HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG



Trương Thị Thu Hà

NGHIÊN CỨU PHƯƠNG PHÁP TRA CỨU ẢNH ĐA PHƯƠNG THỨC SỬ DỤNG MÔ HÌNH TỰ CHÚ Ý KẾT HỢP MÃ HÓA BĂM SÂU

ĐỀ ÁN TỐT NGHIỆP THẠC SĨ KỸ THUẬT

(Theo định hướng ứng dụng)

HÀ NỘI – NĂM 2025

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG



Trương Thị Thu Hà

NGHIÊN CỨU PHƯƠNG PHÁP TRA CỨU ẢNH ĐA PHƯƠNG THỨC SỬ DỤNG MÔ HÌNH TỰ CHÚ Ý KẾT HỢP MÃ HÓA BĂM SÂU

Chuyên ngành: Khoa học máy tính

Mã số: 8.48.01.01

ĐỀ ÁN TỐT NGHIỆP THẠC SĨ KỸ THUẬT

(Theo định hướng ứng dụng)

NGƯỜI HƯỚNG DẪN KHOA HỌC

TS. ĐÀO THỊ THÚY QUỲNH

HÀ NỘI – NĂM 2025

LỜI CAM ĐOAN

Tôi cam đoan đây là công trình nghiên cứu của riêng tôi.

Các số liệu, kết quả nêu trong đề án tốt nghiệp là trung thực và chưa từng được ai công bố trong bất kỳ công trình nào khác.

Hà Nội, tháng 12 năm 2025

Tác giả đề án



Trương Thị Thu Hà

LỜI CẢM ƠN

Lời đầu tiên, em xin bày tỏ lòng biết ơn sâu sắc và chân thành nhất tới Quý Thầy, Cô giáo – những người đã trực tiếp tham gia giảng dạy và truyền đạt kiến thức trong suốt quá trình em học tập tại Học viện Công nghệ Bưu chính Viễn thông, đặc biệt là các Thầy, Cô thuộc Khoa Công nghệ Thông tin 1 và Khoa Đào tạo sau đại học. Những kiến thức nền tảng và chuyên sâu mà Thầy, Cô trang bị là yếu tố tiên quyết giúp em có thể thực hiện công trình nghiên cứu này.

Đặc biệt, em xin gửi lời tri ân sâu sắc nhất tới Tiến sĩ Đào Thị Thúy Quỳnh, người đã nhận lời hướng dẫn cho em. Cô không chỉ định hướng cho em một đề tài nghiên cứu có ý nghĩa khoa học và thực tiễn, mà còn dành nhiều thời gian, tâm huyết để tận tình hướng dẫn, chỉ bảo, đồng hành và tạo mọi điều kiện thuận lời trong suốt quá trình em thực hiện đề án. Sự hỗ trợ chuyên môn sâu sắc và những lời khuyên quý báu của Cô là động lực và kim chỉ nam giúp em hoàn thành đề án này.

Em cũng xin gửi lời cảm ơn tới gia đình, bạn bè và đồng nghiệp đã luôn động viên, khích lệ và chia sẻ khó khăn, tạo điều kiện tốt nhất để em tập trung hoàn thành chương trình học tập và nghiên cứu của mình.

Em đã nỗ lực hết sức, làm việc nghiêm túc và vận dụng những kiến thức, phương pháp luận đã được học để hoàn thành đề án. Tuy nhiên, do còn hạn chế về kinh nghiệm nghiên cứu cũng như giới hạn về mặt thời gian, đề án chắc chắn không thể tránh khỏi những thiếu sót và hạn chế.

Kính mong Quý Thầy, Cô đóng góp những ý kiến quý báo để đề án của em được hoàn thiện hơn nữa.

Em xin trân trọng cảm ơn!

MỤC LỤC

LỜI CAM ĐOAN i
LỜI CẢM ƠN ii
MỤC LỤC iii
DANH MỤC THUẬT NGỮ, CHỮ VIẾT TẮT v
DANH SÁCH BẢNG vii
DANH SÁCH HÌNH VẼ viii
MỞ ĐẦU 1
1. Lý do chọn đề tài 1
2. Tổng quan về vấn đề nghiên cứu 3
3. Mục đích nghiên cứu 4
4. Đối tượng và phạm vi nghiên cứu 5
5. Phương pháp nghiên cứu 6
NỘI DUNG 7
CHƯƠNG 1: GIỚI THIỆU VỀ TRA CỨU ẢNH 7
1.1. Tổng quan về tra cứu ảnh 7
1.1.1. Khái niệm tra cứu ảnh 7
1.1.2. Phân loại tra cứu ảnh 9
1.2. Tổng quan về tra cứu ảnh đa phương thức: 11
1.2.1. Khái niệm và đặc điểm của truy xuất ảnh đa phương thức 11
1.2.2. Đặc trưng dữ liệu đầu vào (hình ảnh và văn bản) 12
1.2.3. Các thách thức trong việc kết nối đặc trưng hình ảnh và văn bản 15
1.3. Các nghiên cứu liên quan 17
1.4. Kết luận chương 1 20
CHƯƠNG 2: ĐỀ XUẤT MÔ HÌNH TRA CỨU ẢNH ĐA PHƯƠNG THỨC 22
2.1. Phân tích yêu cầu và định hướng mô hình 22
2.1.1. Yêu cầu về hiệu năng và độ chính xác 22
2.1.2. Tính mở rộng và khả năng áp dụng thực tế 23

2.2. Bộ mã hóa hình ảnh 25
2.2.1. Khối Transformer (Transformer Block) 25
2.2.2. Mô đun tương tác (Interaction Module) 30
2.3. Bộ mã hóa văn bản 34
2.3.1. Mã hóa văn bản bằng CLIP 34
2.3.2. Vấn đề bất đối xứng giữa hai dòng đặc trưng 35
2.3.3. Mô đun ánh xạ văn bản (Text Projection Module) 37
2.4. Kiến trúc tổng thể của mô hình đề xuất 38
2.4.1. Ánh xạ không gian đặc trưng chung 41
2.4.2. Mô đun mã hóa băm 42
2.4.3. Hàm mất mát 45
2.4.4. Chiến lược huấn luyện kết hợp hai miền dữ liệu 48
2.5. Kết luận chương 2 51
CHƯƠNG 3: XÂY DỰNG ỨNG DỤNG 52
3.1. Nền tảng và công cụ triển khai 52
3.2. Xây dựng và xử lý dữ liệu thử nghiệm 53
3.3. Phát triển mô hình tra cứu ảnh 56
3.4. Thử nghiệm và đánh giá 58
3.5. Kết luận chương 3 60
KẾT LUẬN VÀ KIẾN NGHỊ 62
1. Các kết quả đạt được của đề án tốt nghiệp 62
2. Nhận xét, đề xuất, khuyến nghị 63
3. Hạn chế 63
4. Hướng nghiên cứu tiếp theo 64
CÔNG BỐ KHOA HỌC 65
DANH MỤC CÁC TÀI LIỆU THAM KHẢO 65

DANH MỤC THUẬT NGỮ, CHỮ VIẾT TẮT

Ký hiệu, chữ viết tắt	Tiếng Anh	Tiếng Việt
SADH	Self-Attention and Deep Hashing model	Mô hình tự chú ý kết hợp mã hóa băm sau
VLP	Vision-Language Pretrained model	Mô hình tiền huấn luyện thị giác–ngôn ngữ
CLIP	Contrastive Language-Image Pre-Training model	Mô hình tiền huấn luyện ngôn ngữ–hình ảnh theo phương pháp tương phản
ViT	Vision-Transformer model	Mô hình bộ biến đổi thị giác
CNN	Convolutional Neural Network	Mạng nơ-ron tích chập
CvT	Convolutional Vision Transformer model	Mô hình bộ biến đổi thị giác kết hợp tích chập
CoAtNet	Convolution and Attention Network	Mạng kết hợp Tích chập và Cơ chế Chú ý
ViLT	Vision-Language Transformer model	Mô hình Transformer cho tác vụ thị giác–ngôn ngữ
BLIP	Bootstrapped Language-Image Pretraining model	Mô hình tiền huấn luyện ngôn ngữ–hình ảnh theo phương pháp "tự khởi tạo"
SSAH	Self-supervised Adversarial Hashing model	Mô hình băm tự giám sát theo cơ chế đối nghịch
DCH	Deep Cauchy Hashing model	Mô hình hashing học sâu dựa trên hàm mất mát Cauchy
CBIR	Content-Based Image Retrieval	Tra cứu ảnh dựa vào nội dung

TBIR	Text-Based Image Retrieval	Tra cứu ảnh dựa vào văn bản
SigLIP	Sigmoid loss for Language Image Pre-training	Hàm mất mát sigmoid dùng để huấn luyện ánh xạ hình ảnh–văn bản
DSH	Deep Supervised Hashing model	Mô hình băm học sâu có giám sát
DPSH	Deep Pairwise-Supervised Hashing model	Mô hình băm học sâu với giám sát theo cặp
LSH	Locality-Sensitive Hashing model	Mô hình băm nhạy cục bộ
MHSA	Multi-head self-attention	Cơ chế tự chú ý đa đầu
FFN	Feed Forward Network	Mạng nơ-ron truyền thẳng
LayerNorm	Layer Normalization	Chuẩn hóa theo lớp
BPE	Byte-Pair Encoding	Mã hóa theo cặp byte
MLP	Multi-layer perceptron	Mạng kết nối đầy đủ nhiều lớp
mAP	Mean Average Precision	Độ chính xác trung bình

DANH SÁCH BẢNG

Bảng 3.1: Nền tảng và công cụ triển khai 533
Bảng 3.2: Tốc độ học theo từng mô đun 5757
Bảng 3.3: Kết quả thực nghiệm đánh giá theo độ đo mAP (%) 59

DANH SÁCH HÌNH VẼ

Hình 2.1: Kiến trúc tổng thể của mô hình SADH 400

Hình 3.1: Hình minh họa bộ dữ liệu MS COCO Captions..... 544
Hình 3.2: Hình minh họa bộ dữ liệu Flickr30k 555

MỞ ĐẦU

1. Lý do chọn đề tài

Trong thời đại công nghệ thông tin bùng nổ, trên nền tảng mạng internet đã rất phát triển, con người đã và đang tạo ra một lượng lớn dữ liệu mỗi ngày. Việc nghiên cứu nhằm xử lý dữ liệu thô dưới dạng văn bản, hình ảnh, âm thanh, video đang rất được quan tâm trên thế giới. Trong đó phải kể đến lĩnh vực xử lý ảnh đang phát triển không ngừng, và thu hút rất nhiều sự quan tâm của các nhà nghiên cứu bởi tính ứng dụng cao của nó. Cùng với sự phát triển của các thiết bị lưu trữ, khối lượng hình ảnh được lưu lại để phục vụ cho nhu cầu sử dụng sau này ngày càng lớn, nhu cầu tra cứu ảnh là tất yếu. Đặc biệt, nhu cầu tra cứu ảnh dựa trên mô tả văn bản ngày càng trở nên phổ biến trong các ứng dụng thực tế như tìm kiếm hình ảnh trên Internet, y học, thương mại, quản lý và giám sát an ninh. Điều này thúc đẩy sự phát triển của các phương pháp tra cứu ảnh đa phương thức.

Tra cứu ảnh đa phương thức (multimodal image retrieval) đóng vai trò quan trọng trong nhiều ứng dụng thực tế như tìm kiếm hình ảnh theo văn bản, quản lý nội dung đa phương tiện, hệ thống đề xuất và các ứng dụng web quy mô lớn. Sự xuất hiện của các mô hình tiền huấn luyện thị giác–ngôn ngữ đã tạo ra bước ngoặt trong lĩnh vực này. Đặc biệt, CLIP [1] cho thấy khả năng học biểu diễn chung giữa hình ảnh và văn bản thông qua huấn luyện tương phản trên tập dữ liệu quy mô hàng trăm triệu cặp ảnh–chú thích. CLIP cung cấp embedding giàu thông tin ngữ nghĩa và đạt hiệu quả vượt trội ở nhiều tác vụ như zero-shot classification, image-text retrieval và semantic matching [1],[2].

Mặc dù hiệu năng biểu diễn của CLIP rất mạnh, mô hình này vẫn gặp hạn chế đáng kể khi áp dụng trong các hệ thống tìm kiếm ảnh quy mô lớn. Cụ thể, đại diện embedding của CLIP thường là vector liên tục kích thước lớn (512–1024 chiều), yêu cầu chi phí tính toán cao khi đo độ tương đồng bằng cosine và tiêu tốn bộ nhớ đáng kể để lưu trữ hàng triệu embedding [3]. Điều này khiến tốc độ truy vấn không đáp

ứng được yêu cầu thời gian thực (real-time) trong các hệ thống sản phẩm. Một hướng tiếp cận quan trọng để giải quyết vấn đề này là deep hashing, cho phép mã hóa các embedding liên tục thành mã nhị phân nén, từ đó tăng tốc truy vấn lên hàng chục lần nhờ sử dụng khoảng cách Hamming [4],[5]. Tuy nhiên, các phương pháp deep hashing truyền thống thường chỉ hoạt động trên dữ liệu đơn phương thức, đòi hỏi nhãn mạnh và không tận dụng được đặc tính đa phương thức mạnh mẽ của các mô hình VLP hiện đại như CLIP. Điều này đặt ra nhu cầu về một mô hình hybrid kết hợp tốt giữa khả năng liên kết ngữ nghĩa của mô hình VLP và tốc độ truy vấn của hashing.

Xuất phát từ nhu cầu thực tiễn và nền tảng công nghệ nêu trên, đề tài "Nghiên cứu phương pháp tra cứu ảnh đa phương thức sử dụng mô hình tự chú ý kết hợp mã hóa băm sâu" (SADH – Self-Attention and Deep Hashing model) được lựa chọn nhằm mục tiêu tận dụng sức mạnh tổng hợp của hai hướng tiếp cận hiện đại: khả năng học biểu diễn mạnh mẽ của mô hình tự chú ý và tính hiệu quả về tốc độ cũng như lưu trữ của mã hóa băm. Mô hình này không chỉ kết hợp được điểm mạnh trong trích xuất đặc trưng hình ảnh của mô hình Transformer và CNN mà còn khai thác được tính khái quát mạnh mẽ của các mô hình pre-trained CLIP và Resnet. Nghiên cứu này không chỉ mang tính thời sự trong lĩnh vực trí tuệ nhân tạo và thị giác máy tính, mà còn có tiềm năng ứng dụng rộng rãi trong thực tiễn.

Để trình bày nội dung một cách hệ thống, khoa học và nhất quán, đề án được tổ chức thành ba chương chính như sau:

Chương 1: Giới thiệu về tra cứu ảnh

Chương này trình bày các khái niệm, cơ sở lý thuyết và mô hình nền tảng liên quan đến tra cứu ảnh và tổng quan các hướng nghiên cứu liên quan đến tra cứu ảnh đa phương thức, mô hình tự chú ý, và mã hóa băm sâu.

Chương 2: Đề xuất mô hình tra cứu ảnh đa phương thức

Chương này trình bày chi tiết mô hình tra cứu ảnh được đề xuất, bao gồm kiến trúc kỹ thuật, chiến lược huấn luyện và cơ chế học biểu diễn kết hợp mã hóa băm sâu.

Chương 3: Xây dựng ứng dụng

Chương này mô tả quá trình hiện thực hóa mô hình đề xuất, từ xây dựng dữ liệu đến huấn luyện, đánh giá và kiểm tra hiệu quả thực tế của hệ thống tra cứu ảnh đa phương thức.

2. Tổng quan về vấn đề nghiên cứu

Tra cứu ảnh đa phương thức là một nhánh nghiên cứu quan trọng trong lĩnh vực trí tuệ nhân tạo, thị giác máy tính và xử lý ngôn ngữ tự nhiên. Vấn đề cốt lõi của bài toán này là xây dựng một hệ thống cho phép người dùng truy vấn hình ảnh thông qua văn bản mô tả , trong đó dữ liệu đầu vào và đầu ra thuộc về hai miền biểu diễn khác nhau.

Truy xuất đa phương thức đặt ra nhiều thách thức, chủ yếu là do sự khác biệt về mặt ngữ nghĩa, cấu trúc và phân bố của các loại dữ liệu (hình ảnh và văn bản). Để giải quyết vấn đề này, các hướng tiếp cận hiện đại thường tập trung vào việc học biểu diễn đặc trưng trong một không gian chung, nơi các mẫu dữ liệu từ các miền khác nhau nhưng mang cùng nội dung sẽ được ánh xạ gần nhau. Đây chính là nền tảng cho các mô hình học biểu diễn liên phương thức.

Trong những năm gần đây, mô hình tự chú ý, đặc biệt là kiến trúc Transformer, đã chứng minh được hiệu quả vượt trội trong việc trích xuất và học đặc trưng ngữ nghĩa từ cả hình ảnh và văn bản. Các mô hình như CLIP (Contrastive Language-Image Pretraining), ViLT (Vision-and-Language Transformer), và BLIP (Bootstrapping Language-Image Pretraining) đã tạo ra bước tiến lớn trong việc huấn luyện mô hình đa phương thức dựa trên bộ dữ liệu quy mô lớn.

Tuy nhiên, khi triển khai trong thực tế, vấn đề hiệu suất và khả năng mở rộng vẫn là rào cản đáng kể, đặc biệt là đối với các hệ thống phải xử lý cơ sở dữ liệu có quy mô lớn. Trong bối cảnh đó, kỹ thuật mã hóa băm sâu được quan tâm nhiều nhờ khả năng biểu diễn dữ liệu bằng các mã nhị phân ngắn gọn, giúp giảm chi phí lưu trữ và tăng tốc độ truy vấn mà vẫn giữ được độ chính xác chấp nhận được. Khi kết hợp

với các kiến trúc học sâu như CNN hoặc Transformer, kỹ thuật băm sâu có thể học được các mã nhị phân mang tính phân biệt cao, phù hợp với mục tiêu truy xuất.

Việc tích hợp mô hình tự chú ý với mã hóa băm sâu trong bài toán truy xuất ảnh đa phương thức là một hướng nghiên cứu mới mẻ, nhằm khai thác đồng thời hai ưu điểm: khả năng học biểu diễn đặc trưng mạnh mẽ và phù hợp giữa các miền dữ liệu của mô hình tự chú ý, và hiệu quả truy xuất cao của mã hóa băm. Tuy nhiên, thách thức đặt ra là làm thế nào để huấn luyện một hệ thống có thể đồng thời tối ưu hóa biểu diễn ngữ nghĩa liên phương thức và sinh ra các mã băm nhị phân hiệu quả.

Do đó, việc nghiên cứu một phương pháp mới kết hợp hai hướng tiếp cận này là cần thiết và có ý nghĩa cả về mặt lý thuyết lẫn thực tiễn, đặc biệt trong bối cảnh các hệ thống tìm kiếm thông minh đang ngày càng phát triển và được ứng dụng rộng rãi.

3. Mục đích nghiên cứu

Mục đích của đề tài là nghiên cứu và đề xuất một phương pháp hiệu quả cho bài toán tra cứu ảnh đa phương thức bằng cách kết hợp mô hình tự chú ý (self-attention) – tiêu biểu là các kiến trúc dựa trên Transformer – với kỹ thuật mã hóa băm sâu (deep hashing). Cụ thể, đề tài hướng đến việc xây dựng một hệ thống có khả năng:

- Học biểu diễn đặc trưng đa phương thức giữa hình ảnh và văn bản trong cùng một không gian đặc trưng tiềm ẩn, thông qua việc khai thác sức mạnh của mô hình tự chú ý trong việc xử lý và kết nối thông tin giữa hai miền dữ liệu khác nhau.
- Sinh mã băm nhị phân hiệu quả từ các biểu diễn đặc trưng này, đảm bảo độ phân biệt cao giữa các mẫu dữ liệu khác nhau, đồng thời giúp giảm thiểu chi phí lưu trữ và tăng tốc độ truy xuất trong hệ thống tra cứu.
- Tối ưu hóa độ chính xác truy xuất ảnh từ văn bản mô tả, cũng như khả năng mở rộng của hệ thống khi làm việc với tập dữ liệu lớn và đa dạng.

Về tổng thể, đề tài không chỉ hướng tới việc cải thiện độ chính xác và hiệu suất trong bài toán tra cứu ảnh đa phương thức, mà còn góp phần mở rộng khả năng

ứng dụng của các mô hình học sâu hiện đại trong các hệ thống truy xuất thông tin thông minh. Kết quả nghiên cứu có thể làm nền tảng cho các ứng dụng thực tiễn như tìm kiếm ảnh thông minh, gợi ý nội dung, quản lý dữ liệu đa phương tiện, và các hệ thống trợ lý ảo.

4. Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu của đề tài là các phương pháp tra cứu ảnh đa phương thức, trong đó tập trung vào việc:

- Nghiên cứu cơ chế học biểu diễn đặc trưng đa phương thức giữa văn bản và hình ảnh bằng mô hình tự chú ý, tiêu biểu là kiến trúc Transformer.
- Tìm hiểu và áp dụng kỹ thuật mã hóa băm sâu nhằm biểu diễn các đặc trưng đa phương thức dưới dạng mã nhị phân phục vụ cho truy vấn nhanh và hiệu quả.
- Đánh giá khả năng kết hợp giữa hai thành phần trên để xây dựng một hệ thống tra cứu ảnh có độ chính xác cao và hiệu suất truy vấn tốt.

Phạm vi nghiên cứu của đề tài được giới hạn trong những nội dung sau:

- Chỉ tập trung vào tra cứu ảnh với dữ liệu đầu vào bao gồm hình ảnh và văn bản mô tả, chưa mở rộng sang các loại dữ liệu khác như âm thanh hay video.
- Tìm hiểu và áp dụng các kiến trúc tự chú ý phổ biến, như Transformer hoặc các biến thể gần đây được thiết kế cho dữ liệu hình ảnh và văn bản.
- Áp dụng mã hóa băm sâu dưới hình thức học có giám sát, với đầu ra là các mã nhị phân có độ dài cố định.
- Không đi sâu vào tối ưu hóa hệ thống phần cứng hay các kỹ thuật triển khai phân tán trên quy mô lớn; thay vào đó, đề tài tập trung vào phương diện mô hình hóa và hiệu quả truy xuất ở cấp độ thuật toán và kiến trúc mạng.

5. Phương pháp nghiên cứu

Để thực hiện hóa mục tiêu nghiên cứu, đề tài sử dụng kết hợp các phương pháp định tính và định lượng, cụ thể như sau:

- Phân tích tài liệu: Tổng hợp và phân tích các tài liệu học thuật, công trình nghiên cứu gần đây liên quan đến các chủ đề: tra cứu ảnh đa phương thức, mô hình tự chú ý, học biểu diễn đa phương thức và kỹ thuật mã hóa băm sâu. Từ đó, xây dựng cơ sở lý thuyết vững chắc cho đề tài và xác định hướng tích hợp mô hình phù hợp.
- Khảo sát và đánh giá các mô hình hiện có: Tiến hành khảo sát và thử nghiệm một số mô hình nổi bật trong lĩnh vực học đa phương thức như CLIP, ViLT, SAH, DCH... để đánh giá hiệu năng và xác định những điểm mạnh, điểm yếu của từng phương pháp. Việc này giúp làm rõ khoảng trống nghiên cứu và cơ hội cải tiến thông qua tích hợp các kỹ thuật mới.
- Thiết kế và xây dựng mô hình đề xuất: Đề xuất kiến trúc mô hình kết hợp giữa mô hình tự chú ý cho việc học biểu diễn đặc trưng liên phương thức và kỹ thuật mã hóa băm sâu để ánh xạ dữ liệu vào không gian nhị phân. Mô hình sẽ được thiết kế sao cho vừa học tốt biểu diễn ngữ nghĩa giữa ảnh – văn bản, vừa đảm bảo hiệu suất truy xuất cao nhờ mã băm.
- Phát triển nguyên mẫu hệ thống: Xây dựng nguyên mẫu mô hình tra cứu ảnh đa phương thức bằng Python, sử dụng các thư viện học sâu như PyTorch hoặc TensorFlow. Quá trình này bao gồm xử lý dữ liệu đầu vào, huấn luyện mô hình, sinh mã băm và xây dựng công cụ truy vấn ảnh.
- Thử nghiệm và đánh giá hiệu quả mô hình: Thực hiện các kịch bản thử nghiệm trên các tập dữ liệu chuẩn để đánh giá hiệu năng hệ thống dựa trên các chỉ số về độ chính xác và thời gian truy xuất ảnh. Đồng thời so sánh với các mô hình cơ sở để chứng minh tính hiệu quả của phương pháp đề xuất.

NỘI DUNG

CHƯƠNG 1: GIỚI THIỆU VỀ TRA CỨU ẢNH

Trong bối cảnh khối lượng dữ liệu hình ảnh ngày càng gia tăng với tốc độ chưa từng có, bài toán tra cứu ảnh đã trở thành một chủ đề nghiên cứu quan trọng trong lĩnh vực thị giác máy tính và truy xuất thông tin. Khả năng tìm kiếm và truy xuất hình ảnh dựa trên nội dung hoặc mô tả văn bản đóng vai trò then chốt trong nhiều ứng dụng thực tế như thương mại điện tử, hệ thống đề xuất, tìm kiếm bằng thị giác, phân tích dữ liệu đa phương thức, và lưu trữ số. Chương này trình bày những nền tảng lý thuyết cần thiết cho việc nghiên cứu bài toán tra cứu ảnh và tra cứu ảnh đa phương thức, bao gồm các khái niệm cơ bản, phân loại, đặc trưng dữ liệu, và các hướng nghiên cứu liên quan. Những nội dung này đóng vai trò làm cơ sở lý luận để phát triển mô hình đề xuất ở các chương tiếp theo.

1.1. Tổng quan về tra cứu ảnh

1.1.1. Khái niệm tra cứu ảnh

Tra cứu ảnh là quá trình tìm kiếm và thu hồi những hình ảnh liên quan nhất từ một cơ sở dữ liệu dựa trên thông tin truy vấn do người dùng cung cấp. Khác với cách tiếp cận truyền thống phụ thuộc vào siêu dữ liệu – chẳng hạn như tên tệp, thẻ mô tả do con người gán, hoặc thông tin phân loại – tra cứu ảnh hiện đại tập trung vào việc phân tích trực tiếp nội dung thị giác của ảnh để đưa ra kết quả tìm kiếm chính xác hơn. Bản chất của tra cứu ảnh nằm ở việc chuyển đổi hình ảnh thành các biểu diễn số đủ mạnh để mô hình có thể so sánh, đo đạc mức độ tương đồng và từ đó truy xuất những hình ảnh phù hợp nhất.

Trong thực tế, tra cứu ảnh được triển khai thông qua hai bước chính: biểu diễn ảnh và đo độ tương đồng. Bước đầu tiên đòi hỏi hệ thống phải xây dựng được vector đặc trưng phản ánh thông tin thị giác quan trọng như màu sắc, cấu trúc, hình dạng, bố cục đối tượng và đặc trưng ngữ nghĩa. Các đặc trưng này có thể được trích xuất bằng các thuật toán thủ công hoặc bằng mô hình học sâu, chẳng hạn như CNN hoặc

Vision Transformer (ViT), nơi mỗi ảnh được ánh xạ vào một điểm trong không gian có chiều cao. Bước thứ hai sử dụng các thước đo tương đồng như khoảng cách Euclidean, cosine hoặc Hamming để đánh giá mức độ giống nhau giữa truy vấn và các ảnh trong cơ sở dữ liệu. Sự kết hợp của hai bước này quyết định chất lượng và hiệu quả của toàn bộ hệ thống tra cứu.

Một khía cạnh quan trọng của tra cứu ảnh là yêu cầu mô hình phải có khả năng phân biệt giữa các loại thông tin thị giác. Một ảnh có thể chứa nhiều đối tượng, nhiều khung cảnh và tồn tại trong những điều kiện nhiễu như ánh sáng thay đổi, góc chụp bất thường, che khuất một phần hoặc biến dạng hình học. Vì vậy, việc xây dựng biểu diễn sao cho vừa ổn định trước những thay đổi tự nhiên của ảnh, vừa nhạy bén với sự khác biệt quan trọng giữa các ảnh là nhiệm vụ cốt lõi của lĩnh vực này. Các phương pháp hiện đại dựa trên học sâu thường hướng đến việc học biểu diễn ngữ nghĩa ở mức cao, nơi hình ảnh không chỉ được mô tả bằng đặc trưng thị giác đơn thuần mà còn được gắn với các khái niệm ý nghĩa, giúp mô hình hiểu sâu hơn về nội dung của ảnh.

Tra cứu ảnh không chỉ mang ý nghĩa về mặt kỹ thuật mà còn có ý nghĩa thực tiễn sâu rộng. Trong thương mại điện tử, người dùng có thể tìm sản phẩm bằng hình ảnh chụp từ điện thoại; trong y tế, bác sĩ tìm kiếm các hình ảnh X-quang tương tự để hỗ trợ chẩn đoán; trong an ninh, hệ thống giám sát truy tìm khuôn mặt hoặc đối tượng trong tập dữ liệu lớn. Các ứng dụng này yêu cầu mô hình vừa chính xác, vừa có khả năng mở rộng khi dữ liệu tăng lên. Điều này đặt ra nhu cầu nghiên cứu các kỹ thuật biểu diễn hiệu quả, tối ưu hóa tốc độ truy vấn và giảm chi phí lưu trữ.

Như vậy, tra cứu ảnh là một bài toán tổng hợp, đòi hỏi sự kết hợp giữa biểu diễn đặc trưng, mô hình hóa ngữ nghĩa và tối ưu hóa hiệu năng truy xuất. Đây cũng chính là nền tảng cho các hệ thống tra cứu ảnh đa phương thức ở những phần sau, nơi mô hình không chỉ xử lý thông tin thị giác mà còn phải học cách liên kết chúng với văn bản tự nhiên.

1.1.2. Phân loại tra cứu ảnh

Việc phân loại các hệ thống tra cứu ảnh giúp định hướng rõ ràng cho quá trình thiết kế mô hình, đồng thời đưa ra lựa chọn tối ưu về biểu diễn, thuật toán và kiến trúc phù hợp cho từng ứng dụng cụ thể. Trong lịch sử phát triển, tra cứu ảnh đã được tiếp cận theo nhiều hướng khác nhau, tùy thuộc vào loại truy vấn, cách biểu diễn đặc trưng và mục tiêu ứng dụng. Một số phân loại nổi bật bao gồm tra cứu dựa trên nội dung thị giác, tra cứu dựa trên văn bản, và các hệ thống tra cứu ảnh đa phương thức,... Phân loại tra cứu ảnh không chỉ giúp xác định các thách thức đặc thù của từng loại hệ thống mà còn mở đường cho những mô hình kết hợp - nơi ưu điểm của nhiều kỹ thuật được hội tụ để giải quyết các bài toán thực tiễn phức tạp hơn.

Tra cứu ảnh dựa trên nội dung – CBIR (Content-Based Image Retrieval)

CBIR là cách tiếp cận cổ điển và nền tảng nhất, trong đó hình ảnh truy vấn được so khớp trực tiếp với hình ảnh trong cơ sở dữ liệu thông qua các đặc trưng thị giác. Các hệ thống CBIR đầu tiên sử dụng đặc trưng thủ công như histogram màu, texture, hoặc mô tả hình dạng; tuy nhiên, những đặc trưng này thường mang tính cục bộ, kém bền vững và không nắm bắt được ý nghĩa trừu tượng. Sự xuất hiện của học sâu, đặc biệt là CNN và Vision Transformer, đã tạo ra bước tiến lớn cho CBIR, khi các biểu diễn học được có khả năng mã hóa ngữ nghĩa toàn cục của hình ảnh. Trong CBIR, người dùng cung cấp một hình ảnh mẫu, và hệ thống tìm ảnh tương tự dựa trên các thước đo trong không gian đặc trưng. Đây là hướng nghiên cứu phù hợp cho các bài toán tìm kiếm theo hình thức thị giác thuần túy.

Tra cứu ảnh dựa trên văn bản – TBIR (Text-Based Image Retrieval)

TBIR mở rộng khả năng truy xuất bằng cách cho phép người dùng mô tả hình ảnh cần tìm bằng ngôn ngữ tự nhiên. Trong TBIR, mô hình phải đảm nhiệm hai nhiệm vụ: hiểu nội dung văn bản và liên kết văn bản với hình ảnh tương ứng trong cơ sở dữ liệu. Trước đây, TBIR chủ yếu dựa vào từ khóa hoặc nhãn gắn thủ công, khiến hệ thống bị giới hạn bởi chất lượng và độ đầy đủ của nhãn. Các phương pháp hiện đại học biểu diễn văn bản thông qua mô hình Transformer và ánh xạ chúng vào một

không gian so sánh chung với đặc trưng hình ảnh. TBIR đặc biệt quan trọng trong những ứng dụng mà hình ảnh khó mô tả bằng mẫu trực quan nhưng dễ mô tả bằng câu văn – chẳng hạn tìm ảnh dựa trên hành động, bối cảnh, hoặc mối quan hệ giữa các đối tượng.

Tra cứu ảnh đa phương thức (Multimodal Image Retrieval)

Tra cứu ảnh đa phương thức là sự mở rộng tự nhiên của CBIR và TBIR, khi hệ thống cần xử lý đồng thời hai loại dữ liệu mang bản chất hoàn toàn khác nhau: hình ảnh và văn bản. Khác với CBIR – chỉ dựa vào đặc trưng thị giác và TBIR – chỉ dựa vào biểu diễn ngôn ngữ, các hệ thống đa phương thức học không gian biểu diễn dùng chung (joint embedding space), trong đó hình ảnh và văn bản có thể so sánh trực tiếp. Đây là bước tiến lớn nhờ sự phát triển của các mô hình liên kết thị giác – ngôn ngữ (vision–language models) như CLIP, ALIGN và SigLIP. Các hệ thống này không chỉ thực hiện tra cứu ảnh bằng văn bản và tra cứu văn bản bằng ảnh mà còn hỗ trợ nhiều tác vụ liên quan như phân loại không giám sát, gợi ý nội dung hoặc phân tích tương tác đa phương thức.

Phân loại theo dạng đặc trưng và phương pháp so khớp

Bên cạnh cách phân loại theo truy vấn, các hệ thống tra cứu ảnh còn được phân chia theo dạng đặc trưng sử dụng. Các đặc trưng thủ công phù hợp cho những hệ thống nhỏ hoặc bài toán đơn giản, trong khi đặc trưng học sâu đóng vai trò chủ đạo trong các hệ thống hiện đại nhờ khả năng biểu diễn ngữ nghĩa. Một phân loại quan trọng nữa là mức độ nhị phân hóa đặc trưng. Các phương pháp hashing, đặc biệt là deep hashing, cho phép chuyển embedding liên tục thành mã nhị phân, giúp giảm chi phí lưu trữ và tăng tốc độ truy vấn trong các cơ sở dữ liệu lớn. Điều này đặc biệt hữu ích cho cả CBIR và TBIR, cũng như cho tra cứu đa phương thức.

Phân loại theo cấu trúc mô hình

Dựa trên kiến trúc mô hình, ta có thể phân loại hệ thống thành nhóm dựa trên CNN, nhóm dựa trên Transformer, nhóm mô hình kết hợp (hybrid CNN–attention), và nhóm mô hình thị giác–ngôn ngữ tiền huấn luyện (vision–language foundation

models). Mỗi nhóm có ưu điểm riêng: CNN giỏi đặc trưng cục bộ, Transformer mạnh ở mối quan hệ toàn cục, còn mô hình đa phương thức có khả năng liên kết ngữ nghĩa giữa các modal.

1.2. Tổng quan về tra cứu ảnh đa phương thức:

1.2.1. Khái niệm và đặc điểm của truy xuất ảnh đa phương thức

Tra cứu ảnh đa phương thức là một hướng nghiên cứu phát triển từ bài toán tra cứu ảnh truyền thống, trong đó hệ thống phải xử lý đồng thời nhiều dạng dữ liệu khác nhau, phổ biến nhất là hình ảnh và văn bản. Khác với tra cứu dựa trên nội dung thị giác chỉ sử dụng thông tin thị giác, và tra cứu dựa trên văn bản chỉ khai thác thông tin ngôn ngữ, tra cứu ảnh đa phương thức yêu cầu mô hình học được một không gian biểu diễn chung, nơi đặc trưng của các phương thức có thể được so sánh một cách trực tiếp và có ý nghĩa ngữ nghĩa. Đây chính là điểm khác biệt quan trọng tạo nên giá trị và sức mạnh của tra cứu đa phương thức trong các ứng dụng thực tế.

Về bản chất, bài toán này đặt ra mục tiêu xây dựng sự tương thích giữa các dạng thông tin có bản chất hoàn toàn khác nhau. Hình ảnh biểu hiện thông tin dưới dạng tín hiệu không gian liên tục, trong khi văn bản là chuỗi ký hiệu rời rạc được tổ chức theo cấu trúc cú pháp và mang ý nghĩa trừu tượng. Sự khác biệt về bản chất dữ liệu khiến tra cứu ảnh đa phương thức trở thành một bài toán thách thức và cần đến các mô hình có khả năng hiểu ngữ nghĩa sâu. Các mô hình hiện đại thường sử dụng hai bộ mã hóa riêng biệt cho hình ảnh và văn bản, sau đó ánh xạ cả hai loại đặc trưng vào cùng một không gian thông qua các ràng buộc tương thích ngữ nghĩa—thường dưới dạng "contrastive learning" hoặc "cross-modal alignment".

Một đặc điểm nổi bật của tra cứu ảnh đa phương thức là tính linh hoạt trong tương tác với người dùng. Thay vì yêu cầu người dùng cung cấp hình ảnh mẫu khó có sẵn như trong CBIR, tra cứu đa phương thức cho phép sử dụng mô tả văn bản tự nhiên, từ đó làm tăng mức độ thân thiện và khả dụng trong nhiều tình huống thực tế. Ví dụ, người dùng có thể tìm kiếm hình ảnh dựa trên mô tả "một chú mèo đang ngồi trên bàn cạnh cửa sổ", điều mà hệ thống CBIR không thể xử lý trực tiếp nếu không

có ảnh mẫu tương ứng. Nhờ khả năng diễn giải ngôn ngữ, tra cứu đa phương thức có thể hiểu các yếu tố không xuất hiện rõ ràng trong thị giác, chẳng hạn hành vi, bối cảnh hoặc mối quan hệ giữa các đối tượng.

Một đặc tính quan trọng khác là khả năng tổng hợp và khái quát hóa ngữ nghĩa. Các hệ thống tra cứu ảnh đa phương thức, đặc biệt là các mô hình thị giác–ngôn ngữ tiền huấn luyện VLP (Vision–Language Pretrained), được huấn luyện trên tập dữ liệu rất lớn với các cặp ảnh – văn bản phong phú. Điều này giúp mô hình học được không chỉ đặc trưng thị giác và ngôn ngữ độc lập mà còn hiểu cách chúng gắn kết với nhau trong đời sống thực tế. Nhờ đó, không gian biểu diễn chung được xây dựng có khả năng khái quát tốt ngay cả với các dữ liệu chưa từng xuất hiện trong quá trình huấn luyện, cho phép mô hình hoạt động hiệu quả trong môi trường truy vấn mở.

Ngoài ra, tra cứu ảnh đa phương thức có tính chất mở rộng mạnh mẽ, cho phép kết hợp thêm các modal khác như âm thanh, dữ liệu cảm biến hoặc thông tin ngữ cảnh. Tuy nhiên, việc tích hợp mỗi modal mới đòi hỏi mô hình phải mở rộng không gian biểu diễn và đảm bảo các modal duy trì được tính tương thích ngữ nghĩa. Đây chính là lý do các mô hình dựa trên cơ chế tự chú ý và contrastive learning trở thành nền tảng quan trọng trong lĩnh vực này.

1.2.2. Đặc trưng dữ liệu đầu vào (hình ảnh và văn bản)

Trong bài toán tra cứu ảnh đa phương thức, hai loại dữ liệu trung tâm là hình ảnh và văn bản mô tả. Cùng hướng tới việc biểu đạt thông tin về một đối tượng hoặc bối cảnh, nhưng hai modal này lại khác nhau căn bản về cấu trúc, cách mã hóa và mức độ trừu tượng. Vì vậy, trước khi xây dựng một không gian biểu diễn thống nhất, cần phân tích rõ đặc điểm riêng của từng modal để lựa chọn chiến lược mô hình hóa phù hợp.

Dữ liệu hình ảnh

Hình ảnh số là tín hiệu liên tục trong không gian hai chiều, chứa lượng lớn thông tin về màu sắc, biên dạng, kết cấu, hình khối và bố cục tổng thể của cảnh. Các

đặc trưng thị giác không tồn tại rời rạc mà được tổ chức theo cấu trúc không gian: vị trí, kích thước và mối quan hệ giữa các vùng ảnh đóng vai trò quan trọng trong việc giải mã nội dung. Tuy nhiên, dữ liệu ảnh trong thực tế thường chịu ảnh hưởng của nhiều yếu tố gây nhiễu như thay đổi ánh sáng, góc chụp, độ phân giải, hiện tượng che khuất hoặc phông nền phức tạp, khiến việc trích xuất đặc trưng ổn định trở nên khó khăn.

Các mô hình học sâu hiện đại xử lý vấn đề này bằng cách sử dụng những backbone mạnh cho thị giác. Mạng CNN như ResNet [7] khai thác cơ chế convolution để học đặc trưng cục bộ, hình thành dần các đặc trưng trung gian và khái niệm ở tầng sâu. Trong khi đó, Vision Transformer (ViT) [6] chia ảnh thành các patch cố định và xử lý chúng như một chuỗi token trong kiến trúc Transformer, sử dụng self-attention để mô hình hóa quan hệ toàn cục giữa các vùng ảnh. Các kiến trúc lai như CvT [11] và CoAtNet [12] kết hợp ưu điểm của CNN (inductive bias cục bộ) và tự chú ý (khả năng nắm bắt quan hệ dài hạn), từ đó thu được đặc trưng vừa chi tiết vừa khái quát.

Đầu ra của các backbone này thường là véc-tơ liên tục có số chiều lớn (từ vài trăm đến hàng nghìn), mã hóa đồng thời thông tin cục bộ và ngữ nghĩa mức cao. Tuy nhiên, các embedding hình ảnh vẫn mang đậm tính hình học và phân phối thống kê đặc trưng của miền thị giác, nên không tự nhiên "ăn khớp" với không gian biểu diễn của ngôn ngữ.

Dữ liệu văn bản

Văn bản mô tả ảnh là một chuỗi token rời rạc (từ, cụm từ) được tổ chức theo cú pháp ngôn ngữ tự nhiên. Khác với ảnh, văn bản không phản ánh trực tiếp các đặc trưng mức thấp như màu sắc hay texture, mà tập trung thể hiện các khái niệm trong miền trừu tượng hơn: danh từ chỉ đối tượng, động từ mô tả hành động, tính từ thể hiện thuộc tính, và đôi khi là cả thái độ, bối cảnh hoặc ý nghĩa hàm ẩn mà ảnh không thể hiện rõ. Vì mang tính tóm tắt và chọn lọc, văn bản thường chỉ nêu ra một phần nội dung của ảnh thay vì tái tạo toàn bộ.

Các mô hình mã hóa văn bản hiện đại dựa trên kiến trúc Transformer, điển hình là text encoder trong CLIP [1] hoặc ALIGN [2]. Chúng ánh xạ chuỗi token sang embedding liên tục và học cấu trúc câu, quan hệ ngữ nghĩa thông qua cơ chế self-attention. Nhờ đó, mỗi câu chú thích được biểu diễn bởi một véc-tơ gọn, nhưng chứa lượng thông tin ngữ nghĩa đáng kể về nội dung mô tả tổng thể.

So với hình ảnh, văn bản có cấu trúc hình thức ổn định hơn, nhưng lại chịu ảnh hưởng mạnh của tính chủ quan và phong cách diễn đạt. Cách lựa chọn từ ngữ, mức độ chi tiết, giọng điệu, hay thói quen ngôn ngữ của người chú thích đều có thể tạo ra sai lệch giữa nội dung thực của ảnh và nội dung được mô tả. Vì vậy, mặc dù embedding văn bản giàu ngữ nghĩa, nó không phải lúc nào cũng trùng khớp hoàn toàn với đặc trưng thị giác của ảnh tương ứng.

Sự dị thể (heterogeneity) giữa hai modal

Hình ảnh và văn bản khác nhau không chỉ ở dạng biểu diễn (liên tục và rời rạc) mà còn ở phân phối thống kê, mức độ trừu tượng và loại thông tin mang theo. Hình ảnh thiên về cấu trúc không gian và các đặc trưng mức thấp, trong khi văn bản tập trung vào biểu đạt các khái niệm ngữ nghĩa mức cao. Sự chênh lệch này tạo ra một "modal gap" rõ rệt, khiến việc đưa hai dạng dữ liệu vào cùng một không gian embedding trở nên không tầm thường.

Trong thực tế, một câu mô tả có thể quy chiếu đồng thời tới nhiều vùng ảnh khác nhau; ngược lại, một ảnh có thể chứa vô số chi tiết thị giác mà văn bản không đề cập. Văn bản cũng có thể chứa thông tin ngữ cảnh nằm ngoài những gì có thể quan sát trực tiếp trong ảnh (chẳng hạn cảm xúc, mục đích hoặc bối cảnh thời gian – địa điểm). Điều này đòi hỏi không gian biểu diễn chung phải vừa đủ linh hoạt để chứa được các khái niệm trừu tượng của ngôn ngữ, vừa đủ nhạy để phản ánh những khác biệt quan trọng trong miền thị giác.

Vai trò của embedding chung trong việc kết nối hai modal

Để thu hẹp khoảng cách giữa hai modal dị thể, các mô hình đa phương thức như CLIP [1], ALIGN [2] và SigLIP [13] học một không gian embedding chung,

trong đó embedding ảnh và embedding văn bản tương ứng được kéo gần nhau, còn các cặp không tương ứng bị đẩy xa. Các mô hình này thường sử dụng contrastive learning để tối ưu hóa sự tương thích giữa hai không gian, nhờ đó thiết lập được mối quan hệ ngữ nghĩa chung giữa hình ảnh và văn bản.

Khi đã có không gian chung, hệ thống có thể thực hiện tra cứu dựa trên những độ đo đơn giản như khoảng cách cosine cho embedding thực, hoặc khoảng cách Hamming trong trường hợp embedding được lượng tử hóa thành mã băm nhị phân [3][4][5]. Không gian embedding chung vì vậy đóng vai trò cầu nối giữa hai dạng biểu diễn khác biệt, là nền tảng cho các hệ thống tra cứu ảnh đa phương thức vừa chính xác vừa hiệu quả về tính toán. Tuy nhiên, chất lượng của không gian này phụ thuộc trực tiếp vào cách mô hình hóa đặc trưng ban đầu của ảnh và văn bản, do đó việc phân tích và thiết kế hợp lý cho từng phương thức là bước khởi đầu quan trọng để xây dựng một hệ thống hoàn chỉnh.

1.2.3. Các thách thức trong việc kết nối đặc trưng hình ảnh và văn bản

Việc kết nối hai phương thức dị thể – hình ảnh và văn bản – là trọng tâm của các hệ thống tra cứu ảnh đa phương thức và đồng thời cũng là một trong những thách thức lớn nhất của lĩnh vực này. Dù nhiều mô hình thị giác–ngôn ngữ hiện đại như CLIP [1], ALIGN [2] hay SigLIP [13] đã đạt được thành tựu đáng kể trong việc ánh xạ ảnh và văn bản vào cùng một không gian biểu diễn, vẫn tồn tại những vấn đề cơ bản khiến quá trình này phức tạp và thiếu ổn định. Những thách thức này bắt nguồn từ sự khác biệt bản chất giữa hai dạng dữ liệu, từ tính chất của các mô hình mã hóa, cho đến sự nhiễu loạn vốn có của dữ liệu thực tế.

Một vấn đề quan trọng đầu tiên là khoảng cách biểu giữa hai loại dữ liệu. Hình ảnh được biểu diễn dưới dạng tín hiệu hai chiều liên tục giàu thông tin hình học, trong khi văn bản là chuỗi token rời rạc mang cấu trúc logic và ngữ nghĩa. Các bộ mã hóa hình ảnh, gồm mạng CNN truyền thống như ResNet [7] hoặc các mô hình dựa trên Transformer như ViT [6] và CoAtNet [12], học được các biểu diễn rất khác so với các bộ mã hóa văn bản dựa trên Transformer [1][2]. Do đó, ngay cả khi hai dữ liệu

mô tả cùng một nội dung, biểu diễn của chúng có thể nằm ở hai phân phối hoàn toàn khác biệt trong không gian embedding. Điều này buộc mô hình phải sử dụng các chiến lược điều chỉnh như "contrastive learning" hoặc "projection layers" nhằm thu hẹp khoảng cách này.

Thách thức tiếp theo nằm ở sự không đối xứng về thông tin giữa hình ảnh và văn bản. Văn bản mô tả không phải lúc nào cũng phản ánh đầy đủ hoặc chính xác nội dung ảnh. Người chú thích thường chỉ đề cập đến các đối tượng hoặc hành động nổi bật, bỏ qua nhiều chi tiết thị giác mức thấp, chẳng hạn như texture, pattern hoặc bố cục tổng thể. Ngược lại, văn bản đôi khi bổ sung thông tin không xuất hiện trực tiếp trong ảnh, chẳng hạn mô tả cảm xúc ("một buổi chiều bình yên") hay ý định ngữ nghĩa ("đang chuẩn bị khởi hành"). Độ chênh này khiến việc ánh xạ trở thành bài toán nhiều–một hoặc một–nhiều phức tạp, đặc biệt khi mô hình phải nén biểu diễn vào dạng mã băm nhị phân [3][4][5].

Một thách thức sâu hơn là sự lệch mức trừu giữa hai loại dữ liệu. Hình ảnh chứa nhiều lớp thông tin khác nhau, từ đặc trưng mức thấp (biên, màu, kết cấu) đến đặc trưng mức trung (hình dạng đối tượng) và mức cao (mối quan hệ giữa các đối tượng trong bối cảnh). Trong khi đó, văn bản thường mô tả trực tiếp ở mức ngữ nghĩa cao: hành động, quan hệ, cảm xúc hoặc các khái niệm trừu tượng. Kết quả là biểu diễn hình ảnh có xu hướng mang tính chi tiết cục bộ, còn văn bản thiên về tổng quát hóa ngữ nghĩa. Nếu mô hình không kết hợp hiệu quả đặc trưng cục bộ và toàn cục – như cách tiếp cận của các mô hình hybrid CNN–ViT [11][12] – thì véc-tơ ảnh và véc-tơ văn bản tương ứng có thể không hội tụ về cùng một vùng trong không gian embedding, khiến chất lượng truy vấn giảm đáng kể.

Ngoài ra, dữ liệu thực tế thường đi kèm với nhiễu và sự thiếu nhất quán (real-world noise). Hình ảnh có thể bị ảnh hưởng bởi điều kiện chụp như ánh sáng, chuyển động, che khuất hoặc chất lượng thấp; trong khi văn bản mô tả có thể bị sai ngữ pháp, diễn đạt mơ hồ hoặc được viết theo phong cách chủ quan của người dùng. Các mô hình huấn luyện trên dữ liệu quy mô lớn như CLIP [1] và ALIGN [2] đã phần nào

thích ứng với dữ liệu nhiễu, nhưng điều này vẫn gây ảnh hưởng đáng kể đến việc học không gian biểu diễn ổn định. Trong bối cảnh mà biểu diễn phải được lượng tử hóa thành mã nhị phân [3][5], sự dao động nhỏ trong embedding liên tục có thể dẫn đến những thay đổi lớn về bit, làm giảm hiệu quả truy vấn.

Cuối cùng, dù các mô hình đa phương thức hiện đại đã tiến rất xa, chúng vẫn gặp hạn chế trong việc cân bằng giữa tính biểu diễn mạnh mẽ và khả năng nén hiệu quả. Embedding liên tục từ CLIP [1] hoặc SigLIP [13] rất giàu ngữ nghĩa nhưng khó chuyển đổi trực tiếp sang mã băm ổn định mà không làm mất thông tin quan trọng. Bên cạnh đó, các mô hình như ViT [6] còn hạn chế trong việc xử lý chi tiết cục bộ, dẫn đến xu hướng bỏ sót các yếu tố quan trọng mà văn bản mô tả đề cập đến. Những thách thức này là động lực thúc đẩy sự ra đời của các phương pháp lai, như mô hình đề xuất SADH, nhằm kết hợp ưu điểm của các kiến trúc CNN và Transformer cùng với cơ chế băm sâu để tạo ra không gian biểu diễn đa phương thức vừa giàu ngữ nghĩa, vừa tối ưu cho truy vấn tốc độ cao.

1.3. Các nghiên cứu liên quan

Nghiên cứu về tra cứu ảnh đa phương thức đã phát triển mạnh trong những năm gần đây, nhờ sự kết hợp giữa học sâu, mô hình đa modal và các phương pháp tối ưu hóa tìm kiếm quy mô lớn. Các công trình liên quan có thể chia thành bốn hướng tiếp cận chính: mô hình thị giác–ngôn ngữ tiền huấn luyện, (2) hợp nhất đặc trưng đa phương thức, (3) mô hình trích xuất đặc trưng hình ảnh lai giữa CNN và ViT (Hybrid Visual Encoders), và (4) các phương pháp hashing và deep hashing cho truy vấn tốc độ cao. Những hướng nghiên cứu này tạo thành nền tảng vững chắc cho sự phát triển của các hệ thống tra cứu ảnh hiện đại và đóng vai trò quan trọng trong việc định hình mô hình SADH được đề xuất trong đề án này.

Các mô hình thị giác–ngôn ngữ tiền huấn luyện

Một trong những hướng nghiên cứu có ảnh hưởng lớn nhất đến tra cứu ảnh đa phương thức là các mô hình thị giác–ngôn ngữ được tiền huấn luyện trên dữ liệu quy mô web. CLIP của Radford và các công sự [1] đánh dấu bước ngoặt quan trọng, sử

dụng hơn 400 triệu cặp ảnh–văn bản để học không gian biểu diễn chung thông qua contrastive learning. Mô hình này không chỉ đạt hiệu quả vượt trội trong phân loại ảnh zero-shot mà còn tạo nên tiêu chuẩn mới cho các bài toán retrieval hai chiều (image-to-text, text-to-image).

Tiếp nối CLIP, ALIGN [2] mở rộng tính hiệu quả của contrastive learning bằng cách khai thác dữ liệu web chưa được làm sạch, chứng minh rằng quy mô dữ liệu lớn có thể bù đắp cho nhiễu trong nhãn. Gần đây hơn, SigLIP [13] cải thiện chất lượng embedding bằng cách loại bỏ softmax trong hàm mất mát, chuyển sang tối ưu hóa cặp, giúp mô hình phân biệt tốt hơn các mẫu tương tự gần nhau. Bên cạnh đó, các mô hình như ViLT hay BLIP cũng tập trung vào học biểu diễn đa modal, trong đó BLIP sử dụng cơ chế bootstrapped captioning để tạo nhãn văn bản chất lượng cao hơn, còn ViLT loại bỏ hoàn toàn backbone CNN/Vision Transformer nhằm giảm chi phí tính toán.

Nhìn chung, các mô hình VLP mang đến bước tiến lớn trong việc thiết lập liên kết ngữ nghĩa giữa ảnh và văn bản, nhưng các mô hình này vẫn gặp hạn chế ở chi phí suy luận cao, khiến việc sử dụng trực tiếp cho tra cứu quy mô lớn trở nên khó khăn.

Các mô hình hợp nhất đặc trưng đa modal

Một nhánh nghiên cứu khác tập trung vào cách kết hợp thông tin giữa hai modal thay vì chỉ ánh xạ độc lập như CLIP. Các mô hình fusion như FashionVLP [19], ViLT hoặc BLIP khai thác cross-attention hoặc self-attention đa modal để cho phép hai dạng dữ liệu tác động qua lại trong nhiều tầng xử lý. Điều này giúp mô hình hiểu sâu hơn mối quan hệ giữa đặc trưng thị giác và đặc trưng ngôn ngữ.

Tuy vậy, các mô hình fusion thường có chi phí tính toán lớn và yêu cầu xử lý embedding dạng float trong suốt pipeline, khiến tốc độ truy vấn bị giảm đáng kể khi áp dụng trên cơ sở dữ liệu lớn. Do đó, các phương pháp hợp nhất đặc trưng dù biểu diễn tốt nhưng chưa phải lựa chọn lý tưởng cho các hệ thống yêu cầu tốc độ truy vấn cao và tiêu thụ bộ nhớ thấp.

Các mô hình trích xuất đặc trưng hình ảnh kết hợp CNN và ViT (Hybrid Visual Encoders)

Sự kết hợp giữa CNN và ViT đã trở thành xu hướng mạnh mẽ trong thị giác máy tính nhờ khả năng bổ sung ưu điểm lẫn nhau. CNN, tiêu biểu như kiến trúc ResNet [7], có khả năng học đặc trưng cục bộ ổn định thông qua các bộ lọc tích chập – phù hợp để phân tích cạnh, hoa văn và kết cấu. Ngược lại, Vision Transformer (ViT) [6] sử dụng self-attention giúp mô hình hóa mối quan hệ toàn cục trong ảnh, cho phép học ra các khái niệm cấp cao và cấu trúc ngữ nghĩa rộng.

Tuy nhiên, ViT thuần túy lại thiếu inductive bias cục bộ, làm giảm hiệu suất khi dữ liệu không đủ lớn hoặc chi tiết nhỏ đóng vai trò quan trọng. Để giải quyết điều này, nhiều mô hình hybrid đã được đề xuất. CvT [11] thêm convolution vào các bước tạo token hoặc tính toán attention để tăng khả năng tập trung cục bộ. CoAtNet [12] kết hợp hai dạng layer (convolution và attention) một cách có hệ thống để đạt hiệu năng tốt trên nhiều kích thước dữ liệu. Ngoài ra, nhiều mô hình dual-branch kết hợp trực tiếp ResNet với ViT để thu được các đặc trưng đa dạng và phong phú hơn.

Mặc dù các mô hình hybrid này đã chứng minh hiệu quả vượt trội trong phân loại ảnh, nhưng việc áp dụng chúng vào tra cứu đa phương thức hoặc băm sâu vẫn chưa được khai thác đầy đủ - đây chính là điểm mới mà SADH hướng tới.

Các phương pháp hashing và deep hashing

Hashing là công nghệ quan trọng trong các hệ thống retrieval tốc độ cao, cho phép ánh xạ dữ liệu độ nổi (floating-point embeddings) sang mã nhị phân có thể so sánh nhanh thông qua khoảng cách Hamming. Khảo sát của Wang và các công sự [3] tổng hợp nhiều phương pháp hashing truyền thống như LSH, ITQ, cho thấy tính hiệu quả về tốc độ và độ mở rộng khi xử lý cơ sở dữ liệu hàng triệu mẫu.

Với sự phát triển của học sâu, các mô hình deep hashing như DSH [5], DPSH hoặc HashNet đã cải thiện đáng kể khả năng bảo toàn ngữ nghĩa trong không gian nhị phân. Một số mô hình như DHC và HybridHash [18] đã bước đầu kết hợp embedding từ mô hình pretrained như CLIP với hashing, giúp cải thiện accuracy so với hashing

truyền thống. Tuy nhiên, các phương pháp này còn hạn chế ở chỗ chỉ sử dụng embedding từ ViT hoặc CNN đơn lẻ và chưa khai thác triệt để mối tương hỗ giữa hai modal ảnh–văn bản trước khi lượng tử hóa. Điều này dẫn đến khoảng cách thực nghiệm so với các mô hình embedding mạnh mẽ như CLIP, đặc biệt khi yêu cầu độ chính xác cao nhưng tốc độ truy vấn phải được đảm bảo.

Khoảng trống nghiên cứu và vai trò của mô hình SADH

Từ tổng quan bốn nhánh nghiên cứu trên, có thể thấy một khoảng trống rõ rệt trong lĩnh vực:

- Các mô hình VLP học biểu diễn đa phương thức mạnh nhưng tốn kém khi truy vấn
- Các mô hình hybrid CNN–ViT giàu đặc trưng nhưng chưa được khai thác cho tra cứu đa phương thức.
- Các mô hình hashing có tốc độ truy vấn nhanh nhưng chưa tích hợp hiệu quả đặc trưng mạnh từ mô hình thị giác hiện đại.

Mô hình SADH được đề xuất nhằm lấp đầy khoảng trống này bằng cách kết hợp đặc trưng toàn cục của ViT với đặc trưng cục bộ của CNN thông qua cơ chế fusion dựa trên self-attention, cùng với việc sử dụng module băm sâu để lượng tử hóa embedding thành mã nhị phân hiệu quả. Nhờ đó, SADH hướng đến xây dựng một hệ thống tra cứu ảnh đa phương thức vừa mạnh về biểu diễn ngữ nghĩa, vừa tối ưu về tốc độ, phù hợp với yêu cầu triển khai thực tế trên kho dữ liệu lớn.

1.4. Kết luận chương 1

Chương 1 đã trình bày nền tảng lý thuyết và bối cảnh nghiên cứu liên quan đến tra cứu ảnh và tra cứu ảnh đa phương thức. Trước hết, chương đã phân tích khái niệm và các cách phân loại hệ thống tra cứu ảnh, từ đó làm rõ sự khác biệt giữa truy xuất dựa trên đặc trưng thị giác, truy xuất dựa trên văn bản và truy xuất ảnh đa phương thức. Những phân tích này cho thấy bản chất phức tạp của dữ liệu hình ảnh và ngôn

ngữ, đồng thời nhấn mạnh tầm quan trọng của việc xây dựng không gian biểu diễn chung để kết nối hai phương thức vốn khác biệt mạnh về cấu trúc và thống kê.

Tiếp theo, chương đã trình bày sâu về đặc trưng của dữ liệu đầu vào, bao gồm hình ảnh – với tính liên tục, giàu cấu trúc và phụ thuộc mạnh vào yếu tố không gian – và văn bản mô tả – với tính rời rạc, cô đọng nhưng giàu ngữ nghĩa. Sự khác biệt giữa hai phương thức dẫn tới khoảng cách biểu diễn, đặt ra các thách thức lớn trong việc ánh xạ chúng vào cùng một không gian embedding phục vụ cho truy xuất. Phân tích này cho thấy rằng mọi mô hình đa phương thức đều phải đối mặt với cả vấn đề biểu diễn cục bộ lẫn vấn đề ngữ nghĩa trừu tượng.

Trên cơ sở đó, chương đã tổng hợp và phân tích các hướng nghiên cứu liên quan, bao gồm mô hình thị giác–ngôn ngữ tiền huấn luyện (như CLIP, ALIGN, SigLIP), các mô hình hợp nhất đặc trưng đa modal, mô hình hybrid CNN–ViT cho thị giác, và các phương pháp hashing hiện đại. Việc so sánh các hướng tiếp cận này cho thấy rằng mặc dù đã có nhiều thành tựu quan trọng, vẫn còn tồn tại khoảng trống nghiên cứu: chưa có mô hình nào kết hợp đầy đủ sức mạnh của ViT, CNN, cơ chế tự chú ý, và deep hashing để vừa đảm bảo độ chính xác cao trong không gian đa modal, vừa đáp ứng tốc độ truy vấn lớn của các hệ thống tìm kiếm quy mô hàng triệu ảnh.

Những kết luận này tạo tiền đề rõ ràng cho việc phát triển mô hình SADH được đề xuất trong luận văn. Các phân tích ở Chương 1 đóng vai trò như khung lý thuyết giúp định hình mục tiêu, hướng tiếp cận và các thành phần kỹ thuật sẽ được trình bày chi tiết trong Chương 2 về phương pháp đề xuất.

CHƯƠNG 2: ĐỀ XUẤT MÔ HÌNH TRA CỨU ẢNH ĐA PHƯƠNG THỨC

Chương 2 trình bày mô hình tra cứu ảnh đa phương thức SADH (Self-Attention and Deep Hashing) được đề xuất nhằm đáp ứng đồng thời yêu cầu về độ chính xác ngữ nghĩa và tốc độ truy vấn cao. Trên cơ sở các hạn chế đã nêu ở Chương 1, chương này mô tả định hướng thiết kế mô hình, bao gồm yêu cầu về hiệu năng, khả năng mở rộng và tính ứng dụng thực tế.

Tiếp theo, chương giới thiệu chi tiết các thành phần của mô hình, bắt đầu từ bộ mã hóa hình ảnh với các khối Transformer, mô-đun tương tác và hàm mất mát dành cho miền thị giác. Sau đó là bộ mã hóa văn bản, bao gồm lớp embedding, positional encoding và Transformer encoder.

Phần cuối chương trình bày kiến trúc tổng thể của mô hình SADH, trong đó mô tả không gian đặc trưng chung, tầng mã hóa băm nhị phân và chiến lược huấn luyện kết hợp hai miền dữ liệu. Những nội dung này tạo nền tảng quan trọng cho quá trình đánh giá và thực nghiệm ở Chương 3.

2.1. Phân tích yêu cầu và định hướng mô hình

2.1.1. Yêu cầu về hiệu năng và độ chính xác

Trong các hệ thống tra cứu ảnh đa phương thức quy mô lớn, hai yêu cầu quan trọng nhất quyết định mức độ khả thi của mô hình là độ chính xác truy vấn và hiệu năng xử lý. Độ chính xác phản ánh khả năng mô hình hiểu và ánh xạ mối quan hệ giữa ảnh và văn bản trong không gian biểu diễn chung. Đây là yếu tố cốt lõi đảm bảo rằng các ảnh được trả về thực sự phù hợp với ý nghĩa ngữ nghĩa của truy vấn văn bản. Các mô hình biểu diễn đa modal hiện đại như CLIP đạt hiệu suất cao trong việc học mối tương quan giữa ảnh và văn bản, tuy nhiên khi áp dụng trực tiếp vào retrieval, các embedding có kích thước lớn và ở dạng số thực khiến quá trình so khớp trở nên tốn kém và khó mở rộng.

Bên cạnh độ chính xác, hệ thống cũng cần đạt hiệu năng suy luận cao để có thể xử lý truy vấn trong thời gian ngắn, đặc biệt khi cơ sở dữ liệu có hàng triệu ảnh. Phương pháp so sánh embedding dạng float thông qua cosine similarity thường tiêu tốn nhiều tài nguyên và không đáp ứng yêu cầu thời gian thực trong các hệ thống triển khai thực tế. Do đó, việc lượng tử hóa embedding thành mã nhị phân thông qua deep hashing là cần thiết nhằm tối ưu hóa tốc độ truy vấn. Sử dụng mã nhị phân cho phép tính toán khoảng cách Hamming nhanh hơn hàng chục đến hàng trăm lần so với phép tính vector thông thường, đồng thời giảm đáng kể chi phí bộ nhớ.

Tuy nhiên, tốc độ chỉ thật sự có ý nghĩa khi mô hình vẫn duy trì được độ chính xác ngữ nghĩa. Điều này đặt ra yêu cầu mô hình phải học được biểu diễn giàu thông tin trước khi lượng tử hóa, đồng thời hạn chế mất mát thông tin khi chuyển sang dạng nhị phân. Để đáp ứng yêu cầu này, mô hình cần tận dụng được các đặc trưng toàn cục do Vision Transformer cung cấp và các đặc trưng cục bộ do CNN học được, sau đó hợp nhất chúng một cách hiệu quả nhằm tăng khả năng phân biệt trong không gian embedding.

Từ những yêu cầu trên, mô hình cần được thiết kế sao cho vừa duy trì độ chính xác cao trong ánh xạ ngữ nghĩa, vừa tối ưu hóa tốc độ truy vấn và chi phí bộ nhớ, đảm bảo tính khả thi khi triển khai trên hệ thống quy mô lớn. Đây chính là nền tảng định hướng cho việc xây dựng kiến trúc SADH trong các mục tiếp theo.

2.1.2. Tính mở rộng và khả năng áp dụng thực tế

Một trong những yêu cầu quan trọng của hệ thống tra cứu ảnh đa phương thức hiện đại là khả năng mở rộng theo quy mô dữ liệu trong khi vẫn duy trì hiệu năng ổn định. Khi kho dữ liệu tăng từ hàng trăm nghìn lên hàng triệu hoặc hàng chục triệu ảnh, mô hình cần đảm bảo khả năng truy vấn nhanh, chi phí lưu trữ thấp và năng lực xử lý ổn định. Các phương pháp truyền thống dựa trên embedding dạng float thường gặp khó khăn khi kích thước bộ nhớ và độ phức tạp tính toán tăng tuyến tính theo số lượng mẫu dữ liệu. Vì vậy, nhu cầu sử dụng mã hóa băm nhị phân là rất quan trọng, bởi mã nhị phân có kích thước cố định, nén hiệu quả, và hỗ trợ truy vấn bằng khoảng

cách Hamming với chi phí tính toán cực thấp, giúp mô hình dễ dàng mở rộng mà không làm giảm tốc độ hoặc gây quá tải bộ nhớ.

Bên cạnh đó, khả năng áp dụng thực tế của một hệ thống retrieval còn gắn liền với tính linh hoạt khi triển khai trên nhiều môi trường. Mô hình cần phù hợp cả trong hạ tầng máy chủ mạnh lẫn các hệ thống giới hạn tài nguyên như thiết bị biên hoặc dịch vụ thời gian thực. Những mô hình thị giác–ngôn ngữ lớn như CLIP, ALIGN hay SigLIP đã chứng minh hiệu quả rất cao về biểu diễn đa modal, nhưng chi phí tính toán lớn của chúng gây khó khăn cho việc sử dụng trực tiếp trong các hệ thống truy vấn tốc độ cao. Do vậy, việc thiết kế kiến trúc mô hình SADH theo hướng tận dụng các mô hình pre-train mạnh mẽ nhưng chỉ tinh chỉnh một phần nhỏ là yếu tố quan trọng. Cách tiếp cận này vừa khai thác được tri thức đã học ở quy mô web, vừa giảm chi phí huấn luyện và đảm bảo mô hình đủ nhẹ để triển khai thực tế.

Khả năng tận dụng các mô hình pre-train còn hỗ trợ tốt cho tính mở rộng, bởi việc cập nhật hoặc mở rộng kho dữ liệu không đòi hỏi huấn luyện lại từ đầu. Nhờ đó, hệ thống có thể thích ứng nhanh với các miền dữ liệu mới mà chỉ cần tinh chỉnh (fine-tuning) các tầng cuối hoặc thêm module học mới như hashing module. Đây là lợi thế quan trọng cho các ứng dụng thực tế trong thương mại điện tử, tìm kiếm nội dung, quản lý thư viện ảnh, hoặc các hệ thống khuyến nghị đa phương thức.

Một khía cạnh khác của tính áp dụng thực tế là tương thích với dữ liệu không đồng nhất. Ảnh có thể đến từ nhiều nguồn khác nhau với chất lượng, ánh sáng, góc chụp và phong cách khác nhau; tương tự, mô tả văn bản có thể thay đổi về độ chi tiết và phong cách diễn đạt. Do đó, mô hình cần duy trì khả năng trích xuất đặc trưng ổn định và linh hoạt. Việc kết hợp đặc trưng toàn cục của ViT và đặc trưng cục bộ của CNN, đồng thời tận dụng kiến thức của mô hình pre-train, giúp biểu diễn cuối cùng vừa bền vững trước nhiễu, vừa mang tính khái quát cao.

Cuối cùng, từ góc độ vận hành, một hệ thống tra cứu thực tế yêu cầu dễ dàng tích hợp, cập nhật và mở rộng chỉ mục. Mã hóa nhị phân giúp đơn giản hóa cấu trúc indexing, giảm tải bộ nhớ và cho phép áp dụng hiệu quả các kỹ thuật tìm kiếm như

hashing tables hoặc bitwise operations trên phần cứng phổ thông. Điều này đảm bảo mô hình có thể triển khai ngay trên các hệ thống sản xuất mà không cần hạ tầng tính toán chuyên biệt.

Tổng hợp lại, tính mở rộng và khả năng áp dụng thực tế của một mô hình retrieval phụ thuộc vào ba yếu tố: biểu diễn gọn nhẹ, dễ truy vấn; khả năng tận dụng tri thức giàu ngữ nghĩa từ các mô hình pretrain; khả năng triển khai linh hoạt trên nhiều hạ tầng khác nhau. Đây chính là các tiêu chí định hướng cho kiến trúc SADH được trình bày trong các mục tiếp theo của chương.

2.2. Bộ mã hóa hình ảnh

Bộ mã hóa hình ảnh đóng vai trò trung tâm trong mô hình SADH, chịu trách nhiệm trích xuất biểu diễn thị giác giàu thông tin từ ảnh đầu vào trước khi được hợp nhất và mã hóa băm. Để tận dụng ưu điểm của cả đặc trưng toàn cục và đặc trưng cục bộ, mô hình sử dụng kiến trúc kết hợp giữa ViT và nhánh CNN. ViT cung cấp khả năng mô hình hóa quan hệ toàn cục và học các khái niệm trừu tượng. CNN bổ sung chú ý cục bộ thông qua các bộ lọc tích chập. Sự kết hợp này giúp mô hình thu nhận thông tin thị giác ổn định, đa dạng và phù hợp với các tình huống trong thực tế.

2.2.1. Khối Transformer (Transformer Block)

Trong mô hình SADH, bộ mã hóa hình ảnh được xây dựng trên kiến trúc Vision Transformer, một trong những kiến trúc thị giác hiện đại dựa trên self-attention được đề xuất bởi Dosovitskiy và các công sự [6]. Khác với CNN, vốn sử dụng phép tích chập cục bộ để học đặc trưng, ViT xử lý ảnh bằng cách biểu diễn ảnh dưới dạng chuỗi token tương tự như ngôn ngữ tự nhiên, cho phép mô hình học các quan hệ toàn cục một cách trực tiếp và hiệu quả. Cách tiếp cận này tận dụng khả năng mô hình hóa dài hạn và sự linh hoạt của Transformer, vốn đã chứng minh hiệu quả vượt trội trong NLP [1].

Chia patch và ánh xạ tuyến tính (Patch Embedding)

Ảnh đầu vào $I \in \mathbb{R}^{H\times W\times 3}$ được chia thành các patch không chồng lấn có kích thước cố định, ví dụ 16×16 như trong CLIP ViT-B/16 [1]. Mỗi patch được làm phẳng và ánh xạ thông qua một lớp tuyến tính:

$$x_p = W_p \cdot Flatten(P_i) + b_p \quad (2.1)$$

Trong đó, $x_p \in \mathbb{R}^D$ là vector embedding của patch, D = 512 với CLIP ViT-B/16. Bước này tương đương với việc thực hiện phép tích chập kernel 16×16 stride 16, như giải thích trong [6]. Việc chuyển ảnh sang chuỗi patch token giúp Transformer xử lý ảnh theo cấu trúc tuần tự tương tự như chuỗi từ trong NLP.

Thêm thông tin vị trí (Positional Embedding)

Do Transformer không có inductive bias về không gian, vị trí của các patch phải được cung cấp thông qua positional embedding [6]:

$$z_0 = \left[x_{cls}, x_p^1, x_p^2, \ldots, x_p^N\right] + E_{pos} \quad (2.2)$$

Trong đó:

x_{cls} là token đặc biệt đóng vai trò tóm tắt toàn ảnh,

E_{pos} là embedding vị trí được học.

Điều này đảm bảo mô hình duy trì quan hệ không gian giữa các vùng ảnh, tương đương vai trò của kernel locality trong CNN.

Cơ chế tự chú ý đa đầu (Multi-Head Self-Attention)

Trong khối Transformer của bộ mã hóa hình ảnh, cơ chế tự chú ý đa đầu (multi-head self-attention – MHSA) là thành phần then chốt giúp mô hình học được quan hệ giữa các vùng khác nhau trong ảnh ở mức toàn cục. Cho chuỗi đầu vào gồm các vector đặc trưng patch $\{x_i\}_{i=1}^N$, mỗi vector $x_i \in \mathbb{R}^D$ được ánh xạ tuyến tính sang ba không gian khác nhau: truy vấn (query), khóa (key) và giá trị (value):

$$q_i = W_Q x_i, k_i = W_K x_i, v_i = W_V x_i\,, \quad (2.3)$$

Trong đó W_Q, W_K, $W_V \in \mathbb{R}^{D\times d}$ là các ma trận trọng số học được, d là chiều không gian của từng head attention. Tập hợp các vector này tạo thành ma trận Q, K, $V \in \mathbb{R}^{N\times d}$. Cơ chế self-attention tính trọng số chú ý giữa mọi cặp patch thông qua scaled dot-product attention:

$$Attention(Q, K, V) = soft\max\left(\frac{QK^T}{\sqrt{d}}\right)V. \quad (2.4)$$

Ma trận $\frac{QK^T}{\sqrt{d}}$ biểu diễn mức độ tương quan giữa từng cặp patch; phép chia cho $\sqrt{d}$ giúp ổn định giá trị trước softmax, tránh gradient quá nhỏ hoặc quá lớn khi d tăng. Softmax theo từng hàng tạo ra phân phối xác suất chú ý, cho biết mỗi patch đang "tập trung" vào những patch nào quan trọng nhất. Như vậy, biểu diễn đầu ra tại mỗi vị trí không chỉ phụ thuộc vào chính nó, mà là tổ hợp tuyến tính của toàn bộ các giá trị v_j, được cân bằng bởi trọng số chú ý.

Trong tự chú ý đa đầu, thay vì chỉ có một bộ (W_Q, W_K, W_V), mô hình sử dụng h song song: $\{(W_Q^{(h)}, W_K^{(h)}, W_V^{(h)})\}_{h=1}^{H}$. Mỗi head tính attention độc lập:

$$head_h = Attention(Q^{(h)}, K^{(h)}, V^{(h)}), \quad (2.5)$$

sau đó các head được nối lại và chiếu về không gian ban đầu:

$$MHSA(X) = W_O \cdot Concat(head_1, \ldots, head_H), \quad (2.6)$$

trong đó W_O là ma trận chiếu đầu ra. Việc dùng nhiều head cho phép mô hình học đồng thời nhiều kiểu quan hệ khác nhau giữa các vùng ảnh: có head tập trung vào biên và texture, head khác tập trung vào cấu trúc lớn, head khác nữa chú ý đến mối quan hệ giữa đối tượng chính và nền. Điều này giúp biểu diễn cuối cùng của ảnh vừa giàu thông tin cục bộ, vừa giàu thông tin ngữ cảnh toàn cục.

Self-attention có hai điểm mạnh quan trọng đối với bài toán tra cứu ảnh đa phương thức. Thứ nhất, nó cho phép mỗi patch nhìn toàn bộ ảnh ngay từ các tầng đầu, nhờ đó có thể mô hình hóa các quan hệ xa (long-range dependencies) như "một người đứng trước tòa nhà", "con vật nhỏ ở góc ảnh đang nhìn về phía trung tâm", vốn

rất quan trọng khi khớp với mô tả văn bản. Thứ hai, các trọng số chú ý được tính phụ thuộc vào nội dung (data-dependent), nên mô hình có khả năng "tự chọn" những vùng quan trọng với ngữ nghĩa mô tả và bỏ qua vùng nhiễu. Điều này làm cho embedding hình ảnh thu được phù hợp hơn với không gian ngôn ngữ, là nền tảng cho việc ánh xạ ảnh–văn bản vào cùng một không gian đặc trưng chung.

Cuối cùng, vì cơ chế self-attention cũng được sử dụng trong bộ mã hóa văn bản (Transformer encoder), việc dùng MHSA ở cả hai modal giúp giảm khoảng cách kiến trúc giữa ảnh và văn bản. Khi đưa embedding ảnh từ ViT qua các bước hợp nhất (interaction module) và mã hóa băm, mô hình SADH đã sở hữu một biểu diễn thị giác đã được "chuẩn hóa" về mặt kiến trúc với biểu diễn ngôn ngữ, từ đó hỗ trợ hiệu quả hơn cho việc học không gian nhị phân chung phục vụ truy vấn tốc độ cao.

Mạng lan truyền tiến (Feed-Forward Network – FFN)

Trong mỗi khối Transformer, sau tầng MHSA là mạng lan truyền tiến (Feed-Forward Network – FFN). Đây là thành phần đóng vai trò tăng cường khả năng biểu diễn phi tuyến, giúp mô hình học được các quan hệ trừu tượng và phức tạp hơn từ đầu vào. FFN được áp dụng độc lập cho từng token (patch) theo dạng "vị trí-theo-vị-trí" (position-wise), nghĩa là mỗi token được xử lý bởi cùng một mạng con nhưng không chia sẻ thông tin giữa các vị trí. Cấu trúc FFN bao gồm hai lớp tuyến tính có hàm kích hoạt phi tuyến ở giữa:

$$FFN(x) = W_2\sigma(W_1x + b_1) + b_2, \tag{2.7}$$

trong đó $W_1 \in \mathbb{R}^{D\times D'}$, $W_2 \in \mathbb{R}^{D\times D'}$ là các ma trận trọng số, còn σ thường được chọn là GELU để tăng tính ổn định và hiệu quả học. FFN mở rộng chiều không gian tạm thời từ D sang D′ (thường gấp 2–4 lần), rồi chiếu ngược về chiều ban đầu, từ đó cung cấp khả năng mô hình hóa phi tuyến mạnh hơn.

Khác với tầng attention, vốn tập trung vào việc tổng hợp quan hệ giữa các token, FFN giúp xử lý nội dung bên trong từng token độc lập. Nhờ đó, biểu diễn cuối cùng của mỗi patch ảnh không chỉ phản ánh mối quan hệ với các patch khác (từ

MHSA), mà còn được tinh chỉnh bởi các phép biến đổi phi tuyến cục bộ. Điều này đặc biệt quan trọng với ViT, vốn không có inductive bias mạnh như CNN, nên cần một cơ chế bổ sung để tăng độ biểu diễn của mô hình. Kết hợp attention và FFN tạo ra một pipeline hai giai đoạn: (1) tổng hợp thông tin toàn cục; (2) khuếch đại và tái cấu trúc đặc trưng token, giúp embedding cuối cùng giàu ngữ nghĩa và ổn định hơn trước khi đưa sang các bước hợp nhất và mã hóa băm trong mô hình SADH.

Chuẩn hóa và kết nối tắt (LayerNorm & Residual Connection)

Hai cơ chế quan trọng giúp Transformer ổn định trong huấn luyện và duy trì khả năng truyền gradient là Layer Normalization (LayerNorm) và kết nối tắt (Residual Connection). Trong mỗi tầng Transformer, cả block attention và FFN đều được bao bọc bởi kết nối tắt và chuẩn hóa theo cấu trúc:

$$x' = x + MHSA\big(LN(x)\big), \tag{2.8}$$

$$x'' = x' + FFN\big(LN(x')\big). \tag{2.9}$$

Ở đây, LayerNorm được áp dụng trước khi đưa vào MHSA/FFN, còn phần đầu ra được cộng lại với đầu vào ban đầu qua đường tắt. Thiết kế này, gọi là pre-norm, đã được chứng minh mang lại sự ổn định cao hơn trong quá trình tối ưu cho các mô hình sâu.

Residual connection giúp gradient lan truyền hiệu quả xuyên suốt nhiều tầng, tránh hiện tượng tiêu biến gradient khi mô hình có độ sâu lớn. Điều này đặc biệt quan trọng với ViT, vốn thường có 12–24 tầng encoder. Khi có residual, mỗi tầng học phần "phần bù" cần thiết thay vì học toàn bộ ánh xạ, giúp mô hình hội tụ nhanh hơn và tránh overfitting.

LayerNorm, ngược lại, chuẩn hóa từng token theo chiều đặc trưng, giúp giảm sự dao động trong phân phối đầu vào giữa các batch và giúp mô hình ổn định khi học. Khác với BatchNorm của CNN, LayerNorm hoạt động tốt cho dữ liệu dạng chuỗi và không phụ thuộc vào kích thước batch, do đó phù hợp với kiến trúc Transformer của cả hai modal: hình ảnh và văn bản.

Sự kết hợp giữa LayerNorm và residual tạo ra cơ chế vừa ổn định vừa linh hoạt: attention giữ thông tin toàn cục, FFN học phi tuyến cục bộ, còn residual đảm bảo thông tin ban đầu được bảo toàn và khuếch đại đúng mức. Trong mô hình SADH, đặc biệt khi thực hiện fine-tuning một phần ViT, cơ chế này giúp tránh làm "hỏng" embedding gốc từ pretraining CLIP, đồng thời vẫn cho phép mô hình thích nghi với nhiệm vụ học hashing đa phương thức.

2.2.2. Mô đun tương tác (Interaction Module)

Trong kiến trúc SADH, mô đun tương tác giữ vai trò trung tâm trong việc kết hợp hai nguồn đặc trưng khác nhau: đặc trưng toàn cục từ bộ mã hóa hình ảnh CLIP ViT, và đặc trưng cục bộ từ nhánh CNN. Trong khi phần backbone ViT nắm bắt cấu trúc toàn cảnh và quan hệ dài hạn trên toàn bộ bức ảnh, CNN lại có khả năng mô hình hóa chi tiết cục bộ như cạnh, kết cấu, màu sắc hoặc hình dạng nhỏ—những yếu tố thường bị làm mờ khi xử lý ảnh thành patch trong ViT. Vì vậy, mô đun tương tác gồm hai bước: (1) trích xuất đặc trưng tăng cường bằng CNN; (2) kết hợp hai nguồn đặc trưng này thông qua cơ chế MHSA nhằm học trọng số kết hợp tối ưu.

Nhánh CNN – tăng cường đặc trưng cục bộ

Nhánh CNN được cài đặt bằng kiến trúc ResNet-50 pretrained trên ImageNet. Mục tiêu của nhánh này là bổ sung đặc trưng cục bộ mà ViT có thể bỏ sót do quá trình chia ảnh thành patch làm mất cấu trúc không gian liên tục.

Ảnh đầu vào $I \in \mathbb{R}^{3\times224\times224}$ được đưa qua ResNet-50 (loại bỏ tầng fully-connected):

$$F_{cnn} = E_{cnn}(I) \in \mathbb{R}^{2048}. \tag{2.10}$$

Sau đó, véc-tơ này được chiếu về không gian 512 chiều để đồng nhất với đặc trưng từ ViT:

$$f_{cnn} = W_c F_{cnn}, \tag{2.11}$$

với $W_c \in \mathbb{R}^{512\times2048}$.

Nhánh CNN đóng vai trò quan trọng vì nó giữ được thông tin cấu trúc cục bộ mà ViT xử lý chưa tối ưu; đồng thời cải thiện khả năng phân biệt của hệ thống; và cung cấp kênh đặc trưng bổ sung giúp mô đun hợp nhất tự chú ý học tương tác chính xác hơn. Kết quả là đặc trưng hình ảnh trong SADH đại diện tốt hơn cho cả thông tin toàn cục và cục bộ.

Kết hợp đặc trưng bằng Self-Attention

Sau khi thu được hai nguồn đặc trưng: $f_{vit}, f_{cnn} \in \mathbb{R}^{512}$, chúng được ghép thành một chuỗi hai token $X = \begin{bmatrix} f_{vit} \\ f_{cnn} \end{bmatrix}$. Khối AttentionFusion áp dụng MHSA lên chuỗi này để học quan hệ giữa hai đặc trưng, sử dụng công thức (2.4). Cơ chế multi-head cho phép mô hình học nhiều kiểu tương tác khác nhau:

- Head 1: Ưu tiên đặc trưng toàn cục (bố cục, ngữ nghĩa) của vit
- Head 2: Ưu tiên thông tin chi tiết của CNN
- Head 3: Học quan hệ tương hỗ giữa chúng
- Head 4–8: Học các kiểu tương tác phức tạp hơn

Kết quả của attention được kết hợp với đầu vào bằng residual connection:

$$X' = LayerNorm(X + MHSA(X)). \tag{2.12}$$

Sau đó, token được truyền qua mạng FFN hai lớp:

$$X'' = LayerNorm(X' + FFN(X')), \tag{2.13}$$

tạo ra biểu diễn đã được kết hợp đầy đủ. Cuối cùng, mô hình thực hiện phép gộp trung bình:

$$f_{fused} = \frac{1}{2}\sum_{i=1}^{2} X''_i. \tag{2.14}$$

Đây là biểu diễn cuối cùng của ảnh dùng cho module hashing.

Tính chất nổi bật của mô đun tương tác

Mô đun tương tác là một trong những thành phần quan trọng nhất của SADH, bởi nó đảm nhiệm nhiệm vụ học cách kết hợp tối ưu giữa hai nguồn đặc trưng mang bản chất khác nhau: đặc trưng toàn cục từ ViT và đặc trưng cục bộ từ CNN. Đặc trưng nổi bật đầu tiên của mô đun này là khả năng tự động học trọng số kết hợp thông qua cơ chế tự chú ý đa. Khác với các phương pháp ghép nối truyền thống vốn gán trọng số cố định cho từng nguồn đặc trưng, self-attention cho phép mô hình đánh giá mức độ đóng góp của từng token (ViT hoặc CNN) tùy theo nội dung ảnh. Nhờ đó, các trường hợp ảnh giàu thông tin cấu trúc sẽ ưu tiên CNN, trong khi ảnh chứa đối tượng phức tạp và nhiều ngữ cảnh sẽ ưu tiên ViT.

Đặc tính nổi bật thứ hai là khả năng giảm sự phân kỳ của hai dạng đặc trưng. Do ViT xử lý ảnh dưới dạng patch embedding còn CNN xử lý qua lọc convolution, phân phối đặc trưng thu được có thể chênh lệch đáng kể. Self-attention, cùng với LayerNorm và residual, đóng vai trò điều hòa và kéo hai dạng đặc trưng này về cùng một không gian, từ đó giúp cho biểu diễn hợp nhất ổn định hơn và dễ được mã hóa bởi tầng hashing.

Đặc tính quan trọng tiếp theo nằm ở khả năng mô hình hóa quan hệ phi tuyến và đa chiều giữa hai nguồn đặc trưng. Mỗi head trong cơ chế multi-head attention học một kiểu tương tác khác nhau, chẳng hạn: một head tập trung vào quan hệ tương quan trực tiếp giữa hình dạng và texture, head khác chú ý đến sự nhất quán giữa bố cục toàn cục và chi tiết cục bộ, các head còn lại có thể học các vùng ngữ nghĩa hoặc kiểu biến đổi ánh sáng–bối cảnh. Nhờ đó, mô đun tương tác xây dựng được một biểu diễn giàu ngữ nghĩa hơn, kết hợp cả khả năng phân biệt chi tiết và khả năng trừu tượng hóa toàn cảnh.

Cuối cùng, mô đun này cho phép thích nghi linh hoạt với nhiều loại ảnh và điều kiện thu nhận, từ ảnh tự nhiên nhiều texture đến ảnh chứa đối tượng rõ nét và ít

chi tiết nền. Điều này giúp SADH đạt hiệu suất ổn định trên các tập dữ liệu có độ đa dạng cao.

Vai trò của mô đun tương tác trong kiến trúc SADH

Trong tổng thể kiến trúc SADH, mô đun tương tác đóng vai trò như "cầu nối thông minh" giữa hai miền đặc trưng, đảm bảo rằng thông tin cục bộ từ CNN và thông tin toàn cục từ ViT không bị xử lý tách biệt mà được liên kết một cách có ý nghĩa. Không chỉ đơn thuần ghép hai véc-tơ, module này học cách thức kết hợp tối ưu dựa trên cấu trúc mẫu dữ liệu, từ đó tạo ra một biểu diễn hợp nhất f_{fused} giàu tính phân biệt.

Mô đun tương tác cũng đóng vai trò quyết định trong việc định dạng lại không gian đặc trưng trước khi mã hóa băm. Bởi tầng hashing yêu cầu đặc trưng đầu vào tương đối "phẳng", ổn định và có mức độ phân bố đồng đều nhằm đảm bảo mã băm nhị phân dễ lượng tử hóa, mô đun tương tác sẽ chuẩn hóa và tái cấu trúc đặc trưng thông qua self-attention, FFN và LayerNorm. Điều này làm giảm sai lệch giữa hai dạng đặc trưng và hạn chế tình trạng "mất cân bằng bit" hoặc "mã băm không ổn định".

Ngoài ra, mô đun tương tác là thành phần đóng góp mạnh mẽ vào khả năng mở rộng và tính tổng quát của SADH. Vì mô đun này không phụ thuộc vào cấu trúc cụ thể của CNN hay ViT, nó có thể dễ dàng tích hợp với các backbone khác nhau—chẳng hạn ResNet101, ConvNeXt, ViT-L/14, hoặc các mô hình hybrid như CoAtNet. Điều này tạo ra một kiến trúc linh hoạt, dễ điều chỉnh cho nhiều bài toán thực tế.

Cuối cùng, mô đun tương tác góp phần trực tiếp vào hiệu quả truy vấn khi kết hợp với tầng hashing. Việc có một biểu diễn hợp nhất nhất quán, giàu ngữ nghĩa giúp giảm đáng kể sai số khi ánh xạ sang mã nhị phân, từ đó cải thiện độ chính xác trong các tác vụ image-to-text, text-to-image và image-to-image. Nhờ vậy, SADH đạt được

sự cân bằng giữa tốc độ truy vấn và độ chính xác—mục tiêu quan trọng của các hệ thống tra cứu ảnh quy mô lớn.

2.3. Bộ mã hóa văn bản

2.3.1. Mã hóa văn bản bằng CLIP

Trong mô hình SADH, mã hóa văn bản được kế thừa nguyên vẹn từ text encoder của CLIP, một kiến trúc Transformer thuộc nhóm mô hình thị giác–ngôn ngữ quy mô lớn, được huấn luyện trên tập dữ liệu ghép cặp ảnh–văn bản lên tới hàng trăm triệu mẫu [1]. Kiến trúc này được xem là nhân tố quan trọng giúp liên kết giữa modal ngôn ngữ và modal thị giác, bởi nó sản sinh ra biểu diễn văn bản giàu thông tin ngữ nghĩa và phù hợp với không gian embedding chung.

Quá trình mã hóa văn bản trong CLIP bắt đầu bằng thao tác token hóa sử dụng Byte-Pair Encoding (BPE), một phương pháp phân đoạn chuỗi ký tự thành các token con nhằm giảm kích thước từ vựng nhưng vẫn duy trì được tính biểu đạt của ngôn ngữ. Mỗi token sau khi được biểu diễn bằng embedding được cộng với một positional embedding cố định. Do Transformer không có cơ chế tự nhiên để xử lý vị trí, việc bổ sung positional encoding giúp mô hình phân biệt được trật tự của token trong câu, từ đó duy trì được cấu trúc cú pháp và quan hệ ngữ nghĩa liên vị trí.

Sau bước nhúng, chuỗi vector đi qua nhiều tầng Transformer encoder giống kiến trúc gốc của Vaswani và các công sự [6]. Mỗi tầng bao gồm hai khối: MHSA và FFN. Self-attention cho phép mỗi token "quan sát" toàn bộ phần còn lại của câu, tạo điều kiện cho mô hình học được các phụ thuộc dài hạn và các tương tác ngữ nghĩa phức tạp. Ví dụ, tính nhất quán giữa chủ ngữ đầu câu và động từ ở cuối câu hay quan hệ giữa danh từ chỉ đối tượng và cụm mô tả thuộc tính đều được mô hình hóa sâu qua các phép tính $\frac{QK^T}{\sqrt{d}}$. Trong khi đó, feed-forward network đóng vai trò tinh chỉnh biểu diễn ở từng vị trí, bổ sung tính phi tuyến và tăng khả năng mô hình hóa đặc trưng trừu tượng.

Điểm then chốt của CLIP là mô hình được huấn luyện bằng contrastive learning, trong đó embedding của văn bản phải gần với embedding của ảnh tương ứng và xa với các ảnh khác. Điều này buộc text encoder học một không gian biểu diễn mang tính "hình ảnh hóa" (vision-aware semantic space). Khác với các mô hình ngôn ngữ thuần túy, biểu diễn của CLIP không chỉ lưu giữ ngữ nghĩa nội tại của văn bản mà còn phản ánh cách văn bản mô tả nội dung thị giác. Đây là lý do khiến embedding văn bản của CLIP trở thành lựa chọn phù hợp cho các mô hình tra cứu ảnh đa phương thức.

Tuy nhiên, embedding văn bản do text encoder của CLIP sinh ra chủ yếu ở mức khái niệm. Chúng mô tả thông tin liên quan đến đối tượng, hành động và ngữ cảnh nhưng không chứa các đặc trưng cục bộ như hình dạng chi tiết hay kết cấu bề mặt như trong ảnh. Điều này dẫn đến sự khác biệt đáng kể về phân phối so với embedding hình ảnh, đặc biệt là khi nhánh hình ảnh của SADH còn được tăng cường bởi CNN và fusion layer. Do đó, mặc dù CLIP cung cấp biểu diễn văn bản mạnh, sự bất cân xứng này vẫn cần được xử lý thông qua Text Projection Module ở phần sau để đảm bảo hai dòng thông tin có thể được ánh xạ vào cùng một không gian hashing ổn định.

2.3.2. Vấn đề bất đối xứng giữa hai dòng đặc trưng

Trong các hệ thống tra cứu ảnh đa phương thức, một yêu cầu quan trọng là biểu diễn của ảnh và văn bản phải được ánh xạ vào cùng một không gian ngữ nghĩa chung để có thể so sánh trực tiếp bằng những độ đo tương đồng đơn giản như khoảng cách cosine hoặc khoảng cách Hamming. Tuy nhiên, do bản chất khác nhau của hai phương thức, sự bất đối xứng trong phân phối đặc trưng là một vấn đề phổ biến và có ảnh hưởng đáng kể đến hiệu suất truy vấn. Mô hình SADH cũng gặp vấn đề tương tự khi hai dòng đặc trưng—nhánh hình ảnh và nhánh văn bản—mang các đặc tính cấu trúc và ngữ nghĩa rất khác nhau.

Nhánh hình ảnh trong SADH được thiết kế dưới dạng một pipeline phức tạp bao gồm bộ mã hóa ViT của CLIP, nhánh CNN bổ sung cho thông tin cục bộ, và một

mô-đun hợp nhất dựa trên self-attention. Quá trình này làm giàu đặc trưng hình ảnh bằng nhiều cấp độ thông tin: từ các chi tiết mức thấp như cạnh, texture cho đến quan hệ toàn cục và khái niệm trừu tượng. Việc kết hợp nhiều tầng biểu diễn như vậy khiến embedding hình ảnh vừa đa dạng, vừa giàu chiều sâu ngữ nghĩa. Đồng thời, quá trình fine-tuning của ViT và CNN còn làm cho phân phối đặc trưng hình ảnh thay đổi theo task, tạo ra một không gian embedding có động lực học phức tạp.

Ngược lại, nhánh văn bản chỉ dựa trên text encoder của CLIP và thường được giữ nguyên (frozen) trong toàn bộ quá trình huấn luyện. Lý do của sự lựa chọn này là text encoder của CLIP đã được huấn luyện trên một tập dữ liệu khổng lồ và nắm giữ khả năng biểu diễn ngôn ngữ mạnh, do đó việc fine-tuning có thể làm mất đi cấu trúc ngữ nghĩa tối ưu. Tuy nhiên, việc giữ nguyên hoàn toàn nhánh văn bản dẫn đến việc embedding văn bản không thể thích nghi với sự biến đổi của embedding hình ảnh trong quá trình học, tạo ra sự lệch pha giữa hai không gian biểu diễn.

Sự bất đối xứng biểu hiện rõ ràng ở ba khía cạnh. Thứ nhất, mức độ trừu tượng của hai dòng đặc trưng khác nhau: văn bản tập trung vào khái niệm và mô tả nội dung, trong khi hình ảnh bao gồm cả những chi tiết mà văn bản không đề cập. Thứ hai, phân phối thống kê của hai embedding không tương thích: nhánh hình ảnh trải qua nhiều phép biến đổi phi tuyến và được tinh chỉnh liên tục, trong khi nhánh văn bản duy trì phân phối cố định. Cuối cùng, độ phong phú về ngữ nghĩa của hai modal không đồng đều; embedding hình ảnh có thể mang thông tin không được mô tả trong văn bản, còn embedding văn bản có thể chứa thông tin ngữ cảnh mà hình ảnh không thể hiện.

Những khác biệt này dẫn đến một modal gap đáng kể, khiến mô hình khó căn chỉnh (align) hai biểu diễn một cách tự nhiên. Trong bối cảnh hashing, sự bất đối xứng còn nghiêm trọng hơn: nếu phân phối đặc trưng ảnh và văn bản không tương thích trước khi đi vào các tầng băm, các mã nhị phân sinh ra sẽ lệch nhau, làm tăng khoảng cách Hamming giữa các cặp ảnh–văn bản tương ứng và từ đó làm giảm đáng kể hiệu suất truy vấn.

Do đó, việc xử lý bất đối xứng giữa hai dòng đặc trưng không chỉ là yêu cầu kỹ thuật mà còn là điều kiện tiên quyết để mô hình SADH đạt được hiệu quả tối ưu. Giải pháp được đề xuất trong công trình này là Mô đun ánh xạ văn bản, một khối học được dùng để điều chỉnh embedding văn bản sao cho phù hợp với không gian biểu diễn của hình ảnh, qua đó thu hẹp modal gap và cải thiện khả năng ánh xạ đa phương thức. Phần tiếp theo mô tả chi tiết cơ chế hoạt động và đóng góp của mô-đun này.

2.3.3. Mô đun ánh xạ văn bản (Text Projection Module)

Để giải quyết triệt để vấn đề bất đối xứng giữa hai dòng đặc trưng nêu ở mục trước, mô hình SADH bổ sung một mô-đun mới có vai trò điều chỉnh embedding văn bản trước khi đưa vào tầng mã hóa băm. Mô-đun này, gọi là Mô đun ánh xạ văn bản, được thiết kế như một bộ biến đổi phi tuyến nhằm chiếu/ánh xạ (project) các biểu diễn văn bản từ CLIP sang không gian đặc trưng tương thích với embedding hình ảnh đã được tăng cường bởi ViT, CNN và mô-đun tương tác. Mục tiêu của mô-đun này không phải là cải thiện khả năng ngôn ngữ của CLIP—vốn đã rất mạnh nhờ quá trình pretraining quy mô lớn—mà là tái định hình phân phối embedding văn bản để thu hẹp khoảng cách biểu diễn giữa hai dòng đặc trưng trước khi thực hiện hashing.

Về mặt kiến trúc, mô đun ánh xạ văn bản được triển khai dưới dạng một mạng MLP ba tầng, kết hợp tuyến tính, phi tuyến (GELU) và chuẩn hóa lớp (LayerNorm). Cách thiết kế này giúp tăng tính biểu đạt mà không phá vỡ cấu trúc ngữ nghĩa vốn có của embedding văn bản. Kiến trúc nhiều tầng cho phép mô-đun học được các biến đổi phức tạp trong không gian liên tục, góp phần điều chỉnh độ lan truyền, độ phân tán và tương quan giữa các chiều của vector ngôn ngữ. Từ góc nhìn học biểu diễn, đây là quá trình hiệu chỉnh embedding (latent space calibration), nhằm đưa phân phối văn bản trở nên tương thích với phân phối của hình ảnh sau khi đã được tăng cường bằng thông tin cục bộ và toàn cục.

Một đặc điểm quan trọng trong thiết kế của mô đun ánh xạ văn bản là trọng số residual có thể học được, cho phép mô hình kiểm soát mức độ pha trộn giữa embedding văn bản ban đầu và embedding sau biến đổi. Nếu residual weight tiến gần

0, mô hình chủ yếu giữ nguyên biểu diễn văn bản gốc; nếu tiến gần 1, mô hình ưu tiên biểu diễn đã được chiếu mạnh vào không gian hình ảnh. Cơ chế này giúp duy trì sự ổn định trong giai đoạn đầu của quá trình huấn luyện, khi các trọng số của hệ thống chưa tối ưu, đồng thời cho phép hệ thống tự điều chỉnh để đạt trạng thái cân bằng giữa bảo toàn thông tin ngữ nghĩa và thích nghi với không gian biểu diễn hợp nhất.

Tác động của mô đun ánh xạ văn bản không chỉ nằm ở việc biến đổi embedding mà còn góp phần giảm hiện tượng lệch phân phối (distributional shift) giữa hai phương thức. Khi kết hợp với thành phần Alignment Loss trong hàm mất mát, mô-đun này giúp điều chỉnh cả trung bình và phương sai của embedding văn bản sao cho gần với embedding hình ảnh. Điều này đặc biệt quan trọng trong mô hình hashing, nơi đầu ra là mã nhị phân: nếu phân phối của hai dòng đặc trưng lệch nhau, các bit sinh ra sẽ không đồng nhất, làm giảm khả năng khớp giữa ảnh và văn bản trong không gian Hamming.

Nhìn chung, mô đun ánh xạ văn bản đóng vai trò cầu nối quan trọng trong SADH. Nó đảm bảo rằng sự phong phú của nhánh hình ảnh—vốn được mở rộng bởi các khối ViT, CNN và module tương tác—không bị phá vỡ bởi sự cố định hóa của embedding văn bản. Nhờ mô-đun này, hai dòng đặc trưng được đưa về cùng phân phối tiệm cận, giúp tầng hashing hoạt động ổn định và nâng cao hiệu suất truy vấn ảnh đa phương thức. Đây là một trong những đóng góp quan trọng của mô hình SADH so với các hệ thống chỉ sử dụng embedding nguyên bản từ CLIP mà không có bước điều chỉnh bổ sung.

2.4. Kiến trúc tổng thể của mô hình đề xuất

Mô hình được đề xuất là một hệ thống tra cứu ảnh đa phương thức (multimodal image retrieval) với mục tiêu ánh xạ hình ảnh và văn bản vào một không gian đặc trưng chung, từ đó cho phép truy vấn ảnh bằng mô tả văn bản một cách chính xác và hiệu quả. Toàn bộ kiến trúc được thiết kế xoay quanh ba ý tưởng cốt lõi: khai thác mô hình tiền huấn luyện mạnh, tăng cường khả năng liên kết liên phương thức bằng cơ chế chú ý, và tối ưu tốc độ truy xuất thông qua mã hóa băm nhị phân.

Về dòng xử lý hình ảnh, mô hình sử dụng bộ mã hóa thị giác của CLIP dựa trên ViT để trích xuất đặc trưng toàn cục, phản ánh ngữ cảnh và bố cục tổng thể của ảnh. Song song với đó, một nhánh CNN (ResNet-50 tiền huấn luyện trên ImageNet) được bổ sung nhằm khai thác các đặc trưng cục bộ như chi tiết, kết cấu và hình dạng. Hai loại đặc trưng này mang tính bổ sung: ViT mạnh về ngữ nghĩa toàn cục, trong khi CNN hiệu quả trong việc nắm bắt thông tin không gian chi tiết.

Các đặc trưng toàn cục và cục bộ sau đó được đưa vào Interaction Module, nơi MHSA được sử dụng để học cách kết hợp hai nguồn thông tin một cách thích nghi. Thay vì gán trọng số cố định, mô-đun này cho phép mô hình tự động quyết định mức độ quan trọng của từng nhánh đặc trưng trong từng ngữ cảnh ảnh cụ thể. Nhờ đó, biểu diễn hình ảnh sau hợp nhất giàu thông tin hơn và có khả năng liên kết tốt hơn với biểu diễn văn bản.

Đối với dòng xử lý văn bản, mô hình sử dụng trực tiếp bộ mã hóa văn bản của CLIP, vốn dựa trên Transformer encoder. Văn bản đầu vào được ánh xạ thành embedding từ vựng, kết hợp với positional encoding để bảo toàn thứ tự từ, sau đó được xử lý qua nhiều lớp self-attention nhằm học biểu diễn ngữ nghĩa ở mức câu. Đặc trưng văn bản thu được phản ánh tốt ý nghĩa ngữ cảnh của câu truy vấn, nhưng về mặt cấu trúc vẫn đơn giản hơn so với đặc trưng hình ảnh đã qua hợp nhất nhiều nguồn.

Sự khác biệt này dẫn tới vấn đề bất đối xứng giữa hai dòng đặc trưng: đặc trưng ảnh giàu thông tin hơn đặc trưng văn bản, gây khó khăn cho việc ánh xạ và so khớp trong không gian chung. Để khắc phục, mô hình đưa vào Text Projection Module, một mạng MLP nhiều lớp với kết nối tắt, nhằm biến đổi đặc trưng văn bản sao cho phân bố của nó tiệm cận với đặc trưng ảnh đã hợp nhất. Mô-đun này vừa mở rộng khả năng biểu diễn của văn bản, vừa giữ lại thông tin ngữ nghĩa cốt lõi thông qua cơ chế residual.

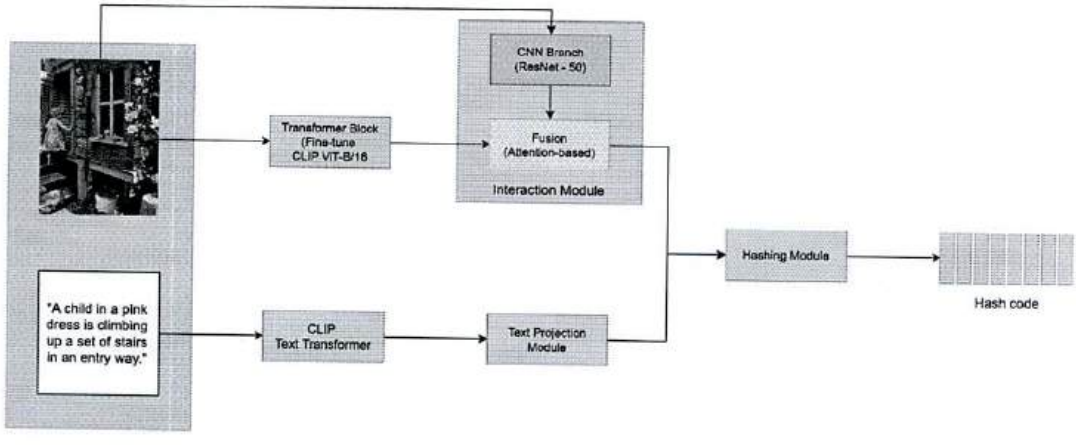
Sau khi cả hình ảnh và văn bản được ánh xạ vào cùng một không gian đặc trưng, mô hình sử dụng module mã hóa băm sâu (Deep Hashing Module) để chuyển

các vector liên tục thành mã băm nhị phân có độ dài cố định. Quá trình này giúp giảm đáng kể chi phí lưu trữ và cho phép tính toán độ tương đồng nhanh bằng khoảng cách Hamming, rất phù hợp cho các hệ thống truy xuất quy mô lớn.

Việc huấn luyện mô hình được dẫn dắt bởi hàm mất mát tổng hợp, bao gồm: (i) hàm mất mát đối sánh tương phản (contrastive loss) để kéo gần các cặp ảnh–văn bản đúng và đẩy xa các cặp không tương ứng; (ii) hàm mất mát lượng tử hóa nhằm ép các giá trị đầu ra tiến gần ±1; (iii) hàm cân bằng bit để tránh hiện tượng sụp bit; và (iv) hàm alignment loss nhằm đảm bảo phân bố đặc trưng của hai miền dữ liệu là tương thích. Các thành phần này phối hợp giúp mô hình vừa chính xác về mặt ngữ nghĩa, vừa hiệu quả về mặt truy xuất.

Tổng thể, mô hình hình thành một kiến trúc thống nhất, trong đó mỗi thành phần đảm nhiệm một vai trò rõ ràng: CLIP cung cấp nền tảng biểu diễn mạnh, Interaction Module tăng cường liên kết thị giác, Text Projection giải quyết bất đối xứng liên phương thức, và Deep Hashing tối ưu hiệu suất truy xuất. Sự kết hợp này cho phép mô hình đạt được sự cân bằng tốt giữa độ chính xác và tốc độ, đáp ứng yêu cầu của các hệ thống tra cứu ảnh đa phương thức trong thực tế.

Các thành phần và pipeline của mô hình SADH được minh họa trong Hình 2.1.



Hình 2.1: Kiến trúc tổng thể của mô hình SADH

2.4.1. Ánh xạ không gian đặc trưng chung

Trong các hệ thống truy xuất ảnh đa phương thức, việc xây dựng một không gian biểu diễn chung cho hình ảnh và văn bản là bước then chốt quyết định hiệu năng toàn bộ mô hình. Do hai modal khác nhau về bản chất tín hiệu—hình ảnh là dữ liệu liên tục, giàu cấu trúc không gian, trong khi văn bản là tín hiệu rời rạc, mang tính trừu tượng cao—việc so sánh trực tiếp hai dạng dữ liệu này là không khả thi nếu không có một phương pháp ánh xạ phù hợp. Mục tiêu của không gian đặc trưng chung là cung cấp một miền biểu diễn thống nhất, trong đó các cặp ảnh–văn bản tương ứng được đưa về các vector gần nhau, còn các cặp không tương ứng được đẩy ra xa.

Trong mô hình SADH, quá trình ánh xạ này được xây dựng dựa trên hai nhánh mã hóa độc lập cho hình ảnh và văn bản, sau đó hội tụ trong một không gian trung gian trước khi được mã hóa băm. Đối với hình ảnh, đặc trưng được kết hợp từ hai nguồn: đặc trưng toàn cục thu được từ Vision Transformer của CLIP, và đặc trưng cục bộ trích xuất từ nhánh CNN dựa trên ResNet-50. Hai nguồn thông tin này phản ánh những khía cạnh khác nhau của nội dung hình ảnh—quan hệ toàn cục giữa các vùng và chi tiết thị giác nhỏ—và được hợp nhất thông qua Attention Fusion nhằm tạo ra một representation giàu ngữ nghĩa và ổn định. Quá trình hợp nhất này đóng vai trò quan trọng, bởi các đặc trưng toàn cục và cục bộ không nằm cùng không gian thống kê, và self-attention giúp mô hình tự động học cách phân bổ trọng số phù hợp giữa hai loại thông tin khi ánh xạ về không gian chung.

Trong khi đó, văn bản được mã hóa bằng bộ mã hóa ngôn ngữ của CLIP, vốn đã được huấn luyện đa phương thức trên quy mô lớn nhằm học sự tương ứng giữa mô tả ngôn ngữ và nội dung thị giác. Tuy nhiên, do đường đi của đặc trưng văn bản ngắn hơn và không có phần học bổ sung như nhánh CNN trong pipeline của ảnh, sự bất đối xứng giữa hai modal dễ dẫn đến chênh lệch phân bố trong không gian biểu diễn. Vì vậy, SADH sử dụng Text Projection Module để điều hòa đặc trưng văn bản trước khi đưa vào không gian chung. Mô-đun này không chỉ thay đổi hình dạng biểu diễn

mà còn chuẩn hóa và làm giàu đặc trưng, giúp chúng phù hợp hơn với phân bố của đặc trưng ảnh sau hợp nhất.

Quá trình ánh xạ không dừng lại ở việc đưa hai loại vector về cùng số chiều; SADH áp dụng chuẩn hóa L2 nhằm đưa tất cả vector lên hypersphere đơn vị. Điều này đảm bảo rằng quan hệ ngữ nghĩa chủ yếu được mô tả thông qua hướng của vector thay vì độ lớn, từ đó giúp phép đối sánh ảnh–văn bản trở nên ổn định hơn. Chính việc chuẩn hóa này đóng vai trò như một "cầu nối hình học", giúp giảm đáng kể sự khác biệt nội tại giữa hai không gian biểu diễn nguyên thủy.

Một cải tiến quan trọng khác là Alignment Loss, được thiết kế để tối ưu hóa phân bố đặc trưng của hai modal về mức độ toàn cục. Trong khi contrastive loss tập trung thiết lập quan hệ gần–xa giữa các cặp dữ liệu cụ thể, alignment loss xem xét trung bình và phương sai của toàn bộ batch để giảm sự khác biệt giữa hai phân bố. Nhờ đó, SADH không chỉ học quan hệ cặp đôi mà còn điều chỉnh quan hệ tổng thể giữa hai miền đặc trưng, giảm thiểu hiện tượng lệch miền (modality shift) vốn là một trong những nguyên nhân chính làm suy giảm hiệu năng của các mô hình hashing đa phương thức.

Kết quả của toàn bộ quá trình này là sự hình thành một không gian đặc trưng chung mà trong đó hình ảnh và văn bản có thể được so khớp trực tiếp, đồng thời các vector thu được cũng có cấu trúc thuận lợi cho việc lượng tử hóa thành mã nhị phân. Khi ánh xạ được học tốt, sự tương đồng trong không gian liên tục được bảo toàn sau khi mã hóa băm, giúp truy vấn dựa trên khoảng cách Hamming trở nên nhanh chóng mà vẫn đảm bảo độ chính xác cao. Đây chính là nền tảng cho khả năng truy xuất hiệu quả ở quy mô lớn của mô hình SADH.

2.4.2. Mô đun mã hóa băm

Mô đun mã hóa băm (Hashing Module) đóng vai trò cầu nối giữa không gian biểu diễn liên tục giàu ngữ nghĩa và không gian nhị phân rời rạc phục vụ cho truy vấn tốc độ cao. Trong mô hình SADH, mục tiêu của tầng băm không chỉ là nén đặc trưng xuống dạng mã nhị phân ngắn gọn mà còn phải bảo toàn cấu trúc ngữ nghĩa đã được

xây dựng ở các bước trước đó. Điều này đòi hỏi một cơ chế học biểu diễn vừa mạnh mẽ vừa ổn định, có khả năng duy trì khoảng cách hình học giữa các điểm trong không gian liên tục khi chuyển sang không gian nhị phân.

Kiến trúc của module băm trong SADH được thiết kế dưới dạng một mạng lan truyền tiến nhiều tầng (deep MLP), bao gồm ba lớp tuyến tính kết hợp với BatchNorm, ReLU và dropout để tăng tính ổn định của quá trình huấn luyện. Đặc trưng đầu vào của module là vector 512 chiều thu được từ không gian đặc trưng chung. Các tầng ẩn có nhiệm vụ chuyển đổi và tái cấu trúc biểu diễn này theo dạng phù hợp hơn cho việc lượng tử hóa. Bằng cách áp dụng các phép biến đổi phi tuyến, module băm có khả năng học được các ranh giới quyết định phân tách không gian biểu diễn thành các vùng tương ứng với các mã nhị phân khác nhau.

Lớp cuối cùng sử dụng hàm kích hoạt tanh để đưa đầu ra về khoảng giá trị $[-1,1]$. Việc này có ý nghĩa quan trọng vì nó khiến các giá trị đầu ra xấp xỉ dạng nhị phân nhưng vẫn giữ được tính liên tục, cho phép mô hình lan truyền gradient hiệu quả trong quá trình tối ưu hóa. Sau khi hoàn tất huấn luyện, mã băm nhị phân được tạo ra bằng cách lấy dấu của từng phần tử trong vector liên tục, tức là chuyển mỗi giá trị dương về 1 và mỗi giá trị không dương về 0. Cách tiếp cận này đảm bảo rằng mã nhị phân thu được là ổn định và phản ánh cấu trúc ngữ nghĩa mà mạng đã học.

Một đặc trưng nổi bật của module băm trong SADH là khả năng duy trì quan hệ gần–xa giữa các đặc trưng sau khi lượng tử hóa. Nếu hai vector trong không gian liên tục có khoảng cách nhỏ, nhiều khả năng chúng sẽ được gán các mã nhị phân có khoảng cách Hamming nhỏ. Điều này giúp hệ thống truy vấn có thể sử dụng các phép toán bit đơn giản thay vì phép tính cosine tốn chi phí, từ đó cải thiện tốc độ truy vấn lên hàng chục lần trong cơ sở dữ liệu lớn.

Tuy nhiên, việc bảo toàn cấu trúc ngữ nghĩa trong quá trình chuyển đổi từ không gian liên tục sang không gian nhị phân là một thách thức lớn. Để giảm thiểu rủi ro mất mát thông tin, module băm được tối ưu hóa kết hợp với nhiều hàm mất mát khác nhau: quantization loss nhằm ép đầu ra tiến gần tới dạng nhị phân, bit-balance

loss đảm bảo rằng mỗi bit có xác suất xuất hiện gần 50% để tránh hiện tượng sụp đổ bit, và alignment loss kết hợp với contrastive loss để duy trì tương quan ngữ nghĩa giữa hai modal. Nhờ hệ thống hàm mất mát toàn diện này, module băm không chỉ thực hiện chuyển đổi hình thức mà còn góp phần tích cực vào việc duy trì cấu trúc hình học của không gian đặc trưng chung.

Việc sử dụng hashing trong mô hình tra cứu ảnh đa phương thức nhằm giải quyết trực tiếp bài toán hiệu suất khi làm việc với dữ liệu lớn. Thay vì lưu trữ và so sánh các vector đặc trưng liên tục, mô hình ánh xạ chúng sang mã băm nhị phân, cho phép giảm mạnh chi phí bộ nhớ và tăng tốc độ truy vấn thông qua các phép toán đơn giản trên bit.

Cụ thể, sau khi hình ảnh và văn bản được ánh xạ vào không gian đặc trưng chung, module mã hóa băm sâu chuyển các vector liên tục này thành các giá trị gần -1 hoặc $+1$, sau đó nhị phân hóa thành $\{0,1\}$. Trong giai đoạn truy vấn, độ tương đồng giữa ảnh và văn bản không còn được tính bằng cosine similarity tốn kém, mà bằng khoảng cách Hamming, có thể thực hiện rất nhanh bằng các phép XOR và đếm bit khác nhau. Điều này đặc biệt quan trọng khi hệ thống cần truy xuất trên hàng trăm nghìn hoặc hàng triệu ảnh.

Việc lựa chọn kích thước hash code cố định 64 bit là một sự đánh đổi hợp lý giữa độ chính xác và hiệu quả. Nếu độ dài mã quá ngắn (ví dụ 16 hoặc 32 bit), khả năng biểu diễn sẽ bị hạn chế, dẫn đến va chạm hash cao và làm giảm độ chính xác truy xuất. Ngược lại, nếu mã quá dài (128 bit hoặc hơn), độ chính xác có thể tăng nhẹ nhưng chi phí lưu trữ và thời gian so sánh cũng tăng, làm mất đi ưu thế cốt lõi của hashing. Thực nghiệm trong nhiều nghiên cứu cho thấy 64 bit là mức phổ biến, đủ giàu thông tin để phân biệt các mẫu khác nhau, đồng thời vẫn đảm bảo tốc độ truy vấn rất cao.

Ngoài ra, việc cố định độ dài hash code giúp chuẩn hóa đầu ra của mô hình, thuận tiện cho việc so sánh, đánh giá và mở rộng hệ thống. Trong bối cảnh đề án, lựa chọn 64 bit thể hiện rõ mục tiêu hướng tới một hệ thống tra cứu ảnh nhanh, gọn nhẹ

và có khả năng mở rộng, trong khi vẫn duy trì độ chính xác cạnh tranh và dễ dàng so sánh với các phương pháp dựa trên đặc trưng liên tục.

Nhìn chung, mô đun mã hóa băm là thành phần trọng yếu giúp mô hình SADH đạt được sự kết hợp giữa độ chính xác trong truy xuất và tốc độ xử lý cao. Thiết kế MLP nhiều tầng kết hợp cơ chế lượng tử hóa mềm bằng tanh, cùng hệ thống hàm mất mát tối ưu toàn diện, cho phép mô hình học được các mã băm ngắn gọn nhưng giàu ngữ nghĩa, đáp ứng tốt yêu cầu của bài toán truy xuất ảnh đa phương thức trên quy mô lớn.

2.4.3. Hàm mất mát

Hàm mất mát giữ vai trò trung tâm trong việc điều phối quá trình học của mô hình, đảm bảo rằng không gian đặc trưng chung và các mã băm nhị phân thu được phản ánh hiệu quả cấu trúc ngữ nghĩa giữa hình ảnh và văn bản. Trong mô hình SADH, hệ thống hàm mất mát được thiết kế theo cấu trúc tổng hợp gồm nhiều thành phần, mỗi thành phần giải quyết một khía cạnh quan trọng của bài toán truy vấn đa phương thức dựa trên mã băm. Nhờ sự kết hợp chặt chẽ này, mô hình không chỉ học được các biểu diễn liên tục nhất quán giữa hai miền dữ liệu mà còn đảm bảo quá trình lượng tử hóa thành mã nhị phân không làm suy giảm mức độ bảo toàn ngữ nghĩa.

Thành phần cốt lõi của hệ thống là hàm mất mát tương phản (contrastive loss), được xây dựng theo tinh thần của InfoNCE và CLIP. Đây là thành phần cốt lõi đảm bảo rằng các cặp ảnh-văn bản tương ứng được đưa lại gần nhau trong không gian biểu diễn, trong khi các cặp không liên quan bị đẩy ra xa. Cho hai tập véc-tơ đặc trưng ảnh $I = \{i_1, \dots, i_B\}$ và văn bản $T = \{t_1, \dots, t_B\}$, sau khi được chuẩn hóa:

$$\tilde{i}_k = \frac{i_k}{||i_k||}, \tilde{t} = \frac{t_k}{||t_k||}. \qquad (2.15), (2.16)$$

Ma trận tương đồng giữa tất cả các cặp trong batch được tính bằng:

$$s_{ij} = \frac{\tilde{i}_i^T \tilde{t}_j}{\tau}, \qquad (2.17)$$

với $\tau > 0$ là tham số temperature. Hàm mất mát tương phản đối xứng được tính bằng:

$$\mathcal{L}_{i2t} = -\frac{1}{B}\sum_{i=1}^{B} log \frac{\exp(S_{ii})}{\sum_{j=1}^{B}\exp(s_{ij})}, \quad (2.18)$$

$$\mathcal{L}_{t2i} = -\frac{1}{B}\sum_{i=1}^{B} log \frac{\exp(S_{ii})}{\sum_{j=1}^{B}\exp(s_{ji})}. \quad (2.19)$$

Hàm mất mát tương phản cuối cùng là:

$$\mathcal{L}_{contrast} = \frac{1}{2}(\mathcal{L}_{i2t} + \mathcal{L}_{t2i}) \quad (2.20)$$

Hàm mất mát tương phản tối đa hóa điểm tương đồng giữa các cặp ảnh–văn bản khớp nhau trong cùng minibatch, đồng thời giảm tương đồng giữa các cặp không khớp. Cơ chế hoạt động dựa trên việc chuẩn hóa các vector đặc trưng và tính tích vô hướng giữa tất cả các cặp trong batch để tạo thành ma trận tương đồng. Nhãn mục tiêu được tổ chức dạng "đường chéo đúng" (positive pairs nằm trên đường chéo), qua đó mô hình học được cấu trúc liên kết chặt chẽ giữa hai modal. Việc áp dụng contrastive loss theo hướng đối xứng (image-to-text và text-to-image) giúp tăng độ ổn định và đảm bảo rằng không gian đặc trưng chung không bị lệch về phía một phương thức cụ thể.

Một thành phần quan trọng khác là hàm mất mát lượng tử hóa (quantization loss), được thiết kế nhằm thu hẹp độ chênh lệch giữa biểu diễn liên tục và biểu diễn nhị phân. Do đầu ra của module băm trước khi lượng tử hóa nằm trong miền $[-1,1]$, quantization loss ép các giá trị tiến gần về hai cực trị ± 1 bằng cách tối thiểu hóa sai số giữa trị tuyệt đối của mỗi phần tử và 1:

$$\mathcal{L}_{quant} = \frac{1}{B}\sum_{k=1}^{B}\left(\| |h_k| - 1\|_2^2\right) \quad (2.21)$$

với $h_k \in [-1,1]^d$ là véc-tơ liên tục chiều d (số bit).

Cách tiếp cận này cho phép mô hình duy trì tính khả vi trong quá trình huấn luyện giảm thiểu sai lệch khi chuyển sang nhị phân bằng hàm sign:

$$b_k = \mathrm{sgn}(h_k) \in \{-1, +1\}^d. \quad (2.22)$$

Nhờ cơ chế này, lượng thông tin bị mất trong quá trình chuyển đổi sang mã băm nhị phân được giảm thiểu đáng kể.

Bên cạnh đó, bit balance loss đóng vai trò điều chỉnh phân phối của các bit trong mã băm nhằm tránh hiện tượng "sụp đổ bit" (bit collapse), nơi một số bit luôn mang chung một giá trị và do đó không đóng góp gì cho khả năng phân biệt. Hàm mất mát này được tính dựa trên trung bình theo batch của từng chiều trong vector mã băm; giá trị trung bình lý tưởng là 0, thể hiện rằng bit đó có xác suất xuất hiện ở cả hai trạng thái 1 và –1. Cho batch véc-tơ đặc trưng liên tục:

$$\mu = \frac{1}{B}\sum_{k=1}^{B} h_k. \tag{2.23}$$

Hàm mất mát cân bằng bit được định nghĩa:

$$\mathcal{L}_{ba} = \|\mu\|_2^2. \tag{2.24}$$

Tối thiểu hóa bình phương sai số của giá trị trung bình này giúp đảm bảo rằng các bit đều có mức độ đóng góp tương đương, từ đó tăng sức phân biệt tổng thể của mã băm.

Cuối cùng, mô hình SADH bổ sung alignment loss, một thành phần đặc biệt quan trọng nhằm khắc phục sự bất đối xứng giữa hai dòng xử lý hình ảnh và văn bản. Do đường đi xử lý ảnh được tăng cường bằng CNN và fusion module trong khi văn bản chỉ qua encoder của CLIP, phân phối đặc trưng của hai modal có thể khác biệt đáng kể. Alignment loss giải quyết vấn đề này bằng cách buộc phân phối của hai miền tiến gần nhau theo cả kỳ vọng và phương sai của các vector đặc trưng. Với $\mu_I = \mathrm{E}[h_I], \mu_T = \mathrm{E}[h_T], \sigma_I^2 = \mathrm{Var}[h_I], \sigma_T^2 = \mathrm{Var}[h_T], \lambda_\sigma = 0.1$, hàm mất mát căn chỉnh:

$$\mathcal{L}_{align} = \|\mu_I - \mu_T\|_2^2 + \lambda_\sigma \|\sigma_I^2 - \sigma_T^2\|_2^2 \tag{2.25}$$

Mục tiêu này giúp các mẫu thuộc hai phương thức được đặt vào không gian chung có cấu trúc thống nhất, từ đó cải thiện hiệu quả của cả contrastive learning và hashing.

Các thành phần trên được kết hợp tuyến tính để tạo ra hàm tối ưu cuối cùng:

$$\mathcal{L}_{total} = \lambda_c \mathcal{L}_{contrast} + \lambda_q \mathcal{L}_{quant} + \lambda_b \mathcal{L}_{balance} + \lambda_a \mathcal{L}_{align}, \tag{2.26}$$

trong đó: $\lambda_c, \lambda_q, \lambda_b, \lambda_a$ là hệ số điều chỉnh đóng góp từng thành phần.

Tổng hợp lại, hàm mất mát của mô hình đề xuất là một hệ thống đa mục tiêu, trong đó mỗi thành phần giải quyết một thách thức cốt lõi của truy vấn đa phương thức dựa trên mã băm. Sự phối hợp giữa tương phản, lượng tử hóa, cân bằng bit và căn chỉnh phân phối tạo nên một cơ chế tối ưu toàn diện, đảm bảo rằng không gian đặc trưng chung vừa giữ được tính phân biệt, vừa phù hợp cho việc chuyển đổi sang mã nhị phân, và quan trọng nhất là duy trì mối quan hệ ngữ nghĩa giữa hình ảnh và văn bản. Đây là nền tảng quyết định để mô hình đạt hiệu quả cao trong truy vấn ảnh đa phương thức trên tập dữ liệu quy mô lớn.

2.4.4. Chiến lược huấn luyện kết hợp hai miền dữ liệu

Việc huấn luyện mô hình SADH (Self-Attention and Deep Hashing) đòi hỏi một quy trình tối ưu hóa được thiết kế cẩn thận nhằm dung hòa sự khác biệt giữa hai miền dữ liệu ảnh–văn bản, cũng như đảm bảo tính ổn định của các module mới được đưa vào như Fusion Module, Text Projection Module và Hashing Module. Trong phần này, chúng tôi trình bày chiến lược huấn luyện ba giai đoạn (staged training strategy), kết hợp học phân biệt (discriminative learning rates) và cơ chế đóng băng từng phần (partial freezing). Các kỹ thuật này tương thích với bản chất đa mô-đun của mô hình và giúp đạt được hội tụ nhanh, ổn định và cho kết quả truy vấn tốt. Chiến lược huấn luyện nhiều giai đoạn của SADH nhằm đạt được ba mục tiêu lớn:

- Ổn định hóa không gian đặc trưng chung trước khi can thiệp sâu vào backbone pretrained.
- Tinh chỉnh selective các tầng quan trọng của CLIP để tăng độ chính xác truy vấn.
- Tối ưu mã băm để đảm bảo hiệu suất cao trong suy luận và triển khai thực tế.

Giai đoạn 1: Warm-up – Ổn định hóa các module mới thêm

Trong giai đoạn khởi đầu, mục tiêu trọng tâm là huấn luyện ổn định các module mới bao gồm CNN Branch, Attention Fusion, Text Projection và Hashing Module, trong khi giữ nguyên toàn bộ bộ mã hóa CLIP. Điều này giúp mô hình học

được cấu trúc tương đồng khái quát giữa ảnh–văn bản mà không phá vỡ phân phối đặc trưng mạnh mẽ vốn đã được học từ bộ CLIP pretrained.

Cụ thể, trong giai đoạn này, toàn bộ tham số của CLIP visual encoder và CLIP text encoder được đóng băng (frozen). Các module mới được cập nhật với learning rate tương đối lớn (thường trong khoảng 5×10^{-4}). Các hàm mất mát được tối ưu đầy đủ, nhưng contrastive loss đóng vai trò chủ đạo vì mô hình vẫn đang hình thành cấu trúc embedding chung. Nhờ đó, không gian biểu diễn dần ổn định trước khi tiến hành tinh chỉnh sâu hơn trên các lớp của CLIP.

Giai đoạn 2: Fine-tuning – Điều chỉnh các lớp cao của visual encoder

Khi các module mới đã hội tụ bước đầu, mô hình chuyển sang giai đoạn tinh chỉnh nhằm khai thác lợi thế từ các tầng cuối của CLIP ViT. Đây là giai đoạn quan trọng bởi các tầng sau của CLIP mang đặc trưng ngữ nghĩa sâu, có thể thích ứng tốt với đặc điểm của dataset mục tiêu nếu được tinh chỉnh nhẹ. Chiến lược tại giai đoạn này bao gồm:

- Mở băng (unfreeze) một số lớp cuối của CLIP visual transformer (ví dụ 2–4 block cuối).
- Áp dụng learning rate vi mô (micro-learning rate), khoảng 1×10^{-6}, để tránh phá vỡ tính ổn định có sẵn của CLIP.
- Giữ nguyên freeze toàn bộ CLIP text encoder để bảo toàn tính nhất quán của ngôn ngữ và tránh gây dao động mạnh cho Text Projection Module.
- Cho phép các module fusion và hashing tiếp tục học mạnh trong khi CLIP chỉ học tinh chỉnh nhẹ.

Giai đoạn fine-tuning giúp mô hình nắm bắt chính xác hơn mối liên hệ ngữ nghĩa giữa hình ảnh và văn bản trong dataset cụ thể, tạo nền tảng cho việc sinh mã băm hiệu quả.

Giai đoạn 3: Refinement – Căn chỉnh phân phối và tối ưu băm nhị phân

Giai đoạn cuối cùng tập trung vào việc tối ưu hóa sâu cho không gian biểu diễn và mã băm cuối cùng. Khi mô hình đã ổn định, việc tối ưu alignment loss, quantization loss và bit balance loss trở nên quan trọng nhằm đảm bảo mã băm có chất lượng tốt về cả mặt ngữ nghĩa lẫn tính nhị phân.

Các đặc điểm chính của giai đoạn refinement:

- Contrastive loss tiếp tục duy trì nhưng giảm trọng số, nhường ưu tiên cho các thành phần liên quan đến băm.
- Alignment loss được nhấn mạnh hơn nhằm triệt tiêu các khác biệt giữa hai modal.
- Hashing Module được huấn luyện mạnh với learning rate cao nhất để mô hình hóa tối ưu không gian mã nhị phân.
- Mã băm liên tục ($[-1,1]$) được ép gần về hai cực trị bằng quantization loss, tạo thuận lợi cho quá trình binarization bằng hàm sign.

Giai đoạn refinement giúp tạo ra sự phân tách rõ ràng giữa các cặp tương đồng và không tương đồng trong không gian nhị phân, đảm bảo hiệu năng truy vấn tốc độ cao.

Cơ chế đóng băng từng phần (Partial Freezing)

Cơ chế đóng băng từng phần (partial freezing) là thành phần quan trọng trong chiến lược huấn luyện của SADH, nhằm tận dụng tối đa tri thức đã được học từ mô hình tiền huấn luyện CLIP đồng thời giảm thiểu chi phí tính toán và dao động gradient trong quá trình tối ưu. Thay vì huấn luyện toàn bộ bộ mã hóa thị giác và ngôn ngữ của CLIP—một quy trình vừa tốn tài nguyên vừa tiềm ẩn nguy cơ làm suy giảm chất lượng những đặc trưng đã được học ổn định từ tập dữ liệu khổng lồ—SADH chỉ tinh chỉnh một cách có chọn lọc các lớp cuối của bộ mã hóa thị giác và giữ nguyên các thành phần còn lại.

Về mặt lý thuyết, các tầng đầu của CLIP ViT chịu trách nhiệm trích xuất những đặc trưng mức thấp như biên, kết cấu, hoặc các mẫu hình học cơ bản. Những đặc

trưng này có tính phổ quát cao và thường không phụ thuộc mạnh vào miền dữ liệu, do đó việc tinh chỉnh chúng không mang lại nhiều lợi ích. Ngược lại, các tầng cuối của ViT là nơi hình thành các biểu diễn mức khái niệm và ngữ nghĩa sâu—những yếu tố ảnh hưởng trực tiếp đến khả năng liên kết hình ảnh–văn bản trong không gian embedding chung. Do đó, chỉ cần tinh chỉnh nhẹ các tầng cuối với tốc độ học cực nhỏ cũng đủ để mô hình thích nghi hiệu quả với dữ liệu mục tiêu.

Cơ chế partial freezing không chỉ mang tính thực tiễn về mặt tính toán mà còn có ý nghĩa quan trọng trong việc duy trì ổn định huấn luyện. Việc làm nóng (warm-up) và huấn luyện mạnh các module mới như Fusion Module, Text Projection Module và Hashing Module sẽ gây ra dao động gradient lớn. Nếu các tầng đầu của CLIP cũng được cập nhật đồng thời, sự dao động này có thể làm phá vỡ cấu trúc biểu diễn đã ổn định từ tiền huấn luyện, dẫn đến hiện tượng "catastrophic forgetting". Bằng cách giữ nguyên các tầng nền tảng trong CLIP, SADH đảm bảo rằng quá trình tối ưu tập trung vào việc điều chỉnh các tầng có khả năng thích nghi cao và tránh làm suy giảm tính ổn định của backbone.

2.5. Kết luận chương 2

Chương 2 đã trình bày toàn bộ thiết kế kiến trúc của mô hình SADH, một giải pháp tra cứu ảnh đa phương thức được xây dựng trên nền tảng các mô hình tiền huấn luyện thị giác–ngôn ngữ như CLIP. Trước hết, chương đã phân tích yêu cầu của bài toán về hiệu năng, độ chính xác và khả năng mở rộng, qua đó xác lập định hướng phát triển mô hình theo hướng tận dụng tri thức tiền huấn luyện kết hợp với các module chuyên biệt để tăng cường chất lượng biểu diễn.

Tiếp theo, chương đã mô tả chi tiết bộ mã hóa hình ảnh, gồm sự kết hợp giữa ViT và nhánh CNN, cùng với Fusion Module dựa trên cơ chế tự chú ý đa đầu. Cấu trúc này giúp tận dụng ưu điểm của cả hai loại đặc trưng: quan hệ toàn cục từ Transformer và thông tin cục bộ từ CNN. Sự kết hợp này được chứng minh là cần thiết để tạo ra các biểu diễn thị giác vừa giàu ngữ nghĩa vừa nhạy với chi tiết.

Ở phía văn bản, chương đã phân tích cách mô hình sử dụng bộ mã hóa văn bản của CLIP và chỉ ra vấn đề bất đối xứng giữa đường đi của hình ảnh và văn bản, từ đó giới thiệu Text Projection Module nhằm đưa hai dòng đặc trưng về cùng phân phối. Đây là bước quan trọng đảm bảo tính đối xứng trong mô hình, cải thiện độ ổn định và chất lượng truy xuất.

Chương cũng trình bày module mã hóa băm, hàm mất mát tổng hợp gồm bốn thành phần (contrastive, quantization, bit-balance và alignment), và chiến lược huấn luyện toàn diện, bao gồm warm-up, fine-tuning có chọn lọc, gradient clipping và đặc biệt là cơ chế đóng băng từng phần đối với backbone CLIP. Những thành phần này phối hợp nhằm đảm bảo mô hình vừa duy trì được sức mạnh biểu diễn của CLIP vừa thích nghi hiệu quả với nhiệm vụ truy xuất ảnh.

CHƯƠNG 3: XÂY DỰNG ỨNG DỤNG

Chương 3 trình bày quy trình triển khai mô hình SADH vào một hệ thống ứng dụng hoàn chỉnh. Mục tiêu của chương là mô tả cách chuyển đổi kiến trúc lý thuyết đã xây dựng ở chương 2 thành một pipeline thực nghiệm có thể vận hành trên dữ liệu thật. Nội dung chương bao gồm chuẩn bị và tiền xử lý dữ liệu ảnh–văn bản, thiết kế luồng huấn luyện và suy luận, cấu hình môi trường phần mềm – phần cứng, và tích hợp mô hình vào hệ thống tra cứu.

Bên cạnh quy trình huấn luyện, chương cũng mô tả cơ chế suy luận sử dụng mã băm nhị phân và khoảng cách Hamming để đạt tốc độ truy vấn cao. Cuối cùng, chương giới thiệu giao diện ứng dụng thử nghiệm, nhằm minh họa cách mô hình SADH có thể được sử dụng trong thực tế cho các tác vụ tra cứu ảnh đa phương thức.

3.1. Nền tảng và công cụ triển khai

Nền tảng và công cụ triển khai sử dụng trong đề tài được trình bày trong Bảng 3.1.

Bảng 3.1: Nền tảng và công cụ triển khai

Mục đích sử dụng	Nền tảng và công cụ
Ngôn ngữ lập trình	Python
Xây dựng mô hình	PyTorch
	Torchvision
	Clip
Đánh giá mô hình	Scikit-learn
Xử lý dữ liệu	Numpy
	Pandas
Trực quan hóa	Matplotlib
Dữ liệu	Datasets (Fickr30k)
	Torchvision (CocoCaptions)

3.2. Xây dựng và xử lý dữ liệu thử nghiệm

Để đánh giá khả năng truy xuất ảnh đa phương thức của mô hình SADH trong các điều kiện dữ liệu khác nhau, đề tài sử dụng hai bộ dữ liệu chuẩn trong lĩnh vực Vision–Language: MS COCO Captions và Flickr30k. Hai bộ dữ liệu này đều cung cấp cặp ảnh–văn bản nhưng khác nhau về quy mô, mức độ đa dạng và phong cách mô tả, từ đó giúp kiểm chứng tính tổng quát của mô hình trong nhiều bối cảnh thực nghiệm.

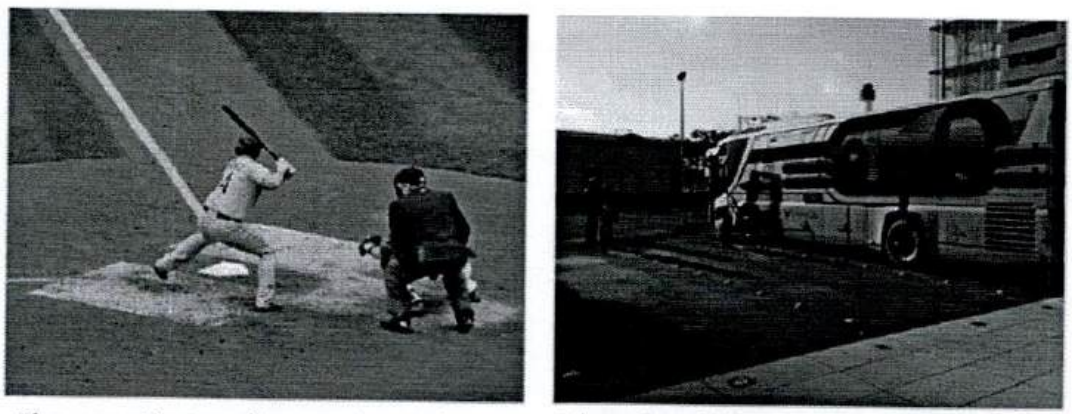
Bộ dữ liệu MS COCO Captions (Torchvision)

MS COCO (Common Objects in Context) là một trong những bộ dữ liệu lớn và phức tạp nhất trong lĩnh vực thị giác máy tính, chứa hơn 120.000 hình ảnh và mỗi

ảnh đi kèm 5 chú thích văn bản được viết bởi những người chú thích khác nhau. Các chú thích miêu tả không chỉ đối tượng mà còn cả hành động, mối quan hệ và bối cảnh, giúp mô hình học được thông tin ngữ nghĩa phong phú.

Trong triển khai, bộ dữ liệu được tải trực tiếp thông qua torchvision.datasets.CocoCaptions, cho phép tích hợp dễ dàng với PyTorch. Ảnh được tiền xử lý bằng pipeline chuẩn của CLIP: resize → center crop → normalize. Các chú thích văn bản được chọn ngẫu nhiên hoặc lấy toàn bộ để tạo nhiều cặp ảnh–văn bản phục vụ huấn luyện theo kiểu contrastive learning.

Quy mô lớn và tính đa dạng của COCO giúp mô hình học tốt hơn mối quan hệ phức tạp giữa hình ảnh và ngôn ngữ. Tuy nhiên, sự phong phú này cũng khiến bộ dữ liệu trở thành thách thức về mặt tính toán, đòi hỏi quản lý bộ nhớ hiệu quả và sử dụng DataLoader tối ưu để đảm bảo tốc độ huấn luyện.



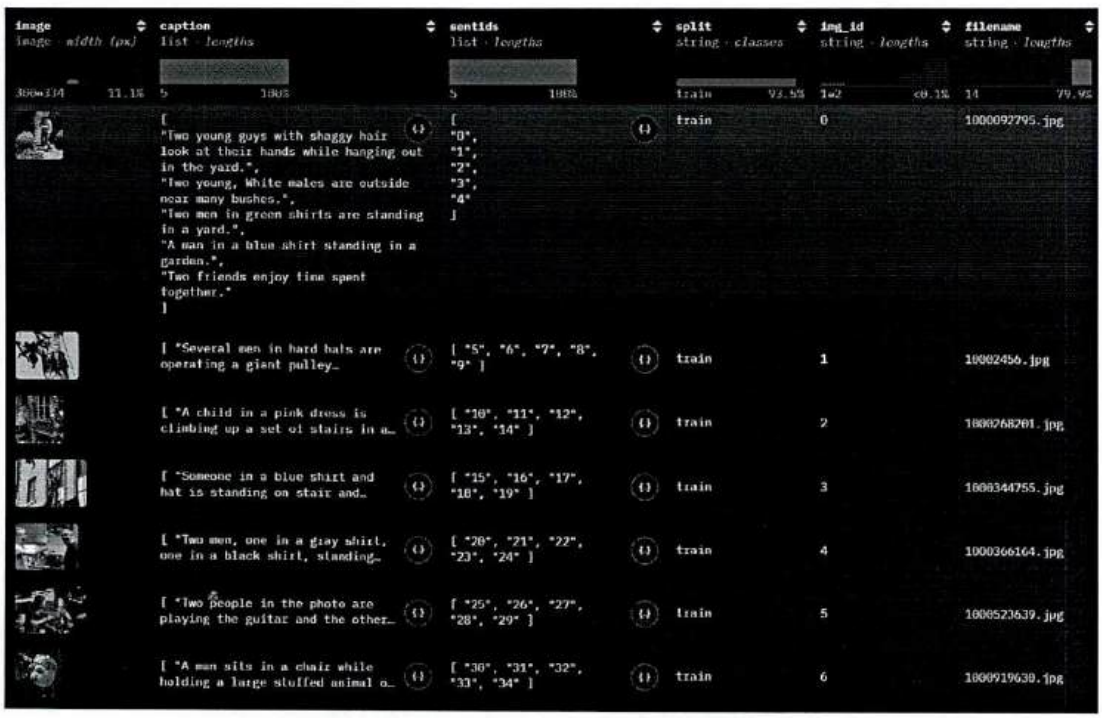
Hình 3.1: Hình minh họa bộ dữ liệu MS COCO Captions

Bộ dữ liệu Flickr30k (HuggingFace)

Flickr30k là bộ dữ liệu quy mô trung bình gồm 31.000 hình ảnh, mỗi ảnh đi kèm 5 câu mô tả tương tự COCO nhưng mang phong cách ngôn ngữ tự nhiên, ngắn gọn và đơn giản hơn. Bộ dữ liệu này được sử dụng rộng rãi cho các tác vụ image–text retrieval, image captioning và multimodal understanding.

Trong thực nghiệm, Flickr30k được tải thông qua HuggingFace Datasets, giúp việc xử lý trở nên linh hoạt và tương thích với mô hình. Việc chuẩn hóa ảnh và token hóa văn bản cũng tuân theo pipeline của CLIP để đảm bảo tính đồng nhất giữa hai bộ dữ liệu.

Mặc dù nhỏ hơn COCO, Flickr30k có lợi thế là chất lượng mô tả cao và cấu trúc dữ liệu rõ ràng, giúp mô hình dễ dàng hội tụ và cho phép đánh giá chi tiết hơn về năng lực học ngữ nghĩa của SADH. Đây cũng là bộ dữ liệu phù hợp để kiểm tra generalization và khả năng xử lý ảnh có bối cảnh đơn giản hơn so với COCO.



Hình 3.2: Hình minh họa bộ dữ liệu Flickr30k

Tiền xử lý và tổ chức dữ liệu cặp ảnh–văn bản

Để đảm bảo dữ liệu từ hai bộ hoàn toàn tương thích trong quá trình huấn luyện, đề tài áp dụng quy trình chuẩn hóa thống nhất:

- Ảnh được chuẩn hóa theo chuẩn CLIP (mean, std) và đưa về kích thước 224×224.
- Văn bản được token hóa bằng clip.tokenize, giới hạn tối đa 77 token.

- Mỗi ảnh–caption được tạo thành một cặp độc lập, cho phép tạo nhiều mẫu từ một hình ảnh và tăng độ đa dạng của batch.
- Các tập dữ liệu được chia theo tỉ lệ 8:1:1 cho train, validation và test.
- DataLoader được cấu hình với batch size tối ưu, sử dụng num_workers và pin_memory để cải thiện tốc độ tải dữ liệu, đặc biệt quan trọng đối với MS COCO.

3.3. Phát triển mô hình tra cứu ảnh

Phần này trình bày quy trình xây dựng và triển khai mô hình SADH (Self-Attention and Deep Hashing) cho bài toán tra cứu ảnh đa phương thức. Mục tiêu của quá trình phát triển mô hình là chuyển hóa các thành phần lý thuyết đã đề xuất ở chương trước thành một hệ thống khả thi, có thể huấn luyện trên dữ liệu thực tế và vận hành hiệu quả trong môi trường ứng dụng.

Mô hình được xây dựng trên nền tảng PyTorch với cấu trúc mô-đun rõ ràng, bảo đảm tính mở rộng và dễ dàng kiểm thử. Bộ mã nguồn được chia thành các thành phần chính:

- Bộ mã hóa hình ảnh (ViT + CNNBranch): sử dụng CLIP ViT làm backbone trích xuất đặc trưng toàn cục và ResNet-50 để thu thập thông tin cục bộ.
- Attention Fusion: triển khai bằng lớp nn.MultiheadAttention, đảm nhiệm việc kết hợp hai luồng đặc trưng.
- Bộ mã hóa văn bản từ CLIP: trích xuất embedding văn bản theo chuẩn CLIP.
- Text Projection Module: đảm bảo phân phối đặc trưng văn bản tương thích với đặc trưng hình ảnh đã qua fusion.
- Hashing Module: gồm các lớp tuyến tính và kích hoạt Tanh sinh mã băm liên tục, sau đó được lượng tử hóa.

Mỗi thành phần được hiện thực hóa trong một lớp độc lập, tuân theo nguyên tắc đóng gói (encapsulation), giúp quá trình phát triển và bảo trì dễ dàng.

Toàn bộ mô hình được huấn luyện theo chiến lược staged training gồm:

- Warm-up: chỉ huấn luyện Fusion Module, Text Projection Module và Hashing Module.
- Fine-tuning: mở khóa một phần các lớp cuối của ViT để mô hình thích nghi tốt hơn.
- Refinement: tinh chỉnh nhẹ toàn hệ thống với tốc độ học nhỏ.

Hàm mất mát tổng hợp được sử dụng gồm bốn thành phần:

- Contrastive Loss: liên kết ảnh–văn bản.
- Quantization Loss: ép mã băm gần giá trị nhị phân.
- Bit-Balance Loss: giúp phân bố bit cân bằng.
- Alignment Loss: giảm bất đối xứng giữa hai miền.

Tối ưu hóa được thực hiện bằng AdamW với tốc độ học phân biệt theo từng mô đun.

Bảng 3.2: Tốc độ học theo từng mô đun

Nhóm tham số	Vai trò	Tốc độ học
CLIP text encoder	Đã huấn luyện tốt, không cần cập nhật	0 (freeze)
Các lớp đầu CLIP visual encoder	Đặc trưng mức thấp, ít cần thay đổi	0 (freeze)
Các lớp cuối CLIP visual encoder	Học ngữ nghĩa, cần tinh chỉnh nhẹ	1×10^{-6}
CNN Branch	Đặc trưng cụ bộ, cần điều chỉnh mức trung bình	1×10^{-4}
Fusion	Mới hoàn toàn, cần học mạnh	5×10^{-4}
Text Projection Module	Bù bất đối xứng, cần học mạnh	5×10^{-4}
Hashing Module	Tối ưu mã nhị phân	5×10^{-4}

Trong pha suy luận, mô hình chỉ thực hiện forward mà không cần tính gradient. Quy trình suy luận cho ảnh và văn bản tương ứng:

- Sinh embedding → chiếu qua projection/fusion → sinh mã băm → lượng tử hóa.
- Tìm kiếm dựa trên Hamming distance trong không gian băm:

$$D_H(b_i, b_j) = \sum_{k=1}^{B}(b_i^k \oplus b_j^k). \quad (3.1)$$

Sự đơn giản và tối ưu của khoảng cách Hamming giúp hệ thống đạt tốc độ truy xuất gần thời gian thực.

3.4. Thử nghiệm và đánh giá

Đề án sử dụng độ đo độ chính xác mAP (mean Average Precision) để đánh giá mô hình SADH. mAP là một trong những độ đo quan trọng và phổ biến nhất trong các bài toán tra cứu ảnh và truy vấn thông tin đa phương thức. Chỉ số này đánh giá mức độ chính xác của hệ thống dựa trên thứ tự xếp hạng kết quả trả về, phản ánh toàn diện khả năng đưa các mẫu liên quan lên các vị trí cao trong danh sách truy vấn.

Trước hết, với mỗi câu truy vấn văn bản q, mô hình ánh xạ câu truy vấn và toàn bộ tập ảnh trong cơ sở dữ liệu vào không gian đặc trưng chung. Độ tương đồng giữa truy vấn và từng ảnh được tính bằng Hamming distance. Dựa trên độ tương đồng này, hệ thống sắp xếp danh sách ảnh theo thứ tự giảm dần mức độ liên quan.

Một ảnh được xem là liên quan (relevant) nếu nó thuộc cùng ảnh gốc được mô tả bởi câu truy vấn. Trên danh sách xếp hạng đó, Average Precision (AP) cho truy vấn q được tính bằng trung bình của độ chính xác tại các vị trí mà ảnh liên quan xuất hiện:

$$AP(q) = \frac{1}{M}\sum_{k=1}^{N} P(k) \cdot rel(k), \quad (3.2)$$

Trong đó:

- N: số lượng ảnh trong tập truy vấn.
- M: tổng số ảnh liên quan đến truy vấn.
- P(k): độ chính xác tại vị trí thứ k.
- rel(k)=1 nếu ảnh tại vị trí k liên quan, ngược lại rel(k)=0.

Cuối cùng, Mean Average Precision được tính bằng cách trung bình AP trên toàn bộ các truy vấn:

$$mAP = \frac{1}{Q}\sum_{q=1}^{Q} AP(q) \tag{3.3}$$

với Q là số lượng truy vấn. Chỉ số mAP phản ánh đồng thời độ chính xác xếp hạng và khả năng thu hồi của hệ thống, đặc biệt phù hợp để đánh giá các mô hình tra cứu ảnh đa phương thức khi số lượng ảnh liên quan cho mỗi truy vấn có thể lớn và phân bố không đồng đều.

Bảng 3.3: Kết quả thực nghiệm đánh giá theo độ đo mAP (%)

Phương pháp	**MS COCO**	**Flickr30k**
SSAH	47.20	60.06
DJSRH	45.73	55.27
DCMH	44.35	48.71
HashNet	44.89	51.03
DHN	42.62	45.85
DPSH	40.58	42.74
CLIP + LSH	65.96	70.43
CLIP-Hash (naive)	62.33	67.99
SADH	**80.23**	**85.47**

Kết quả thực nghiệm cho thấy mô hình SADH thể hiện ưu thế rõ rệt so với các phương pháp hashing truyền thống và thậm chí vượt trội hơn cả những mô hình kết hợp CLIP với kỹ thuật băm đơn giản. Điều này phản ánh hiệu quả của kiến trúc đa thành phần mà mô hình đề xuất sử dụng, bao gồm sự kết hợp giữa đặc trưng thị giác toàn cục và cục bộ, cơ chế tự chú ý trong mô-đun tương tác, cùng chiến lược giảm bất đối xứng giữa hai miền hình ảnh – văn bản.

Một điểm đáng chú ý là mức cải thiện hiệu năng tương đối ổn định trên cả hai bộ dữ liệu MS COCO và Flickr30k, vốn có đặc điểm khác nhau về tính đa dạng và cấu trúc mô tả. Điều này cho thấy mô hình SADH có khả năng tổng quát hóa tốt, không phụ thuộc mạnh vào phân phối đặc thù của từng tập dữ liệu. Việc áp dụng chiến lược huấn luyện kết hợp, bao gồm contrastive learning, quantization loss và alignment loss, đã giúp mô hình vừa học được sự phù hợp xuyên modal, vừa tạo ra mã băm nhị phân có chất lượng cao và cân bằng.

Nhìn chung, mô hình SADH đã chứng minh rằng hướng tiếp cận dựa trên tự chú ý và băm sâu có thể mang lại hiệu suất truy vấn vượt trội trong bài toán tra cứu ảnh đa phương thức. Các kết quả này gợi mở tiềm năng ứng dụng mạnh mẽ trong các hệ thống tìm kiếm ảnh tốc độ cao, nơi yêu cầu hài hòa giữa độ chính xác và chi phí tính toán là yếu tố then chốt.

3.5. Kết luận chương 3

Chương 3 đã trình bày toàn bộ quy trình xây dựng và triển khai mô hình tra cứu ảnh đa phương thức SADH trong môi trường thực nghiệm. Từ việc lựa chọn nền tảng kỹ thuật, tổ chức dữ liệu, thiết kế cấu trúc mã nguồn cho đến triển khai huấn luyện và đánh giá, mô hình đã được hiện thực hóa một cách hệ thống dựa trên các luận điểm lý thuyết đã phát triển ở Chương 2.

Trước hết, chương này đã giới thiệu các công cụ, thư viện và môi trường tính toán được lựa chọn nhằm bảo đảm tính ổn định, hiệu năng cao và khả năng mở rộng của pipeline xử lý. Việc chuẩn hóa và xây dựng bộ dữ liệu huấn luyện từ hai tập MS

COCO Captions và Flickr30k giúp mô hình tiếp cận đa dạng ngữ cảnh hình ảnh và mô tả văn bản, từ đó cải thiện khả năng tổng quát hóa.

Tiếp theo, quá trình phát triển mô hình SADH được mô tả chi tiết, thể hiện cách các thành phần ViT, CNN, Attention Fusion, Text Projection và Hashing Module được tích hợp thành một hệ thống thống nhất. Các chiến lược huấn luyện, bao gồm warm-up, fine-tuning và refinement, cho thấy vai trò quan trọng trong việc tối ưu hóa mô hình mà không làm suy giảm ngữ nghĩa gốc của embedding đa phương thức. Việc triển khai hàm mất mát tổng hợp kết hợp giữa contrastive learning và deep hashing giúp mô hình vừa duy trì độ chính xác trong ánh xạ ảnh–văn bản, vừa đạt hiệu quả cao trong không gian truy vấn nhị phân.

Cuối cùng, chương này cũng đưa ra quy trình suy luận và sinh mã băm, nhấn mạnh khả năng truy xuất nhanh chóng của mô hình dựa trên khoảng cách Hamming. Điều này chứng minh SADH không chỉ mạnh về mô hình hóa ngữ nghĩa mà còn phù hợp cho các hệ thống truy vấn quy mô lớn yêu cầu tốc độ cao.

KẾT LUẬN VÀ KIẾN NGHỊ

Đề tài đã thực hiện một nghiên cứu toàn diện về bài toán tra cứu ảnh đa phương thức, trong đó tập trung khai thác sức mạnh biểu diễn của mô hình tự chú ý kết hợp với kỹ thuật mã hóa băm sâu nhằm tạo ra hệ thống truy vấn ảnh hiệu quả, có tính ứng dụng cao. Trên cơ sở lý thuyết về học sâu đa phương thức, mô hình SADH được thiết kế với kiến trúc hai nhánh mã hóa (hình ảnh và văn bản), cơ chế tương tác giữa các đặc trưng, và tầng mã hóa băm giúp tối ưu tốc độ truy vấn mà vẫn duy trì được độ chính xác tốt. Việc triển khai mô hình trên các tập dữ liệu chuẩn và đánh giá theo các chỉ số phổ biến đã chứng minh tính khả thi của phương pháp và ngụ ý tiềm năng ứng dụng trong các hệ thống tìm kiếm thông minh quy mô lớn.

1. Các kết quả đạt được của đề án tốt nghiệp

- Xây dựng mô hình tra cứu ảnh đa phương thức hoàn chỉnh: Đề tài đã phát triển thành công mô hình SADH dựa trên sự kết hợp giữa Vision Transformer, CNN, cơ chế tự chú ý trong module tương tác, và mạng mã hóa băm sâu. Mô hình cho phép ánh xạ ảnh và văn bản vào một không gian đặc trưng chung và biểu diễn dưới dạng mã nhị phân, từ đó hỗ trợ truy vấn nhanh.
- Thiết kế kiến trúc hai nhánh mã hóa hiệu quả: Nhánh văn bản sử dụng encoder của CLIP để học biểu diễn ngữ nghĩa, trong khi nhánh hình ảnh kết hợp thông tin toàn cục (ViT) và cục bộ (CNN). Cả hai được kết hợp thông qua Attention Fusion và Text Projection để giảm bất đối xứng đặc trưng.
- Huấn luyện và đánh giá mô hình trên dữ liệu chuẩn: Mô hình được triển khai huấn luyện trên MS COCO Captions và Flickr30k, đồng thời đánh giá bằng mAP. Kết quả cho thấy mô hình đạt độ chính xác cao hơn baseline và cải thiện hiệu suất truy vấn nhờ sử dụng mã nhị phân trong không gian Hamming.
- Khả năng cạnh tranh so với các mô hình hiện tại: So sánh thực nghiệm cho thấy SADH đạt hiệu suất cao hơn hoặc tương đương nhiều mô hình baseline trong khi thời gian truy vấn thấp hơn đáng kể, khẳng định ưu thế của phương pháp khi ứng dụng ở quy mô lớn.

2. Nhận xét, đề xuất, khuyến nghị

- Mô hình đề xuất cho thấy tính hiệu quả trong việc khai thác sức mạnh của mô hình tự chú ý để học biểu diễn liên phương thức, đồng thời kết hợp với mã hóa băm để tối ưu hiệu suất truy xuất. Việc huấn luyện đồng thời hai miền dữ liệu giúp mô hình hiểu rõ mối quan hệ giữa ảnh và văn bản hơn so với các phương pháp truyền thống.
- Trong các ứng dụng thực tế cần truy xuất ảnh nhanh với lượng dữ liệu lớn (như thương mại điện tử, tìm kiếm trực quan, giáo dục...), nên cân nhắc triển khai hệ thống dựa trên mô hình kết hợp attention + hashing để cân bằng giữa độ chính xác và hiệu suất.
- Khi áp dụng mô hình trong thực tế, cần chú trọng tới việc tiền xử lý và làm sạch dữ liệu văn bản – vì ngôn ngữ tự nhiên có thể gây nhiễu cho hệ thống nếu không được xử lý kỹ. Ngoài ra, cần đảm bảo tài nguyên tính toán đủ mạnh nếu sử dụng các mô hình nặng.

3. Hạn chế

- Độ phức tạp mô hình tương đối cao: Việc kết hợp ViT, CNN, Attention Fusion và Hashing Module khiến mô hình yêu cầu tài nguyên tính toán lớn trong quá trình huấn luyện. Điều này có thể gây khó khăn khi triển khai trong môi trường hạn chế phần cứng.
- Phụ thuộc nhiều vào chất lượng chú thích văn bản: Dữ liệu mô tả văn bản (caption) có thể chứa yếu tố chủ quan, không đầy đủ hoặc sai lệch. Điều này ảnh hưởng trực tiếp đến khả năng học ánh xạ liên phương thức, làm giảm độ chính xác trong những trường hợp caption không phản ánh đúng nội dung ảnh.
- Chưa triển khai hoàn chỉnh các chiều truy xuất ngược: Đề tài mới tập trung vào text-to-image retrieval; khả năng image-to-text retrieval hoặc các ứng dụng sinh mô tả vẫn chưa được khai thác đầy đủ.
- Chưa thử nghiệm trên dữ liệu tiếng Việt hoặc đa ngôn ngữ: Mặc dù mô hình có tiềm năng áp dụng đa ngôn ngữ, đề tài chưa tiến hành kiểm thử hiệu quả

trên các tập dữ liệu bản địa như caption tiếng Việt, điều này hạn chế khả năng triển khai vào thực tiễn trong bối cảnh Việt Nam.

4. Hướng nghiên cứu tiếp theo

- Mở rộng sang truy xuất hai chiều (bi-directional retrieval): Phát triển mô hình hỗ trợ image-to-text retrieval giúp hệ thống đa dụng hơn và phục vụ rộng rãi cho nhiều bài toán, bao gồm tìm kiếm, phân loại, đề xuất nội dung và hỗ trợ người dùng.
- Nghiên cứu các phương pháp học không giám sát hoặc bán giám sát: Mục tiêu là tận dụng khối lượng dữ liệu không nhãn lớn trên Internet để giảm chi phí chú thích, đồng thời tăng khả năng tổng quát hóa.
- Xây dựng kiến trúc mã hóa băm động (dynamic hashing): Cho phép mô hình tự điều chỉnh chiều dài mã băm dựa trên độ phức tạp của dữ liệu hoặc yêu cầu ứng dụng, tối ưu hóa dung lượng lưu trữ và tốc độ truy vấn.
- Mở rộng hỗ trợ đa ngôn ngữ, đặc biệt là tiếng Việt: Tích hợp bộ mã hóa văn bản có khả năng xử lý tiếng Việt hoặc các ngôn ngữ khác, từ đó phát triển ứng dụng tra cứu ảnh dành cho người dùng bản địa.
- Triển khai thử nghiệm trên hệ thống thực tế quy mô lớn: Tập trung xây dựng giao diện người dùng, API phục vụ truy vấn trực tiếp và thử nghiệm trong các lĩnh vực như thương mại điện tử, giáo dục, hoặc công cụ tìm kiếm hình ảnh theo mô tả.

CÔNG BỐ KHOA HỌC

- Trương Thị Thu Hà, Nguyễn Văn Tiến (2025). "Tra cứu ảnh đa phương thức sử dụng mô hình tự chú ý và mã hóa băm sâu", Tạp chí Khoa học Công nghệ Thông tin và Truyền Thông (JSTIC), đang phản biện (under review).

DANH MỤC CÁC TÀI LIỆU THAM KHẢO

Tài liệu tiếng Anh

[1] A. Radford, J. Wook Kim, C. Hallacy et al., "Learning Transferable Visual Models From Natural Language Supervision," in Proc. ICML, 2021.

[2] C. Jia, Y. Yang, Y. Chen et al., "Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision," in Proc. ICML, 2021.

[3] J. Wang, W. Liu, S. Kumar and S. Chang, "Learning to Hash for Indexing Big Data—A Survey," IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 4, pp. 769–790, 2018.

[4] W. Li, S. Wang and W. Kang, "Feature Learning Based Deep Supervised Hashing With Pairwise Labels," in Proc. IJCAI, 2016.

[5] H. Liu, R. Ji, Y. Yang and X. Li, "Deep Supervised Hashing for Fast Image Retrieval," in Proc. CVPR, 2016.

[6] A. Dosovitskiy, L. Beyer, A. Kolesnikov et al., "An Image Is Worth 16×16 Words: Transformers for Image Recognition at Scale," in Proc. ICLR, 2021.

[7] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," in Proc. CVPR, 2016.

[8] H. Wu, B. Peng, Z. Li et al., "CvT: Introducing Convolutions to Vision Transformers," in Proc. ICCV, 2021.

[9] Z. Dai, H. Liu, Q. Le et al., "CoAtNet: Marrying Convolution and Attention for All Data Sizes," in Proc. NeurIPS, 2021.

[10] S. Zhai, M. Tambe, A. Bulat et al., "SigLIP: Simple Language-Image Pre-training with Sigmoid Loss," arXiv preprint arXiv:2303.15343, 2023.

[11] C. Li, C. Deng, N. Li, W. Liu, X. Gao, and D. Tao, "Self-Supervised Adversarial Hashing Networks for Cross-Modal Retrieval," in Proc. CVPR, 2018.

[12] S. Su, Z. Zhong, and C. Zhang, "Deep Joint-Semantics Reconstructing Hashing for Large-Scale Unsupervised Cross-Modal Retrieval," in Proc. ICCV, 2019.

[13] Q.-Y. Jiang and W.-J. Li, "Deep Cross-Modal Hashing," in Proc. CVPR, 2017.

[14] Z. Cao, M. Long, J. Wang, and P. S. Yu, "HashNet: Deep Learning to Hash by Continuation," in Proc. ICCV, 2017.

[15] H. Zhu, M. Long, J. Wang, and Y. Cao, "Deep Hashing Network for Efficient Similarity Retrieval," in Proc. AAAI, 2016.

[16] W.-J. Li, S. Wang, and W.-C. Kang, "Feature Learning Based Deep Supervised Hashing With Pairwise Labels," in Proc. IJCAI, 2016.

[17] A. Gionis, P. Indyk, and R. Motwani, "Similarity Search in High Dimensions via Hashing," in Proc. VLDB, 1999.

[18] Chao He, Hongxi Wei (2024), "HybridHash: Hybrid Convolutional and Self-Attention Deep Hashing for Image Retrieval", ICMR '24: Proceedings of the 2024 International Conference on Multimedia Retrieval, pp. 824 - 832.

[19] Sonam Goenka, Zhaoheng Zheng, Ayush Jaiswal, Rakesh Chada, Yue Wu, Varsha Hedau, Pradeep Natarajan (2022), "FashionVLP: Vision Language Transformer for Fashion Retrieval with Feedback", IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Tài liệu Internet

[20] https://www.geeksforgeeks.org/computer-vision/mean-average-precision-map-in-computer-vision/. Truy cập ngày 01 tháng 12 năm 2025.

BẢN CAM ĐOAN

Tôi cam đoan đã thực hiện việc kiểm tra mức độ tương đồng nội dung đề án/luận văn/luận án qua phần mềm Turnitin một cách trung thực và đạt kết quả mức độ tương đồng 17% toàn bộ nội dung đề án/luận văn/luận án. Bản đề án/luận văn/luận án kiểm tra qua phần mềm là bản cứng đề án/luận văn/luận án đã nộp để bảo vệ trước hội đồng. Nếu sai tôi xin chịu các hình thức kỷ luật theo quy định hiện hành của Học viện.

Hà Nội, ngày 10 tháng 12 năm 2025

HỌC VIÊN CAO HỌC/NCS

(Ký và ghi rõ họ tên)



Trương Thị Thu Hà

17% Tính tương đồng nói chung

Tổng cộng của tất cả các kết quả trùng khớp, bao gồm cả các nguồn trùng lặp, cho mỗi c...

Đã lọc khỏi Báo cáo

- Mục lục tham khảo
- Văn bản được trích dẫn
- Kết quả trùng khớp nhỏ (ít hơn 8 từ)

Nguồn hàng đầu

13% Nguồn Internet
8% Ấn bản
6% Bài tập được nộp (bài của học sinh)

HỌC VIÊN CAO HỌC/NCS	GIẢNG VIÊN HƯỚNG DẪN
(Ký và ghi rõ họ tên)	*(Ký và ghi rõ họ tên)*
Trương Thị Thu Hà	Đào Thị Thúy Quỳnh

BỘ KHOA HỌC VÀ CÔNG NGHỆ
HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

CỘNG HOÀ XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc

BÁO CÁO GIẢI TRÌNH
SỬA CHỮA, HOÀN THIỆN ĐỀ ÁN TỐT NGHIỆP THẠC SĨ

Họ và tên học viên: Trương Thị Thu Hà

Chuyên ngành: Khoa học máy tính

Khóa: 2024 đợt I

Tên đề tài: Nghiên cứu phương pháp tra cứu ảnh đa phương thức sử dụng mô hình tự chú ý kết hợp mã hóa băm sâu

Người hướng dẫn khoa học: TS. Đào Thị Thúy Quỳnh

Ngày bảo vệ: 22/01/2026

Các nội dung học viên đã sửa chữa, bổ sung trong đề án tốt nghiệp theo ý kiến đóng góp của Hội đồng chấm đề án tốt nghiệp:

TT	Ý kiến hội đồng	Sửa chữa của học viên
1	Rà soát, làm rõ và cập nhật mô hình và các thành phần.	Đã bổ sung phân tích mô hình SADH và các thành phần ở chương 2, mục 2.4. Kiến trúc tổng thể của mô hình đề xuất, từ trang 38 đến trang 40.
2	Rà soát, làm rõ và cập nhật tiêu chí, phương pháp và kết quả đánh giá.	Đã giải thích rõ việc sử dụng chỉ số mAP để đánh giá mô hình ở chương 3, mục 3.4. Thử nghiệm và đánh giá, trang 58, 59.
3	Rà soát, làm rõ và cập nhật việc sử dụng Hashing và kích thước.	Đã trình bày chi tiết về việc sử dụng Hashing và kích thước mã băm ở chương 2, mục 2.4.2. Mô đun mã hóa băm, trang 44, 45.
4	Rà soát về lỗi trình bày, lỗi chính tả	Đã chỉnh sửa một số lỗi chính tả, bỏ khoảng trống thừa.

Hà Nội, ngày 03 tháng 02 năm 2026

Ký xác nhận của

CHỦ TỊCH HỘI ĐỒNG CHẤM ĐỀ ÁN	THƯ KÝ HỘI ĐỒNG	NGƯỜI HƯỚNG DẪN KHOA HỌC	HỌC VIÊN
GS.TS. Từ Minh Phương	TS. Đào Ngọc Phong	TS. Đào Thị Thúy Quỳnh	Trương Thị Thu Hà

BỘ KHOA HỌC VÀ CÔNG NGHỆ
HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

CỘNG HOÀ XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc

BIÊN BẢN
HỌP HỘI ĐỒNG CHẤM ĐỀ ÁN TỐT NGHIỆP THẠC SĨ

Căn cứ quyết định số Quyết định số 3299/QĐ-HV ngày 26 tháng 12 năm 2025 của Giám đốc Học viện Công nghệ Bưu chính Viễn thông về việc thành lập Hội đồng chấm đề án tốt nghiệp thạc sĩ. Hội đồng đã họp vào hồi ..9.... giờ .30. phút, ngày 22 tháng 01 năm 2025 tại Học viện Công nghệ Bưu chính Viễn thông để chấm đề án tốt nghiệp thạc sĩ cho:

Học viên: **Trương Thị Thu Hà**

Tên đề án tốt nghiệp: **Nghiên cứu phương pháp tra cứu ảnh đa phương thức sử dụng mô hình tự chú ý kết hợp mã hóa băm sâu**

Ngành: **Khoa học máy tính** Mã số: **8.48.01.01**

Các thành viên của Hội đồng chấm đề án tốt nghiệp có mặt: .05../ 05

TT	HỌ VÀ TÊN	TRÁCH NHIỆM TRONG HĐ	GHI CHÚ
1	**GS.TS. Từ Minh Phương**	Chủ tịch	
2	**TS. Đào Ngọc Phong**	Thư ký	
3	**TS. Đỗ Thanh Hà**	Phản biện 1	
4	**TS. Nguyễn Thanh Tuyên**	Phản biện 2	
5	**PGS.TS. Nguyễn Linh Giang**	Uỷ viên	

Các nội dung thực hiện:

1. Chủ tịch Hội đồng điều khiển buổi họp. Công bố quyết định của Giám đốc Học viện Công nghệ Bưu chính Viễn thông về việc thành lập Hội đồng chấm đề án tốt nghiệp thạc sĩ.
2. Người hướng dẫn khoa học hoặc thư ký đọc lý lịch khoa học và các điều kiện bảo vệ đề án tốt nghiệp của học viên (Có bản lý lịch khoa học và kết quả các môn học cao học của học viên kèm theo).
3. Học viên trình bày tóm tắt đề án tốt nghiệp.
4. Phản biện 1 đọc nhận xét (có văn bản kèm theo)
5. Phản biện 2 đọc nhận xét (có văn bản kèm theo)
6. Các câu hỏi của thành viên Hội đồng:

1/ Tại sao sử dụng self attention để kết nối các mô hình?
2/ Làm sao để dự đoán được trường của hình ảnh và văn bản, có thể trường khác nhau?
3/ Việc lựa chọn kích thước hàm băm?
4/ Có khả năng thích ứng khi dữ liệu thay đổi và khả năng dự liệu lớn?
5/ Số dung hàm kích hoạt để xấp xỉ như thế nào?

7. Trả lời của học viên:

1/ Do sự khác nhau về đặc trưng của hình ảnh và văn bản
2/ Sử dụng self Attention để giải quyết vấn đề do sự khác biệt vector đầu ra
3/ Chạy kịch thaza 64 bit thì thực nghiệm
4/ Có khả năng thích ứng về mô hình
5/ Sử dụng để ánh xạ vector đặc trưng

8. Thư ký đọc nhận xét về quá trình thực hiện đề án tốt nghiệp của học viên *(có văn bản kèm theo).*

9. Hội đồng họp riêng:

- Bầu Ban kiểm phiếu:
  1. Trưởng Ban kiểm phiếu: Đào Ngọc Phong
  2. Uỷ viên Ban kiểm phiếu: Nguyễn Thanh Tuyên
  3. Uỷ viên Ban kiểm phiếu: Đỗ Thanh Hà
- Hội đồng chấm đề án tốt nghiệp bằng bỏ phiếu kín.
- Ban kiểm phiếu làm việc:
- Trưởng Ban kiểm phiếu báo cáo kết quả kiểm phiếu *(có Biên bản họp Ban kiểm phiếu kèm theo)*
- **Điểm trung bình của đề án tốt nghiệp:** 9.0

Kết luận:

1. Các nội dung cần chỉnh sửa, hoàn thiện sau bảo vệ đề án tốt nghiệp:

1/ Rà soát, làm rõ và cập nhật về
- MS hình vẽ các thành phần
- Tiền chú, phương pháp và kết quả đánh giá
- Việc sử dụng Hashing và kịch thaza
2/ Tiếp thu đầy đủ ý kiến Hội đồng và 2 Thầy/Cô phản biện
3/ Rà soát lại trình bày, lỗi chính tả

2. Đề nghị Học viện công nhận (hoặc không) và cấp bằng (hoặc không) thạc sĩ cho học viên: Công nhận và cấp bằng

3. Đề án tốt nghiệp có thể phát triển thành đề tài nghiên cứu cho NCS: không

Buổi làm việc kết thúc vào 10h15 cùng ngày.

Chủ tịch	**Thư ký**
GS.TS. Từ Minh Phương	**TS. Đào Ngọc Phong**

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM

Độc lập – Tự do – Hạnh phúc

BẢN NHẬN XÉT ĐỀ ÁN TỐT NGHIỆP THẠC SĨ

(Dùng cho người phản biện)

Tên đề tài đề án tốt nghiệp: Nghiên cứu phương pháp tra cứu ảnh đa phương thức sử dụng mô hình tự chú ý kết hợp mã hóa băm sâu.

Chuyên ngành: Khoa học máy tính.

Mã chuyên ngành: 8.48.01.01.....

Họ và tên học viên: Trương Thị Thu Hà.....

Họ và tên người nhận xét: Đỗ Thanh Hà .....

Học hàm, học vị: Tiến sĩ .....

Chuyên ngành: Tin học .....

Cơ quan công tác: Khoa Trí tuệ nhân tạo, Học viện công nghệ bưu chính viễn thông .....

Số điện thoại: 0913378868 .....E-mail: dothanhha@ptit.edu.vn

NỘI DUNG NHẬN XÉT

I/ Cơ sở khoa học và thực tiễn, tính cấp thiết của đề tài:

Trong thời đại Big Data, việc tìm kiếm hình ảnh bằng ngôn ngữ tự nhiên trên quy mô hàng tỷ ảnh là một thách thức lớn về cả độ chính xác lẫn tốc độ. Đề tài đã xác định đúng "điểm nghẽn" của các mô hình hiện đại như CLIP: embedding số thực quá nặng và chậm. Việc nghiên cứu mã hóa băm sâu (Deep Hashing) để đưa về mã nhị phân là hướng đi rất thiết thực và có tính thời sự cao.

Kết quả nghiên cứu có khả năng ứng dụng trực tiếp vào các hệ thống tìm kiếm thương mại điện tử, quản lý kho ảnh số, hoặc các công cụ tìm kiếm thông minh yêu cầu phản hồi thời gian thực.

II/ Nội dung của đề án tốt nghiệp, các kết quả đã đạt được:

Về nội dung:

Chương 1: Đề án trình bày tốt các cập nhật về các mô hình Vision-Language (VLP) như CLIP, SigLIP và các kỹ thuật Transformer. Việc phân tích rõ khoảng cách giữa ảnh và văn bản cho thấy hiểu về bản chất vấn đề.

Chương 2: Trình bày Mô hình đề xuất SADH

- Kiến trúc lai (Hybrid Visual Encoder): Kết hợp ResNet-50 và ViT thông qua *Attention Fusion* giúp khắc phục nhược điểm mất chi tiết của ViT thuần túy.
- Việc bổ sung MLP 3 tầng để điều chỉnh phân phối embedding văn bản giúp thu hẹp khoảng cách với nhánh ảnh, tạo điều kiện tối ưu cho bước băm nhị phân.
- Việc kết hợp đồng thời *Contrastive Loss, Quantization Loss, Bit-balance Loss* và đặc biệt là *Alignment Loss* thể hiện quy trình tối ưu hóa rất bài bản và chặt chẽ.
- Phương pháp huấn luyện: Chiến lược huấn luyện 3 giai đoạn (Warm-up, Fine-tuning, Refinement) là cách tiếp cận phù hợp để huấn luyện các mô hình lớn mà không làm mất đi tri thức đã học (catastrophic forgetting).

Về kết quả đạt được

- Kết quả mAP trên MS COCO (~80.23%) và Flickr30k (~85.47%) là tốt, vượt xa các phương pháp hashing truyền thống (SSAH, HashNet) và các cách kết hợp CLIP thông thường.
- Đề án chứng minh được ưu thế của khoảng cách Hamming, cho phép truy xuất gần như tức thời trên không gian mã băm, giải quyết triệt để bài toán hiệu năng.

Hạn chế

- Mô hình kết hợp cả ViT, CNN và Hashing Module đòi hỏi năng lực tính toán rất lớn để huấn luyện, điều này có thể gây khó khăn cho việc triển khai ở các đơn vị có hạ tầng hạn chế.
- Đồ án mới chỉ thực hiện trên tiếng Anh (từ bộ CLIP gốc). Việc chưa thử nghiệm trên tiếng Việt là một hạn chế nhỏ về tính bản địa hóa.
- Đồ án chưa trình bày chi tiết sự đánh đổi (trade-off) giữa độ dài mã băm (64, 128, 256 bit) đối với độ chính xác mAP.

III/ Những vấn đề cần giải thích thêm:

1. Về mô đun Interaction: Tại sao bạn chọn ghép nối đặc trưng CNN và ViT bằng MHSA thay vì các phương pháp đơn giản hơn như cộng (addition) hoặc nối (concatenation) vector? MHSA mang lại lợi ích cụ thể nào trong việc lọc nhiễu giữa hai nguồn đặc trưng này

2. Về Alignment Loss: Giải thích rõ hơn làm thế nào việc ép trung bình và phương sai của hai phân phối gần nhau lại giúp cải thiện chất lượng mã băm nhị phân cuối cùng?

3. Về tính ổn định của Bit: Trong thực tế, làm thế nào để đảm bảo mã băm sinh ra không bị thay đổi quá lớn khi hình ảnh đầu vào chỉ bị nhiễu nhẹ (ví dụ: thay đổi độ sáng hoặc xoay ảnh nhẹ)?

4. Về khả năng triển khai: Với mô hình SADH, nếu áp dụng cho một kho dữ liệu 100 triệu ảnh, quy trình "indexing" (đánh chỉ mục) nên như thế nào để tối ưu hóa bộ nhớ RAM của máy chủ?

IV/ Kết luận:

Đồ án của học viên Trương Thị Thu Hà là một công trình nghiên cứu nghiêm túc, có tính sáng tạo. Đồng ý cho phép học viên bảo vệ đề án tốt nghiệp.

Ngày.....tháng.1.....năm 2026

NGƯỜI NHẬN XÉT

Đỗ Thanh Hà

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM

Độc lập – Tự do – Hạnh phúc

BẢN NHẬN XÉT ĐỀ ÁN TỐT NGHIỆP THẠC SĨ

(Dùng cho người phản biện)

- **Tên đề án tốt nghiệp:** Nghiên cứu phương pháp tra cứu ảnh đa phương thức sử dụng mô hình tự chú ý kết hợp mã hóa băm sâu.
- **Chuyên ngành:** Khoa học máy tính.
- **Mã chuyên ngành:** 8.48.01.01.
- **Họ và tên học viên:** Trương Thị Thu Hà.
- **Người hướng dẫn khoa học:** TS. Đào Thị Thúy Quỳnh.
- **Người nhận xét phản biện:** TS. Nguyễn Thanh Tuyên

NỘI DUNG NHẬN XÉT

I. Cơ sở khoa học và thực tiễn, tính cấp thiết của đề tài

- **Tính cấp thiết:** Trong kỷ nguyên dữ liệu lớn, việc tìm kiếm thông tin giữa các miền dữ liệu khác nhau (như văn bản và hình ảnh) trở thành một nhu cầu thiết yếu. Thử thách lớn nhất là thu hẹp khoảng cách ngữ nghĩa giữa các loại dữ liệu không đồng nhất. Đề tài tập trung vào mô hình tự chú ý kết hợp mã hóa băm sâu (SADH – Self-Attention and Deep Hashing model) là hướng đi hiện đại, giúp giải quyết cả vấn đề độ chính xác lẫn tốc độ tra cứu.
- **Cơ sở khoa học:** Đề tài áp dụng các kỹ thuật học sâu tiên tiến (Deep Learning) như Transformer và Deep Hashing để tối ưu hóa việc biểu diễn dữ liệu đa phương thức.

II. Nhận xét chi tiết nội dung đề án theo các chương

- **Chương 1 - Tổng quan về tra cứu đa phương thức:** Học viên đã trình bày rõ ràng về bài toán tra cứu ảnh, các đặc trưng từ vựng và cấu trúc dữ liệu liên quan. Phần tổng quan cho thấy sự hiểu biết về các mô hình hiện có và nhu cầu tự động hóa trong phân tích dữ liệu.
- **Chương 2 - Mô hình tự chú ý và mã hóa băm sâu:** Chương này khái quát hóa tốt sự khác biệt giữa học máy truyền thống và học sâu. Việc trình bày cơ chế tự chú ý để học các biểu diễn phi tuyến tính là một điểm cộng lớn, giúp bắt được nhiều loại thông tin ngữ nghĩa phức tạp.
- **Chương 3 - Thực nghiệm và đánh giá:** Học viên đã mô tả chi tiết môi trường thử nghiệm và quá trình thu thập dữ liệu. Việc sử dụng các độ đo

như Accuracy, Precision, Recall và F1-score để đánh giá mô hình là hoàn toàn phù hợp và đúng tiêu chuẩn khoa học.

III. Những vấn đề cần giải thích thêm hoặc chỉnh sửa

- **Lý giải về hàm mất mát (Loss Function):** Đề nghị học viên làm rõ cách thiết kế hàm mất mát để đảm bảo tính bảo toàn cấu trúc tương đồng giữa các miền dữ liệu khác nhau (Cross-modal contrastive loss và cách kết hợp giữa thành phần tự chú ý và thành phần mã hóa băm trong hàm mục tiêu để tối ưu hóa quá trình huấn luyện.
- **Độ dài mã băm (Hash bit length):** Tại sao học viên lại chọn các độ dài mã băm cụ thể (ví dụ: 16, 32, 64 bits)? Sự đánh đổi giữa độ chính xác (mAP) và tốc độ/dung lượng lưu trữ khi thay đổi độ dài mã băm như thế nào?
- **Tài liệu tham khảo:** Cần rà soát lại định dạng trích dẫn theo đúng quy định của Học viện (Quyết định số 550/QĐ-HV). Một số nguồn tài liệu Internet cần được ghi rõ ngày truy cập và bổ sung vào danh mục chính thức.
- **Hình vẽ:** Một số hình vẽ kiến trúc mô hình trong Chương 3 cần được chú thích rõ hơn các thành phần (layers) để người đọc dễ theo dõi luồng dữ liệu.
- **Thuật ngữ và viết tắt:** Đề nghị rà soát và bổ sung đầy đủ danh mục các thuật ngữ viết tắt trong báo cáo (như CNN, ViT, mAP) để đảm bảo tính tường minh.
- **Văn phong học thuật:** Một số đoạn văn cần tránh sử dụng đại từ nhân xưng không phù hợp với văn phong học thuật và sửa các lỗi chính tả nhỏ trong báo cáo.
- **Tài liệu tham khảo:** Cần trình bày lại danh mục tài liệu tham khảo theo đúng quy định của Học viện và bổ sung đầy đủ các nguồn trích dẫn cho hình vẽ, bảng biểu.

IV. Câu hỏi dành cho học viên trong buổi bảo vệ

1. Cơ chế "Self-attention" đóng vai trò cụ thể như thế nào trong việc xử lý sự không đồng nhất giữa dữ liệu hình ảnh và văn bản?
2. Lý do bạn chọn độ dài mã băm cụ thể trong các thực nghiệm là gì? Độ dài này ảnh hưởng như thế nào đến sự cân bằng giữa Precision và Recall?
3. Làm thế nào để mô hình của bạn có thể thích ứng với các mối đe dọa hoặc sự thay đổi của dữ liệu mới trong tương lai?
4. Trong quá trình huấn luyện, làm thế nào bạn giải quyết vấn đề đạo hàm của hàm băm (vốn là hàm rời rạc không có đạo hàm trực tiếp)? Bạn đã sử dụng hàm kích hoạt nào để xấp xỉ?

5. Mô hình đề xuất có khả năng mở rộng (scalability) như thế nào khi áp dụng vào các hệ thống tra cứu thực tế với hàng triệu hình ảnh?

V. Kết luận

Đề án của học viên Trương Thị Thu Hà có cấu trúc chặt chẽ, áp dụng các kỹ thuật học sâu hiện đại để giải quyết bài toán thực tiễn quan trọng. Các kết quả đạt được có giá trị khoa học và độ tin cậy.

Đồng ý cho phép học viên bảo vệ đề án tốt nghiệp trước Hội đồng.

Hà Nội, ngày 22 tháng 01 năm 202...

NGƯỜI NHẬN XÉT

TS. Nguyễn Thanh Tuyên