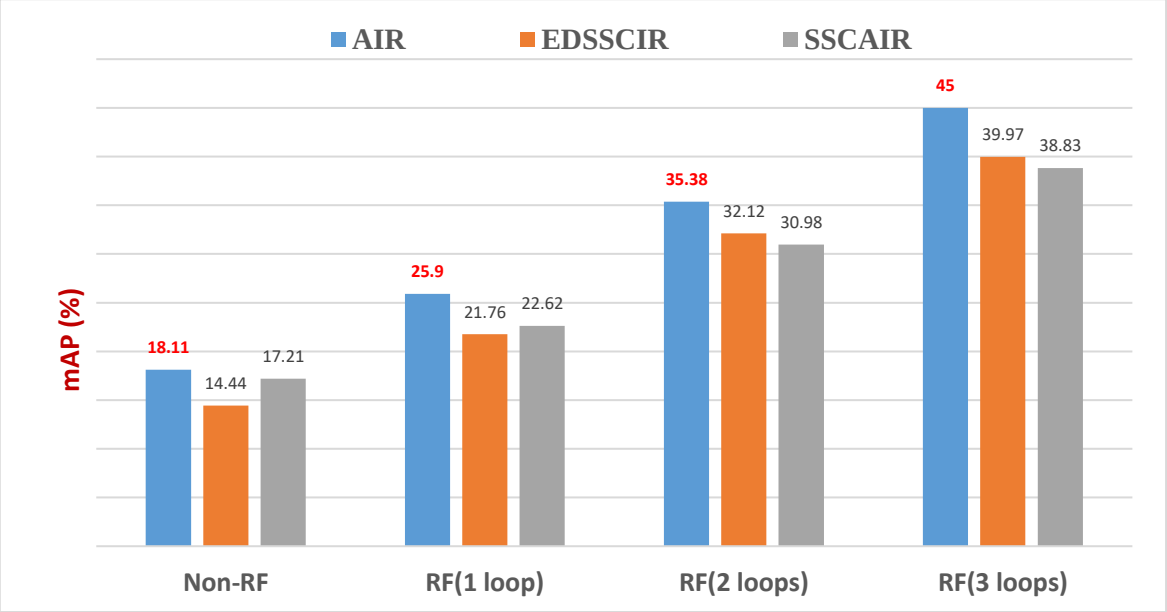
Kiến trúc mạng trong [163] với các cấu hình Shortcut (con-decon) và NoP_ shortcut (con-decon) cho kết quả thấp hơn kiến trúc mạng đề xuất (bao gồm Shortcut và NoP_ shortcut). Lý do của việc này là bởi vì mặc dù cả hai cấu hình Shortcut (con-decon) và NoP_ shortcut (con-decon) của kiến trúc mạng trong [163] đều dùng shortcut connections, nhưng chúng sử dụng shortcut connections đối xứng. Kiến trúc mạng sử dụng shortcut connections đối xứng phù hợp cho loại bỏ nhiễu trong ảnh hơn là tra cứu ảnh. Lý do của việc này là bởi vì autoencoder dùng shortcut connections đối xứng giúp loại bỏ nhiễu trong ảnh (có thể dẫn đến loại bỏ một số chi tiết ảnh),

điều này cũng dẫn đến đặc trưng được trích rút có thể loại đi một số chi tiết ảnh có thể mang ngữ nghĩa quan trọng của ảnh. Do đó, so sánh hai véc tơ đặc trưng của hai ảnh có thể không phản ánh độ giống nhau của hai ảnh gốc ban đầu, điều này có thể ảnh hưởng đến độ chính xác tra cứu ảnh. Trong số 5 cấu hình, hai cấu hình trong kiến trúc mạng được đề xuất cho kết quả cao nhất trên toàn bộ 20, 40, và 60 layers, chứng tỏ một điều rằng việc sử dụng shortcut connections không đối xứng vào autoencoder để tạo ra các mạng sâu autoencoder cho đối sánh ảnh là hiệu quả trên tập CIFAR-100. Hình 3.9, thể hiện kiến trúc mạng ở 40 lớp có hiệu quả cao hơn, tuy nhiên, khi số lớp tăng lên là 60, thì độ chính xác của các cấu hình lại giảm đi bởi vì khả năng cao là cỡ của tập dữ liệu huyến luyện ở đây là nhỏ nên không đủ để huấn luyện hiệu quả trên các mô hình quá sâu.

Dựa vào thực nghiệm ở trên, cho thấy rằng số lớp tối ưu cho kiến trúc mạng autoencoder với shortcut connections là 40. Để chứng tỏ tính hiệu quả của phương pháp được đề xuất cho tra cứu ảnh, luận án thực nghiệm phương pháp với RF trên cấu hình mạng này như sau:

Sau khi người dùng cung cấp các RF, các phương pháp Baseline, AIR, EDSSCIR (Encoder-Decoder with Symmetric Skip Connection for Image Retrieval) trong [163], và SSCAIR (Feature extraction using self-supervised convolutional autoencoder for content based image retrieval) [164] được áp dụng để phân hạng lại các ảnh trong CSDL. Luận án chọn phương pháp EDSSCIR cho so sánh hiệu năng, nó sử dụng đặc trưng được học từ autoencoder tích chập với connection bỏ qua đối xứng, là bởi vì muốn minh chứng hiệu quả của đặc trưng được học bởi phương pháp được đề xuất trong tra cứu ảnh với RF.

Hình 3.10 chỉ ra mAP của ba phương pháp gồm AIR, EDSSCIR, và SSCAIR cho ba lần lặp phản hồi đầu tiên. Hình 3.10 cho thấy rằng, phương pháp AIR được đề xuất thực hiện tốt hơn hai phương pháp còn lại trên tất cả các lần lặp và không lặp. Hiệu năng khi sử dụng phản hồi của phương pháp đề xuất là tốt hơn đáng kể so với khi không sử dụng phản hồi, nó chỉ ra rằng các RF được người dùng cung cấp là rất hữu ích trong cải tiến hiệu năng tra cứu. AIR thực hiện tốt hơn EDSSCIR là bởi vì AIR thu được biểu diễn đặc trưng tốt.



Hình 3.10. So sánh hiệu năng (dưới dạng mAP) của bốn phương pháp cho ba lần lặp đầu tiên

Luận án đã thực hiện việc đo thời gian TB của một truy vấn trên tập CSDL CIFAR-100. Thời gian thực hiện truy vấn theo số vòng lặp phản hồi có thể được thấy trên Bảng 3.4.

Bảng 3.4. Thời gian thực hiện truy vấn của AIR trên CIFAR-100.

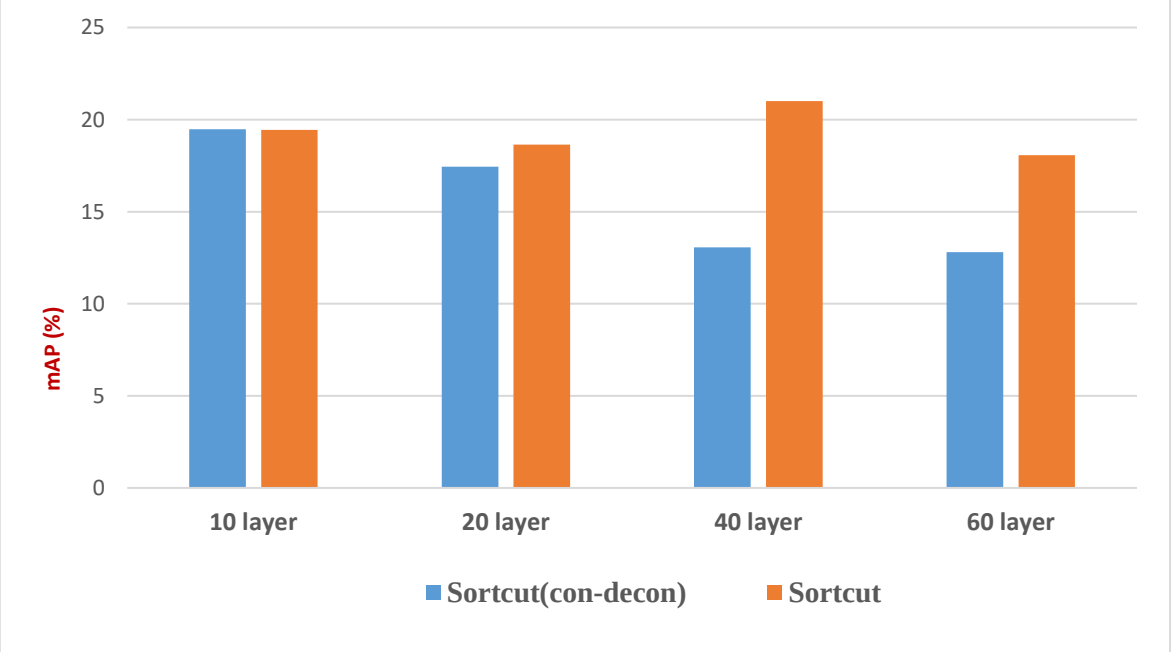
Vòng lặp phản hồi	Thời gian trung bình cho 1 truy vấn của AIR với cấu hình		
	Shortcut (con-decon) (s)	Shortcut (s)	Classic (s)
Không có phản hồi	0.2449	0.2650	0.2335
Vòng lặp thứ nhất	25.5623	28.1375	24.0926
Vòng lặp thứ hai	26.2186	28.9882	24.4392
Vòng lặp thứ ba	27.2913	29.1830	24.5538

Qua Bảng 3.4 cho thấy rằng, thời gian thực hiện truy vấn của phương pháp đề xuất AIR với cấu hình Classic là thấp nhất trên tất cả các lần lặp phản hồi và không phản hồi. Phương pháp sử dụng cấu hình Shortcut có thời gian cao hơn sử dụng cấu hình Shortcut(con-decon) một chút (xấp xỉ 2s). Nguyên nhân của việc sử dụng cấu hình Shortcut cao hơn một chút so với sử dụng cấu hình Shortcut(con-decon) là bởi nó phải dành thời gian để tính thêm kết nối tắt. Trong khi đó, phương pháp sử dụng cấu hình Classic không phải tính thêm kết nối tắt.

3.4.2. Các kết quả trên tập dữ liệu ảnh Corel

Quan sát trên Hình 3.11 cho thấy rằng, số lớp tối ưu của kiến trúc mạng autoencoder cho tra cứu ảnh trên tập COREL mà trên hai kiến trúc là 40 lớp.

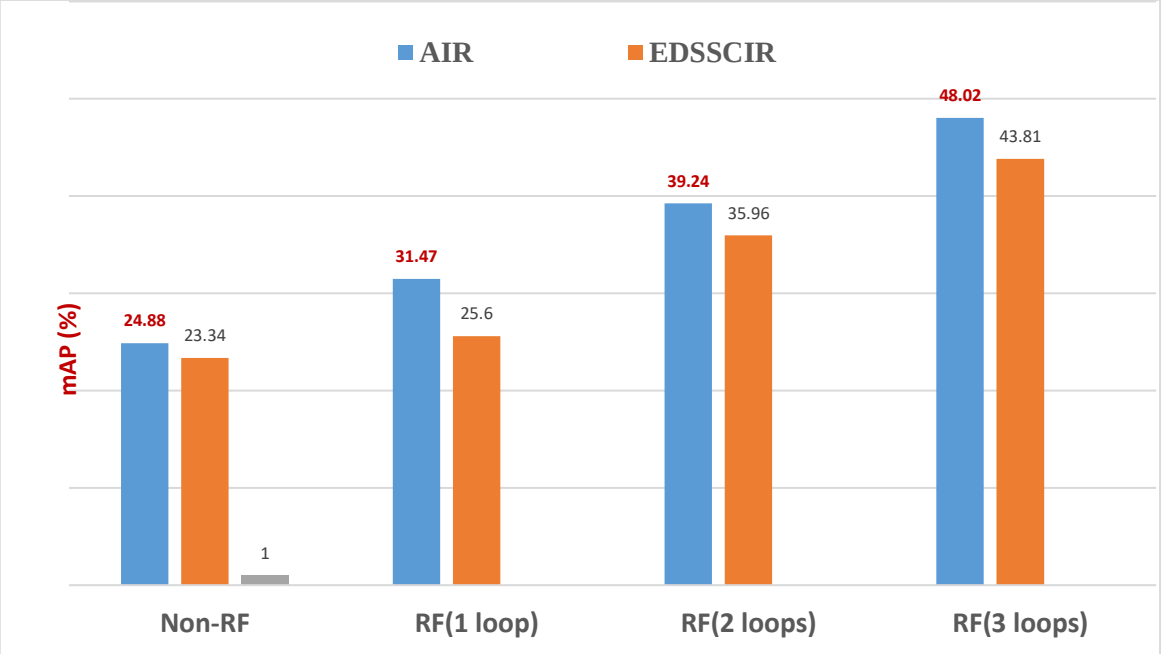
Kiến trúc mạng trong [163] với cấu hình Shortcut(con-decon) cho kết quả thấp hơn kiến trúc mạng được đề xuất (Shortcut). Lý do của việc này là bởi vì mặc dù cấu hình Shortcut(con-decon) của kiến trúc mạng trong [163] dùng shortcut connections, nhưng chúng sử dụng shortcut connections đối xứng. Kiến trúc mạng sử dụng shortcut connections đối xứng phù hợp cho loại bỏ nhiễu trong ảnh hơn là tra cứu ảnh. kiến trúc mạng được đề xuất cho kết quả cao nhất trên 40, và 60 lớp, riêng 20 lớp là cho kết quả xấp xỉ nhau. Điều này cho thấy rằng việc sử dụng shortcut connections không đối xứng vào autoencoder để tạo ra các mạng sâu autoencoder cho đối sánh ảnh là hiệu quả trên tập COREL.



Hình 3.11. Kết quả tra cứu ảnh theo các độ sâu khác nhau của mạng autoencoder trên tập COREL

Hình 3.12 chỉ ra mAP của ba phương pháp gồm Baseline (Non-RF), AIR, EDSSCIR cho ba lần lặp phản hồi đầu tiên. Có thể nhận thấy rằng phương pháp Baseline cho ra kết quả độ chính xác thấp nhất. Nguyên nhân của việc này là do phương pháp Baseline không sử dụng cơ chế học, mà chỉ tính khoảng cách Euclide giữa véc tơ đặc trưng của ảnh truy vấn và ảnh trong CSDL. Phương pháp được đề xuất AIR thực hiện tốt hơn hai phương pháp còn lại trên tất cả các lần lặp. Phương pháp AIR có hiệu năng tốt hơn đáng kể so với Baseline, nó chỉ ra rằng các RF được

người dùng cung cấp là rất hữu ích trong cải tiến hiệu năng tra cứu. AIR thực hiện tốt hơn EDSSCIR là bởi vì AIR thu được biểu diễn đặc trưng tốt.



Hình 3.12. So sánh hiệu năng (dưới dạng mAP) của ba phương pháp cho ba lần lặp đầu tiên

Bảng 3.5. Thời gian thực hiện truy vấn của AIR trên COREL

Vòng lặp phản hồi	Thời gian trung bình cho 1 truy vấn của AIR với cấu hình		
	Shortcut(con-decon) (s)	Shortcut (s)	Classic (s)
Không có phản hồi	0.1289	0.1468	0.0457
Vòng lặp thứ nhất	5.5781	5.5734	4.8175
Vòng lặp thứ hai	5.6410	5.6508	4.8858
Vòng lặp thứ ba	5.8743	5.8919	4.8108

Bảng 3.5 cũng thể hiện rằng, thời gian thực hiện truy vấn của phương pháp đề xuất AIR với cấu hình Classic là thấp nhất trên tất cả các lần lặp phản hồi và không phản hồi. Phương pháp sử dụng cấu hình Shortcut có thời gian ngang với sử dụng cấu hình Shortcut(con-decon) (chênh nhau xấp xỉ 0.02s). Nguyên nhân của việc phương pháp sử dụng cấu hình Shortcut cao hơn một chút so với sử dụng cấu hình Shortcut(con-decon) là bởi nó phải dành thời gian để tính thêm kết nối tắt. Trong khi đó, phương pháp sử dụng cấu hình Classic không phải tính thêm kết nối tắt.

3.5. Kết luận Chương 3

Trong chương này, luận án trình bày một phương pháp hiệu quả cho tra cứu ảnh. Phương pháp này đã khắc phục được hai vấn đề: *thứ nhất*, khả năng phân biệt hạn chế của các phương pháp đã có và *thứ hai*, giảm nhẹ vấn đề vanishing/exploding gradients và quá trình hội tụ nhanh. Mô hình mạng nơ ron tích chập sâu autoencoder được tận dụng để học các biểu diễn đặc trưng hiệu quả cho tra cứu ảnh thông qua việc sử dụng shortcut connections trong kiến trúc autoencoder. Mô hình học này được sử dụng vào việc tạo ra các biểu diễn đặc trưng của các ảnh CSDL. Trên cơ sở các biểu diễn đặc trưng này, Nghiên cứu sinh đã thiết kế một cơ chế học RF sử dụng SVM để tận dụng các mẫu có nhãn từ phản hồi của người dùng. Các mẫu huấn luyện thu được từ cơ chế RF, được cung cấp cho bộ phân lớp SVM giúp tăng cường khả năng học các đặc trưng phân biệt mà được dùng cho tra cứu ảnh. Kết quả của phương pháp này đã thu được các danh sách phân hạng có chất lượng tốt, vừa khắc phục được sự thiếu hụt các mẫu có nhãn vừa tận dụng được ưu điểm của các mạng nơ ron sâu.

Các kết quả thực nghiệm trên tập CIFAR-100 với 60,000 ảnh đã minh chứng rằng phương pháp được đề xuất sinh ra kết quả có độ chính xác cao hơn và thời gian huấn luyện mô hình trích rút đặc trưng thấp hơn một số phương pháp hiện nay (thời gian tra cứu của phương pháp đề xuất sử dụng cấu hình Shortcut connections cao hơn hai cấu hình Shortcut(con-decon) và Classic là không đáng kể). Các đóng góp chính của chương này đã được Nghiên cứu sinh công bố trong các công trình [**CT1, CT3**].

KẾT LUẬN

Nghiên cứu về "***Tra cứu ảnh dựa vào nội dung với học biểu diễn và giảm chiều dữ liệu***" là một hướng tiếp cận mang tính thời sự và thực tiễn cao đối với bài toán CBIR. Những năm gần đây, đã có nhiều công trình nghiên cứu về kỹ thuật học máy cho bài toán CBIR với RF được công bố, giúp cải thiện độ chính xác tra cứu của hệ thống tra cứu ảnh. Tuy nhiên nó vẫn còn bộc lộ nhiều khó khăn, thách thức cần được giải quyết hoặc cải tiến để có hiệu năng tốt hơn. Trong đó điển hình là bài toán nâng cao tốc độ và độ chính xác tra cứu của hệ thống tra cứu ảnh trên CSDL ảnh lớn, dữ liệu có cỡ lớp nhỏ, cỡ mẫu nhỏ và chiều cao. Để giải quyết vấn đề này, luận án đã tiến hành nghiên cứu cơ sở lý thuyết về CBIR, cơ chế RF và vấn đề giảm khoảng trống ngữ nghĩa trong CBIR, đồng thời tập trung khảo sát, phân tích và đánh giá một số công trình nghiên cứu ở trong nước và trên thế giới, từ đó đề xuất hai giải pháp sử dụng học máy (đặc biệt là học sâu) vào quá trình tra cứu ảnh với RF, giúp cải thiện được độ chính xác và tốc độ tra cứu, đồng thời thu hẹp khoảng trống ngữ nghĩa đối với các bài toán được đặt ra ở trên.

1. Các kết quả đạt được của luận án

Luận án đã đạt được một số kết quả chính như sau:

(1) Đề xuất được phương pháp tra cứu ảnh SDAIR. Phương pháp này kết hợp mô hình trích rút đặc trưng quan trọng dựa trên phương pháp RSLDA với mô hình phân lớp trong hệ thống CBIR nhằm cải tiến độ chính xác và thời gian truy vấn. Sự khác biệt giữa phương pháp đề xuất và các hệ thống tra cứu ảnh hiện có là phương pháp đề xuất đưa ra một mô hình linh hoạt mà không chỉ gắn với mô hình học hoặc một độ đo tương tự nào đó. SDAIR cũng có cơ chế bổ sung mẫu dương vào tập huấn luyện một cách tự động mà không đòi hỏi số lượng mẫu dương lớn. Ngoài ra, nó có thể phục vụ đồng thời đối với hai nhiệm vụ: lựa chọn tập đặc trưng quan trọng và bổ sung mẫu huấn luyện dương. Kết quả thực nghiệm trên tập CSDL ảnh cho thấy phương pháp đề xuất có thể cải thiện độ chính xác và thời gian truy vấn cho bài toán tra cứu ảnh với RF trong trường hợp cỡ mẫu nhỏ, cỡ lớp nhỏ và dữ liệu có chiều cao. [CT4, CT2]

(2) Đề xuất phương pháp tra cứu ảnh AIR dựa trên 3 thành phần: Huấn luyện bán giám sát bằng mạng nơ ron tích chập autoencoder, trích rút đặc trưng ảnh và phân lớp SVM trong RF nhằm cải tiến độ chính xác và thời gian truy vấn. Mạng nơ ron tích

chập autoencoder được tận dụng để học các biểu diễn đặc trưng hiệu quả cho tra cứu ảnh thông qua việc sử dụng shortcut connections trong kiến trúc autoencoder. Mô hình học này được sử dụng vào việc tạo ra các biểu diễn đặc trưng của ảnh CSDL. Trên cơ sở các biểu diễn đặc trưng này, luận án đã thiết kế một cơ chế học RF sử dụng máy véc tơ hỗ trợ SVM để tận dụng các mẫu có nhãn từ phản hồi của người dùng. Các mẫu huấn luyện thu được từ cơ chế RF, được cung cấp cho bộ phân lớp SVM đã làm tăng cường khả năng học các đặc trưng phân biệt mà được dùng cho tra cứu ảnh. Kết quả của phương pháp này đã thu được các danh sách phân hạng có chất lượng tốt, vừa khắc phục được sự thiếu hụt các mẫu có nhãn vừa tận dụng được ưu điểm của các mạng nơ ron sâu. Thực nghiệm được thực hiện trên tập CIFAR-100 với 60,000 ảnh, cho thấy rằng phương pháp AIR được đề xuất tạo ra các kết quả có độ chính xác cao hơn một số phương pháp hiện nay. [CT1, CT3]

Bằng việc phân tích lý thuyết và thực nghiệm trên các tập CSDL ảnh chuyên nghiệp mà các nhà nghiên cứu trong nước và quốc tế khuyên dùng. Kết quả của nghiên cứu đã chỉ ra rằng hai phương pháp mới được đề xuất trong luận án có nhiều ưu điểm hơn so với các phương pháp tiêu biểu hiện có. Điều này khẳng định tính đúng đắn và khả năng nâng cao hiệu quả tra cứu của các phương pháp đề xuất, đồng thời thực hiện được mục tiêu đã đề ra trong luận án.

2. Định hướng phát triển

Mặc dù kết quả bước đầu của luận án đã khẳng định được một số nghiên cứu quan trọng về lý luận khoa học và thực tiễn trong sử dụng kỹ thuật học máy vào quá trình CBIR với RF, nhưng luận án vẫn còn một số vấn đề cần được nghiên cứu, cải tiến và phát triển tiếp. Trong tương lai, định hướng phát triển tiếp theo của luận án có thể thực hiện theo các hướng nghiên cứu sau:

(1) Tiếp tục nghiên cứu và khắc phục những vấn đề còn tồn tại của các phương pháp đã đề xuất trong luận án.

(2) Nghiên cứu tận dụng các thành tựu của học máy hiện đại như mô hình Vision Transformer, mạng nơ ron tích chập đồ thị và cơ chế học truyền để nâng cao hiệu năng tra cứu.

(3) Triển khai các phương pháp đã đề xuất vào việc giải quyết các lớp bài toán trong thực tiễn, đặc biệt là các bài toán phức tạp, có sử dụng dữ liệu hình ảnh với độ chính xác cao, thuộc các lĩnh vực khác nhau như quân sự, y học, giáo dục, du lịch, dự báo thời tiết, ….

DANH MỤC CÔNG TRÌNH CÔNG BỐ LIÊN QUAN ĐẾN LUẬN ÁN

Trong nước:

[CT1] An Hong Son, Nguyen Huu Quynh, Dao Thi Thuy Quynh, Cu Viet Dung, "Deep Learning o f Image Representations with Con volutional Neural Networks Autoencoder for Image Re trieval with Relevance Feedback", *Journal onInformation Technologies and Communications*, Vol. 2022, No. 1, pp. 17-24 (ISSN: 1859-3534,

Quốc tế:

[CT2] An Hong Son, Dao Thi Thuy Quynh, Nguyen Huu Quynh, "Stuck Query Point Processing O f Multi point Query Fo r Image Re trieval With Relevance Feedback", *Journal ofInformation Hiding & Multimedia Signal Processing*, Vol. 12, No. 2, pp. 42-55, June 2021. (ISSN:2073-4212/2073-4239; **SCOPUS).**

[CT3] An Hong Son, Nguyen Huu Quynh, Cu Viet Dung, Dao Thi Thuy Quynh, Ngo Quoc Tao, "Learning Binary Codes for Fast Image Retrieval with Sparse Discriminant Analysis and Deep Autoencoders", *Intelligent Data Analysis*, Vol. 27, No. 3, pp. 809-831, 2023 (ISSN: 1088-467X/1571-4128; **SCIE).**

[CT4] An Hong Son, Nguyen Huu Quynh, Cu Viet Dung, Dao Thi Thuy Quynh, Ngo Quoc Tao, "Improving image retrieval effectiveness via sparse discriminant analysis", *Multimedia Tools and Applications*, March 2023, pp.1-24 (ISSN: 1380-7501/1573-7721; **SCIE).**

DANH MỤC TÀI LIỆU THAM KHẢO

1. M. Alkhawlani, M. Elmogy, H. El Bakry (2015), *Text-based, content-based, and semantic-based image retrievals: A survey,* Int. J. Comput. Inf. Technol, 4(01), pp. 5866.
2. A. Krizhevsky, I. Sutskever, & G.E. Hinton, (2017, May). ImageNet classification with deep convolutional neural networks. Communications of the ACM, 60(6), 84-90.
3. F. Ali and A. Hashem (2020, June). *Content Based Image Retrieval (CBIR) by statistical methods*. Baghdad Science Journal, 17 (2(SI)), 694.
4. C. Bai, J. Chen, L. Huang, K. Kpalma, & S. Chen, (2018, January). *Saliency-based multi-feature modeling for semantic image retrieval*. Journal of Visual Communication and Image Representation, 50, 199-204.
5. R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image retrieval: ideas, influences, and trends of the new age," ACM Computing Surveys, vol. 40, no. 2, pp. 1-60, May 2008.
6. A.H. Halawani, A. Teynor, L. Setia, G. Brunner, & C.I. Retrieval, (2006, January). Fundamentals and Applications of Image Retrieval: An Overview. Image (Rochester, N.Y.), 14-23.
7. M.J.J. Ghrabat, G. Ma, I.Y. Maolood, S.S. Alresheedi, & Z.A. Abduljabbar, (2019, December). An effective image retrieval based on optimized genetic algorithm utilized a novel SVM-based convolutional neural network classifier. Human-centric Computing and Information Sciences, 9(1), 31.
8. M. Bansal & M. Kumar, (2020, February). *2D object recognition techniques: State-of-the-art work.* Archives of Computational Methods in Engineering, 28 (3), 1147-1161.
9. U. Mittal, S. Srivastava, & P. Chawla, "Review of different techniques for object detection using deep learning," in Proceedings of the Third International Conference on Advanced Informatics for Computing Research - ICAICR '19, New York, USA, 2019, pp. 1-8.
10. L. Piras, & G. Giacinto, G. (2017, September). Information fusion in content based image retrieval: A comprehensive overview. Information Fusion, 37, 50-60.
11. D. Zhang, M.M. Islam, & G. Lu, (2012, January). A review on automatic image annotation techniques. Pattern Recognition, 45(1), 346-362.
12. M. Lew, N. Sebe, C. Djeraba and R. Jain, *Content-based Multimedia Information Retrieval: State of the Art and Challenges,* ACM Transactions on Multimedia Computing, Communications, and Applications, pp. 1-19, 2006.

13. F. Baig, Z. Mehmood, M. Rashid, M.A. Javid, A. Rehman, T. Saba, & A. Adnan (2020, March). *Boosting the performance of the BoVW model using SURF-CoHOGbased sparse features with relevance feedback for CBIR*. Iranian Journal of Science and Technology, Transactions of Electrical Engineering, 44(1), 99-118.

14. D.G. Low, (2004). Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis, 60(2), 91-110.

15. X. Duanmu, "Image retrieval using color moment invariant," in 2010 Seventh International Conference on Information Technology: New Generations, 2010, pp. 200-203. Las Vegas, Nevada.

16. J. Huang, S.R. Kumar, M. Mitra, W.J. Zhu, & R. Zabih, "Image indexing using color correlograms," in Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on, 1997, pp. 762-768. San Juan, PR.

19. N. Shrivastava, & V. Tyagi, (2015, August). An efficient technique for retrieval of color images in large databases. Computers & Electrical Engineering, 46, 314-327.

20. A. Alzu'bi, A. Amira, & N. Ramzan, (2015, October). *Semantic content-based image retrieval: A comprehensive study*. Journal of Visual Communication and Image Representation, 32, 20-54.

23. R.M. Hawlick, (2017). Statistical and structural approaches to texture. Advances in Intelligent Systems and Computing, 459(5), 1-644.

25. D.P. Tian, (2013). A review on image feature extraction and representation techniques. International Journal Multimedia Ubiquitous Engineering, 8(4), 385-395.

27. T. Suk, & J. Flusser, (2011, September). Affine moment invariants generated by graph method. Pattern Recognition, 40(2), 2047-2056.

28. M. Naeem, R. Ashraf, N. Ali, M. Ahmad, & M.A. Habib, "Bottom up approach for better requirements elicitation," in Proceedings of the International Conference on Future Networks and Distributed Systems - ICFNDS '17, New York, USA, 2017, 1305, pp. 1-4.

29. Y.D. Chun, N.C. Kim, & I.H. Jang, (2008, October). Content-based image retrieval using multiresolution color and texture features. IEEE Transactions on Multimedia, 10(6), 1073-1084.

30. G.A. Montazer, & D. Giveki, (2015, September). Content based image retrieval system using clustered scale invariant feature transforms. Optik (Stuttg), 126(18), 1695-1699.

31. H. Bay, A. Ess, T. Tuytelaars, & L. Van Gool, (2008, June). *Speeded-Up Robust Features (SURF)*. Computer Vision and Image Understanding, 110(3), 346-359.

32. Z. Zhao, X. He, D. Cai, L. Zhang, W. Ng, and Y. Zhuang. Graph regularized feature selection with data reconstruction. IEEE Transactions on Knowledge and Data Engineering, 28(3):689-700, 2015.

33. R. O. Duda, P. E. Hart, and D. G. Stork. Pattern classification. John Wiley & Sons, 2012.

35. I. Kononenko, E. Šimec, and M. Robnik-Šikonja. Overcoming the myopia of inductive learning algorithms with relieff. Applied Intelligence, 7(1):39–55, 1997.

36. M. Bennasar, Y. Hicks, and R. Setchi. Feature selection using joint mutual information maximisation. Expert Systems with Applications, 42(22):8520–8532, 2015.

37. L. Song, A. Smola, A. Gretton, K. M. Borgwardt, and J. Bedo. Supervised feature selection via dependence estimation. In Proceedings of the 24th International Conference on Machine Learning, pages 823–830, 2007.

38. X. He, D. Cai, and P. Niyogi. Laplacian score for feature selection. Advances in Neural Information Processing Systems, 18:507–514, 2005.

40. S. Hijazi. Semi-supervised Margin-based Feature Selection for Classification. PhD thesis, Université du Littoral Côte d'Opale; Université Libanaise, école doctorale .. . , 2019.

41. J. Wen, X. Fang, J. Cui, L. Fei, K. Yan, Y. Chen, and Y. Xu. Robust sparse linear discriminant analysis. IEEE Transactions on Circuits and Systems for Video Technology, 29(2):390-403, 2018.

42. F. Dornaika and A. Khoder. Linear embedding by joint robust discriminant analysis and inter-class sparsity. Neural Networks, 127:141-159, 2020.

43. A. Khoder and F. Dornaika. *A hybrid discriminant embedding with feature selection: application to image categorization*. Applied Intelligence, pages 1-17, 2020.

44. A. Tharwat, T. Gaber, A. Ibrahim, and A. E. Hassanien. *Linear discriminant analysis: A detailed tutorial*. AI Communications, 30(2):169-190, 2017.

45. L. I. Smith. A tutorial on principal components analysis. 2002.

46. D. J. Hand. Classifier technology and the illusion of progress. Statistical Science, pages 1-14, 2006.

47. Z. Lai, Y. Xu, Z. Jin, and D. Zhang. Human gait recognition via sparse discriminant projection learning. IEEE Transactions on Circuits and Systems for Video Technology, 24(10):1651–1662, 2014.

48. D. Arthur, & S. Vassilvitskii, (2007, January). *K-means++: The advantages of careful seeding*. Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms, 07-09, 1027-1035.

49. M. Yousuf, Z. Mehmood, H.A. Habib, T. Mahmood, T. Saba, A. Rehman, & M. Rashid, (2018). A novel technique based on visual words fusion analysis of sparse features for effective content-based image retrieval. Mathematical Problems in Engineering, 2018, 1-13.

50. Z. Mehmood, F. Abbas, T. Mahmood, M.A. Javid, A. Rehman, & T. Nawaz, (2018, December). Contentbased image retrieval based on visual words fusion versus features fusion of local and global features. Arabian Journal for Science and Engineering, 43(12), 7265-7284.

51. C. Cortes, & V. Vapnik, (1995, September). Support-vector networks. Machine Learning, 20(3), 273-297.

52. M. Garg, & G. Dhiman, (2020, June). A novel content-based image retrieval approach for classification using GLCM features and texture fused LBP variants. Neural Computing & Applications, 33(4), 1311-1328.

53. A. Vedaldi, & A. Zisserman, "Sparse kernel approximations for efficient classification and detection," in 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 2012, pp. 2320-2327.

55. I. Basheer, & M. Hajmeer, (2000, December). *Artificial neural networks: Fundamentals, computing, design, and application*, Journal of Microbiological Methods, 43(1), 3-31.

56. R. Ashraf, K. Bashir, A. Irtaza, & M. Mahmood, (2015, May). *Content based image retrieval using embedded neural networks with bandletized regions*. Entropy, 17(6), 3552-3580.

57. H. Yoon, C-S. Park, J.S. Kim, & Baek, (2013, January). Algorithm learning based neural network integrating feature selection and classification. Expert Systems with Applications, 40(1), 231-241.

58. J. Wan, "Deep learning for content-based image retrieval," In Proceedings of the ACM International Conference on Multimedia - MM '14, New York, USA, 2014, pp. 157-166.

59. A. Voulodimos, N. Doulamis, A. Doulamis, & E. Protopapadakis, (2018). Deep learning for computer vision: A brief review. Computational Intelligence and Neuroscience, 2018, 1-13.

60. S. Chaudhry, & R. Chandra, (2016, October). Unconstrained face detection from a mobile source using convolutional neural networks. Lecture Notes in Computer Science, 9948, 567-576. Including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics.

61. I. Gogul, & V.S. Kumar, “Flower species recognition system using convolution neural networks and transfer learning,” in 2017 Fourth International Conference on Signal Processing, Communication and Networking (ICSCN), 2017, no. March, pp. 1-6. Chennai, India.

62. A. Patil, & M. Rane, (2021). Convolutional neural networks: an overview and its applications in pattern recognition. In Smart Innovation, Systems and Technologies, 195(Insights into Imaging), 21-30.

63. A. Alzu’bi, A. Amira, & N. Ramzan, (2017, August). *Content-based image retrieval with compact deep convolutional features*. Neurocomputing, 249, 95-105.

64. M. Tzelepi, & A. Tefas, (2018, January). Deep convolutional learning for content based image retrieval. Neurocomputing, 275, 2467–2478.

65. Q. Zheng, X. Tian, M. Yang, & H. Wang, (2019). Differential learning: A powerful tool for interactive content-based image retrieval. Engineering Letters, 27(1), 202-215.

66. K. Simonyan, & A. Zisserman, (2014, September). Very deep convolutional networks for large-scale image recognition. 3rd International Conference Learning Represent ICLR 2015 - Conference Track Proceedings, 1-14.

67. A. Sezavar, H. Farsi, & S. Mohamadzadeh, (2019, August). Content-based image retrieval by combining convolutional neural networks and sparse representation. Multimedia Tools and Applications, 78(15), 20895-20912.

68. K. He, X. Zhang, S. Ren, & J. Sun, (2016). Deep Residual Learning for Image Recognition. Computer Vision and Pattern Recognition (pp.770-778). IEEE.

69. L. Breiman. Bagging predictors. Machine Learning, 24(2):123–140, 1996.

70. H. Lu and R. Mazumder. Randomized gradient boosting machine. SIAM Journal on Optimization, 30(4):2780-2808, 2020.

71. L. Deng and J. C. Platt. Ensemble deep learning for speech recognition. In Fifteenth Annual Conference of the International Speech Communication Association, 2014.

72. M. J. Laan. van der, eric c. polley, and alan e. hubbard. “super learner”. Statistical Applications in Genetics and Molecular Biology, 6, 2007.

74. D. Benkeser, C. Ju, S. Lendle, and M. van der Laan. Online cross-validationbased ensemble learning. Statistics in Medicine, 37(2):249-260, 2018.

75. M. M. Davies and M. J. Van Der Laan. Optimal spatial prediction using ensemble machine learning. The International Journal of Biostatistics, 12(1):179-201, 2016.

77. Yong Rui and Thomas Huang. *Optimizing learning in image retrieval*. In Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662), volume 1, pages 236-243. IeEE, 2000.

78. V. Tyagi, (2017). Similarity measures and performance evaluation. In Content-based image retrieval (pp. 63-83). Springer Singapore.

79. P. Howarth, & S. Rüger, (2005). Fractional distance measures for content-based image retrieval. 447-456.

80. O. Pele, & M. Werman, (2010). The quadratic-chi histogram distance family. Lecture Notes in Computer Science, 6312 (2), 749-762. Including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics.

81. P.E. Forssén, & D.G. Lowe, (2007). Shape descriptors for maximally stable extremal regions. Proceedings / IEEE International Conference on Computer Vision, 0-7.

82. D.R. Martin, C.C. Fowlkes, & J. Malik, (2004). Learning to detect natural image boundaries using local brightness, color, and texture cues. IEEE Transactions on Pattern Analysis and Machine Intelligence, 26(5), 530-549.

83. M. Varma, & A. Zisserman, (2009). CUReT1. IEEE Transactions on Pattern Analysis and Machine Intelligence, 31(11), 2032-2047.

84. M.J. Swain, & D.H. Ballard, (1991). Color indexing. International Journal of Computer Vision, 7(1), 11-32. https://doi.org/10.1007/BF00130487.

85. G.N. Lance, & W.T. Williams, (1967). Mixed-data classificatory programs i - agglomerative systems. Australian Computer Journal, 1(1), 15-20.

86. H.S. Rui, D.S. Ruger, H. Liu, & Z. Huang, "Dissimilarity measures for content-based image retrieval," In 2008 IEEE International Conference on Multimedia and Expo, Hannover, Germany, 2008, pp. 1365-1368.

87. S.P. Rana, M. Dey, & P. Siarry, (2019, January). Boosting content based image retrieval performance through integration of parametric & nonparametric approaches. Journal of Visual Communication and Image Representation, 58(3), 205-219.

88. A. Raza, H. Dawood, H. Dawood, S. Shabbir, R. Mehboob, & A. Banjar (2018). Correlated primary visual texton histogram features for content base image retrieval. IEEE Access, 6, 46595-46616.

89. R. Wang, X. Wang, S. Kwong, C. Xu (2017), Incorporating diversity and informativeness in multiple-instance active learning. IEEE Trans Fuzzy Syst 25(6):1460-1475.

90. A. Ponomarev, H.S. Nalamwar, I. Babakov, C.S. Parkhi, & G. Buddhawar, (2016, February). Content-based image retrieval using color, texture and shape features. Key Engineering Materials, 685, 872-876.

91. P. Srivastava, & A. Khare, (2017, January). Integration of wavelet transform, local binary patterns and moments for content-based image retrieval. Journal of Visual Communication and Image Representation, 42, 78-103.

92. M. Sajjad, A. Ullah, J. Ahmad, N. Abbas, S. Rho, & S.W. Baik, (2018, February). Integrating salient colors with rotational invariant texture features for image representation in retrieval systems. Multimedia Tools and Applications, 77(4), 4769-4789.

93. L.K. Pavithra, & T. Sree Sharmila, (2019, December). An efficient seed points selection approach in dominant color descriptors (DCD). Cluster Computing, 22(4), 1225-1240.

94. R. Ashraf, M. Ahmedm, S. Jabbar, S. Khalid, A. Ahmad, S. Din, & G. Jeon, (2018, March). *Content based image retrieval by using color descriptor and discrete Wavelet transform*. Journal of Medical Systems, 42 (3), 44.

95. C.E. Jacobs, A. Finkelstein, & D.H. Salesin, "Fast multiresolution image querying," in Proceedings of the 22nd annual conference on Computer graphics and interactive techniques - SIGGRAPH '95, New York, USA, 1995, pp. 277-286.

96. R. Ashraf, M. Ahmed, U. Ahmad, M.A. Habib, S. Jabbar, & K. Naseer, (2020, April). *MDCBIR-MF: Multimedia data for content-based image retrieval by using multiple features*. Multimedia Tools and Applications, 79(13-14), 8553-8579.

97. M.K. Alsmadi, (2020, April). *Content-based image retrieval using color, shape and texture descriptors and features*. Arabian Journal for Science and Engineering, 45(4), 3317-3330.

98. N. Tadi Bani, & S. Fekri-Ershad, (2019, August). Contentbased image retrieval based on combination of texture and colour information extracted in spatial and frequency domains. Electronic Library, 37(4), 650-666.

99. Z. Zhao, Q. Tian, H. Sun, X. Jin, & J. Guo, (2016, January). Content based image retrieval scheme using color, texture and shape features. International Journal of Signal Processing Image Processing Pattern Recognition, 9(1), 203-212.

100. B.S. Phadikar, A. Phadikar, & G.K. Maity, (2018, May). Content-based image retrieval in DCT compressed domain with MPEG-7 edge descriptor and genetic algorithm. Pattern Analysis and Applications, 21(2), 469-489.

101. Vu Van Hieu, *Nghiên cứu một số kỹ thuật phân hạng trong tra cứu ảnh dựa vào nội dung*, Hoc vien Khoa hoc va Cong nghe-Vien Han lam KH&CN Viet Nam, (Luận án tiến sĩ, 2017).

102. Đao Thi Thuy Quynh, *Nâng cao độ chính xác tra cứu ảnh dựa vào nội dung sử dụng kỹ thuật điều chỉnh trọng số hàm khoảng cách*, Hoc vien Khoa hoc va Cong nghe-Vien Han lam KH&CN Viet Nam, (Luận án tiến sĩ, 2019).

103. Cu Viet Dung, *Nghiên cứu phát triển một số thuật toán tra cứu ảnh dựa vào khái niệm mức cao sử dụng kỹ thuật học sâu*, Hoc vien Khoa hoc va Cong nghe-Vien Han lam KH&CN Viet Nam, (Luận án tiến sĩ, 2022).

104. D. Tao, X. Tang, X. Li, and X. Wu, *Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28, no. 7, (2007), pp. 1088-1099.

105. H. Muller, W. Muller, D.M. Squire, S. Marchand-Maillet, T. Pun (2001), *Performance evaluation in content-based image retrieval: overview and proposals.* Pattern recognition letters, 22(5), pp. 593-601.

106. I.M. Hameed, S.H. Abdulhussain, and B.M. Mahmmod, *Content-based image retrieval: A review of recent trends.* Cogent Engineering, 2021. 8(1): p. 1927469.

107. Z. Lai, J. Bao, H. Kong, M. Wan, & G. Yang, (2020). *Discriminative low-rank projection for robust subspace learning*. International Journal of Machine Learning and Cybernetics, pages 1-14, 2020.

108. Huu Quynh Nguyen, Dung Cu Viet, Quynh Dao Thi Thuy. "Semantic class discriminant projection for image retrieval with relevance feedback." *Multimedia Tools and Applications* 80.10 (2021): 15351-15376.

109. X. Wang, R. Wang, C. Xu (2018), Discovering the relationship between generalization and uncertainty by incorporating complexity of classification. IEEE Trans Cybern 48(2):703-715.

110. High-dimensional data analysis: the curses and blessings of dimensionality. In: AMS conference on math challenges of the 21st century, pp 1-33.

111. M. Kirby and L. Sirovich, “Application of the karhunen-loeve procedure for the characterization of human faces,” IEEE Transactions on Pattern analysis and Machine intelligence, vol. 12, no. 1, pp. 103-108, Jan. 1990.

112. Z. Fan, Y. Xu, and D. Zhang, “Local linear discriminant analysis framework using sample neighbors,” *IEEE Transactions on Neural Networks*, vol. 22, no. 7, pp. 1119-1132, Jul. 2011.

113. X. Fang, Y. Xu, X. Li, Z. Lai, and W. K. Wong, “Learning a nonnegative sparse graph for linear regression,” *IEEE Transactions on Image Processing*, vol. 24, no. 9, pp. 2760-2771, Sep. 2015.

114. Z. Qiao, L. Zhou, and J. Z. Huang, “Sparse linear discriminant analysis with applications to high dimensional low sample size data,” *Iaeng International Journal of Applied Mathematics*, vol. 39, no. 1, pp. 48-60, Jan. 2009.

115. Li, Jing, Nigel Allinson, Dacheng Tao, and Xuelong Li. "Multitraining support vector machine for image retrieval." *IEEE Transactions on Image Processing* 15, no. 11 (2006): 3597-3601.

116. L. Qiao, S. Chen, and X. Tan, “Sparsity preserving projections with applications to face recognition,” *Pattern Recognition*, vol. 43, no. 1, pp. 331-341, Jan. 2010.

117. J. Ye and T. Xiong, “Null space versus orthogonal linear discriminant analysis,” in *International Conference on Machine Learning*, Jun. 2006, pp. 1073-1080.

118. J. Yang, D. Zhang, X. Yong, and J.-y. Yang, “Two- dimensional discriminant transform for face recognition,” *Pattern recognition*, vol. 38, no. 7, pp. 1125-1129, Jul. 2005.

119. S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin, “Graph embedding and extensions: A general framework for dimensionality reduction,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 1, pp. 40–51, Jan. 2007.

120. T. Zhang, D. Tao, and J. Yang, “Discriminative locality alignment,” in *European Conference on Computer Vision*, Oct. 2008, pp. 725-738.

121. Y. Zhou and S. Sun, “Manifold partition discriminant analysis,” *IEEE Transactions on Cybernetics*, vol. 47, no. 4, pp. 830-840, 2017.

122. X. Li, W. Hu, H. Wang, and Z. Zhang, “Linear discriminant analysis using rotational invariant l1 norm,” *Neurocomputing*, vol. 73, no. 13-15, pp. 2571–2579, 2010.

123. H. Wang, X. Lu, Z. Hu, and W. Zheng, “Fisher discriminant analysis with l1-norm,” *IEEE Transactions on Cybernetics*, vol. 44, no. 6, pp. 828-842, Jun. 2014.

124. L. Clemmensen, T. Hastie, D. Witten, and B. Ersboll, “Sparse discriminant analysis,” *Technometrics*, vol. 53, no. 4, pp. 406-413, Apr. 2011.

125. X. Zhang, D. Chu, and R. C. Tan, “Sparse uncorrelated linear discriminant analysis for undersampled problems,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 7, pp. 1469-1485, 2016.

126. Z. Li, J. Liu, J. Tang, and H. Lu, “Robust structured subspace learning for data representation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 10, pp. 2085–2098, Oct. 2015.

127. H. Tao, C. Hou, F. Nie, Y. Jiao, and D. Yi, “Effective discriminative feature selection with nontrivial solution.” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 4, pp. 796–808, Apr. 2015.

128. M. Dorfer, R. Kelz, and G. Widmer, “Deep linear discriminant analysis,” in *International Conference on Learning Representations*, 2015, pp. 1-13.

129. R.O. Duda, P.E. Hart, and D.G. Stork, Pattern Classification, second ed. Wiley-Interscience, 2000.

130. S. Xiang, F. Nie, G. Meng, C. Pan, & C. Zhang, (2012). Discriminative least squares regression for multiclass classification and feature selection. *IEEE Transactions on Neural Networks and Learning Systems*, *23*(11), 1738-1754.

131. H. Zou, T. Hastie, & R. Tibshirani, (2006). Sparse principal component analysis. Journal of Computational and Graphical Statistics, 15(2), 265-286.

132. S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, “Distributed optimization and statistical learning via the alternating direction method of multipliers,” *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1-122, 2011.

133. Olivier Chapelle. Training a support vector machine in the primal. Neural Computation, 19(5):1155-1178, 2007.

134. T. Deselaers, D. Keysers, and H. Ney, *“Features for image retrieval: an experimental comparison”*, Information Retrieval, vol. 11, no. 2, pp. 77-107, 2008.

135. G. Sumbul, J. Kang, & B. Demir, (2021). Deep learning for image search and retrieval in large remote sensing archives. Deep Learning for the Earth Sciences: A Comprehensive Approach to Remote Sensing, Climate Science, and Geosciences, 150-160.

136. A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, “*Cnn features off-the-shelf: an astounding baseline for recognition*,” in CVPR workshops, 2014, pp. 806-813.

137. Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, no. 7553, pp. 436-444, 2015.

138. F. Amato, A. Lopez, E. M. Pena-Mendez, P. Vanhara, A. Hampl, and J. Havel, *“Artificial neural networks in medical diagnosis,”* J Appl Biomed, vol. 11, pp. 47-58, 2013.

139. J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, “Gated feedback recurrent neural networks,” in ICML, 2015, pp. 2067-2075.

140. L. Zheng, Y. Yang, and Q. Tian, “Sift meets cnn: A decade survey of instance retrieval,” IEEE TPAMI, vol. 40, no. 5, pp. 1224–1244, 2017.

141. A. Krizhevsky and G.E. Hinton, *Using very deep autoencoders for content-based image retrieval*, in ESANN. Citeseer, (2011).

142. J. Zhang, S. Shan, M. Kan, X. Chen, Coarse-to-fineauto-encoder networks (cfan) for real-time face alignment. In: European Conference on Computer Vision. (2014).

143. Dubey, Shiv Ram. "A decade survey of content based image etrieval using deep learning." IEEE Transactions on Circuits and Systems for Video Technology 32.5 (2021): 2687-2704.

144. M. Chen, P. Zhou, G. Fortino, "Emotion Communication System," IEEE Access, IEEE Access, DOI: 10.1109/ACCESS.2016.2641480, 2016.

145. Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1891-1898,2014.

146. C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," IEEE transactions on pattern analysis and machine intelligence, vol. 35, no. 8, pp. 1915-1929,2013.

147. H. Su, Z. Yin, S. Huh, T. Kanade, and J. Zhu, "Interactive cell segmentation based on active and semi-supervised learning," IEEE transactions on medical imaging, vol. 35, no. 3, pp. 762-777, 2016.

148. K. Kamnitsas, C. Ledig, V. Newcombe, S. Joanna, D. Andrew, K. David, R. Daniel, and G. Ben, "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation," in Medical Image Analysis, vol. 37, pp. 61-78, 2017.

149. G. E. Hinton, "Training products of experts by minimizing contrastive divergence," Neural computation, vol. 14, no. 8,pp.1771-1800,2002.

150. G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," Science, vol. 313, no. 5786, pp. 504-507, 2006.

151. P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in ACM Proceedings of the 25th international conference on Machine learning, pp. 1096-1103, 2008.

152. D. Kumar, A. Wong, and D. A. Clausi, "Lung nodule classification using deep features in ct images," in 12th. IEEE Conference on Computer and Robot Vision (CRV), pp. 133-138, 2015.

153. M. Kallenberg, K. Petersen, M. Nielsen, A. Y. Ng, P. Diao, C. Igel, C. M. Vachon, K. Holland, R. R. Winkel, N. Karssemeijer et al., *"Unsupervised deep learning applied to breast density segmentation and mammographic risk scoring"*, IEEE transactions on medical imaging, vol. 35, no. 5, pp. 1322-1331, 2016.

154. Q. Li, W. Cai, and D. D. Feng, "Lung image patch classification with automatic feature learning," in 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 6079-6082, 2013.

155. Fangxiang Feng, Xiaojie Wang, Ruifan Li, and Ibrar Ahmad. 2015. Correspondence autoencoders for cross-modal retrieval. ACM Transactions on Multimedia Computing, Communications, and Applications 12, 1s, Article 26 (Oct. 2015), 22 pages.

156. Chun Chet Tan. 2008. Autoencoder Neural Networks: A Performance Study Based on Image Recognition, Reconstruction and Compression. Ph.D. Dissertation. Multimedia University.

157. Hao Xue, Like Xue, and Feng Su. 2015. Multimodal music mood classification by fusion of audio and lyrics. In Proceedings of International Conference on MultiMedia Modeling. Springer, 26-37.

158. Minmin Chen, Kilian Q. Weinberger, Fei Sha, and Yoshua Bengio. 2014. Marginalized denoising auto-encoders for nonlinear representations. In Proceedings of International Conference on Machine Learning. 1476–1484.

159. X. Liu, M. Wang, Z.J. Zha, & R. Hong, (2019). Cross-modality feature learning via convolutional autoencoder. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 15(1s), 1-20.

160. D. Tao, X. Tang, X. Li, and Y. Rui, “Direct kernel biased discriminant analysis: a new content-based image retrieval relevance feedback algorithm”, IEEE Transactions on Multimedia, Nol. 8, No. 4, pp. 716-727, 2006.

161. Chen, Yaxiong, Xiaoqiang Lu, and Xuelong Li. "Supervised deep hashing with a joint deep network." Pattern Recognition 105 (2020): 107368.

162. C. Deng, E. Yang, T. Liu, J. Li, W. Liu, & D. Tao, (2019). Unsupervised semantic-preserving adversarial hashing for image search. IEEE Transactions on Image Processing, 28(8), 4032-4044.

163. Mao, Xiaojiao, Chunhua Shen, and Yu-Bin Yang. "Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections." Advances in neural information processing systems 29 (2016).

164. Siradjuddin, Indah Agustien, Wrida Adi Wardana, and Mochammad Kautsar Sophan. "Feature extraction using self-supervised convolutional autoencoder for content based image retrieval." In 2019 3rd International Conference on Informatics and Computational Sciences (ICICoS), pp. 1-5. IEEE, 2019.

165 Pawar, Aashay. "Evaluation of autoencoder for CBIR system in deep learning" In 2020 IEEE 17th India Council International Conference (INDICON), pp. 1-4. IEEE, 2020.

166 Lu Minh Phuc & Tran Cong An (2017). *Tìm kiếm ảnh theo nội dung và ngữ nghĩa*. Tap chi Khoa hoc, Truong ĐH Can Tho. So chuyen đe: CNTT. tr. 58-64.

167 Nguyen Thi Uyen Nhi, Van The Thanh (2021), *Một phương pháp trích xuất đặc trưng cho bài toán tìm kiếm ảnh*, Tap chi Khoa hoc & cong nghe, Truong Đai hoc Khoa học, Đai hoc Hue, Vol. 18(1), tr. 33-46.

168 Nguyen Thi Huyen, Tran Thi Thu Huyen, Vu Thi Luu (2021). *Tìm kiếm ảnh theo nội dung dựa trên mạng nơron tích chập và phương pháp sinh mã nhị phân*. Tap chi Khoa hoc Nong nghiep Viet Nam, J. Agri. Sci, Vol. 19, No. 4: 497-506.