HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG



Nguyễn Văn Khánh

NGHIÊN CỨU PHƯƠNG PHÁP TRUY XUẤT ẢNH DỰA TRÊN NGỮ NGHĨA SỬ DỤNG MÔ HÌNH GRAPHORMER

ĐỀ ÁN TỐT NGHIỆP THẠC SĨ KỸ THUẬT

(Theo định hướng ứng dụng)

HÀ NỘI – NĂM 2026

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG



Nguyễn Văn Khánh

NGHIÊN CỨU PHƯƠNG PHÁP TRUY XUẤT ẢNH DỰA TRÊN NGỮ NGHĨA SỬ DỤNG MÔ HÌNH GRAPHORMER

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 2026

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.

Nguyễn Văn Khánh

LỜI CẢM ƠN

Trước hết, em xin gửi lời cảm ơn chân thành đến TS. Đào Thị Thúy Quỳnh – người đã tận tình hướng dẫn, chỉ dẫn và hỗ trợ em trong suốt quá trình thực hiện đề án. Nhờ sự tận tâm, những góp ý quý giá và kinh nghiệm chuyên môn của cô, em đã hoàn thành đề tài này một cách tốt nhất có thể.

Em cũng xin gửi lời cảm ơn đến các thầy cô trong khoa/ngành đã giảng dạy, chia sẻ kiến thức và tạo điều kiện học tập, nghiên cứu. Cuối cùng, em xin cảm ơn gia đình và các bạn đã luôn động viên, hỗ trợ em trong quá trình thực hiện đề án.

Mục lục

MỞ ĐẦU 1

1. Lý do lựa chọn đề tài 1

2. Tổng quan về vấn đề nghiên cứu 2

3. Mục đích nghiên cứu 3

4. Đối tượng và phạm vi nghiên cứu 4

5. Phương pháp nghiên cứu 6

6. Đóng góp của đề án 7

7. Cấu trúc đề án 8

NỘI DUNG 10

CHƯƠNG I: CƠ SỞ LÝ LUẬN 10

1. Tổng quan về tra cứu ảnh 10

2. Bài toán tra cứu ảnh dựa trên ngữ nghĩa 11

3. Các phương pháp thử nghiệm và đánh giá 12

4. Các nghiên cứu trước đây 13

5. Convolutional Neural Network (CNN) 15

6. Các mô hình đồ thị 16

6.1. Graph Convolutional Network (GCN) 17

6.2. Graph Attention Network (GAT) 18

6.3. Graphormer 19

7. Tiểu kết chương 21

CHƯƠNG II: PHƯƠNG PHÁP NGHIÊN CỨU ĐỀ XUẤT 23

1. Trích xuất đặc trưng hình ảnh bằng CNN 23

2. Biểu diễn đồ thị 24

3. Học biểu diễn bằng Graphormer 27

4. Huấn luyện và tối ưu hóa mô hình 29
5. Truy xuất nhanh dựa trên cụm 32
6. Tổng kết và đánh giá 35
7. Tiểu kết chương 35
CHƯƠNG 3: THỬ NGHIỆM VÀ ĐÁNH GIÁ 37
1. Mục tiêu thực nghiệm 37
2. Chi tiết cài đặt 37
3. Tập dữ liệu 38
4. Đánh giá hiệu quả mô hình 39
5. Phân tích kết quả 39
6. Tiểu kết chương 49
KẾT LUẬN 51
1. Kết quả đạt được 51
2. Nhận xét và đề xuất 52
3. Hướng phát triển tiếp theo 52
DANH MỤC TÀI LIỆU THAM KHẢO 54

DANH MỤC CÁC THUẬT NGỮ, CHỮ VIẾT TẮT		
Viết tắt	**Tiếng Anh**	**Tiếng Việt**
SBIR	Semantic Based Image Retrieval	Tra cứu ảnh dựa trên ngữ nghĩa
mAP	Mean Average Precision	Độ chính xác trung bình
TBIR	Text Based Image Retrieval	Tra cứu ảnh dựa trên văn bản
CBIR	Content Based Image Retrieval	Tra cứu ảnh dựa trên nội dung
CNN	Convolutional Neural Network	Mạng nơ-ron tích chập
ViT	Vision Transformer	
CLIP	Contrastive Language–Image Pre-training	Tiền huấn luyện đối sánh ngôn ngữ – hình ảnh
GNN	Graph Neural Network	Mạng nơ-ron đồ thị
GCN	Graph Convolutional Network	Mạng tích chập trên đồ thị
GAT	Graph Attention Network	Mạng chú ý trên đồ thị
Resnet	Residual Network	Mạng hồi tiếp
PyramidNet	Pyramid Residual Network	Mạng hồi tiếp hình kim tự tháp
PCFR	Proxy-Cluster Fast Retrieval	Truy xuất nhanh dựa trên cụm đại diện
FC	Fully Connected layer	Lớp kết nối đầy đủ

DANH SÁCH BẢNG

Bảng 3.1: So sánh hiệu suất trên tập dữ liệu CIFAR-100. 39
Bảng 3.2: So sánh hiệu suất trên tập dữ liệu CUB-200. 40
Bảng 3.3: Bảng so sánh kết quả khi thực hiện truy xuất trong không gian k cụm trên tập dữ liệu CIFAR-100. 42
Bảng 3.4: Bảng so sánh kết quả khi thực hiện truy xuất trong không gian k cụm trên tập dữ liệu CUB-200. 42
Bảng 3.5: So sánh kết quả của theo từng thành phần của mô hình. 47
Bảng 3.6: So sánh giữa mô hình đề xuất và phương pháp học Proxy Anchor truyền thống. 48

DANH SÁCH HÌNH VẼ

Hình 2.1: Kiến trúc tổng thể của mô hình được đề xuất. 23
Hình 2.2: Hoạt động của các hàm mất mát. 29
Hình 2.3: Sơ đồ hoạt động của pha truy xuất. 33
Hình 3.1: Trực quan hóa các embedding được học bởi mô hình trên tập dữ liệu CIFAR-100. 44
Hình 3.2: Trực quan hóa truy xuất trên tập dữ liệu CIFAR-100. 45
Hình 3.3: Trực quan hóa truy xuất trên tập dữ liệu CUB-200-2011. 46

MỞ ĐẦU

1. Lý do lựa chọn đề tài

Sự phát triển bùng nổ của công nghệ thông tin đã dẫn đến lượng dữ liệu hình ảnh tăng nhanh về cả quy mô lẫn tính đa dạng. Các kho lưu trữ hiện đại không chỉ chứa hình ảnh đời sống và thương mại mà còn tích hợp dữ liệu giám sát, hình ảnh y tế chuyên sâu và nhiều nguồn đặc thù khác. Trước thực tế này, nhu cầu về các hệ thống truy xuất thông minh trở nên cấp bách, bởi các kỹ thuật truyền thống dựa trên metadata hoặc đặc trưng thị giác cơ bản như màu sắc và hình dạng không còn đáp ứng được hiệu quả tìm kiếm. Rào cản lớn nhất chính là khoảng cách ngữ nghĩa – sự bất đồng nhất giữa dữ liệu thô của hình ảnh mà máy tính xử lý và khái niệm trừu tượng mà con người tư duy. Hệ quả là các hệ thống hiện hành thường thất bại trong việc nắm bắt ý định người dùng, dẫn đến kết quả tìm kiếm thiếu trực quan và kém chính xác về mặt ý nghĩa.

Đối diện với áp lực về nhu cầu tìm kiếm chuẩn xác và hiệu suất cao, việc nghiên cứu các giải pháp truy xuất ảnh dựa trên ngữ nghĩa đang trở thành nhiệm vụ tiên quyết. Một hệ thống tra cứu lý tưởng ngày nay không chỉ bị ràng buộc bởi tiêu chuẩn khắt khe về độ chính xác nhận dạng, mà còn phải đảm bảo độ trễ thấp nhất có thể, ngay cả khi đối mặt với một lượng lớn dữ liệu và tần suất truy vấn dày đặc. Đây chính là bài toán khó trong thiết kế hệ thống, thúc đẩy sự ra đời của các kiến trúc mới giúp tối ưu hóa đồng thời cả chất lượng tìm kiếm lẫn tài nguyên tính toán.

Trọng tâm nghiên cứu là phát triển một phương pháp tra cứu hình ảnh ngữ nghĩa hiện đại dựa trên Graphormer. Kiến trúc này cho phép hệ thống phân tích và mã hóa các tương quan phức tạp giữa các đối tượng, góp phần cải thiện chất lượng vector đặc trưng và khả năng biểu diễn thông tin ngữ nghĩa. Điểm đột phá nằm ở việc tích hợp chiến lược phân cụm dữ liệu, giúp giới hạn không gian tìm kiếm. Cơ chế này không chỉ đẩy nhanh tốc độ phản hồi mà còn duy trì độ chính xác ở mức cao, đảm bảo kết quả trả về vừa logic vừa sát với ý định người dùng.

Nghiên cứu không chỉ đóng góp về mặt lý thuyết mà còn đem lại giá trị ứng dụng đáng kể. Mô hình có tiềm năng triển khai rộng rãi trong các lĩnh vực quan trọng như thương mại điện tử, y tế số, giám sát an ninh và quản lý dữ liệu đa phương tiện. Việc áp dụng các kỹ thuật truy xuất hiện đại này là bước nền quan trọng để phát triển các hệ thống tìm kiếm thế hệ mới, với khả năng phản hồi nhanh hơn, độ chính xác cao hơn và trải nghiệm người dùng thuận tiện hơn. Trong bối cảnh dữ liệu hình ảnh ngày càng phức tạp, hướng tiếp cận này mở ra cơ hội mới cho việc xây dựng các nền tảng tra cứu semantic hiệu quả và thực tiễn.

2. Tổng quan về vấn đề nghiên cứu

Trong lĩnh vực truy xuất hình ảnh, các phương pháp đã trải qua nhiều giai đoạn phát triển khác nhau. Ở giai đoạn đầu, các hệ thống chủ yếu dựa trên metadata hoặc chú thích thủ công để phục vụ truy vấn. Tuy nhiên, cách tiếp cận này phụ thuộc nhiều vào yếu tố con người, thiếu tính mở rộng và khó đảm bảo tính nhất quán khi khối lượng dữ liệu ngày càng tăng. Để khắc phục hạn chế đó, các phương pháp truy xuất ảnh dựa trên nội dung (Content-Based Image Retrieval – CBIR) đã được đề xuất, trong đó hình ảnh được biểu diễn thông qua các đặc trưng thị giác cấp thấp như màu sắc, kết cấu và hình dạng.

Mặc dù CBIR giúp tự động hóa quá trình truy xuất, các đặc trưng thị giác cơ bản vẫn chưa phản ánh đầy đủ ý nghĩa ngữ nghĩa của hình ảnh. Sự chênh lệch giữa biểu diễn số của máy tính và nhận thức của con người, được gọi là khoảng cách ngữ nghĩa, tiếp tục là thách thức lớn trong các hệ thống truy xuất. Kết quả là nhiều hệ thống cho ra những hình ảnh có độ tương đồng về mặt hình thức nhưng không phù hợp với ngữ cảnh hoặc mục đích tìm kiếm của người dùng.

Những năm gần đây, sự phát triển của học sâu đã mở ra hướng tiếp cận mới cho bài toán truy xuất ảnh ngữ nghĩa. Các mô hình mạng nơ-ron sâu, đặc biệt là Transformer và các biến thể của nó, cho phép học được các biểu diễn đặc trưng giàu ngữ nghĩa hơn từ dữ liệu lớn. Tuy vậy, phần lớn các mô hình hiện tại vẫn tập trung vào việc xử lý hình ảnh như các thực thể độc lập, chưa khai thác đầy đủ mối quan hệ cấu trúc và tương tác giữa các đối tượng bên trong ảnh.

Song song với bài toán biểu diễn ngữ nghĩa, vấn đề hiệu suất truy vấn trên tập dữ liệu lớn cũng đặt ra nhiều thách thức. Khi quy mô dữ liệu tăng nhanh, việc so khớp trực tiếp trên toàn bộ không gian đặc trưng dẫn đến chi phí tính toán cao và độ trễ lớn. Các kỹ thuật tăng tốc như lập chỉ mục, tìm kiếm láng giềng gần nhất và phân cụm dữ liệu đã được đề xuất nhằm thu hẹp không gian tìm kiếm, song việc cân bằng giữa tốc độ và độ chính xác vẫn là một bài toán chưa có lời giải tối ưu.

Từ những phân tích trên có thể thấy rằng vẫn tồn tại khoảng trống nghiên cứu trong việc kết hợp hiệu quả giữa mô hình biểu diễn ngữ nghĩa mạnh và chiến lược tối ưu hóa truy vấn. Đây chính là cơ sở để đề án tập trung nghiên cứu mô hình Graphormer kết hợp với kỹ thuật phân cụm, nhằm đồng thời nâng cao chất lượng biểu diễn ngữ nghĩa và cải thiện hiệu suất tra cứu trên các tập dữ liệu hình ảnh lớn.

3. Mục đích nghiên cứu

Nội dung chính của nghiên cứu tập trung vào việc phân tích khả năng ứng dụng của kiến trúc Graphormer trong lĩnh vực tra cứu ảnh theo ngữ nghĩa. Mục tiêu đề ra là xác định mức độ thích hợp của mô hình trong việc biểu diễn và so sánh hình ảnh bằng cách tận dụng cấu trúc quan hệ giữa các đối tượng. Để thực hiện mục tiêu này, một quy trình xử lý dữ liệu bài bản đã được thiết lập cho các tập dữ liệu ảnh đơn nhãn, bao gồm các bước tiền xử lý, chuẩn hóa đầu vào, trích xuất đặc trưng và chuyển đổi định dạng để tương thích với cơ chế của Graphormer. Trên cơ sở đó, mô hình được huấn luyện để nắm bắt các đặc trưng ngữ nghĩa sâu, phản ánh các tương tác và mối liên kết giữa các thành phần trong ảnh.

Song song với việc xây dựng mô hình, nghiên cứu tiến hành đánh giá thực nghiệm dựa trên các chỉ số định lượng tiêu chuẩn trong lĩnh vực tra cứu. Quá trình này xem xét hiệu quả của Graphormer trên nhiều phương diện như độ chính xác tìm kiếm, khả năng phân tách ngữ nghĩa giữa các lớp và độ ổn định trên các tập dữ liệu khác nhau. Ngoài ra, yếu tố thời gian xử lý cũng được nghiên cứu thông qua việc áp dụng cơ chế phân cụm để thu hẹp phạm vi tìm kiếm, nhằm cân bằng giữa tốc độ và chất lượng kết quả.

Thông qua quá trình phân tích và thực nghiệm đối chứng, nghiên cứu tập trung đánh giá năng lực vận hành cũng như các điểm hạn chế của Graphormer đối với tác vụ tìm kiếm hình ảnh ngữ nghĩa. Các dữ liệu thu thập được là tiền đề quan trọng để đề xuất các hướng nghiên cứu tiếp theo, bao gồm việc áp dụng trên tập dữ liệu gán nhiều nhãn, kết hợp truy vấn đa phương thức (ảnh - văn bản), hoặc cải thiện hiệu năng để đáp ứng yêu cầu xử lý tức thời. Những phân tích này giúp khẳng định tính thực tiễn và tiềm năng khai thác của kiến trúc Graphormer, đặc biệt khi các kỹ thuật biểu diễn cũ đang dần trở nên kém hiệu quả.

4. Đối tượng và phạm vi nghiên cứu

Định hướng chính của đề án là các phương pháp tra cứu ảnh dựa trên ngữ nghĩa, đáp ứng nhu cầu xử lý lượng dữ liệu hình ảnh ngày càng lớn và đa dạng. Các phương pháp hiện đại không chỉ dừng lại ở việc khai thác đặc trưng sơ cấp (màu sắc, cạnh) như trước đây, mà yêu cầu mô hình phải phân tích được nội dung ở tầng nghĩa, cụ thể là vai trò của đối tượng và mối quan hệ giữa chúng. Trong quy trình này, việc học embedding giữ vị trí trung tâm, bởi nó chi phối khả năng đối sánh giữa các ảnh. Việc sử dụng kiến trúc Graphormer – kết hợp ưu điểm của Transformer và mạng đồ thị – được xem là giải pháp để cải thiện chất lượng biểu diễn. Cách tiếp cận này giúp xử lý tốt cả thông tin chi tiết lẫn các tương tác phức tạp trong ảnh, từ đó hỗ trợ quá trình truy vấn nội dung đạt hiệu quả tốt hơn.

Khuôn khổ của bài toán trong nghiên cứu được xác định cụ thể: xây dựng hệ thống tiếp nhận đầu vào là một hình ảnh đơn lẻ được giả định là được gán một nhãn ngữ nghĩa duy nhất ở mức ảnh. Nghiên cứu không đặt giả thiết về số lượng hay sự phân bố của các đối tượng xuất hiện trong ảnh, mà chỉ quan tâm đến nhãn lớp được gán cho toàn bộ hình ảnh. Dựa trên ảnh truy vấn, hệ thống tiến hành truy vấn tập hợp các hình ảnh có sự tương đồng trong cơ sở dữ liệu, dựa trên khoảng cách trong không gian vector nhúng (embedding). Phạm vi bài toán không mở rộng sang các kịch bản đa nhãn, phát hiện đối tượng hay phân đoạn ảnh, mà tập trung vào bài toán truy xuất hình ảnh dựa trên độ tương đồng ngữ nghĩa ở mức toàn ảnh.

Để đạt được mục đích này, một quy trình khép kín bao gồm bốn giai đoạn tuần tự được thiết lập: (1) sử dụng một CNN backbone để thu thập các đặc trưng thị giác từ ảnh; (2) mô hình hóa các đối tượng và mối liên hệ giữa chúng dưới dạng cấu trúc đồ thị; (3) ứng dụng kiến trúc Graphormer để mã hóa và học các vector đại diện chứa đựng thông tin ngữ nghĩa sâu; và (4) thực hiện đối sánh các vector này để phục vụ công tác tìm kiếm ảnh tương tự.

Một điểm đáng lưu ý trong giới hạn nghiên cứu là hệ thống hoàn toàn không phụ thuộc vào các dữ liệu bổ trợ như mô tả văn bản, siêu dữ liệu (metadata) hay các nhãn ngữ nghĩa chi tiết. Phương pháp thiết lập này nhằm mục đích kiểm chứng rằng kết quả đánh giá mô hình sẽ phản ánh trung thực năng lực học biểu diễn từ dữ liệu hình ảnh thuần túy, loại bỏ các tác động từ những nguồn thông tin phi thị giác. Đây được xem là cấu hình thích hợp để khảo sát chính xác hiệu năng nền tảng của Graphormer trong một môi trường thực nghiệm được kiểm soát chặt chẽ.

Bên cạnh việc thiết lập kiến trúc Graphormer đóng vai trò nền tảng, nghiên cứu cũng chú trọng đến các giải pháp cải thiện thời gian phản hồi nhằm giảm thiểu tải xử lý đối với các tập dữ liệu quy mô lớn. Trong hệ thống này, dữ liệu hình ảnh được chuyển đổi vào không gian vector (embedding) giàu ngữ nghĩa, thể hiện đầy đủ mối liên kết giữa các đối tượng khác xung quanh. Mặc dù vậy, sự gia tăng về số lượng ảnh khiến quy trình so sánh đặc trưng của truy vấn với toàn bộ cơ sở dữ liệu trở nên tốn kém về tài nguyên tính toán. Để khắc phục hạn chế này, kỹ thuật phân cụm vector dựa trên đặc tính ngữ nghĩa được triển khai: các vector có sự tương đồng ngữ nghĩa được tổ chức thành các cụm, giúp giới hạn phạm vi tra cứu. Khi xử lý một truy vấn mới, hệ thống chỉ tập trung quét trên các nhóm tiềm năng nhất, qua đó cắt giảm đáng kể khối lượng tính toán mà vẫn bảo toàn độ chính xác.

Phương pháp tiếp cận này giúp duy trì hiệu quả tìm kiếm ngữ nghĩa ở mức ổn định, đồng thời nâng cao hiệu năng vận hành, hỗ trợ khả năng mở rộng hệ thống mà không làm ảnh hưởng đến chất lượng đầu ra. Sự phối hợp giữa Graphormer và cơ chế phân cụm vector giúp hệ thống vừa nắm bắt được các biểu diễn ngữ nghĩa sâu, vừa đáp ứng nhanh chóng các yêu cầu truy vấn, từ đó gia tăng tính khả thi khi triển khai trong các ứng dụng thực tế.

5. Phương pháp nghiên cứu

Chiến lược tiếp cận của đề án được cấu trúc dựa trên ba thành phần cốt lõi, nhằm đảm bảo sự chặt chẽ và tuân thủ các nguyên tắc khoa học hiện hành. Thành phần đầu tiên tập trung vào công tác nghiên cứu lý thuyết, bao gồm việc tổng hợp, đánh giá và hệ thống hóa các tài liệu về tìm kiếm ảnh ngữ nghĩa, kỹ thuật biểu diễn đặc trưng và các mô hình học sâu trên đồ thị. Nhiệm vụ chính trong giai đoạn này là phân tích sâu các nguyên tắc và đặc tính của Graphormer – biến thể của Transformer trên đồ thị – để tạo lập cơ sở khoa học cho giải pháp đề xuất. Các kết quả phân tích lý thuyết này đóng vai trò quan trọng trong việc thiết kế mô hình, định hình quy trình xử lý và xác định các nhân tố tác động đến hiệu quả tìm kiếm.

Thành phần thứ hai xoay quanh công tác phát triển mô hình và thực hiện kiểm chứng. Chiến lược được định hình dựa trên sự tích hợp chặt chẽ giữa hai yếu tố: (1) các mạng học sâu đảm nhiệm việc trích xuất đặc trưng thị giác cơ bản, giúp định danh cụ thể các đối tượng; và (2) mô hình Graphormer chịu trách nhiệm học các biểu diễn ngữ nghĩa thông qua việc mô hình hóa quan hệ đồ thị. Tại giai đoạn này, nghiên cứu thiết lập một luồng xử lý (pipeline) hoàn chỉnh, bao gồm tiền xử lý dữ liệu, khởi tạo đồ thị, huấn luyện và tạo vector nhúng (embedding) cho mục đích tra cứu.

Bên cạnh đó, kỹ thuật phân cụm embedding cũng được áp dụng nhằm giới hạn không gian tìm kiếm, qua đó tối ưu hóa thời gian phản hồi trong các thử nghiệm với lượng dữ liệu lớn. Quy trình thực nghiệm được xây dựng nhằm đảm bảo tính lặp lại (reproducibility) và tiềm năng phát triển cho các công trình nghiên cứu kế tiếp.

Thành phần then chốt thứ ba trong phương pháp luận là hoạt động kiểm định và đo lường chất lượng mô hình. Sử dụng các chỉ số tiêu chuẩn trong lĩnh vực tìm kiếm hình ảnh như Mean Average Precision (mAP), Recall@K và Precision@K để khảo sát hiệu năng hệ thống dưới các góc độ đa chiều. Việc vận dụng kết hợp các thang đo này giúp định lượng khả năng sắp xếp kết quả, tỷ lệ tìm kiếm thành công cũng như độ tin cậy chung của giải pháp. Quy trình kiểm định được triển khai trên bộ dữ liệu test tách biệt theo một trình tự chuẩn hóa, nhằm duy trì sự minh bạch và

giá trị tham khảo của các kết quả. Những phân tích định lượng này hỗ trợ việc nhận định về độ bền vững, tính chính xác và khả năng triển khai của mô hình trong điều kiện thực tế.

Sự tích hợp đồng bộ của ba yếu tố nêu trên hình thành nên một chiến lược nghiên cứu toàn diện, sở hữu cả cơ sở lý luận chặt chẽ lẫn sự xác thực bằng dữ liệu thực nghiệm. Qua đó, tiềm năng triển khai Graphormer cho nhiệm vụ tra cứu ảnh ngữ nghĩa được minh chứng một cách cụ thể và rõ ràng vào bài toán truy xuất ảnh dựa trên ngữ nghĩa.

6. Đóng góp của đề án

Trọng tâm của đề án là khảo sát tính khả thi khi vận dụng kiến trúc Graphormer vào bài toán Truy xuất ảnh dựa trên ngữ nghĩa (SBIR). Mục đích chính là phân tích năng lực vận hành, triển vọng ứng dụng cũng như các rào cản kỹ thuật của mô hình trong việc học các biểu diễn vector (embedding) từ dữ liệu ảnh đơn nhãn.

Một trong những đóng góp đáng chú ý của nghiên cứu là việc triển khai Graphormer như một công cụ học biểu diễn sâu, cho phép hệ thống nhận diện được các đặc trưng phức tạp và mối liên kết ngữ nghĩa giữa các hình ảnh. Hướng tiếp cận này giúp khắc phục những nhược điểm thường thấy ở các kỹ thuật truyền thống vốn chỉ dựa vào đặc trưng thị giác bề mặt hoặc các vector rời rạc. Song song với đó, nghiên cứu cũng thiết lập một quy trình xử lý và mô hình hóa bài bản, bao gồm các công đoạn từ tiền xử lý dữ liệu, trích xuất đặc trưng cơ bản, đến việc kiến tạo cấu trúc đồ thị và huấn luyện Graphormer, nhằm đảm bảo quá trình học embedding diễn ra ổn định và đạt kết quả tốt.

Bên cạnh đó, công tác định lượng hiệu quả thực nghiệm được triển khai thông qua các thước đo phổ biến trong truy xuất ảnh gồm mAP, Recall@K và Precision@K. Các phân tích này giúp làm rõ mức độ chính xác, khả năng nhận diện ngữ nghĩa và độ ổn định của Graphormer khi vận hành trên các nguồn dữ liệu khác nhau, đồng thời cung cấp số liệu đối chứng với các phương pháp cũ để minh chứng cho triển vọng của mô hình.

Đề án cũng đề cập đến đóng góp trong việc cải thiện tốc độ truy vấn thông qua cơ chế phân cụm vector nhúng. Kỹ thuật này hỗ trợ việc giới hạn vùng tìm kiếm cần thiết, giúp hệ thống duy trì chất lượng kết quả tốt trong khi vẫn đảm bảo hiệu suất vận hành cao trên các hệ thống lưu trữ quy mô lớn.

Tóm lại, thành quả đạt được qua các bước khảo sát và thực nghiệm là cơ sở khoa học quan trọng để thúc đẩy các nghiên cứu kế thừa. Đây cũng là nguồn thông tin hữu ích hỗ trợ công tác hiện thực hóa các giải pháp tra cứu ảnh ngữ nghĩa phục vụ nhu cầu đời sống.

Đề án đã có những đóng góp cụ thể trên cả phương diện lý thuyết và kiểm chứng mô hình, minh họa rõ nét hiệu quả của việc ứng dụng Graphormer vào SBIR. Đồng thời, nghiên cứu cũng xác lập nền tảng và định hướng cho các hoạt động mở rộng phạm vi khảo sát trong thời gian tới.

7. Cấu trúc đề án

Nội dung đề án được trình bày qua ba phần chính, đảm bảo tính tuần tự và mạch lạc. Phần thứ nhất dành để hệ thống hóa các kiến thức cơ sở trong lĩnh vực truy xuất hình ảnh, bao quát từ nguyên lý tìm kiếm theo nội dung ngữ nghĩa, các tiêu chuẩn đánh giá đến việc khảo sát các công trình tiền nhiệm. Thêm vào đó, vai trò của các mô hình học sâu chủ đạo như mạng tích chập (CNN) và các biến thể mạng đồ thị (như GCN, GAT, Graphormer) cũng được thảo luận chi tiết nhằm xây dựng luận cứ cho phương pháp tiếp cận của đề tài. Phần này kết thúc bằng việc tóm lược các vấn đề lý thuyết để làm nền tảng cho các chương sau.

Nội dung kế tiếp sẽ đi sâu vào mô tả tường tận giải pháp kỹ thuật được xây dựng trong khuôn khổ đề án. Chu trình xử lý được thiết lập bao gồm việc thu thập đặc trưng thị giác từ ảnh, biến đổi dữ liệu sang định dạng đồ thị, và áp dụng Graphormer để học các biểu diễn ngữ nghĩa. Bên cạnh đó, quá trình tối ưu hóa mô hình và chiến lược tăng tốc tìm kiếm bằng kỹ thuật phân cụm vector cũng được trình bày. Phần này cũng đưa ra những đánh giá bước đầu về tính khả thi khi triển khai giải pháp sau khi hệ thống đã được hoàn thiện.

Chương tiếp theo tập trung vào công tác thực nghiệm và kiểm chứng, bao gồm việc cấu hình môi trường phần cứng, mô tả đặc thù dữ liệu và xác định các chỉ số đo lường (mAP, Precision@K, Recall@K). Đồng thời, tiến hành luận giải các số liệu ghi nhận được thông qua phân tích định lượng và biểu diễn trực quan không gian vector. Thêm vào đó, các phân tích cắt lớp (ablation study) cũng được thực hiện để định lượng sự đóng góp của từng module thành phần đối với kết quả tra cứu cuối cùng. Mục này sẽ khép lại bằng việc tổng hợp các kết luận cốt lõi từ quá trình khảo sát.

Việc tổ chức nội dung đề án tuân thủ một lộ trình bài bản, chuyển tiếp nhịp nhàng từ các kiến thức nền tảng sang mô hình kỹ thuật và cuối cùng là phân tích hiệu năng. Mục đích là tạo ra một luồng thông tin có trật tự, đảm bảo tính liên kết và dễ tiếp cận cho toàn bộ công trình, giúp người đọc nắm bắt vấn đề từ tổng quát đến chi tiết.

NỘI DUNG

CHƯƠNG I: CƠ SỞ LÝ LUẬN

Chương này cung cấp nền tảng lý thuyết và tổng quan các nghiên cứu liên quan, làm cơ sở cho phương pháp truy xuất hình ảnh theo ngữ nghĩa. Nội dung tập trung vào các khái niệm và kỹ thuật từ CBIR, SBIR truyền thống đến các phương pháp học sâu và không gian embedding, cùng phân tích các mô hình tiêu biểu như CNN, GCN, GAT và Graphormer. Chương cũng đề cập đến các thước đo hiệu suất chuẩn mực và phân tích khoảng trống nghiên cứu, làm cơ sở cho việc ứng dụng Graphormer trong phương pháp đề xuất và các thử nghiệm trong các chương tiếp theo.

1. Tổng quan về tra cứu ảnh

Truy xuất hình ảnh (Image Retrieval) là quá trình tìm kiếm các hình ảnh phù hợp trong một kho dữ liệu dựa trên thông tin truy vấn, có thể là văn bản, từ khóa hoặc ảnh mẫu. Hệ thống hướng tới việc nhận diện các ảnh tương đồng về nội dung thị giác, ngữ cảnh hoặc ý nghĩa. Đây là nhiệm vụ quan trọng trong nhiều ứng dụng như thương mại điện tử, y tế số, giám sát an ninh, giao thông thông minh hay quản lý kho dữ liệu đa phương tiện.

Hai hướng chính hiện nay là TBIR (truy xuất dựa trên văn bản) và CBIR (truy xuất dựa trên nội dung thị giác). TBIR tận dụng sự phong phú của ngôn ngữ, nhưng phụ thuộc vào chất lượng metadata, còn CBIR trích xuất trực tiếp các đặc trưng hình ảnh như màu sắc, hình dạng, kết cấu hoặc embedding từ mạng học sâu. CBIR khắc phục hạn chế metadata nhưng gặp thách thức về khoảng cách ngữ nghĩa giữa đặc trưng trích xuất và khái niệm ngữ nghĩa thực tế, cũng như biến thiên về điều kiện chụp, góc nhìn và chất lượng ảnh.

Sự bùng nổ dữ liệu số tạo ra thêm thách thức về quy mô lưu trữ và chi phí xử lý, đòi hỏi các kỹ thuật tối ưu như giảm chiều dữ liệu, lập chỉ mục thông minh, tìm kiếm láng giềng gần nhất hoặc triển khai hệ thống phân tán. Tổng thể, truy xuất ảnh

là lĩnh vực vừa lý thuyết vừa ứng dụng, với xu hướng hiện đại là kết hợp TBIR và CBIR để tăng độ chính xác và linh hoạt. Phát triển các mô hình nắm bắt ngữ nghĩa sâu cùng tối ưu hóa cho dữ liệu lớn sẽ là yếu tố then chốt để nâng cao hiệu quả hệ thống trong tương lai.

2. Bài toán tra cứu ảnh dựa trên ngữ nghĩa

Truy xuất hình ảnh dựa trên ngữ nghĩa (Semantic-Based Image Retrieval – SBIR) hướng tới việc tìm kiếm và lựa chọn những hình ảnh có nội dung mang tính ý nghĩa tương đồng, thay vì chỉ dựa vào sự giống nhau ở mức đặc trưng thị giác cơ bản. Theo quan điểm của con người, ý nghĩa của một bức ảnh không chỉ được xác định từ các điểm ảnh hay cấu trúc pixel, mà dựa trên khả năng nhận diện các đối tượng quan trọng, các hành động đang diễn ra và tổng thể bối cảnh trong ảnh. Vì vậy, một hệ thống SBIR cần biểu diễn hình ảnh theo cách phản ánh các khái niệm mang ý nghĩa thực sự, thay vì chỉ ghi nhận các mẫu thị giác đơn giản như màu sắc, cạnh hay kết cấu.

Khác với CBIR truyền thống, vốn chủ yếu khai thác các đặc trưng mức thấp hoặc các vector nhúng chưa được điều chỉnh tối ưu cho ngữ nghĩa, SBIR hướng tới xây dựng một không gian biểu diễn giàu thông tin hơn. Trong không gian này, các hình ảnh có nội dung tương đồng được đặt gần nhau, bất chấp những biến đổi về ánh sáng, góc chụp, tỉ lệ hay hình dạng của đối tượng. Cách tiếp cận này giúp thu hẹp sự khác biệt về ngữ nghĩa giữa biểu diễn máy tính và cách con người nhận thức thông tin trực quan. Nhờ vậy, SBIR trở thành giải pháp thích hợp cho những bài toán yêu cầu hiểu sâu nội dung hình ảnh, vượt lên trên việc so sánh các đặc trưng bề mặt — ngay cả trong các trường hợp dữ liệu đơn nhãn, nơi mục tiêu là truy xuất những hình ảnh cùng lớp, mang ý nghĩa tương đồng ở mức khái niệm.

Trong phạm vi nghiên cứu này, bài toán truy xuất hình ảnh được coi là một trường hợp của SBIR, mặc dù dữ liệu được tổ chức theo dạng đơn nhãn. Nhãn lớp tại đây đóng vai trò như tín hiệu ngữ nghĩa chính, biểu thị nội dung trung tâm của từng bức ảnh; vì thế, việc xác định và truy xuất các hình ảnh thuộc cùng nhóm được

xem như một dạng truy xuất dựa trên ý nghĩa, thay vì chỉ dựa trên mức độ tương đồng thị giác thuần túy.

Với cách tiếp cận này, SBIR không đòi hỏi phải mô hình hóa đầy đủ các tầng đối tượng hay các quan hệ phức tạp như trong các hệ thống SBIR đa đối tượng. Thay vào đó, trọng tâm là xây dựng một không gian biểu diễn sao cho các hình ảnh tương đồng về nội dung ngữ nghĩa được đặt gần nhau, còn những ảnh khác biệt về lớp hoặc về ý nghĩa được tách biệt rõ ràng. Khi không gian biểu diễn thể hiện được cấu trúc ngữ nghĩa này, việc truy xuất ảnh trở nên sát với bản chất của SBIR: tìm kiếm dựa trên ý nghĩa được phản ánh trong nhãn lớp và trong đặc trưng của ảnh, hơn là dựa trên hình thức thị giác bề mặt.

Theo quan điểm này, bài toán truy xuất hình ảnh đơn nhãn trong nghiên cứu vẫn phản ánh đầy đủ các đặc trưng cơ bản của SBIR, bởi mục tiêu chính là tổ chức và truy xuất hình ảnh dựa trên ý nghĩa mà nhãn lớp đại diện. Do đó, phương pháp này không chỉ thích hợp với dữ liệu đơn nhãn mà còn phù hợp với triết lý cốt lõi của SBIR, tập trung vào truy xuất dựa trên nội dung ngữ nghĩa thay vì chỉ dựa trên các đặc trưng thị giác cấp thấp.

3. Các phương pháp thử nghiệm và đánh giá

Quá trình thực nghiệm được triển khai trên hai bộ dữ liệu chuẩn là CIFAR-100 [1] và CUB-200-2011 [2] nhằm mục đích kiểm chứng toàn diện năng lực của mô hình đối với tác vụ tra cứu ảnh theo ngữ nghĩa. Quy trình này bao quát trọn vẹn các công đoạn từ trích rút đặc trưng, xây dựng cấu trúc đồ thị cho đến việc học embedding và thực hiện truy vấn.

Về phương diện phương pháp luận, nghiên cứu tận dụng khả năng mạnh mẽ của Graphormer để nâng cao chất lượng biểu diễn hình ảnh; cụ thể, các đặc trưng được trích xuất từ mạng CNN sẽ được chuyển đổi thành cấu trúc đồ thị làm đầu vào cho quá trình huấn luyện. Hiệu suất hệ thống được đánh giá dựa trên các chỉ số chuẩn như Mean Average Precision (mAP), Precision@K (P@K) và Recall@K (R@K), phản ánh độ chính xác, độ tin cậy và khả năng thu hồi thông tin.

Precision@K (P@K) phản ánh tỷ lệ các kết quả truy xuất đúng trong K hình ảnh đầu tiên, được xác định theo công thức:

$$P@K = \frac{1}{K} \sum_{i=1}^{K} rel(i),$$

trong đó $rel(i)$ nhận giá trị 1 nếu hình ảnh ở vị trí i có cùng nhãn với ảnh truy vấn và nhận giá trị 0 trong trường hợp ngược lại.

Recall@K đo lường khả năng thu hồi các hình ảnh liên quan trong K kết quả đầu tiên, được tính bởi:

$$R@K = \frac{\sum_{i=1}^{K} rel(i)}{N},$$

với N là tổng số hình ảnh liên quan trong toàn bộ cơ sở dữ liệu.

Mean Average Precision (mAP) cung cấp đánh giá tổng thể về chất lượng truy xuất bằng cách lấy trung bình độ chính xác tại các vị trí truy xuất đúng và được xác định theo công thức:

$$mAP = \frac{1}{N} \sum_{i=1}^{N} AP_i,$$

trong đó N là số lượng lớp và AP_i là Average Precision của lớp thứ i.

Cuối cùng, kết quả thực nghiệm sẽ được so sánh với các mô hình tham chiếu nhằm làm nổi bật những ưu điểm, hiệu quả và tính khả thi của giải pháp đề xuất.

4. Các nghiên cứu trước đây

Trong lĩnh vực tra cứu hình ảnh, các phương pháp truyền thống chủ yếu tập trung vào việc trích chọn đặc trưng thị giác và xây dựng cơ chế so khớp hiệu quả trên không gian đặc trưng. Với sự phát triển của học sâu, các mô hình mạng nơ-ron tích chập (CNN) [3] đã được sử dụng rộng rãi để học các biểu diễn đặc trưng có khả năng phân biệt cao, thay thế cho các đặc trưng thủ công trước đây. Các vector đặc trưng được trích xuất từ CNN thường được so sánh thông qua các thước đo khoảng cách nhằm xác định mức độ tương đồng giữa các hình ảnh trong bài toán truy xuất.

Gần đây, Vision Transformer (ViT) [4, 5] và các biến thể đã thu hút nhiều sự quan tâm nhờ khả năng mô hình hóa quan hệ dài hạn và ngữ cảnh toàn cục trong ảnh. So với CNN, ViT cho phép khai thác thông tin toàn ảnh một cách trực tiếp thông qua cơ chế attention, từ đó cải thiện chất lượng biểu diễn trong một số kịch bản tra cứu phức tạp. Tuy nhiên, các mô hình transformer thường yêu cầu lượng dữ liệu huấn luyện lớn và chi phí tính toán cao, gây ra thách thức khi triển khai trên các hệ thống truy xuất quy mô lớn hoặc có yêu cầu độ trễ thấp.

Bên cạnh việc nâng cao chất lượng đặc trưng, bài toán hiệu suất truy vấn cũng đóng vai trò then chốt trong tra cứu ảnh quy mô lớn. Việc so khớp trực tiếp các vector đặc trưng có chiều cao dẫn đến chi phí tính toán và bộ nhớ lớn. Để khắc phục vấn đề này, các phương pháp hashing cho truy xuất ảnh đã được đề xuất nhằm ánh xạ vector đặc trưng liên tục sang các mã nhị phân ngắn gọn, cho phép tìm kiếm nhanh thông qua khoảng cách Hamming. Các kỹ thuật hashing học sâu kết hợp CNN hoặc ViT giúp cân bằng giữa độ chính xác truy xuất và hiệu quả lưu trữ – tính toán.

Mặc dù hashing giúp cải thiện đáng kể tốc độ truy vấn và khả năng mở rộng hệ thống, việc lượng tử hóa đặc trưng sang không gian nhị phân có thể gây mất mát thông tin ngữ nghĩa, đặc biệt trong các kịch bản yêu cầu phân biệt tinh tế giữa các lớp tương đồng cao. Do đó, một thách thức quan trọng trong các hệ thống tra cứu ảnh hiện đại là thiết kế biểu diễn đặc trưng vừa giàu ngữ nghĩa vừa phù hợp cho truy vấn nhanh, đảm bảo cân bằng giữa độ chính xác và hiệu suất.

Tổng hợp lại, các nghiên cứu liên quan cho thấy hai hướng tiếp cận chính trong tra cứu hình ảnh hiện nay: (1) trích chọn đặc trưng sâu bằng CNN hoặc Vision Transformer nhằm nâng cao khả năng biểu diễn ngữ nghĩa; và (2) áp dụng các kỹ thuật hashing để tăng tốc truy vấn và mở rộng quy mô hệ thống. Tuy nhiên, việc kết hợp hiệu quả giữa chất lượng đặc trưng và cơ chế truy vấn nhanh vẫn là một vấn đề mở, đặt ra nhu cầu nghiên cứu các phương pháp mới nhằm tối ưu đồng thời cả độ chính xác và hiệu suất trong bài toán tra cứu ảnh.

5. Convolutional Neural Network (CNN)

Convolutional Neural Network (CNN) [3] được xem là trụ cột trong công nghệ thị giác máy tính, nổi bật với khả năng tự tri trích xuất các thuộc tính cốt lõi từ dữ liệu gốc mà không yêu cầu can thiệp thủ công. Hệ thống về kiến trúc tuân theo mô hình nhiều tầng, bao gồm sự phối hợp giữa các tầng tích chập, hàm kích hoạt phi tuyến, tầng giảm mẫu (pooling) và kết thúc bằng các lớp FC. Cơ chế vận hành của tầng tích chập dựa trên việc sử dụng các bộ lọc (kernel) kích thước nhỏ để trượt qua và trích rút thông tin từ các khu vực cục bộ, giúp hệ thống nhạy bén với các chi tiết cơ bản như đường biên hay góc cạnh. Điểm ưu việt của CNN nằm ở kỹ thuật chia sẻ tham số (weight sharing), cho phép cùng một bộ lọc được tái sử dụng trên toàn bộ không gian ảnh, giúp nhận diện đặc trưng bất biến với vị trí và giảm tải đáng kể khối lượng tính toán. Tiếp đó, tầng subsampling đóng vai trò tinh giản chiều dữ liệu, chắt lọc thông tin tinh túy và tăng cường độ bền vững của mô hình trước các nhiễu động. Thông qua việc xếp chồng các lớp theo chiều sâu, CNN hình thành nên khả năng học phân cấp: từ việc nắm bắt các tín hiệu đơn giản ở các tầng đầu đến việc tổng hợp thành các cấu trúc trừu tượng phức tạp ở các tầng sâu, nâng cao khả năng hoạt động vượt trội trong các bài toán nhận dạng và xử lý tài liệu.

Để khắc phục hiện tượng đạo hàm bị triệt tiêu khi gia tăng số tầng của mạng nơ-ron, ResNet [6] được giới thiệu như một cấu trúc CNN cải tiến đặc biệt. Cốt lõi của phương pháp là khối residual, nơi dòng dữ liệu có thể bỏ qua các lớp trung gian thông qua kết nối tắt (skip connection) để hợp nhất với đầu ra của chuỗi xử lý phi tuyến. Chiến lược này khơi thông dòng chảy của gradient qua nhiều tầng mạng, mở đường cho xây dựng các mạng nơ-ron có độ sâu cao (deep networks) mà vẫn đảm bảo khả năng hội tụ. Về mặt tổ chức, ResNet bao gồm chuỗi các khối residual kết hợp với lớp pooling để nén đặc trưng và lớp fully connected để đưa ra dự đoán cuối cùng. Nhờ cơ chế cộng gộp tín hiệu này, mô hình không chỉ duy trì được thông tin nguyên bản mà còn học thêm được các biểu diễn phức tạp, cải thiện hiệu quả một cách rõ rệt so với các kiến trúc CNN truyền thống.

Để nâng cao khả năng học đặc trưng và khắc phục các hạn chế hiện có của các mạng nơ-ron rất sâu, PyramidNet [7]−một biến thể nâng cao của ResNet – đã được phát triển. Điểm sáng tạo của mô hình là việc thay đổi số lượng kênh đặc trưng không theo bậc thang mà tăng trưởng dần dần (tuyến tính) xuyên suốt chiều sâu của mạng. Cách tiếp cận này giúp luồng thông tin được trao đổi một cách liền mạch so với việc tăng kênh đột ngột của ResNet. Bằng cách sử dụng zero-padded skip connections, PyramidNet giải quyết hiệu quả vấn đề kích thước khi số kênh biến thiên, giúp giảm thiểu hiện tượng nút thắt cổ chai và đạt hiệu năng cao hơn trên CIFAR và ImageNet.

Tóm lại, CNN và các biến thể cải tiến như ResNet hay PyramidNet tất cả đều cho thấy sức mạnh vượt trội trong việc rút trích đặc trưng hình ảnh, trong khi CNN cơ bản chủ yếu học các đặc trưng theo cấp bậc từ đơn giản đến trừu tượng, ResNet giải quyết vấn đề đạo hàm triệt tiêu và hỗ trợ xây dựng các mạng sâu hơn, còn PyramidNet tối ưu hóa luồng thông tin bằng cách tăng dần số kênh đặc trưng và sử dụng kết nối tắt thích hợp. Các cải tiến này vừa nâng cao khả năng biểu diễn của mạng, vừa duy trì hiệu suất ổn định và khả năng hội tụ, từ đó tạo tiền đề cho các nhiệm vụ nhận dạng, phân loại và truy xuất hình ảnh trong bối cảnh thị giác máy tính hiện đại.

6. Các mô hình đồ thị

Cấu trúc dữ liệu đồ thị được định nghĩa thông qua tập hợp các đỉnh (nodes) đại diện cho thực thể và các cạnh (edges) phản ánh sự kết nối giữa chúng. Cách tổ chức này mang lại khả năng mô tả sống động các tương tác đa chiều, sự phụ thuộc lẫn nhau hay cấu trúc không gian giữa các phần tử. Đối với bài toán thị giác máy tính, việc chuyển đổi hình ảnh sang dạng đồ thị cho phép hệ thống vượt qua giới hạn của việc nhận diện đơn lẻ để thấu hiểu sâu sắc ngữ cảnh và mối liên hệ giữa các vật thể. Chính vì thế, các mạng nơ-ron đồ thị (GNN) đóng vai trò hạt nhân trong việc khai phá cấu trúc cảnh và nâng cao chất lượng biểu diễn ngữ nghĩa, phục vụ đắc lực cho các tác vụ như phân loại hay truy xuất ảnh. Graph Convolutional Network (GCN) [8] và Graph Attention Network (GAT) [9] là những mô hình tiêu biểu trong

việc tổng hợp thông tin từ các nút lân cận để khai thác quan hệ phức tạp, trong khi Graphormer [10] mở rộng khả năng này bằng cơ chế attention toàn cục, giúp nắm bắt các kết nối tầm xa trong đồ thị.

6.1. Graph Convolutional Network (GCN)

Graph Convolutional Network (GCN) [8] là một kiến trúc mạng nơ-ron được thiết kế đặc biệt để xử lý dữ liệu có cấu trúc đồ thị, nơi các phần tử không nằm trên lưới cố định như hình ảnh hay chuỗi thời gian. Điểm mạnh của GCN nằm ở khả năng học biểu diễn của từng đỉnh không chỉ dựa trên đặc trưng riêng của nó mà còn dựa trên thông tin từ các đỉnh lân cận, nhờ đó mô hình có thể nắm bắt được cấu trúc và mối quan hệ trong toàn bộ đồ thị.

Cốt lõi của GCN là cơ chế lan truyền thông tin giữa các đỉnh. Ở mỗi lớp, mỗi đỉnh sẽ tổng hợp thông tin từ các hàng xóm, kết hợp với đặc trưng ban đầu của chính nó, và từ đó tạo ra một vector biểu diễn mới. Quá trình này được thực hiện lặp đi lặp lại qua nhiều lớp, giúp các embedding không chỉ phản ánh đặc trưng cục bộ mà còn mở rộng ra bối cảnh toàn cục, tạo ra các biểu diễn giàu ngữ nghĩa và có khả năng thể hiện mối quan hệ phức tạp giữa các phần tử trong đồ thị.

Để đảm bảo thông tin lan truyền được ổn định và không bị mất đi trong quá trình tổng hợp từ hàng xóm, GCN sử dụng các cơ chế như self-loop và chuẩn hóa đối xứng. Self-loop giúp mỗi đỉnh giữ lại thông tin gốc của chính nó, tránh bị hòa tan hoàn toàn vào thông tin từ các hàng xóm, trong khi chuẩn hóa đảm bảo rằng các đỉnh có nhiều hoặc ít hàng xóm đều được cân nhắc một cách hợp lý, tránh hiện tượng một số đỉnh chiếm ưu thế hoặc làm nhiễu quá trình học.

Mỗi lớp của GCN tương ứng với một mức độ lan truyền thông tin khác nhau trong đồ thị. Lớp đầu tiên chủ yếu tổng hợp thông tin từ các đỉnh lân cận trực tiếp, trong khi các lớp sâu hơn cho phép đỉnh tiếp nhận bối cảnh xa hơn, mở rộng phạm vi ảnh hưởng qua nhiều bước lặp. Kết quả là embedding cuối cùng không chỉ chứa đặc trưng riêng của từng đỉnh mà còn phản ánh cấu trúc và mối quan hệ trong toàn

đồ thị, tạo ra các biểu diễn toàn diện, có khả năng nắm bắt ngữ nghĩa ở cả cấp độ cục bộ và toàn cục.

Những embedding này sau đó có thể được ứng dụng vào nhiều nhiệm vụ khác nhau, từ phân loại đỉnh, dự báo liên kết, gom cụm, cho đến hỗ trợ trong các mô hình thị giác và xử lý ngôn ngữ. Bản chất của GCN là tạo ra một không gian biểu diễn mà ở đó mỗi đỉnh không còn là một điểm riêng lẻ mà đã chứa thông tin về cấu trúc xung quanh, từ đó nâng cao khả năng hiểu và xử lý dữ liệu có quan hệ phức tạp.

Tóm lại, GCN không chỉ học các đặc trưng riêng lẻ mà còn học cách kết hợp thông tin từ các đỉnh lân cận, tạo ra embedding giàu ngữ nghĩa và ổn định, phù hợp cho các bài toán cần khai thác cấu trúc mối quan hệ trong dữ liệu, đồng thời phục vụ đắc lực cho các ứng dụng như phân loại, truy xuất, dự báo hay khai thác thông tin từ đồ thị phức tạp.

6.2. Graph Attention Network (GAT)

Graph Attention Network (GAT) [11] được xem là một bước phát triển quan trọng so với kiến trúc GCN truyền thống, nhờ khả năng kết hợp cơ chế attention nhằm đánh giá mức độ quan trọng của các kết nối trong đồ thị một cách linh hoạt. Trong khi GCN truyền thống xử lý thông tin từ các đỉnh lân cận dựa trên trọng số cố định từ ma trận kề, GAT cho phép mô hình tự động học được mức độ ảnh hưởng của từng đỉnh lân cận đối với nút trung tâm. Cơ chế này giúp hệ thống phân biệt được những liên kết mang nhiều thông tin hữu ích so với những nút ít đóng góp, từ đó tập trung nguồn lực tính toán vào các quan hệ có ý nghĩa cao hơn.

Một trong những điểm nổi bật của GAT là việc ứng dụng multi-head self attention. Thay vì chỉ có một cách nhìn nhận thông tin từ các hàng xóm, mô hình sử dụng nhiều đầu attention song song để khai thác dữ liệu từ nhiều góc độ khác nhau. Điều này không chỉ làm phong phú khả năng biểu diễn mà còn cải thiện sự ổn định khi huấn luyện, giảm thiểu hiện tượng overfitting hoặc bị chi phối bởi một đầu attention duy nhất. Mỗi đầu attention sẽ học cách phân bổ trọng số khác nhau cho

các đỉnh lân cận, sau đó kết hợp các đầu ra này để tạo thành biểu diễn cuối cùng, giàu thông tin hơn và phản ánh chính xác hơn cấu trúc phức tạp của đồ thị.

Trong quá trình huấn luyện, trạng thái của từng nút được cập nhật dựa trên thông tin tích lũy từ các nút lân cận, nhưng không phải mọi thông tin đều được coi trọng như nhau. Hệ số attention học được giúp điều chỉnh mức độ đóng góp của từng nút, cho phép mô hình tập trung vào những liên kết quan trọng và giảm ảnh hưởng của các nút ít giá trị. Nhờ cơ chế này, GAT không chỉ nắm bắt được đặc trưng cục bộ mà còn tổng hợp thông tin từ toàn đồ thị một cách có chọn lọc, giúp biểu diễn các mối quan hệ phức tạp một cách hiệu quả.

Nhờ những cải tiến này, GAT vượt trội hơn GCN trong việc biểu diễn đặc trưng của dữ liệu đồ thị. Mô hình có khả năng học các mối quan hệ phi tuyến và trọng số hóa thông tin theo cách thích ứng, từ đó nâng cao khả năng phân loại, dự đoán liên kết hoặc bất kỳ nhiệm vụ nào cần nắm bắt ngữ nghĩa và cấu trúc phức tạp. Chính sự linh hoạt và khả năng học chọn lọc này khiến GAT trở thành một công cụ mạnh mẽ, đặc biệt phù hợp trong các bài toán mà mối quan hệ giữa các nút trong đồ thị đóng vai trò quan trọng, chẳng hạn như trong truy xuất hình ảnh dựa trên ngữ nghĩa hay các hệ thống biểu diễn embedding nâng cao.

6.3. Graphormer

Graphormer [10] đánh dấu một bước tiến quan trọng trong lĩnh vực mạng nơ-ron đồ thị, đặc biệt khi so sánh với các kiến trúc GNN truyền thống như GCN hay GAT. Khác với GCN, vốn chủ yếu dựa vào thông tin từ các nút lân cận trực tiếp, và GAT, mặc dù đã áp dụng cơ chế attention nhưng vẫn giới hạn trong việc tập trung vào các hàng xóm gần, Graphormer sử dụng cơ chế self-attention toàn cục. Nhờ đó, bất kỳ cặp nút nào trong đồ thị đều có thể trao đổi thông tin trực tiếp, không phụ thuộc vào cấu trúc ma trận kề. Điều này giúp mô hình không chỉ học được các mối quan hệ cục bộ mà còn khai thác các kết nối tầm xa, từ đó hiểu sâu hơn về cấu trúc tổng thể của đồ thị.

Để duy trì tính nhất quán của thông tin khi áp dụng cơ chế attention, Graphormer kết hợp các kỹ thuật mã hóa đặc biệt như spatial encoding, edge encoding và centrality encoding. Những kỹ thuật này giúp mô hình ghi nhớ vị trí, vai trò và tầm quan trọng của từng nút trong đồ thị, từ đó tạo ra các embedding giàu ngữ nghĩa hơn. Khi áp dụng cho bài toán truy xuất ảnh, mỗi nút trong đồ thị có thể đại diện cho một đối tượng, một vùng hay một đặc trưng quan trọng trong hình ảnh, và cơ chế attention toàn cục cho phép mô hình học cách cân nhắc mối quan hệ giữa tất cả các nút này, ngay cả khi chúng nằm cách xa nhau về mặt không gian trong ảnh.

Một ưu điểm nổi bật khác của Graphormer so với GCN và GAT là khả năng ổn định và linh hoạt trong việc xử lý đồ thị phức tạp. Trong khi GCN dễ bị hạn chế bởi phạm vi lân cận và GAT có thể gặp khó khăn khi số lượng hàng xóm tăng lên, Graphormer giải quyết vấn đề này bằng cách thiết lập các kênh tương tác trực tiếp giữa các nút mà không bị giới hạn, đồng thời vẫn duy trì hiệu quả học tập nhờ vào cơ chế multi-head self attention. Cơ chế này không chỉ tăng cường khả năng chọn lọc thông tin quan trọng mà còn giảm thiểu tác động của các nút ít liên quan, từ đó cải thiện chất lượng embedding và độ chính xác khi truy xuất ảnh.

Trong bối cảnh truy xuất hình ảnh, tiềm năng của Graphormer còn thể hiện rõ khi ảnh được ánh xạ thành đồ thị đối tượng hoặc vùng. Các embedding học được mang thông tin ngữ nghĩa sâu, phản ánh cả mối quan hệ cục bộ và toàn cục giữa các đối tượng trong ảnh. Nhờ vậy, việc truy xuất các hình ảnh có nội dung tương đồng trở nên chính xác hơn, ổn định hơn trước sự biến đổi về bố cục, ánh sáng hay góc chụp. Nhìn chung, Graphormer không chỉ khắc phục những hạn chế của GCN và GAT mà còn mở ra hướng tiếp cận mới cho các bài toán image retrieval dựa trên cấu trúc đồ thị, nhờ khả năng học biểu diễn toàn diện, linh hoạt và giàu ngữ nghĩa.

Mô hình được sử dụng trong nghiên cứu dựa trên kiến trúc Graphormer nhưng đã được đơn giản hóa để phù hợp với bài toán truy xuất ảnh. Cụ thể, các thành phần như spatial encoding, edge encoding và centrality encoding đã được loại bỏ, chỉ giữ lại multi-layer Transformer [12] nhằm học các biểu diễn nút trong đồ thị. Mặc dù

lược giản, mô hình vẫn duy trì khả năng khai thác mối quan hệ giữa các node, từ đó tạo ra embedding giàu ngữ nghĩa, ổn định trước các biến đổi về bố cục hay đặc trưng thị giác. So với các kiến trúc GCN hay GAT, cách tiếp cận này loại bỏ giới hạn về phạm vi lân cận cố định, cho phép thông tin được truyền trực tiếp giữa bất kỳ cặp node nào trong đồ thị, đồng thời giảm bớt độ phức tạp tính toán và dễ triển khai hơn. Nhờ vậy, mô hình vừa giữ được bản chất của các Graph Neural Network trong việc học cấu trúc, vừa tập trung vào việc tối ưu hóa embedding cho truy xuất ảnh dựa trên ngữ nghĩa mà không cần các cơ chế encoding phức tạp vốn không cần thiết trong bài toán đơn nhãn.

7. Tiểu kết chương

Nội dung chương I đã hệ thống hóa các khái niệm và nền tảng lý thuyết liên quan đến bài toán truy xuất hình ảnh theo ngữ nghĩa. Trước hết, bài toán truy xuất hình ảnh được giới thiệu một cách tổng quan, làm rõ đặc điểm của các hướng tiếp cận phổ biến như TBIR và CBIR, đồng thời chỉ ra những hạn chế còn tồn tại khi chỉ dựa trên đặc trưng thị giác mức thấp.

Tiếp đó, bài toán truy xuất hình ảnh dựa trên ngữ nghĩa được phân tích nhằm nhấn mạnh vai trò của không gian biểu diễn giàu ý nghĩa, trong đó các hình ảnh có nội dung tương đồng được đặt gần nhau. Trong phạm vi nghiên cứu, truy xuất ảnh đơn nhãn được xem như một trường hợp đặc biệt của SBIR, với nhãn lớp đóng vai trò là tín hiệu ngữ nghĩa chính để tổ chức và truy xuất dữ liệu.

Bên cạnh đó, các tiêu chí đánh giá hiệu suất thường dùng trong truy xuất ảnh, bao gồm Precision@K, Recall@K và mAP, đã được trình bày nhằm làm rõ cách đo lường độ chính xác, khả năng thu hồi và chất lượng tổng thể của hệ thống. Các bộ dữ liệu chuẩn được sử dụng trong thực nghiệm cũng được giới thiệu để đảm bảo tính khách quan và khả năng so sánh kết quả.

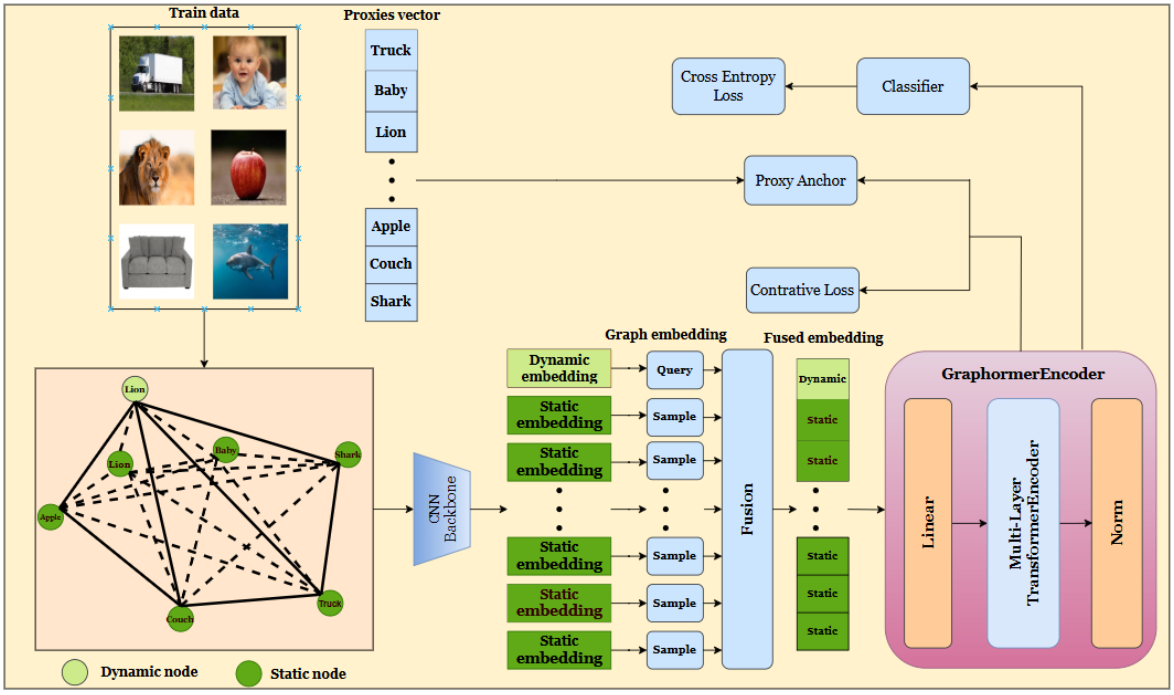
Ngoài ra, các hướng nghiên cứu liên quan đã được tổng hợp, tập trung vào việc trích chọn đặc trưng bằng CNN, Vision Transformer và các kỹ thuật hashing nhằm cân bằng giữa chất lượng biểu diễn và hiệu quả truy vấn. Phân tích ưu – nhược

điểm của các phương pháp này cho thấy nhu cầu về những mô hình biểu diễn vừa giàu ngữ nghĩa vừa phù hợp cho truy xuất ảnh quy mô lớn.

Cuối cùng, các mô hình học sâu nền tảng, bao gồm CNN và các mạng nơ-ron đồ thị như GCN, GAT và Graphormer, đã được giới thiệu nhằm làm rõ cơ sở lý thuyết cho phương pháp đề xuất. Đặc biệt, kiến trúc Graphormer và biến thể đơn giản hóa được sử dụng trong nghiên cứu được xem là tiền đề quan trọng cho việc xây dựng mô hình học biểu diễn phục vụ bài toán truy xuất hình ảnh theo ngữ nghĩa ở các phần tiếp theo.

CHƯƠNG II: PHƯƠNG PHÁP NGHIÊN CỨU ĐỀ XUẤT

Trong chương này, một phương pháp mới được đề xuất để giải quyết bài toán tra cứu ảnh dựa trên ngữ nghĩa với tập dữ liệu đơn nhãn. Phương pháp này dựa trên nòng cốt là Graphormer giúp học được các biểu diễn ngữ nghĩa giữa các lớp, làm giàu thông tin cho đặc trưng trích rút từ CNN backbone. Nhằm giải quyết vấn đề về hiệu năng, đề án sử dụng một phương pháp phân cụm dựa trên vector đại diện lớp (*proxy*) giúp làm giảm không gian tìm kiếm mà vẫn đảm bảo được độ chính xác của kết quả.



Hình 2.1: Kiến trúc tổng thể của mô hình được đề xuất.

Mô hình bao gồm ba thành phần chính: (1) trích xuất đặc trưng hình ảnh bằng mạng CNN để tạo biểu diễn đặc trưng cho từng ảnh; (2) mã hóa ngữ cảnh đồ thị bằng GraphormerEncoder, trong đó mỗi đồ thị gồm một node động (dynamic node) và các node tĩnh (static nodes) đại diện cho các lớp, giúp mô hình hóa cả quan hệ cục bộ và toàn cục giữa các ảnh; và (3) tổ chức không gian embedding thông qua cơ chế proxy-based clustering, nhằm kéo các đặc trưng cùng lớp lại gần nhau và đẩy các lớp khác ra xa nhau.

1. Trích xuất đặc trưng hình ảnh bằng CNN

Tính ưu việt của mạng nơ-ron tích chập (CNN) và các kiến trúc cải tiến như ResNet hay PyramidNet trong việc xây dựng các biểu diễn phân tầng từ dữ liệu thị giác đã được khẳng định qua nhiều nghiên cứu. Điểm mạnh của nhóm mô hình này nằm ở khả năng vận hành song song: vừa ghi nhận chính xác các tín hiệu cục bộ ở mức cơ bản, vừa chắt lọc và tổng hợp chúng thành những đặc trưng ngữ nghĩa có độ trừu tượng cao. Chính vì những đặc tính này, kiến trúc CNN được đề án lựa chọn để đảm nhận vai trò backbone cho nhiệm vụ trích xuất đặc trưng. Đây được xem là bước tiền xử lý then chốt, giúp chuẩn hóa thông tin đầu vào trước khi chuyển giao cho các khối tính toán chuyên sâu hơn.

Chi tiết hơn, quy trình xử lý được bắt đầu bằng việc đưa từng hình ảnh qua mạng CNN backbone nhằm trích xuất vector đặc trưng. Vector này đóng vai trò tổng hợp các thông tin cốt yếu và nội hàm ý nghĩa của dữ liệu thị giác. Kế đến, kỹ thuật chuẩn hóa L2 (L2 normalization) được áp dụng để đồng nhất hóa dữ liệu đầu ra, loại bỏ sự chênh lệch về biên độ giữa các vector, từ đó đảm bảo tính ổn định cho hệ thống. Kết quả thu được là các embedding chứa đựng hàm lượng thông tin cao, đại diện cho những đặc điểm thị giác then chốt, sẵn sàng phục vụ cho công đoạn mô hình hóa tiếp nối bởi Graphormer.

Thông qua chiến lược này, vai trò của CNN không chỉ dừng lại ở việc tạo lập cơ sở hạ tầng cho trích xuất đặc trưng mà còn khai thác tối đa năng lực của các kiến trúc học sâu tiên tiến. Các embedding đầu ra nhờ đó vừa đảm bảo độ sâu ngữ nghĩa, vừa duy trì sự bền vững trước các thay đổi về điều kiện ánh sáng, bố cục hay tỷ lệ vật thể. Đây được xem là bước đệm thiết yếu, tạo môi trường thuận lợi để triển khai các module xử lý kế tiếp trong bài toán tìm kiếm hình ảnh.

2. Biểu diễn đồ thị

Về mặt trực giác, ý tưởng dựa trên giả định rằng mỗi lớp đối tượng trong không gian đặc trưng có thể được biểu diễn bởi một node với một *proxy* đại diện, bao quanh nó là các embedding của các ảnh thuộc lớp tương ứng tạo thành các cụm riêng biệt. Proxy này nằm ở tâm của cụm đóng vai trò như một điểm neo ngữ nghĩa

(semantic anchor), đại diện cho toàn bộ lớp. Dựa trên ý tưởng này, đề án mô hình hóa tập dữ liệu huấn luyện dưới dạng một đồ thị đối tượng, trong đó các node tĩnh (*static nodes*) đại diện cho các lớp cố định, còn node động (*dynamic node*) đại diện cho mẫu ảnh đang xét hoặc ảnh truy vấn. Đặc trưng của mỗi node được biểu diễn bởi vector đặc trưng của ảnh, được trích xuất trực tiếp từ CNN backbone. Quá trình trích xuất này được thực hiện trên toàn bộ ảnh đầu vào, không yêu cầu tách vùng ảnh hay phát hiện đối tượng cục bộ, nhằm đảm bảo tính nhất quán và đơn giản trong biểu diễn đặc trưng, đồng thời phù hợp với nhiệm vụ so khớp đặc trưng. Cấu trúc đồ thị này không chỉ giúp mô hình học được mối tương quan giữa các lớp mà còn cho phép lan truyền thông tin ngữ nghĩa giữa ảnh đầu vào và các đại diện lớp khác nhau, từ đó tạo ra embedding có tính khái quát và phân tách tốt hơn.

Trên cơ sở cấu trúc đồ thị này, mối quan hệ giữa các lớp đối tượng được mã hóa thông qua các cạnh kết nối giữa các node tĩnh, phản ánh mức độ tương đồng hoặc liên quan ngữ nghĩa giữa các lớp trong không gian đặc trưng. Các cạnh này cho phép mô hình khai thác thông tin liên lớp, đặc biệt trong các trường hợp mà ranh giới giữa các lớp có sự chồng lấn hoặc tương đồng cao.

Khi một node động tương ứng với ảnh truy vấn được đưa vào đồ thị, nó được kết nối đồng thời với toàn bộ các node tĩnh. Thông qua cơ chế lan truyền thông tin trên đồ thị, node động không chỉ so khớp trực tiếp với từng đại diện lớp riêng lẻ mà còn gián tiếp tiếp nhận ngữ cảnh ngữ nghĩa toàn cục từ cấu trúc liên kết giữa các lớp. Quá trình này giúp mô hình đánh giá mức độ phù hợp của ảnh truy vấn với từng lớp một cách toàn diện hơn, thay vì chỉ dựa trên khoảng cách cục bộ trong không gian embedding.

Nhờ đó, biểu diễn đặc trưng của node động sau suy luận được điều chỉnh theo cấu trúc ngữ nghĩa của toàn bộ tập lớp, góp phần nâng cao khả năng phân tách giữa các lớp gần nhau và cải thiện tính khái quát của mô hình trong các kịch bản dữ liệu phức tạp hoặc nhiễu.

Cụ thể, xét tập dữ liệu huấn luyện D gồm C lớp đối tượng, với tập nhãn tương ứng $L = \{l_1, l_2, \dots, l_C\}$. Mỗi lớp l_i được biểu diễn bởi một node tĩnh N_i, hình thành nên tập các node tĩnh: $N_s = \{N_1, N_2, \dots, N_C\}$. Mỗi node tĩnh N_i lưu trữ các mẫu ảnh thuộc lớp l_i, và thông qua quá trình huấn luyện, node này dần hội tụ về một biểu diễn đặc trưng trung tâm (*proxy*), phản ánh đặc trưng tổng quát của toàn bộ lớp. Nhờ đó, các node tĩnh mang ý nghĩa phân cụm (clustering representation), trong khi các *proxy* đóng vai trò như tâm của các cụm, cung cấp một đại diện ổn định cho từng lớp trong không gian embedding.

Bên cạnh các node tĩnh, một node động N_0 được định nghĩa sử dụng để biểu diễn ảnh đầu vào trong từng bước huấn luyện hoặc trong pha truy xuất. Node N_0 có thể được xem là một thực thể tạm thời, có nhiệm vụ học cách tương tác và trao đổi thông tin với các node tĩnh thông qua cơ chế chú ý đồ thị (graph attention). Trong giai đoạn huấn luyện, *dynamic node* đóng vai trò học đặc trưng toàn cục bằng cách tổng hợp thông tin từ các *static node* theo trọng số chú ý, phản ánh mức độ tương đồng ngữ nghĩa giữa ảnh hiện tại và từng lớp đối tượng. Trong giai đoạn truy xuất, *dynamic node* được sử dụng để so sánh với các *proxy* nhằm xác định các cụm gần nhất về mặt đặc trưng, qua đó giới hạn phạm vi tìm kiếm trong không gian embedding.

Toàn bộ quá trình trên được mô hình hóa thành một đồ thị có dạng:

$$G = (V, P, E),$$

trong đó:

- $V = \{N_0, N_1, \dots, N_C\}$ là tập các node,
- $P = \{P_1, \dots, P_C\}$ là tập các *proxy* đại diện của các node tương ứng,
- E là tập các cạnh thể hiện mối liên hệ giữa các node.

Mỗi cạnh $e_{ij} \in E$ được thiết lập dựa trên độ tương đồng ngữ nghĩa hoặc đặc trưng giữa hai node $N_i$và N_j, thường được đo bằng cosine similarity, hoặc khoảng cách Euclidean trong không gian biểu diễn, hoặc trọng số trong cơ chế attention. Các

trọng số cạnh này sẽ được cập nhật động trong quá trình huấn luyện thông qua cơ chế self-attention của GraphormerEncoder.

Cấu trúc đồ thị như vậy mang lại hai lợi ích chính. Thứ nhất, nó giúp duy trì mối tương quan ngữ nghĩa toàn cục giữa các lớp thông qua cơ chế truyền thông tin giữa các node, thay vì chỉ học độc lập từng mẫu như các phương pháp CNN truyền thống. Thứ hai, đồ thị phân cụm với các node đại diện lớp cho phép tổ chức lại không gian embedding theo cấu trúc cụm rõ ràng, giúp giảm đáng kể chi phí tìm kiếm trong giai đoạn truy xuất mà không làm suy giảm đáng kể độ chính xác. Nhờ đó, mô hình có thể đạt được sự cân bằng tốt giữa khả năng biểu đạt ngữ nghĩa (semantic expressiveness) và hiệu năng truy xuất (retrieval efficiency).

3. Học biểu diễn bằng Graphormer

Sau khi xây dựng đồ thị $G = (V, E)$ gồm các node tĩnh $\{N_1, N_2, \dots, N_C\}$ và node động N_0 như đã trình bày ở phần trên, bước tiếp theo của mô hình là học biểu diễn đặc trưng ngữ cảnh cho toàn bộ đồ thị thông qua Graphormer.

Khác với các phương pháp dựa trên Graph Attention Network (GAT) [9, 11] chỉ lan truyền thông tin giữa các node lân cận, mô hình sử dụng cơ chế self-attention toàn cục (global self-attention) để mô hình hóa đồng thời mối quan hệ giữa tất cả các node trong đồ thị. Cách tiếp cận này giúp mô hình học được mối tương quan liên cụm toàn cục (inter-cluster correlation), qua đó mỗi node có thể biểu diễn ngữ nghĩa tổng hợp từ toàn bộ không gian đồ thị thay vì chỉ dựa vào kết nối cục bộ.

Cụ thể, với một ảnh đầu vào $I \in N_i$ thuộc lớp l_i, lấy mẫu đúng một ảnh từ mỗi lớp trong tổng số C lớp đối tượng, trong đó ảnh được chọn từ lớp i luôn khác với I. Tập ảnh đầu vào sử dụng để xây dựng đồ thị được định nghĩa như sau:

$$D_I = \{I, I_1, \dots, I_i, \dots, I_C\}.$$

trong đó I_k là ảnh được chọn từ lớp l_k và $I_i \neq I$.

Ảnh truy vấn I được gán vào node động N_0, trong khi các ảnh $I_1, \dots, I_C$ được phân bổ tương ứng vào các node tĩnh $N_1, \dots, N_C$.

Mỗi ảnh trong tập D_I được trích xuất đặc trưng bằng backbone CNN, thu được vector đặc trưng $h_i \in \mathbb{R}^{d_0}$cho từng node $N_i \in V$. Tập hợp đặc trưng của tất cả các node được sắp xếp thành ma trận:

$$H = [h_0, h_1, \dots, h_C]^\top \in \mathbb{R}^{(C+1)\times d_0}.$$

Ma trận H sau đó được đưa vào Graphormer Encoder để thực hiện quá trình học chú ý toàn cục, trong đó mỗi node có thể tương tác trực tiếp với tất cả các node còn lại mà không bị giới hạn bởi cấu trúc kề cận.

Đầu tiên, ma trận H được chiếu vào không gian ẩn bằng một phép biến đổi tuyến tính nhờ ma trận trọng số học được học được $W_L \in \mathbb{R}^{d_0\times d}$, tạo ra ma trận H':

$$H' = HW_L \, \mathbb{R}^{(C+1)\times d}.$$

Ma trận H'sau đó được đưa qua nhiều lớp Transformer Encoder [12], trong đó mỗi lớp attention tính các ma trận truy vấn, khóa và giá trị từ đầu vào như sau:

$$Q = H'W_Q, K = H'W_K, V = H'W_V,$$

với $W_Q, W_K, W_V \in \mathbb{R}^{d\times d}$ là các ma trận trọng số được học.

Cơ chế chú ý toàn cục được định nghĩa bởi:

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^\top}{\sqrt{d}})V.$$

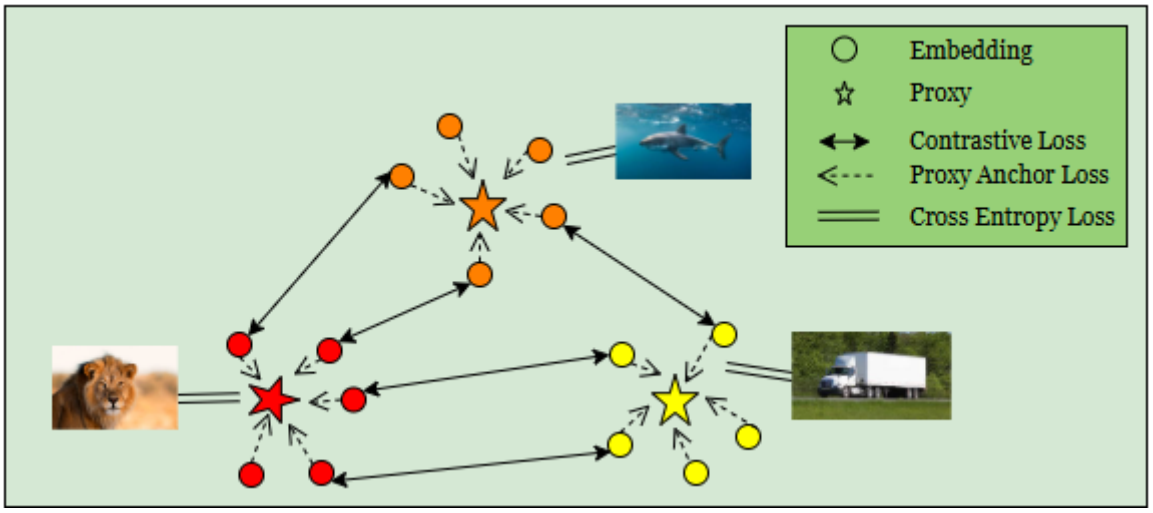
Thông qua cơ chế này, mỗi node có thể "nhìn thấy" và thu nhận thông tin từ toàn bộ các node khác, bao gồm cả những lớp khác cụm. Điều này cho phép mô hình học được phụ thuộc ngữ nghĩa dài hạn (long-range semantic dependency) giữa các đối tượng, điều mà các kiến trúc GAT truyền thống khó đạt được do giới hạn phạm vi kết nối.

Sau khi đi qua nhiều lớp Transformer Encoder, đặc trưng của mỗi node trong đồ thị bao gồm node động N_0và các node tĩnh $N_1, \dots, N_C$ được cập nhật để phản ánh ngữ cảnh từ tất cả các node còn lại. Tiếp theo, các vector đặc trưng của từng node được chuẩn hóa L2, hoàn tất quá trình mã hóa của Graphormer Encoder thu được kết quả cuối cùng là ma trận:

$$h' = [h'_0, h'_1, \dots, h'_C]^\top \in \mathbb{R}^{(C+1)\times d},$$

trong đó h' đại diện cho embedding cập nhật của toàn bộ đồ thị. Mỗi node trong h' mang thông tin ngữ cảnh tổng hợp từ toàn bộ các node còn lại trong đồ thị, phản ánh mối quan hệ giữa các ảnh thuộc các lớp khác nhau trong không gian embedding. Đặc trưng của node động N_0 được cập nhật thành h'_0, mang thông tin ngữ cảnh tổng hợp từ toàn bộ các node tĩnh. Vector h'_0 được xem như đặc trưng ngữ cảnh toàn cục của ảnh, phản ánh mối liên hệ giữa ảnh truy vấn và các lớp đối tượng trong không gian biểu diễn. Đặc trưng này sau đó được sử dụng cho các tác vụ ở pha truy xuất hoặc phân loại, đóng vai trò là embedding cuối cùng của ảnh trong không gian cụm ngữ nghĩa, trong khi các node tĩnh cũng giữ vai trò quan trọng trong việc tính loss và duy trì cấu trúc ngữ nghĩa toàn cục.

4. Huấn luyện và tối ưu hóa mô hình



Hình 2.2: Hoạt động của các hàm mất mát.

Đề án kế thừa ý tưởng kết hợp nhiều hàm mất mát từ những nghiên cứu trước đây [13, 14, 15] để thiết kế một chiến lược đa mất mát, giúp mô hình học được các biểu diễn phong phú và ổn định hơn so với việc chỉ sử dụng một hàm mất mát duy nhất. Mô hình được huấn luyện đa nhiệm dựa trên tổ hợp ba hàm mất mát có mục tiêu bổ trợ lẫn nhau được sử dụng để mô hình đạt được khả năng phân biệt và khái quát tốt hơn trong không gian ngữ nghĩa, gồm: cross-entropy loss cho nhiệm vụ phân

loại, contrastive loss để tăng độ phân tách giữa các cụm khác nhau, và proxy-anchor loss nhằm tăng tính kết dính giữa các mẫu cùng lớp trong không gian biểu diễn.

Proxy-anchor là hàm mất mát chính, đóng vai trò kéo các embedding lại gần với proxy tương ứng (tâm cụm), Contrastive góp phần đẩy các embedding khác lớp ra xa nhau và cross entropy hỗ trợ cho việc phân loại các embedding về đúng lớp của mình.

Thành phần đầu tiên của hàm mất mát là cross-entropy loss, được áp dụng để huấn luyện mô hình dự đoán chính xác nhãn lớp cho từng node trong đồ thị. Sau khi thu được biểu diễn đặc trưng của node h_0 và các node tĩnh h_C, gọi chung là h, mô hình sử dụng một lớp tuyến tính để ánh xạ chúng vào không gian xác suất qua hàm softmax:

$$p(y \mid h) = \text{Softmax}(W_c h + b_c),$$

trong đó W_c và b_c là các tham số học được. Cross-entropy loss được tính như sau:

$$\mathcal{L}_{CE} = -\sum_{i=1}^{C} y_i \log p(y_i \mid h),$$

với y_i là nhãn thật của node tương ứng. Hàm mất mát này hướng dẫn mô hình học cách ánh xạ đặc trưng ảnh vào đúng lớp đối tượng, đóng vai trò nền tảng cho các thành phần mất mát khác.

Mặc dù cross-entropy giúp mô hình phân loại đúng nhãn, nó chưa đủ mạnh để đảm bảo khoảng cách giữa các lớp được duy trì rõ ràng trong không gian biểu diễn. Do đó, contrastive loss được bổ sung nhằm khuyến khích các node thuộc các cụm khác nhau đẩy xa nhau, trong khi các node cùng cụm được kéo lại gần nhau.

Với hai vector đặc trưng h_i và h_j, contrastive loss được định nghĩa như sau:

$$\mathcal{L}_{Con} = \frac{1}{2}\sum_{i,j} \omega_{ij} \begin{cases} D(h_i, h_j)^2, & \text{nếu } y_i = y_j, \\ \max(0, m - D(h_i, h_j))^2, & \text{nếu } y_i \neq y_j, \end{cases}$$

với:

$$\omega_{ij} = 1 + (pos_{weight} - 1) * \frac{\sum_{i,j} 1_{y_i=y_j}}{(C + 1)^2},$$

trong đó:

- h_i, h_j, là các vector đặc trưng của hai node,
- y_i, y_j, là nhãn tương ứng,
- $D(h_i, h_j)$, là khoảng cách Euclidean giữa hai vector,
- m, là ngưỡng biên định trước,
- pos_weight: trọng số của các cặp positive (cùng lớp) để tăng ảnh hưởng của chúng trong quá trình huấn luyện.

Hàm này giúp không gian embedding trở nên có cấu trúc, với ranh giới rõ ràng giữa các cụm ngữ nghĩa, từ đó cải thiện khả năng khái quát của mô hình. Cụ thể, hàm mất mát này học cách đẩy các embedding của các lớp khác nhau ra xa nhau trong không gian đặc trưng, nhằm tạo ra ranh giới rõ ràng giữa các vùng biểu diễn của từng lớp, giúp mô hình phân biệt tốt hơn giữa các khái niệm ngữ nghĩa.

Thành phần thứ ba là proxy-anchor loss [16] , được thiết kế để tăng cường tính tập trung của các điểm ảnh trong cùng một lớp quanh một vector đại diện — gọi là *proxy vector*. Mỗi lớp c được gán một *proxy* $p_c \in \mathbb{R}^d$ học được trong quá trình huấn luyện, đóng vai trò như tâm cụm của lớp đó.

Proxy-anchor loss được định nghĩa như sau:

$$\mathcal{L}_{Proxy} = \frac{1}{|P^+|} \sum_{p_c \in P^+} \log\left(1 + \sum_{h_i \in H_c^+} e^{-\alpha(\mathrm{sim}(h_i, p_c) - m)}\right) + \frac{1}{C} \sum_{p_c \in P} \log\left(1 + \sum_{h_i \in H_c^-} e^{\alpha(\mathrm{sim}(h_i, p_c) + m)}\right)$$

trong đó:

- $\mathrm{sim}(h_i, p_c)$: độ tương đồng cosine giữa embedding $h_i$và proxy p_c.

- P^+: tập các proxy có ít nhất một mẫu cùng lớp trong batch.
- H^+, H^-: lần lượt là tập các embedding cùng lớp và khác lớp với proxy p_c.
- m: margin, xác định biên cách giữa các lớp.
- α: hệ số khuếch đại (scaling factor) điều chỉnh độ sắc nét của phân bố.
- C: tổng số lớp.

Cơ chế này giúp mô hình vừa kéo các mẫu cùng lớp lại gần *proxy* tương ứng, vừa đẩy xa các mẫu khác lớp, từ đó hình thành các cụm rõ ràng trong không gian nhúng.

Ba thành phần trên được kết hợp tuyến tính để hình thành hàm mất mát tổng hợp:

$$\mathcal{L}_{Total} = \lambda_1 \mathcal{L}_{CE} + \lambda_2 \mathcal{L}_{Con} + \lambda_3 \mathcal{L}_{Proxy},$$

với $\lambda_1, \lambda_2, \lambda_3$ là các hệ số điều chỉnh trọng số từng thành phần. Trong quá trình huấn luyện, chiến lược tối ưu hai giai đoạn được sử dụng:

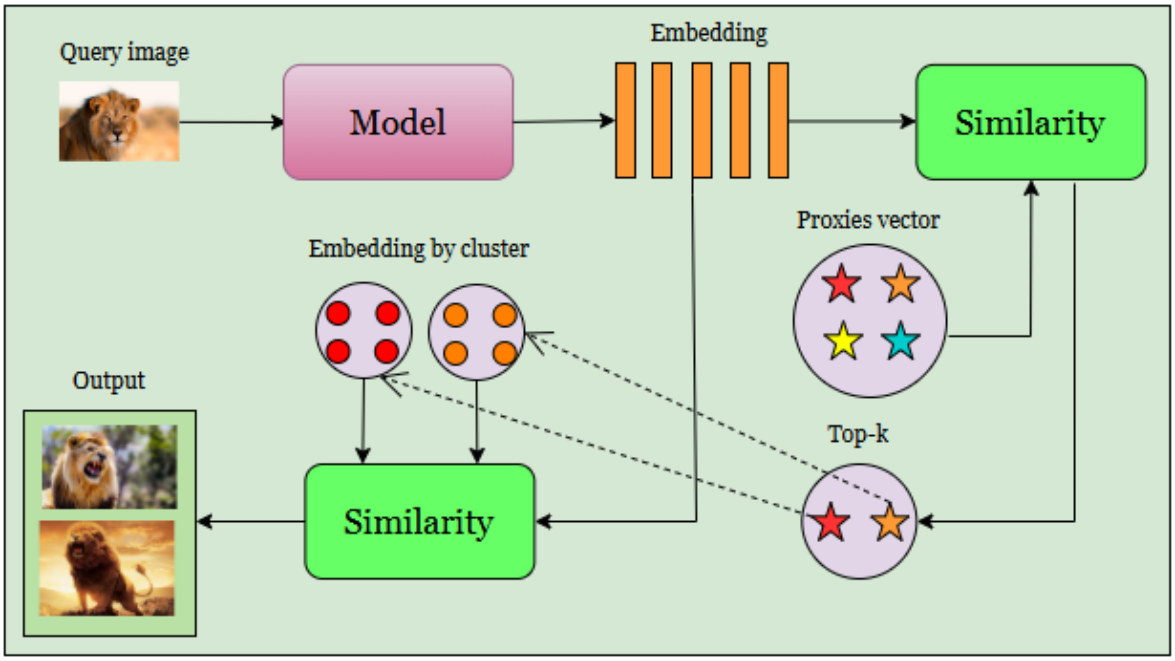
- Ở giai đoạn đầu, trọng số của $\mathcal{L}_{CE}$ được đặt cao hơn để ổn định pha phân loại và giúp mô hình hội tụ nhanh.
- Ở giai đoạn sau, trọng số của $\mathcal{L}_{Con}$và $\mathcal{L}_{Proxy}$ được tăng dần nhằm tối ưu cấu trúc không gian biểu diễn, tạo ra các cụm rõ ràng và bền vững hơn.

Nhờ cơ chế đa hàm mất mát này, mô hình không chỉ học được đặc trưng phân loại chính xác mà còn đạt được biểu diễn ngữ nghĩa có cấu trúc, phản ánh tốt mối quan hệ giữa các lớp và đảm bảo tính ổn định trong quá trình truy xuất ảnh.

5. Truy xuất nhanh dựa trên cụm

Sau khi hoàn tất giai đoạn huấn luyện, pha suy luận (inference phase) được cấu trúc nhằm tối ưu hóa hiệu suất truy xuất mà vẫn bảo toàn độ chính xác ngữ nghĩa toàn cục. Phương pháp này, được gọi là Truy xuất nhanh dựa trên cụm (Proxy-

Cluster Fast Retrieval), tích hợp các đặc trưng đồ thị động cùng các đại diện (*proxy*) của cụm lớp, qua đó giảm thiểu đáng kể độ phức tạp trong tính toán.



Hình 2.3: Sơ đồ hoạt động của pha truy xuất.

Trong pha truy xuất ảnh truy vấn được đưa qua mô hình thu được một embedding, embedding này sẽ được so sánh độ tương đồng với các vector proxy để tìm ra top-k cụm gần nhất, sau đó sẽ thực hiện truy xuất trên tập chứa embedding của k cụm đó.

Cụ thể, với một ảnh truy vấn I, mô hình đầu tiên đưa ảnh qua cùng pipeline trích xuất đặc trưng như trong pha huấn luyện. Đầu tiên, ảnh I được ánh xạ vào node động N_0 của đồ thị. Tại đây, đặc trưng ảnh được trích xuất qua CNN backbone thu được vector $f_I \in \mathbb{R}^d$. Vector này sau đó được đưa qua các tầng Clustered Graph Attention Transformer, tương tác với toàn bộ các node tĩnh $\{N_1, N_2, \ldots, N_C\}$ để cập nhật thành biểu diễn ngữ cảnh đồ thị h_0.

Đặc trưng h_0 này đóng vai trò như đặc trưng ngữ nghĩa toàn cục của ảnh truy vấn, phản ánh mối tương quan giữa ảnh với các lớp đối tượng đã học. Để xác định nhanh các lớp có khả năng chứa ảnh tương đồng, mô hình sử dụng *proxy vectors* đã

học ở pha huấn luyện — mỗi lớp c có một *proxy* $p_c \in \mathbb{R}^d$ đại diện cho trung tâm ngữ nghĩa của lớp đó.

Độ tương đồng giữa ảnh truy vấn và từng *proxy* lớp được tính bằng độ đo cosine:

$$s_c = \frac{h_0 \cdot p_c}{\| h_0 \| \| p_c \|}, c = 1,2,\ldots,C.$$

Các giá trị tương đồng s_c được sắp xếp giảm dần để xác định tập top-k lớp gần nhất với ảnh truy vấn, ký hiệu là $\mathcal{C} = \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k\}$. Bước chọn lọc này đóng vai trò như giai đoạn thu hẹp không gian tìm kiếm, giúp mô hình chỉ cần thực hiện truy xuất chi tiết trên một tập con nhỏ của toàn bộ cơ sở dữ liệu.

Sau khi xác định được các lớp tiềm năng nhất, quá trình truy xuất ảnh chi tiết được thực hiện trên các node tương ứng với $\mathcal{C}$. Mỗi node chứa tập ảnh của một lớp, và các ảnh này đã được biểu diễn dưới dạng đặc trưng f_j được lưu sẵn trong cơ sở dữ liệu. Việc so khớp được thực hiện giữa đặc trưng của ảnh truy vấn h_0 và các đặc trưng f_j trong tập $\mathcal{C}$ bằng độ đo cosine hoặc Euclidean.

Kết quả truy xuất cuối cùng là danh sách ảnh tương tự được sắp xếp theo độ tương đồng giảm dần, trong đó việc giới hạn tìm kiếm trong *top-k* proxy cluster giúp giảm đáng kể độ phức tạp tính toán từ $\mathcal{O}(N)$ xuống còn $\mathcal{O}(N_k)$, với $N_k \ll N$.

Cách tiếp cận PCFR mang lại ba lợi ích chính:

Tốc độ truy xuất cao nhờ giảm không gian tìm kiếm dựa trên thông tin ngữ cảnh toàn cục;

Độ chính xác cao hơn vì *proxy vectors* đã được học sao cho phản ánh trung tâm ngữ nghĩa tối ưu của từng lớp;

Khả năng mở rộng tốt, cho phép hệ thống hoạt động hiệu quả với tập dữ liệu lớn và số lớp đối tượng tăng nhờ phân cụm theo lớp.

Như vậy, PCFR là bước hoàn thiện, cho phép mô hình không chỉ học được biểu diễn ngữ cảnh đồ thị toàn cục mà còn triển khai truy xuất nhanh chóng, chính xác và có khả năng mở rộng trong thực tế.

6. Tổng kết và đánh giá

Kiến trúc đề xuất được thiết lập dựa trên hai nền tảng kỹ thuật cốt lõi. Thứ nhất là cơ chế Học đặc trưng dựa trên bối cảnh đồ thị (Graph-Context Feature Learning). Khác biệt với các phương pháp xử lý ảnh đơn lẻ, mô hình khởi tạo quá trình bằng việc sử dụng backbone CNN để trích xuất đặc trưng cục bộ, sau đó kiến tạo một cấu trúc đồ thị kết hợp giữa nút động (ảnh đầu vào) và các nút tĩnh (đại diện lớp). Cấu trúc này được xử lý qua module Graphformer nhằm khai thác triệt để các liên kết ngữ nghĩa, từ đó tạo ra các vector nhúng (embedding) cho nút động với hàm lượng thông tin ngữ cảnh phong phú hơn.

Thứ hai là chiến lược Truy xuất nhanh dựa trên cụm *proxy* (Proxy-Cluster Fast Retrieval), sử dụng các *proxy* làm đại diện cho từng lớp đối tượng. Quá trình truy vấn bắt đầu bằng việc đối chiếu ảnh với tập *proxy* để xác định *top-k* cụm tiềm năng nhất, giúp thu hẹp không gian tìm kiếm. Giải pháp này tối ưu hóa triệt để tài nguyên tính toán mà vẫn bảo toàn độ chính xác của kết quả.

Để kiểm chứng, mô hình trải qua quy trình đánh giá nghiêm ngặt trên hai tập dữ liệu chuẩn, thực hiện so sánh đối chứng với các công trình tiên tiến nhất về cả tốc độ phản hồi và độ chính xác. Cuối cùng, các thí nghiệm phân tích cắt lớp (ablation study) được tiến hành để xác định rõ đóng góp cụ thể của từng module trong hệ thống.

7. Tiểu kết chương

Nội dung Chương II đã đi sâu phân tích phương pháp luận và kiến trúc hệ thống được kiến nghị. Điểm mấu chốt của đề xuất này nằm ở việc dung hợp mạng CNN với mô hình Graphormer, nhằm mục đích tận dụng triệt để cả đặc tính hình ảnh lẫn các tương quan ngữ nghĩa giữa những phân lớp đối tượng. Quá trình xử lý khởi đầu bằng việc dùng mạng CNN làm backbone để rút trích các tín hiệu thị giác

quan trọng. Tiếp đó, hệ thống tạo một cấu trúc mạng lưới kết hợp các nút động (tương ứng với ảnh đang truy vấn) và các nút tĩnh (đóng vai trò đại diện cho lớp), qua đó chuyển đổi toàn bộ không gian dữ liệu thành dạng đồ thị phân cụm.

Bộ mã hóa Graphormer được triển khai nhằm xây dựng các biểu diễn ngữ cảnh ở phạm vi toàn cục. Thông qua kỹ thuật tự chú ý (self-attention), mỗi nút trong đồ thị có thể giao tiếp trực tiếp với toàn bộ các nút còn lại. Điều này giúp mô hình thâu tóm được các liên kết giữa các nhóm cũng như sự phụ thuộc ngữ nghĩa trong phạm vi rộng, điều mà các mạng GAT thông thường khó thực hiện.

Quá trình huấn luyện sử dụng tổng hợp nhiều hàm mất mát (cross-entropy, contrastive, proxy-anchor) nhằm mục tiêu kép: vừa tối ưu hóa sự tách biệt giữa các nhóm đối tượng, vừa củng cố sự đồng nhất của các vector đặc trưng trong cùng một lớp.

Trong giai đoạn truy xuất, giải pháp "Truy xuất nhanh dựa trên cụm" giúp giới hạn phạm vi tìm kiếm vào danh sách k nhóm triển vọng nhất. Bằng cách sử dụng các vector đại diện (*proxy*) đã được huấn luyện, hệ thống cải thiện đáng kể tốc độ phản hồi mà vẫn bảo toàn độ chính xác. Tóm lại, kiến trúc này vừa tạo ra biểu diễn ngữ cảnh sâu sắc, vừa đảm bảo hiệu năng cao trên dữ liệu lớn.

Chương này đã hoàn thiện quy trình từ bước rút trích thông tin thị giác qua mạng tích chập (CNN), mô hình hóa đồ thị, đến học ngữ cảnh và truy xuất, tạo tiền đề phục vụ quá trình thực nghiệm và đo lường kết quả ở chương sau.

CHƯƠNG 3: THỬ NGHIỆM VÀ ĐÁNH GIÁ

Phần nội dung này tập trung mô tả những thành quả kiểm thử, qua đó đánh giá hiệu năng vận hành của hệ thống truy xuất ảnh được nghiên cứu giới thiệu. Hoạt động đánh giá được thực hiện trên những tập dữ liệu chuẩn hóa, song song với việc so sánh kết quả định lượng cùng các mô hình hiện đại và tiến hành phân tích cắt lớp (ablation study) để xác định mức độ ảnh hưởng của mỗi khối chức năng trong kiến trúc tổng thể.

1. Mục tiêu thực nghiệm

Hệ thống các thử nghiệm được thiết kế trong chương này nhằm mục đích kiểm chứng sâu rộng năng lực của giải pháp tra cứu được xây dựng. Bước đầu tiên là đặt mô hình lên bàn cân so sánh với những kỹ thuật tiên tiến để xác minh khả năng mã hóa đặc trưng trong không gian ngữ nghĩa, qua đó làm rõ tính phân tách giữa các nhóm đối tượng. Song song với đó, chiến lược tìm kiếm tốc độ cao thông qua phân cụm cũng được xem xét kỹ lưỡng, nhằm khẳng định xem việc tổ chức dữ liệu dạng cụm có thực sự tối ưu hóa thời gian phản hồi mà không đánh đổi sự chuẩn xác hay không. Tiếp đến, tiến hành trực quan hóa các vector nhúng để quan sát rõ hơn sự phân bố và ranh giới giữa các nhãn lớp. Vượt lên trên các số liệu thống kê thuần túy, mô hình được đưa vào vận hành trong kịch bản thực tế—thực hiện truy vấn và trình bày các kết quả khớp nhất—nhằm kiểm định chất lượng thông qua cảm quan thị giác. Cuối cùng, phương pháp phân tích cắt lớp (ablation study) được áp dụng để xác định rõ mức độ ảnh hưởng của từng module chức năng đối với hiệu suất toàn hệ thống.

2. Chi tiết cài đặt

Nền tảng thực nghiệm được vận hành trên hệ điều hành Ubuntu 24.04 LTS, tận dụng Python 3.10, framework PyTorch 2.1 cùng các thư viện phụ trợ đi kèm.

Quá trình đào tạo mô hình tập trung vào việc tối ưu hóa tổ hợp ba hàm tổn thất: Cross-Entropy Loss, Proxy-Anchor Loss và Contrastive Loss. Việc huấn luyện diễn ra trong 100 epoch sử dụng thuật toán AdamW [17] với learning rate (tốc độ

học) khởi điểm ở mức 5×10^{-5} và hệ số suy giảm trọng số (weight decay) là 1×10^{-4}; giá trị learning rate sẽ được điều chỉnh giảm 50% sau chu kỳ mỗi 10 epoch.

Về quy trình tạo dữ liệu, mỗi batch trong quá trình huấn luyện được xây dựng bằng cách lấy ngẫu nhiên các mẫu từ tập train để đóng vai trò *dynamic nodes*. Với mỗi *dynamic node*, mô hình tiếp tục lựa chọn một tập ảnh bổ trợ từ dữ liệu huấn luyện làm *static nodes*, nhằm cung cấp ngữ cảnh tham chiếu cần thiết.

Đối với giai đoạn kiểm thử, tập gallery là sự tổng hợp của các *dynamic nodes* (từ cả tập train và test) cùng với *static nodes* lấy từ tập train. trong khi đó, ảnh truy vấn được tạo từ các *dynamic nodes* của tập kiểm tra. Tất cả các truy vấn được đảm bảo không xuất hiện trong gallery.

Sau cùng, hiệu quả của giải pháp được định lượng thông qua thước đo mAP, đại diện cho khả năng truy xuất chính xác và mức độ tách biệt của không gian embedding mà mô hình học được.

3. Tập dữ liệu

Để đánh giá năng lực của mô hình trong các kịch bản phân loại với mức độ phức tạp khác nhau, tiến hành thử nghiệm trên hai bộ dữ liệu chuẩn CIFAR-100 [1] và CUB-200-2011 [2].

Đối với CIFAR-100, tập dữ liệu này cung cấp tài nguyên gồm 60.000 hình ảnh màu (độ phân giải 32×32), được phân bổ đều vào 100 nhóm đối tượng riêng biệt. Mỗi nhóm chứa 600 mẫu, tuân theo tỷ lệ phân chia 500 mẫu cho huấn luyện và 100 mẫu cho kiểm thử. Sự phong phú về chủ đề—trải dài từ động vật, phương tiện giao thông đến các vật dụng đời sống—biến CIFAR-100 thành thước đo lý tưởng nhằm kiểm tra khả năng khái quát hóa của mô hình và nhận diện đa lớp của hệ thống.

Trong khi đó, CUB-200-2011 (Caltech-UCSD Birds) lại đặt ra bài toán nhận diện độ mịn cao (fine-grained) với 11.788 bức ảnh thuộc 200 chủng loại chim. Đặc thù của dữ liệu này là các lớp có ngoại hình rất sát nhau, do đó đặt ra yêu cầu khắt

khe buộc mô hình phải trích xuất được những đặc trưng cực kỳ tinh tế mới có thể phân tách chính xác giữa các loài.

4. Đánh giá hiệu quả mô hình

Hiệu năng của mô hình được đánh giá dựa trên các thước đo chuẩn trong lĩnh vực truy xuất hình ảnh, bao gồm mAP, Precision@K và Recall@K. Trong đó, mAP được xem như chỉ số then chốt phản ánh độ chính xác tổng thể và khả năng phân biệt giữa các vector đặc trưng trong không gian ngữ nghĩa.

Diễn giải kỹ hơn, Precision@K định lượng mật độ các kết quả đúng nhãn xuất hiện trong nhóm K phần tử đầu tiên, còn Recall@K cho biết tỷ lệ bao phủ của các ảnh tìm được trên tổng số dữ liệu liên quan hiện có. Cơ chế so khớp độ tương đồng giữa các vector nhúng được thực hiện dựa trên hàm Cosine Similarity; đây là nền tảng cốt lõi để tính toán các giá trị nêu trên, qua đó giúp thẩm định chất lượng biểu diễn ngữ nghĩa, khả năng phân loại đặc trưng và hiệu suất tra cứu thực tế của giải pháp đề xuất.

5. Phân tích kết quả

Nhằm đánh giá độ chính xác của mô hình, tiến hành thử nghiệm benchmark trên tập dữ liệu CIFAR-100, sử dụng backbone PyramidNet-272 [7], batch size 128 và kích thước embedding 256 để kiểm tra khả năng phân loại đa lớp. Kết quả thu được được trình bày trong bảng dưới đây:

Bảng 3.1: So sánh hiệu suất trên tập dữ liệu CIFAR-100.

Phương pháp	mAP	R@1	R@10	P@1	P@10	P@20	P@100
CCO [18]	-	-	-	64.20	-	-	-
CAM [19]	80.8	-	-	-	-	**79.4**	79.5
RM [20]	~50.10	~64.20	-	-	-	-	-
MaskCon [21]	-	65.52	89.25	-	-	-	-
FGVR [22]	-	73.83	**93.24**	-	-	-	-
Đề xuất	**80.94**	**80.06**	85.14	**78.06**	**78.98**	79.28	**79.78**

Kết quả trong Bảng 3.1 cho thấy mô hình đạt hiệu suất cạnh tranh trên CIFAR-100 so với các phương pháp hiện có. Về chỉ số mAP, mô hình đạt 80.94, vượt qua CAM (80.8) và cao hơn đáng kể so với RM (~50.10). Các chỉ số Precision@K (P@K) cho thấy tính ổn định về độ chính xác của mô hình. Mô hình duy trì một độ chính xác cực kỳ ổn định và ở mức cao khi K tăng lên: 78.98 (tại P@10), 79.28 (tại P@20), và 79.78 (tại P@100). Các kết quả này đều rất cạnh tranh, thậm chí cao hơn CAM (79.5 tại P@100). Điều này có nghĩa là mô hình không chỉ xếp hạng đúng ở top đầu, mà còn trả về một danh sách kết quả ít nhiễu ngay cả khi K tăng lên. Hiệu suất phân biệt đa lớp (multi-class discrimination) tốt này là do thay vì xử lý các lớp một cách độc lập, mô hình học được ngữ cảnh và mối quan hệ tương quan giữa các đặc trưng ảnh (dynamic nodes) và các *proxy* đại diện cho lớp (static nodes). Cơ chế attention cho phép mô hình so sánh và phân biệt một đặc trưng ảnh với nhiều đại diện lớp cùng một lúc, giúp nó hiểu rõ hơn về ranh giới giữa các lớp (inter-class boundaries). Các kết quả này chứng minh rằng mô hình đề xuất có khả năng học các embedding mang tính phân biệt cao và đạt hiệu quả truy hồi đa lớp chính xác trên CIFAR-100.

Tiếp theo, nhằm đánh giá khả năng học các đặc trưng tinh vi và phân biệt các lớp fine-grained của mô hình một thử nghiệm benchmark được thực hiện trên CUB-200-2011. Đối với backbone, mô hình sử dụng mô hình BiT-L [23] (một ResNetV2-152×2 [24]) được pre-train trên ImageNet-21k, với batch size 64 và embedding dimension là 512:

Bảng 3.2: So sánh hiệu suất trên tập dữ liệu CUB-200.

Phương thức	mAP	R@1	R@2	R@4
DFML-V-B [25]	54.0	79.1	86.8	-
LOCORE-base [26]	64.8	78.3	**91.9**	**98.2**
Proxy-ISA [27]	26.8	68.1	-	-
LCR [28]	24.01	49.69	-	-
KAE-Net [29]	70.89	74.2	83.3	89.1

HSE-PA [30]	-	70.6	80.1	87.1
R-MAC [31]	-	56.7	68.4	79.0
Đề xuất	**86.05**	**84.00**	88.23	90.96

Trên benchmark CUB-200-2011, mô hình tiếp tục cho thấy hiệu suất cạnh tranh cao. Về chỉ số mAP, mô hình đạt 86.05, cao hơn +15.16 so với kết quả được báo cáo bởi KAE-Net (70.89). Về chỉ số R@1, mô hình đạt 84.00, cao hơn +4.9 so với DFML-V-B (79.1). Mặc dù LOCORE-base cho thấy R@2 và R@4 cao, các chỉ số mAP (64.8) và R@1 (78.3) của phương pháp này lại thấp hơn đáng kể. Sự khác biệt này do mô hình sử dụng một phương pháp truy xuất dựa trên cụm (cluster-based retrieval), được thiết kế với mục tiêu chính là tối ưu hóa mAP. Mô hình tập trung gom các ảnh cùng lớp vào các cụm cô đặc, giúp các ảnh liên quan nhất (khi tìm đúng cụm) được trả về với thứ hạng rất cao, từ đó đẩy mAP lên cao. Tuy nhiên, cách tiếp cận này có một sự đánh đổi. Hiệu suất phụ thuộc vào việc xác định đúng cụm; nếu xác định không chính xác, mô hình có thể bỏ lỡ các ảnh đúng. Điều này lý giải tại sao các phương pháp khác, như LOCORE, có thể đạt R@K cao hơn ở $K = 2$ và $K = 4$. Các phương pháp đó có thể có không gian tìm kiếm phẳng hơn, cho phép việc tăng K sẽ mở rộng sang các lớp lân cận và truy hồi được các ảnh đúng. Dù vậy, cách tiếp cận này phải đánh đổi bằng một chỉ số mAP (64.8) thấp hơn đáng kể.

Kết quả tổng thể cho thấy mô hình hoạt động hiệu quả trên cả hai tập dữ liệu. Hiệu suất này có thể được lý giải bởi cơ chế của mô hình, vốn tập trung vào việc gom các embedding thuộc cùng một lớp lại gần nhau trong không gian đặc trưng. Cách tiếp cận này được thiết kế để tối ưu hóa Độ chính xác Trung bình (mAP), và dù điều này có thể dẫn đến một sự đánh đổi nhỏ về recall ở các giá trị K lớn hơn, hiệu suất tổng thể vẫn mang tính cạnh tranh cao so với các phương pháp tiên tiến khác. Những phát hiện này cho thấy việc học ngữ cảnh dựa trên đồ thị có thể biểu diễn hiệu quả các mối quan hệ liên lớp (inter-class relationships), qua đó cải thiện chất lượng biểu diễn cho bài toán truy xuất ảnh chi tiết cao. Hiệu quả trên các tập dữ liệu khác nhau cũng cho thấy tính tổng quát hóa và độ ổn định tốt của mô hình.

Để đánh giá hiệu quả của phương pháp truy xuất dựa trên cụm proxy (proxy-cluster retrieval), một thí nghiệm được thực hiện bằng cách thay đổi số lượng cụm k được chọn để tìm kiếm sau khi xác định các proxy gần nhất. Thay vì đo thời gian trực tiếp, mô hình tập trung phân tích sự đánh đổi giữa độ chính xác (mAP, R@K) và mức độ thu hẹp không gian tìm kiếm (là yếu tố quyết định tốc độ về mặt lý thuyết).

Tiếp theo, một thí nghiệm so sánh ba test case trên bộ dữ liệu CIFAR-100 và CUB-200-2011 (với gallery dataset là toàn bộ tập dữ liệu và C là tổng số lớp của tập dữ liệu):

- $k = C$ (Full Search): Tìm kiếm trên toàn bộ cơ sở dữ liệu.
- $k = 3$: Chỉ tìm kiếm trong 3 cụm gần nhất.
- $k = 1$: Chỉ tìm kiếm trong 1 cụm gần nhất.

Tập gallery dataset là toàn bộ cơ sở dữ liệu, không bao gồm ảnh query. Kết quả được trình bày trong Bảng 3.3 và bảng 3.4.

Bảng 3.3: Bảng so sánh kết quả khi thực hiện truy xuất trong không gian k cụm trên tập dữ liệu CIFAR-100.

k cluster	mAP @1	mAP @10	mAP @20	mAP @50	mAP @100	mAP	R@1	R@10
1	**80.09**	80.09	80.09	80.09	80.09	80.09	**80.09**	80.09
3	79.24	**80.90**	**81.01**	**81.23**	**81.38**	80.52	79.63	85.78
$100 = C$	78.29	80.42	80.54	80.84	81.07	**80.78**	79.44	**87.26**

Bảng 3.4: Bảng so sánh kết quả khi thực hiện truy xuất trong không gian k cụm trên tập dữ liệu CUB-200.

k cluster	mAP @1	mAP @2	mAP @4	mAP @8	mAP	R@1	R@2	R@4	R@8
1	**85.80**	85.80	85.80	85.80	**85.80**	**85.80**	85.80	85.80	85.80
3	84.90	**86.44**	**86.71**	**86.51**	85.79	84.90	87.99	89.70	90.94
$200 = C$	84.09	86.15	86.54	86.27	85.18	84.09	**88.21**	**90.63**	**92.39**

Thu hẹp không gian tìm kiếm giúp tăng tốc độ truy xuất: Việc giảm k (từ C xuống 3 hoặc 1) làm giảm đáng kể số lượng vector cần so sánh trong bước tìm kiếm. Bằng cách bỏ qua những cụm không liên quan vốn tốn kém chi phí thậm chí còn gây nhiễu, mô hình chỉ quan tâm tới những cụm có khả năng gần với ảnh truy vấn nhất. Mặc dù có thêm chi phí nhỏ ban đầu để so sánh với các *proxies*, nhưng nó là rất nhỏ so với chi phí tính toán với full search. Hơn nữa độ chính xác của mô hình vẫn được đảm bảo. Điều này cho thấy việc thu hẹp không gian tìm kiếm giúp làm tăng hiệu suất đáng kể mà không ảnh hưởng đến kết quả truy xuất.

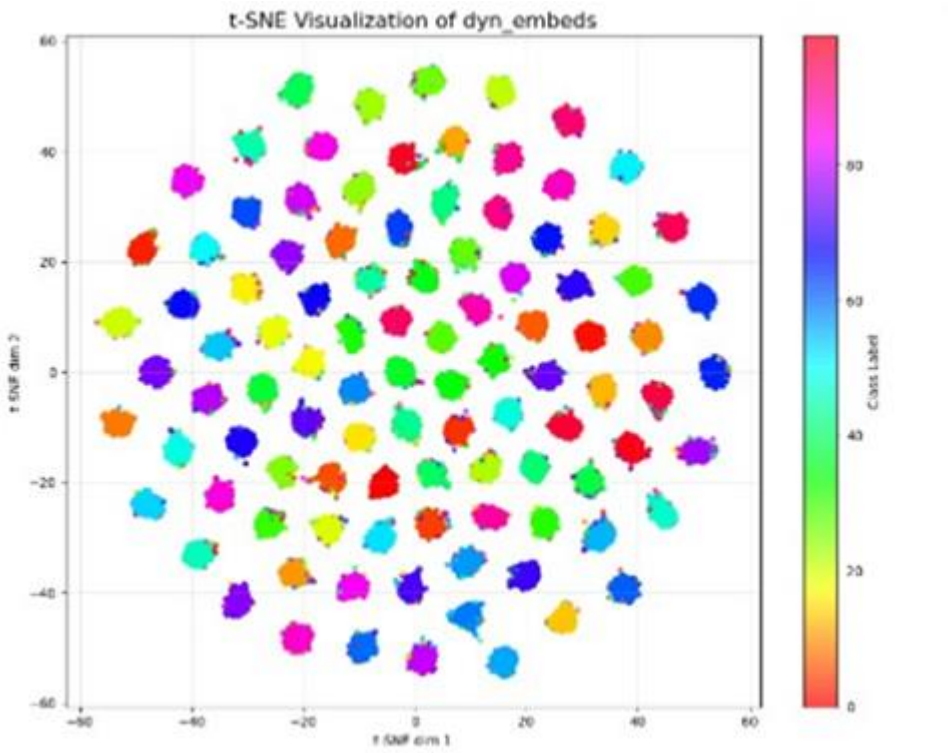
Với k = 1 độ chính xác nhìn chung giảm đáng kể. Điều này là do một số mẫu vẫn có khả năng bị gán sai cụm (đặc biệt với các lớp có đặc trưng ở vùng biên của cụm sẽ sát với đặc trưng của lớp khác hay là những đặc trưng không được đưa lại gần cụm chính xác). Từ đó cho thấy khi chỉ tìm trong 1 cụm đã bỏ lỡ nhiều kết quả liên quan nằm ở các cụm lân cận.

Với k = 3 giúp bảo toàn được độ chính xác. Từ thực nghiệm cho thấy kết quả của việc truy xuất trên 3 cụm cho kết quả rất sát với việc truy xuất toàn bộ gallery:

- mAP tổng thể: Chỉ giảm một lượng rất nhỏ. Mức giảm này gần như không đáng kể mà lại có thể giảm rất nhiều chi phí tính toán.
- mAP@K: Tại các ngưỡng K lớn hơn 1, mAP@K của k = 3 thậm chí còn cao hơn so với $k = C$. Điều này có thể do việc loại bỏ các cụm xa đã giúp giảm nhiễu.
- Recall@K: Cũng chỉ giảm nhẹ, không đáng kể.

Việc chỉ tìm kiếm trong 3 cụm gần nhất đã bảo toàn gần như toàn bộ hiệu suất truy xuất so với tìm kiếm toàn bộ. Điều này chứng tỏ không gian embedding được tổ chức rất tốt, các ảnh liên quan gần như luôn nằm trong 3 cụm gần nhất. Việc chọn k = 3 là điểm cân bằng tối ưu. Nó mang lại lợi ích tăng tốc độ đáng kể (do giảm mạnh không gian tìm kiếm) trong khi không bị giảm đáng kể về độ chính xác. Điều này cho thấy hiệu quả của phương pháp truy xuất dựa trên cụm *proxy*.

Để hiểu rõ hơn tại sao phương pháp truy xuất top-*k* cụm của mô hình hoạt động hiệu quả, một thử nghiệm trực quan hóa không gian embedding mà mô hình đã học được đã được tiến hành. Thử nghiệm thực hiện trực quan các vector đặc trưng (embeddings) của toàn bộ tập dữ liệu CIFAR-100. Mỗi điểm trên biểu đồ đại diện cho một ảnh, và màu sắc của điểm đại diện cho lớp thực tế của ảnh đó, theo thang màu bên phải.



Hình 3.1: Trực quan hóa các embedding được học bởi mô hình trên tập dữ liệu CIFAR-100.

Mỗi điểm đại diện cho một ảnh, và màu sắc biểu thị cho lớp thực tế (ground-truth). Hình ảnh minh họa cách mô hình tổ chức các ảnh tương đồng về ngữ nghĩa (semantically similar) thành các cụm riêng biệt. Hình ảnh cho thấy một cách trực quan chất lượng của không gian embedding do mô hình tạo ra:

- Các embeddings được tổ chức thành các cụm rõ rệt, những embedding cùng lớp được phân bổ sát nhau xoay quanh một tâm cụm chính là *vector proxy* đại diện cho cụm đó.

- Các điểm cùng màu hội tụ rất chặt chẽ vào trong cụm của chúng. Điều này cho thấy mô hình đã học được cách kéo các ảnh có cùng ngữ nghĩa lại gần nhau một cách hiệu quả.

- Các cụm khác nhau được tách biệt rõ ràng với nhau, có khoảng cách đáng kể giữa chúng tránh cho việc bị lẫn lộn giữa các cụm. Điều này chứng tỏ mô hình cũng thành công trong việc đẩy các ảnh có ngữ nghĩa khác nhau ra xa nhau.

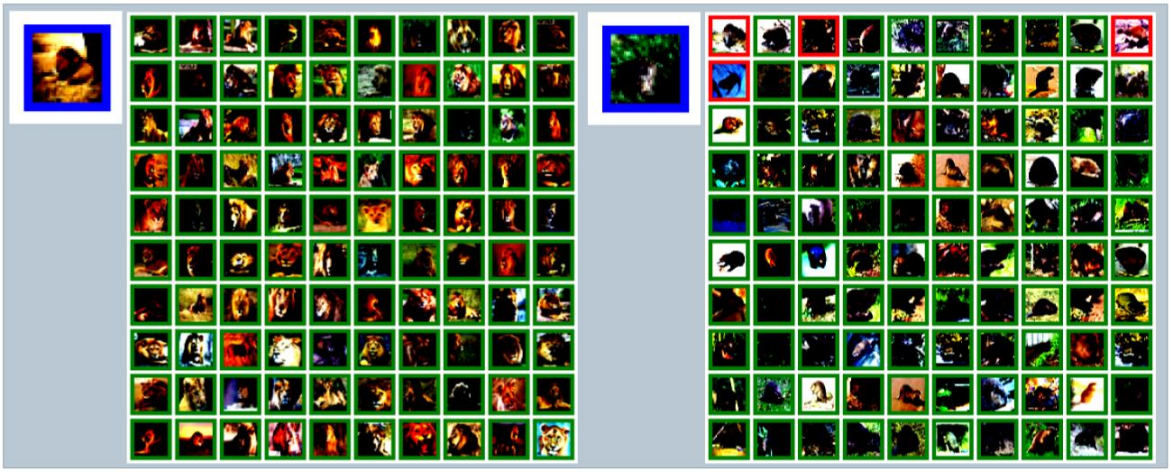
Cấu trúc cụm dày đặc, tách biệt rõ này minh chứng cho sự hiệu quả của mô hình. Lợi ích của việc tổ chức không gian embedding theo cấu trúc cụm:

- Các *proxy* có thể dễ dàng đại diện cho từng lớp.

- Ảnh truy vấn khi được chiếu vào không gian này sẽ rơi vào gần *proxy* của lớp chính xác.

- Việc chỉ cần tìm kiếm trong k cụm gần nhất ($k \ll C$) là đủ để bao phủ các ảnh liên quan giúp bảo toàn độ chính xác trong khi tăng tốc độ đáng kể.

Để cung cấp cái nhìn trực quan hơn về hiệu suất của mô hình, bên cạnh các chỉ số định lượng đã báo cáo, một số ví dụ truy vấn định tính được trình bày trong Hình 3.2 và Hình 3.3. Ảnh truy vấn được đánh dấu bằng viền màu xanh lam. Các kết quả truy xuất đúng (cùng lớp) được hiển thị với viền màu xanh lá, trong khi các kết quả sai (khác lớp) được hiển thị với viền màu đỏ.



Hình 3.2: Trực quan hóa truy xuất trên tập dữ liệu CIFAR-100.

Hình ảnh minh họa hai trường hợp truy vấn đại diện:

- Truy vấn lớp sư tử: Với một ảnh truy vấn thuộc lớp sư tử, mô hình truy xuất 100 kết quả hàng đầu, tất cả (100/100) đều thuộc lớp sư tử đúng.
- Truy vấn lớp hải ly: Với một ảnh truy vấn thuộc lớp hải ly, mô hình truy xuất 100 kết quả hàng đầu, trong đó 96 kết quả đúng và 4 kết quả sai được đánh dấu màu đỏ.

Các kết quả này minh họa khả năng của mô hình trong việc tạo embedding tập trung chặt chẽ và đạt độ Precision@100 cao trên nhiều lớp khác nhau.



Hình 3.3: Trực quan hóa truy xuất trên tập dữ liệu CUB-200-2011.

Hình ảnh minh họa hai trường hợp truy vấn đại diện với độ chi tiết cao, trong đó sự khác biệt giữa các loài được xác định bởi các thuộc tính tinh vi:

- Truy vấn lớp Parakeet Auklet: Mô hình truy xuất 30 kết quả hàng đầu, tất cả (30/30) đều thuộc loài chính xác.
- Truy vấn lớp Orange-crowned Warbler: Mô hình truy xuất 30 kết quả hàng đầu, trong đó 29/30 ảnh thuộc loài chính xác và 1 ảnh sai.

Các kết quả này chứng minh khả năng của mô hình trong việc học và nắm bắt các đặc trưng tinh vi, cho phép mô hình phân biệt chính xác các loài có hình thức tương tự nhau.

Một quá trình phân tích chuyên sâu đã được thực hiện trên tập dữ liệu CIFAR-100 nhằm làm sáng tỏ vai trò của từng yếu tố trong mô hình đề xuất. Thử nghiệm

bao gồm ba kịch bản biến thể riêng biệt, tất cả đều được vận hành trên nền tảng backbone ResNet-50 kinh điển [6]:

- CNN → FC: Sử dụng backbone CNN theo sau bởi lớp Fully Connected để sinh vector đặc trưng.
- CNN → Graph → FC: Sử dụng backbone CNN, sau đó xây dựng một đồ thị tương tự như mô hình đề xuất và đưa qua lớp Fully Connected.
- CNN → Graph → Graphormer: Phiên bản đầy đủ của mô hình được đề xuất.

Kết quả được trình bày trong bảng 3.5:

Bảng 3.5: So sánh kết quả của theo từng thành phần của mô hình.

Phương thức	mAP	P@R
$CNN \rightarrow FC$	55.95	44.35
$CNN \rightarrow Graph \rightarrow FC$	55.96	44.36
$CNN \rightarrow Graph \rightarrow Graphormer$	**67.53**	**64.94**

Phân tích số liệu thực nghiệm dẫn đến những nhận định sau:

- Khi vận hành đơn lẻ ResNet-50 kết hợp với lớp Fully Connected (FC), chỉ số mAP dừng lại ở mức khiêm tốn (55.95). Kết quả này là do mỗi ảnh được mã hóa một cách độc lập, và lớp Fully Connected chỉ đóng vai trò ánh xạ tuyến tính từ không gian đặc trưng sang không gian nhãn. Cách tiếp cận này giả định rằng đặc trưng cục bộ của từng ảnh đã đủ phân biệt, trong khi trên thực tế các ảnh thường chia sẻ những mẫu hình thị giác tương đồng giữa các lớp hoặc tồn tại nhiễu do điều kiện chụp, góc nhìn và bối cảnh. Do đó, embedding thu được có xu hướng chồng lấn giữa các lớp gần nhau, làm suy giảm khả năng truy hồi và xếp hạng.
- Việc chuyển đổi sang kiến trúc $Graph \rightarrow FC$ cũng không tạo ra sự đột phá, phản ánh thực tế rằng hệ thống vẫn đang xử lý các đặc trưng một cách rời rạc, chưa khai thác được sợi dây liên kết ngữ cảnh giữa các nút trong mạng lưới. Mặc dù cấu trúc đồ thị đã được hình thành, nhưng việc các đặc trưng sau đồ thị vẫn được

đưa trực tiếp qua lớp Fully Connected khiến quá trình suy luận cuối cùng diễn ra trên từng nút một cách độc lập. Lớp FC không có cơ chế trao đổi hay lan truyền thông tin giữa các nút, do đó các quan hệ đã được mã hóa ở mức cấu trúc đồ thị không được bảo toàn và khai thác hiệu quả trong không gian embedding. Nói cách khác, đồ thị trong kiến trúc này chỉ đóng vai trò như một phép biến đổi trung gian, trong khi quyết định biểu diễn cuối cùng vẫn mang tính cục bộ, khiến mô hình về bản chất không khác biệt đáng kể so với các phương pháp không sử dụng đồ thị.

- Ngược lại, kiến trúc hoàn chỉnh ghi nhận sự bứt phá mạnh mẽ với mAP đạt 67.53, áp đảo hoàn toàn hai phiên bản rút gọn. Điều này là bằng chứng thép cho thấy việc tích hợp thông tin từ các nút lân cận đã làm giàu đáng kể ngữ nghĩa cho biểu diễn. Qua đó, vai trò then chốt của việc phối hợp Graphormer trong việc thâu tóm các quan hệ toàn cục và liên kết đặc trưng được khẳng định rõ rệt.

- Một điểm nhấn khác là chỉ số Precision@R đạt 64.94, bỏ xa mức ~44.35 của các biến thể nền tảng. Kết quả này phản ánh chất lượng embedding vượt trội với khả năng quy tụ các mẫu cùng loại thành các cụm mật thiết và tách biệt rạch ròi với các lớp khác. Nói cách khác, khả năng khử nhiễu và gom nhóm của mô hình đã được tối ưu hóa, giúp đẩy lùi các mẫu sai lệch ra khỏi không gian đặc trưng của truy vấn.

Để tiếp tục đánh giá hiệu quả của mô hình trong việc tổ chức cụm, tiếp tục so sánh kết quả của mô hình này với phương pháp $CNN \rightarrow Transformer$ kết hợp học metric bằng Proxy Anchor Loss [16]. Cả 2 đều được thực hiện truy xuất trên 3 cụm, kết quả được trình bày trong Bảng 3.6:

Bảng 3.6: So sánh giữa mô hình đề xuất và phương pháp học Proxy Anchor truyền thống.

Phương thức	mAP	P@R
$CNN \rightarrow Transformer$	66.67	60.44
$CNN \rightarrow Graph \rightarrow Graphormer$	**67.61**	**64.94**

Các số liệu thống kê tại Bảng 3.6 minh chứng rằng kiến trúc hoàn chỉnh vượt trội hơn hẳn so với kỹ thuật Proxy Anchor cổ điển trên cả hai thước đo mAP và Precision@R. Thành tựu này bắt nguồn từ hai trụ cột kỹ thuật: (1) việc mã hóa ngữ cảnh bằng cấu trúc đồ thị giúp tăng cường độ giàu thông tin cho đặc trưng, tận dụng triệt để sự tương đồng nội lớp cũng như tính phân tách ngoại lớp; và (2) một chiến lược tối ưu hóa theo hướng đa mục tiêu, trong đó hàm mất mát phụ góp phần củng cố và bổ trợ đáng kể cho Proxy Anchor, giúp quy hoạch không gian đặc trưng khoa học hơn, tăng cường tính phân loại và đảm bảo quá trình hội tụ diễn ra bền vững. Hiệu năng thấp hơn của kiến trúc CNN → Transformer cho thấy rằng việc áp dụng cơ chế self-attention thuần túy trên chuỗi đặc trưng ảnh chưa đủ để mô hình hóa hiệu quả quan hệ cấu trúc giữa các mẫu. Transformer giả định đầu vào có dạng chuỗi tuyến tính và phụ thuộc mạnh vào positional encoding, trong khi các mối quan hệ giữa các ảnh trong bài toán này mang tính phi tuyến và phân tán theo cụm, khó có thể biểu diễn đầy đủ bằng thứ tự tuần tự. Do thiếu ràng buộc cấu trúc tường minh, cơ chế attention có xu hướng phân bổ trọng số dàn trải, làm suy giảm khả năng nhấn mạnh các liên kết ngữ nghĩa quan trọng giữa các mẫu tương đồng.

Tổng kết lại, phân tích cắt lớp (ablation study) khẳng định rằng mỗi mắt xích trong hệ thống đề xuất đều giữ vị trí then chốt đối với việc tối ưu hóa hiệu suất tìm kiếm. Sự hiện diện của tầng Graph, và cụ thể là module Graphormer, hỗ trợ mô hình hiểu và duy trì liên kết ngữ cảnh giữa các mẫu dữ liệu, góp phần gia tăng độ chính xác và mật độ phân nhóm trong không gian biểu diễn. Thực tế này xác thực rằng sự lai ghép giữa học biểu diễn đồ thị và Transformer mang lại ưu thế vượt bậc và lợi ích thiết thực so với các mạng CNN đơn thuần.

6. Tiểu kết chương

Chương này mô tả chi tiết quy trình thực nghiệm và đánh giá hiệu năng của mô hình trong bài toán truy xuất hình ảnh. Các thử nghiệm được tiến hành trên hai bộ dữ liệu chuẩn CIFAR-100 và CUB-200-2011, mô hình đã chứng minh được khả năng học các embedding phân biệt cao, tổ chức không gian đặc trưng thành các cụm rõ ràng và tối ưu hóa hiệu suất truy xuất.

Các số liệu định lượng chứng minh mô hình đạt mức mAP ấn tượng đồng thời duy trì hiệu năng ổn định của Precision@K và Recall@K, cả trên bài toán nhận dạng đa lớp lẫn phân loại tinh vi (fine-grained). Thực nghiệm về truy xuất dựa trên k cụm đã minh chứng hiệu quả của việc thu hẹp không gian tìm kiếm, vừa tăng tốc độ vừa giữ vững độ chính xác. Trực quan hóa embedding và các kết quả truy vấn cụ thể cũng xác nhận khả năng gom nhóm chặt chẽ và phân tách rõ ràng giữa các lớp, đồng thời thể hiện tiềm năng mạnh mẽ của cơ chế attention cùng phương pháp học biểu diễn đồ thị.

Các phân tích cắt lớp (ablation study) minh chứng tầm quan trọng của từng module trong mô hình, đặc biệt là vai trò then chốt của tầng Graphormer trong việc cải thiện chất lượng embedding, tăng độ chính xác và khả năng phân nhóm. So với các phương pháp truyền thống như Proxy Anchor hay $CNN \rightarrow Transformer$, mô hình cho thấy hiệu suất vượt trội nhờ sự kết hợp giữa học biểu diễn dựa trên đồ thị và chiến lược tối ưu hóa đa mục tiêu.

Tổng quan, kết quả thực nghiệm và đánh giá chỉ ra rằng mô hình không chỉ đạt hiệu suất cạnh tranh với các phương pháp hiện đại, mà còn duy trì khả năng tổng quát hóa và sự ổn định trên nhiều tập dữ liệu, từ đó chứng minh tiềm năng ứng dụng thực tiễn trong các hệ thống truy xuất hình ảnh hiệu quả.

KẾT LUẬN

1. Kết quả đạt được

Nghiên cứu này giới thiệu một phương pháp tiếp cận mới cho bài toán truy xuất hình ảnh dựa trên ngữ nghĩa nhằm cải thiện khả năng biểu diễn và khai thác thông tin từ các hình ảnh phức tạp. Phương pháp kết hợp hai thành phần chủ đạo: một mạng CNN để trích xuất đặc trưng hình ảnh đa cấp, và Graphormer để mô hình hóa mối quan hệ ngữ nghĩa giữa các thực thể trong dữ liệu. CNN chịu trách nhiệm nắm bắt các đặc trưng cốt lõi về cấu trúc và nội dung hình ảnh, trong khi Graphormer giúp mô hình nhận biết các mối liên kết giữa các đặc trưng này, từ đó nâng cao khả năng phân tách các lớp và cải thiện chất lượng embedding.

Để triển khai phương pháp, xây dựng một pipeline toàn diện, bao gồm các bước tiền xử lý dữ liệu, trích xuất đặc trưng, xây dựng đồ thị mô tả mối quan hệ giữa các hình ảnh, và tối ưu hóa mô hình thông qua nhiều hàm mất mát. Pipeline này nhằm tạo ra các vector nhúng vừa ổn định vừa giàu thông tin, đồng thời duy trì khả năng tổng quát hóa trên các tập dữ liệu khác nhau. Thêm vào đó, cơ chế lựa chọn node tĩnh và node động được tích hợp để cung cấp ngữ cảnh tham chiếu, từ đó nâng cao độ chính xác trong quá trình truy vấn.

Đánh giá thực nghiệm chứng minh phương pháp được đề xuất mang lại hiệu quả ấn tượng trên các tập dữ liệu chuẩn như CIFAR-100 và CUB-200-2011. Các chỉ số mAP, Precision@K và Recall@K đạt mức nổi bật, trong khi không gian embedding tạo ra thể hiện cấu trúc cụm rõ ràng, phản ánh khả năng cụm các ảnh có ngữ nghĩa tương đồng và tách biệt chúng với các lớp khác. Việc tối ưu hóa tìm kiếm dựa trên cụm cũng giúp giảm đáng kể tài nguyên tính toán mà không làm suy giảm độ chính xác.

Nghiên cứu này đóng góp một số điểm quan trọng:

- Sự kết hợp giữa CNN và Graphormer giúp khai thác đồng thời đặc trưng cục bộ và quan hệ ngữ nghĩa cấp cao.

- Một phương pháp truy vấn tốc độ cao dựa trên proxy-cluster đã được xây dựng, hỗ trợ tái cấu trúc không gian đặc trưng và giới hạn vùng tìm kiếm nhằm cải thiện hiệu suất.

- Các thử nghiệm trên CIFAR-100 và CUB-200 cho thấy mAP đạt xấp xỉ 0.81 và 0.86, minh chứng cho khả năng cân bằng giữa chất lượng truy xuất và tốc độ xử lý.

Tổng quan, nghiên cứu này chứng minh rằng việc kết hợp CNN và Graphormer mang lại mô hình linh hoạt, có khả năng mở rộng và hiệu suất cao trên nhiều bối cảnh dữ liệu khác nhau. Pipeline được đề xuất tạo nền tảng ổn định cho các hệ thống truy xuất hình ảnh theo ngữ nghĩa, phân loại hoặc quản lý dữ liệu hình ảnh. Những kết quả đạt được mở ra cơ hội phát triển mới trong việc ứng dụng các mô hình học sâu kết hợp cấu trúc đồ thị, góp phần nâng cao chất lượng truy xuất ngữ nghĩa và cung cấp cơ sở cho các chiến lược tối ưu hóa tiếp theo.

2. Nhận xét và đề xuất

Nghiên cứu đề xuất một giải pháp truy xuất hình ảnh theo ngữ nghĩa, kết hợp Graph-based representation và Transformer, thể hiện hiệu quả cao trong các bài toán phân loại và tìm kiếm ảnh tương đồng trên CIFAR-100 và CUB-200-2011. Hệ thống không chỉ đạt các thước đo định lượng tốt mà còn tạo ra các cụm đặc trưng rõ ràng, giúp truy xuất trực quan và nhanh chóng. Tuy nhiên, để đảm bảo khả năng tổng quát hóa và ổn định khi triển khai thực tế, cần thử nghiệm trên nhiều nguồn dữ liệu khác nhau và tinh chỉnh cẩn thận các siêu tham số. Bên cạnh đó, việc tối ưu trải nghiệm người dùng và tốc độ phản hồi cũng là yếu tố then chốt, nhằm biến mô hình thành một công cụ truy xuất semantic đáng tin cậy, hiệu quả và dễ tiếp cận cho cả nghiên cứu lẫn ứng dụng thực tiễn.

3. Hướng phát triển tiếp theo

Một phương hướng mở rộng quan trọng của dự án là phát triển hệ thống truy xuất ảnh với khả năng nhận diện nhiều đối tượng kết hợp thông tin văn bản. Ý tưởng

là tích hợp mô hình CLIP (Contrastive Language–Image Pre-Training) [32] nhằm học không gian biểu diễn chung cho cả ảnh và ngôn ngữ. Cụ thể, CLIP sẽ ánh xạ hình ảnh và mô tả tương ứng của chúng vào cùng một không gian embedding. Tiếp theo, kiến trúc Graphormer được áp dụng để xây dựng đồ thị đại diện cho các đối tượng trong hình ảnh cũng như các thành phần ngữ nghĩa trích xuất từ văn bản. Mô hình có thể được cải thiện thêm bằng cách tích hợp các dạng bias cấu trúc tương tự như trong Graphormer truyền thống — chẳng hạn như mã hóa độ trung tâm hoặc khoảng cách không gian — nhằm tăng cường khả năng phân biệt các mối quan hệ phức tạp giữa các đối tượng.

Bên cạnh đó, việc xây dựng một giao diện demo trực quan, cho phép người dùng nhập hình ảnh hoặc mô tả bằng văn bản để nhận về các hình ảnh tương đồng, sẽ minh họa rõ ràng tính khả thi và giá trị ứng dụng thực tiễn của hệ thống đề xuất.

CÔNG BỐ KHOA HỌC

Công bố khoa học của đề án:

- Van Khanh Nguyen, Thi Thuy Quynh Dao, Quynh Nguyen Huu. “Clustered Graph Transformer for Fast and Efficient Image Retrieval” , Computers and Electrical Engineering, under review 2025.

DANH MỤC TÀI LIỆU THAM KHẢO

[1] Krizhevsky, Alex and Hinton, Geoffrey and others, "Learning multiple layers of features from tiny images," 2009.

[2] Wah, Catherine and Branson, Steve and Welinder, Peter and Perona, Pietro and Belongie, Serge, "The caltech-ucsd birds-200-2011 dataset," 2011.

[3] Y. a. B. L. a. B. Y. a. H. P. Lecun, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE,* vol. 86, pp. 2278-2324, 1998.

[4] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929,* 2020.

[5] Z. a. L. Y. a. C. Y. a. H. H. a. W. Y. a. Z. Z. a. L. S. a. G. B. Liu, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021.

[6] He, Kaiming; Zhang, Xiangyu; Ren, Shaoqing; Sun, Jian, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[7] Han, Dongyoon and Kim, Jiwhan and Kim, Junmo, "Deep Pyramidal Residual Networks," *IEEE CVPR,* 2017.

[8] T. Kipf, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907,* 2016.

[9] Veli{\v{c}}kovi{\'c}, Petar and Cucurull, Guillem and Casanova, Arantxa and Romero, Adriana and Lio, Pietro and Bengio, Yoshua, "Graph attention networks," in *arXiv preprint arXiv:1710.10903*, 2017.

[10] Ying, Chengxuan and Cai, Tianle and Luo, Shengjie and Zheng, Shuxin and Ke, Guolin and He, Di and Shen, Yanming and Liu, Tie-Yan, "Do transformers really perform badly for graph representation?," *Advances in neural information processing systems,* vol. 34, pp. 28877--28888, 2021.

[11] Brody, Shaked and Alon, Uri and Yahav, Eran, "How attentive are graph attention networks?," *arXiv preprint arXiv:2105.14491,* 2021.

[12] Vaswani, Ashish and Shazeer, Noam and Parmar, Niki and Uszkoreit, Jakob and Jones, Llion and Gomez, Aidan N and Kaiser, {\L}ukasz and Polosukhin, Illia, "Attention is all you need," *Advances in neural information processing systems,* vol. 30, 2017.

[13] Wang, Xun and Han, Xintong and Huang, Weilin and Dong, Dengke and Scott, Matthew R., "Multi-Similarity Loss With General Pair Weighting for Deep Metric Learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[14] Zheng, Wenzhao and Wang, Chengkun and Lu, Jiwen and Zhou, Jie, "Deep Compositional Metric Learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[15] Sarkar, Rohan and Kak, Avinash, "Dual Pose-invariant Embeddings: Learning Category and Object-specific Discriminative Representations for Recognition and Retrieval," in *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024.

[16] Kim, Sungyeon and Kim, Dongwon and Cho, Minsu and Kwak, Suha, "Proxy anchor loss for deep metric learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020.

[17] Loshchilov, Ilya and Hutter, Frank, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101,* 2017.

[18] Jeong, Yeonwoo and Kim, Yoonsung and Song, Hyun Oh, "End-to-End Efficient Representation Learning via Cascading Combinatorial Optimization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[19] Ghita, Alexandru and Ionescu, Radu Tudor, "Class anchor margin loss for content-based image retrieval," *arXiv preprint arXiv:2306.00630,* 2023.

[20] Seo, Seonguk and Uzunbas, Mustafa Gokhan and Han, Bohyung and Cao, Sara and Lim, Ser-Nam, "Metric compatible training for online backfilling in large-scale retrieval," in *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025.

[21] Feng, Chen and Patras, Ioannis, "Maskcon: Masked contrastive learning for coarse-labelled dataset," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[22] Shi, Guanghui and Liang, Xuefeng and Li, Wenjie and Lin, Xiaoyu, "Learning Separable Fine-Grained Representation via Dendrogram Construction from Coarse Labels for Fine-grained Visual

Recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

[23] Kolesnikov, Alexander and Beyer, Lucas and Zhai, Xiaohua and Puigcerver, Joan and Yung, Jessica and Gelly, Sylvain and Houlsby, Neil, "Big transfer (bit): General visual representation learning," in *European conference on computer vision*, 2020.

[24] He, Kaiming and Zhang, Xiangyu and Ren, Shaoqing and Sun, Jian, "Identity mappings in deep residual networks," in *European conference on computer vision*, 2016.

[25] Wang, Chengkun and Zheng, Wenzhao and Li, Junlong and Zhou, Jie and Lu, Jiwen, "Deep Factorized Metric Learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[26] Xiao, Zilin and Suma, Pavel and Sachdeva, Ayush and Wang, Hao-Jen and Kordopatis-Zilos, Giorgos and Tolias, Giorgos and Ordonez, Vicente, "LOCORE: Image Re-ranking with Long-Context Sequence Modeling," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025.

[27] Li, Aoyu and Sato, Ikuro and Ishikawa, Kohta and Kawakami, Rei and Yokota, Rio, "Informative sample-aware proxy for deep metric learning," in *Proceedings of the 4th ACM International Conference on Multimedia in Asia*, 2022.

[28] Shu, Yangyang and Van den Hengel, Anton and Liu, Lingqiao, "Learning common rationale to improve self-supervised representation for fine-grained visual recognition problems," in *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023.

[29] Moskvyak, Olga and Maire, Frederic and Dayoub, Feras and Baktashmotlagh, Mahsa, "Keypoint-aligned embeddings for image retrieval and re-identification," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2021.

[30] Yang, Bailin and Sun, Haoqiang and Li, Frederick WB and Chen, Zheng and Cai, Jianlu and Song, Chao, "HSE: Hybrid species embedding for deep metric learning," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023.

[31] Devalraju, Dhanunjaya Varma and Sekhar, C Chandra, "Uncertainty-Guided Metric Learning Without Labels," in *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025.

[32] Radford, Alec and Kim, Jong Wook and Hallacy, Chris and Ramesh, Aditya and Goh, Gabriel and Agarwal, Sandhini and Sastry, Girish and Askell, Amanda and Mishkin, Pamela and Clark, Jack and others, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, 2021.