HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Trương Thị Thu Hà

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

Chuyên ngành : Khoa học máy tính

Mã số: 8.48.01.01

TÓM TẮT ĐỀ ÁN TỐT NGHIỆP THẠC SĨ

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 ngàuh : Khoa học máy tính

Mã số: 8.48.01.01

TÓM TẮT ĐỀ ÁN TỐT NGHIỆP THẠC SĨ

HÀ NỘI – NĂM 2025

Đề án tốt nghiệp được hoàn thành tại:

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

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

(Ghi rõ học hàm, học vị)

Phản biện 1: .....TS. Đỗ Thanh Hà.....

Phản biện 2: .....TS. Nguyễn Thanh Tuyên.....

Đề án tốt nghiệp sẽ được bảo vệ trước Hội đồng chấm đề án tốt nghiệp thạc sĩ tại Học viện Công nghệ Bưu chính Viễn thông

Vào lúc: 09. giờ 30. ngày 22. tháng 01. .. năm 2026..

Có thể tìm hiểu đề án tốt nghiệp tại:

- Thư viện của Học viện Công nghệ Bưu chính Viễn thông

MỞ ĐẦU

1. Lý do chọn đề tài

Tra cứu ảnh đa phương thức đóng vai trò quan trọng trong nhiều ứng dụng như tìm kiếm hình ảnh theo văn bản, quản lý nội dung đa phương tiện và hệ thống đề xuất. Mô hình tiền huấn luyện thị giác–ngôn ngữ như CLIP đã tạo bước đột phá bằng cách học biểu diễn chung giữa ảnh và văn bản trên tập dữ liệu lớn, đạt hiệu quả cao trong nhiều tác vụ như zero-shot classification và image-text retrieval. Tuy nhiên, CLIP gặp hạn chế khi áp dụng trong hệ thống quy mô lớn do embedding kích thước lớn gây tốn kém bộ nhớ và tính toán, dẫn đến tốc độ truy vấn chậm. Deep hashing là giải pháp tiềm năng giúp mã hóa embedding thành mã nhị phân, tăng tốc truy vấn nhưng thường chỉ áp dụng cho dữ liệu đơn phương thức và cần nhãn mạnh, chưa tận dụng dược mô hình VLP hiện đại.

Xuất phát từ thực tế đó, đề tài "Nghiên cứu phương pháp tra cứu ảnh đa phương thức sứ đụng mô hình tự chú ý kết hợp mã hóa băm sâu" (SADH) được đề xuất, nhằm kết hợp ưu điểm của mô hình tự chú ý (Transformer, CNN, CLIP, ResNet) và deep hashing để vừa khai thác biểu diễn ngữ nghĩa mạnh mẽ vừa đảm bảo tốc độ truy vấn và hiệu quả lưu trữ. Nghiên cứu này có ý nghĩa quan trọng trong trí tuệ nhân tạo và thị giác máy tính, đồng thời có tiềm năng ứng dụng thực tiễn rộng rãi.

Nội dung đề án được tổ chức thành ba chương chính như sau:

Chương 1: Giới thiệu về tra cứu ảnh

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

Chương 3: Xây dựng úng dụng

2. Tổng quan về vấn đề nghiêu cứu

Tra cứu ảnh đa phương thức là lĩnh vực nghiên cứu quan trọng trong trí tuệ nhân tạo, thị giác máy tính và xử lý ngôn ngữ, tập trung xây dựng hệ thống cho phép truy vấn hình ảnh qua văn bàn. Thách thức chính là sự khác biệt về ngữ nghĩa và cấu trúc giữa hình ành và văn bản, nên các phương pháp hiện đại tập trung học biểu điễn chung cho hai miền dữ liệu này.

Kiến trúc Transformer và các mô hình như CLIP, ViLT, BLIP đã chứng minh hiệu quả vượt trội trong việc học biểu diễn đa phương thức từ bộ dữ liệu lớn. Tuy nhiên, khi áp dụng cho hệ thống quy mô lớn, vấn đề hiệu suất và khả năng mở rộng vẫn còn là rào cản. Kỹ thuật mã hóa băm sâu dược quan tâm vì giúp biểu diễn dữ liệu bằng mã nhị phân gọn nhẹ, giảm chi phí lưu trữ và tăng tốc truy vấn. Việc kết hợp mô hình tự chú ý với mã hóa băm sâu trong tra cứu ảnh đa phương thức là hướng nghiên cứu mới, nhằm tận dụng sức mạnh biểu diễn của Transformer và hiệu quả truy xuất của băm sâu.

Thách thức lớn là làm sao huấn luyện hệ thống tối ưu hóa đồng thời biểu diễn ngữ nghĩa liên phương thức và sinh ra mã bãm hiệu quả. Nghiên cứu này có ý nghĩa lý thuyết và thực tiễn quan trọng, đặc biệt trong bối cảnh các hệ thống tìm kiếm thông minh ngày càng phổ biển vả phát triển.

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). 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: Các phương pháp tra cứu ảnh đa phương thức.

Phạm vi nghiên cứu: Các phương pháp tra cứu ảnh với dữ liệu đầu vào bao gồm hình ảnh và văn bản mô tả sử dụng các kiến trúc tự chú ý kết hợp CNN và mã hóa băm sâu.

5. Phương pháp nghiên cứu

Đề tài sử dụng kết hợp các phương pháp: Phân tích tài liệu, khảo sát và đánh giá các mô hình hiện có, thiết kế và xây dựng mô hinh đề xuất, phát triển nguyên mẫu hệ thống, thử nghiệm và đánh giá hiệu quả mô hình.

NỘI DUNG

CHƯƠNG 1: GIỚI THIỆU VỀ TRA CỨU ẢNH

1.1. Tổng quan về tra cứu àuh:

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. Tra cứu ảnh không chỉ mang ý nghĩa về mặt kỹ thuật mà còn có ý nghĩa thực tiễn sâu rộng, được ứng dụng đa dạng.

1.1.2. Phân loại tra cứu ánh

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 ánh đa phương thức, ..., cụ thể:

Tra cứu ảnh dựa trên nội dung – CBIR: Người đùng cung cấp một hình ánh mẫu, và hệ thống tìm ành tương tự dựa trên các thước đo trong không gian đặc trưng

Tra cứu ảnh dựa trên văn bản – TBIR: Người dùng mô tả hình ảnh cần tìm bằng ngôn ngữ tự nhiên.

Tra cứu ảnh đa phương thức: Người dùng có thể sử dụng nhiều loại dữ liệu để tra cứu ảnh, phổ biến là hình ảnh và văn bản mô tả.

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

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

Tra cứu ảnh đa phương thức là hệ thống xử lý đồng thời nhiều dạng dữ liệu khác nhau, phổ biến nhất là hình ảnh và văn bản. Đặc điểm nổi bật của tra cứu ảnh đa phương thức là tính linh hoạt trong tương tác với người dùng, khả năng tổng hợp và khái quát hóa ngữ nghĩa và có tính chất mở rộng mạnh mẽ.

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

Dữ liệu hình ảnh: 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 cánh, được tổ chức theo cấu trúc không gian, có nhiều nhiễu.

Dữ liệu văn bản: Chuỗi token rời rạc (từ, cụm từ) được tổ chức theo cú pháp ngôn ngữ tự nhiên, thường chỉ nêu ra một phần nội dung ảnh thay vì tái tạo toàn bộ. Vvăn bản có cấu trúc hình thức ổn định hơn, nhưng lại chịu ảnh hưởng mạnh của tính chủ quan và phong cách diễn đạt.

Sự dị thể (heterogeneity) giữa hai loại dữ liệu: Dạng biểu diễn (liên tục và rời rạc), phân phối thống kê, mức độ trừu tượng và loại thông tìn mang theo.

Vai trò của embedding chung trong việc kết nối hai phương thức: 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

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

Các thách thức trong việc kết nối hai phương thức :

- Khoảng cách biểu giữa hai loại dữ liệu.
- Sự không đối xứng về thông tin giữa hình ánh và văn bản.
- Sự lệch mức trừu giữa hai loại dữ liệu
- Dữ liệu thực tế thường đi kèm với nhiễu và sự thiếu nhất.

1.3. Các nghiên cứu liên quan

Các mô hình thị giác–ngôn ngữ tiền huấn luyện: CLIP, ALIGN, SigLIP, ViLT, BLIP.

Các mô hình hợp nhất đặc trưng đa phương thức: FashionVLP, ViLT, BLIP.

Các mô hình trích xuất đặc trưng hình ảnh kết hợp CNN và ViT (Hybrid Visual Encoders): ViT, CvT, CoAtNet.

Các phương pháp hashing: LSH, ITQ, DSH, DPSH, HashNet, DHC.

1.4. Kết luận chương 1

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.

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

2.1. Phâu tích yêu cầu và định hướng mô hình

2.1.1. Yêu cầu về hiệu năng và độ chính xác

Độ chính xác phản ảnh khả năng mô hình hiếu 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. Bên cạnh độ chính xác, hệ thống cũng cấn 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. Tuy nhiên, tốc độ chỉ thật sự có ý nghĩa khi mô hình vẫn duy trì được độ chính xác ngữ nghĩa. Diều này đặt ra yêu cầu mô hình phải 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.

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

Đảm bảo khả năng truy vấn nhanh, chi phí lưu trữ thấp và năng lực xử lý ổn định với kho dữ liệu lớn.

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 tải nguyên như thiết bị biên hoặc dịch vụ thời gian thực.

Khà năng tận dụng các mô hình pre-train 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.

Khả năng tương thích với dữ liệu không đồng nhất: chất lượng, ánh sáng, góc chụp và phong cách khác nhau đối với ảnh; độ chi tiết và phong cách diễn đạt đối với văn bán.

Dễ dàng tích hợp, cập nhật và mở rộng chỉ mục.

2.2. Bộ mã hóa hình ảnh

2.2.1. Khối Transformer (Transformer Block)

Tình chỉnh mô hình: CLIP ViT-B/16:

- Chia patch và ánh xạ tuyến tính (Patch Embedding)
- Thêm thông tin vị trí (Positional Embedding)
- Cơ chế tự chú ý đa đầu (Multi-Head Self-Attention)
- Mạng lan truyền tiến (Feed-Forward Network – FFN)
- Chuẩn hóa và kết nối tắt (LayerNorm & Residual Connection)

2.2.2. Mô đun tương tác (Interaction Module)

Cấu trúc:

- Nhánh CNN – sử dụng Resnet-50 tăng cường đặc trưng cục bộ.
- Kết hợp đặc trưng từ khối Transformer và nhánh CNN bằng multi-head self-attention.

Tính chất nổi bật:

- Khả năng tự động học trọng số kết hợp thông qua cơ chế tự chú ý đa.
- Khả năng giảm sự phân kỳ của hai dạng đặc trưng.
- Khả năng mô hình hóa quan hệ phi tuyến và đa chiều giữa hai nguồn đặc trưng. Cho phép thích nghi linh hoạt với nhiều loại ảnh và điều kiện thu nhận.

2.3. Bộ mã hóa văn bản

2.3.1. Mã hóa văn bản bằng CLIP

Mô hinh đề xuất kế thừa nguyên vẹn từ text encoder của CLIP. Quá trình mã hóa bắt đầu bằng token hóa Byte-Pair Encoding (BPE), cộng positional embedding và đi qua nhiều tầng Transformer với cơ chế Multi-Head Self-

Attention (MHSA) và Feed-Forward Network (FFN), giúp nắm bắt các phụ thuộc dài hạn và quan hệ ngữ nghĩa phức tạp.

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

Yêu cầu quan trọng trong một hệ thống đa phương thức là biểu diễn của ảnh và văn bản phải được ánh xạ vảo cùng một không gian ngữ nghĩa chung để có thể so sánh trực tiếp bằng những độ đo tương đồng đơn giản như khoảng cảch 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. Sự bất đối xứng biểu hiện rõ ràng ở ba khía cạnh: mức độ trừu tượng, phân phối thống kê embedding không tương thích, độ phong phú về ngữ nghĩa không đồng đều.

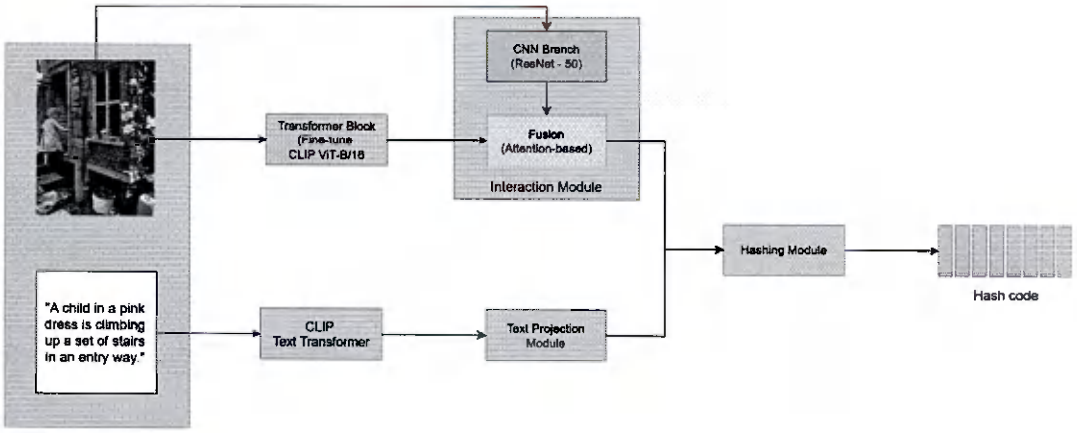
2.3.3. Mô đun ánh xạ văn bản (Text Projection Module)

Để giải quyết triệt để vấn đề bất đối xứng giữa hai dòng đặc trưng, mô hình SADH bổ sung một mô-đun ánh xạ văn bân (Text Projection Module) mới có vai trò điều chỉnh embedding văn bản trước khi đưa vào tầng mã hóa băm. 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), giúp tăng tính biểu đạt mà không phá vỡ cấu trúc ngữ nghĩa vốn có của embedding văn bản.

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. Tác động của mô đun ánh xạ văn bàn không chỉ nằm ở việc biến đổi embedding mà còn góp phần giảm hiện tượng lệch phân phối (distributional shift) giữa hai phương thức.

2.4. Kiến trúc tổng thể của mô hình đề xuất

Các thành phần và pipeline của mô hinh 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

Mô hinh SADH, ánh xạ đầu ra của 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. Trong khi đó, văn bản được mã hóa bằng bộ mã hóa ngôn ngữ của CLIP đã được huấn luyện đa phương thức trên quy mô lớn và Text Projection Module để điều hòa đặc trưng văn bản trước khi đưa vào không gian chung.

Quá trinh 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.

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. Mục tiêu của tầng băm là nén đặc trưng xuống dạng mã nhị phân ngắn gọn và phải bảo toàn cấu trúc ngữ nghĩa dã được xây dựng ở các bước trước đó.

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. Module băm trong SADH có khả năng duy tri quan hệ gần–xa giữa các đặc trưng sau khi lượng tử hóa.

2.4.3. Hàm mất mát

Contrastive loss được xây dựng theo tinh thần của InfoNCE và CLIP. Cho hai tập véc-tơ đặc trưng ảnh $I = \{i_1, \dots, i_B\}$ và văn bàn $T = \{t_1, \dots, t_B\}$, sau khi được chuẩn hóa:

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

(2.16)

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

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

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

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

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

Hàm mất mát tương phản cuối cùng là:

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

Quantization loss được thiết kế nhằm thu hẹp độ chênh lệch giữa biểu diễn liên tục và biểu diễn nhị phân:

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

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 quả trinh huấn luyện giảm thiểu sai lệch khi chuyển sang nhị phân bằng hàm sign:

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

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). Cho batch véc-tơ đặc trưng liên tục:

$$\mu = \frac{1}{B}\sum_{k=1}^{B} h_k. \quad (2.23)$$

Hàm mất mảt cân bằng bit được định nghĩa:

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

Alignment loss là thành phần đặc biệt quan trọng nhằm khắc phục sự bất đối xứng giữa hai dòng xử lý hình ảnh và văn bản. Với $\mu_I = \text{E}[h_I], \mu_T = \text{E}[h_T], \sigma_I^2 = \text{Var}[h_I], \sigma_T^2 = \text{Var}[h_T], \lambda_\sigma = 0.1$, hàm mất mát căn chỉnh:

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

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

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

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

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

Chiến lược huấn luyện nhiều giai đoạn của SADH nhằm đạt được ba mục tiêu lớn:

- Giai đoạn 1: Ổn định hóa không gian đặc trưng chung trước khi can thiệp sâu vào backbone pretrained.
- Giai đoạn 2: Tinh chỉnh selective các tầng quan trọng của CLIP để tăng độ chính xác truy vấn.
- Giai đoạn 3: Tối ưu mã băm để đảm bảo hiệu suất cao trong suy luận và triển khai thực tế.

Cơ chế đóng băng từng phần 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, 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. Cơ chế này 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. Nếu các tầng đầu của CLIP cũng được cập nhật đồng thời, sự dao động này có thể lảm phá vỡ cấu trúc biểu diễn đã ổn định từ tiền huấn luyện, dẫn đến hiện tượng "catastrophic forgetting".

2.5. Kết luận chương 2

Chương 2 trình bày mô hình tra cứu ảnh đa phương thức SADH (Self-Attention and Deep Hashing) được đề xuất, bắt đầu từ bộ mã hóa hình ảnh, sau đó là bộ mã hóa văn bản, cuối cùng là tầng mã hóa băm nhị phân vả chiến lược huấn luyện kết hợp hai miền dữ liệu.

CHƯƠNG 3: XÂY DỰNG ỨNG DỤNG

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

Sử dụng hai bộ dữ liệu chuẩn trong lĩnh vực Vision–Language: MS COCO Captions và Flickr30k:

- **Bộ dữ liệu MS COCO Captions (Torchvision)** chứa hơn 120.000 hình ảnh và mỗi ảnh đi kèm 5 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.

- **Bộ dữ liệu Flickr30k (HuggingFace)** có 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 vả đơn giản hơn.

Để đảm bảo dữ liệu từ hai bộ hoàn toàn tương thích trong quá trình huấn luyện, luận văn 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 ánh vả 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

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 bàn 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 vả kích hoạt Tanh sinh mã băm liên tục, sau đó được lượng tử hóa.

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 hinh thích nghi tốt hơn.
- Refinement: tình 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 dộ 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 distancc 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

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 trong Bảng 3.3. 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.

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.

KẾT LUẬN VÀ KIẾN NGHỊ

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

- Xây dựng mô hình tra cứu ảnh đa phương thức hoàn chỉnh.
- Thiết kế kiến trúc hai nhảnh mã hóa hiệu quả.
- Huấn luyện vả đánh giá mô hình trên dữ liệu chuẩn.
- Khả năng cạnh tranh so với các mô hình hiện tại.

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.
- Trong các ứng dụng thực tế cần truy xuất ảnh nhanh với lượng dữ liệu lớn, 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 xứ lý và lảm sạch dữ liệu văn bản – vì ngôn ngữ tự nhiên có thể gây nhiễu cho hệ thống nếu không được xứ lý kỹ. Ngoài ra, cần đảm bảo tài nguyên tính toán đủ mạnh nếu sử đụng các mô hình nặng.

3. Hạn chế

- Độ phức tạp mô hình tương đối cao.
- Phụ thuộc nhiều vảo chất lượng chú thích văn bản.
- Chưa triển khai hoàn chinh các chiều truy xuất ngược.
- Chưa thử nghiệm trên dữ liệu tiếng Việt hoặc đa ngôn ngữ.

4. Hướng nghiên cứu tiếp theo

- Mở rộng sang truy xuất hai chiều (bi-directional retrieval.
- Xây dựng kiến trúc mã hóa băm động (dynamic hashing).

- Mở rộng hỗ trợ đa ngôn ngữ, đặc biệt là tiếng Việt.
- Triển khai thử nghiệm trên hệ thống thực tế quy mô lớn.