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**

TÓM TẮT ĐỀ ÁN TỐT NGHIỆP THẠC SĨ

HÀ NỘI – NĂM 2026

Đề án tốt nghiệp được hoàn thành tại:

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

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

Phản biện 1: TS. Nguyễn Xuân Thắng

Phản biện 2: PGS.TS. Trần Thị Oanh

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

Vào lúc: 10 giờ 30 ngày 24 tháng 01 năm 2026

Có thể tìm hiểu đề án tốt nghiệp tại:

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

Mục lục

MỞ ĐẦU 1

1. Lý do lựa chọn đề tài 1

2. Tổng quan về vấn đề nghiên cứu 1

3. Mục đích nghiên cứu 1

4. Đối tượng và phạm vi nghiên cứu 2

5. Phương pháp nghiên cứu 2

6. Đóng góp của đề tài 2

7. Cấu trúc luận văn 3

NỘI DUNG 4

CHƯƠNG I: CƠ SỞ LÝ LUẬN 4

1. Tổng quan về tra cứu ảnh 4

2. Bài toán tra cứu ảnh dựa trên ngữ nghĩa 4

3. Các phương pháp thử nghiệm và đánh giá 5

4. Các nghiên cứu trước đây 5

5. Convolutional Neural Network (CNN) 6

6. Các mô hình đồ thị 6

6.1. Graph Convolutional Network (GCN) 6

6.2. Graph Attention Network (GAT) 7

6.3. Graphormer 8

7. Tiểu kết chương 8

CHƯƠNG II: PHƯƠNG PHÁP NGHIÊN CỨU ĐỀ XUẤT 9

1. Trích xuất đặc trưng hình ảnh bằng CNN 9

2. Biểu diễn đồ thị 9

3. Học biểu diễn bằng Graphormer 10

4. Huấn luyện và tối ưu hóa mô hình 11

5. Truy xuất nhanh dựa trên cụm 13

6. Tổng kết và đánh giá 14

7. Tiểu kết chương 15

CHƯƠNG III: THỬ NGHIỆM VÀ ĐÁNH GIÁ 16

1. Mục tiêu thực nghiệm 16

2. Chi tiết cài đặt 16

3. Tập dữ liệu 16

4. Đánh giá hiệu quả mô hình 16

5. Phân tích kết quả 17

6. Tiểu kết chương 19

KẾT LUẬN 21

1. Kết quả đạt được 21

2. Nhận xét và đề xuất 21

3. Hướng phát triển tiếp theo 22

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
SBIR	Semantic Based Image Retrieval	Tra cứu ảnh dựa trên ngữ nghĩa

DANH SÁCH HÌNH VẼ

Hình 2.1: Kiến trúc tổng thể của mô hình được đề xuất. 9
Hình 2.2: Hoạt động của các hàm mất mát. 12
Hình 2.3: Sơ đồ hoạt động của pha truy xuất. 13

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 kéo theo nhu cầu về các hệ thống truy xuất thông minh trở nên cấp bách.

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.

Trọng tâm nghiên cứu là phát triển một phương pháp tra cứu hình ảnh ngữ nghĩa hiện đại dựa trên Graphormer.

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.

2. Tổng quan về vấn đề nghiên cứu

Các hệ thống hiện đại không chỉ chịu áp lực về tính chuẩn xác trong nhận dạng mà còn phải giải quyết bài toán về độ trễ thấp, nhất là trong bối cảnh dữ liệu bùng nổ và nhu cầu tra cứu diễn ra liên tục.

Xuất phát từ thực trạng đó, đề tài tập trung xây dựng một giải pháp tra cứu semantic với hạt nhân là mô hình Graphormer. Kiến trúc này được tận dụng để giải mã và mã hóa các tương tác đa chiều giữa các vật thể trong ảnh, qua đó cường hóa khả năng mô tả ngữ nghĩa và tối ưu hóa giá trị của vector đặc trưng. Thêm vào đó, việc tích hợp kỹ thuật phân cụm giúp hệ thống khu trú vùng tìm kiếm ngay từ đầu, cho phép tăng tốc quy trình truy vấn mà không làm suy hao độ chính xác.

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.

Song song với việc xây dựng mô hình, nghiên cứu tiến hành đánh giá thử nghiệm dựa trên các chỉ số định lượng tiêu chuẩn trong lĩnh vực tra cứu. 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.

4. Đối tượng và phạm vi nghiên cứu

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 và 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).

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.

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.

5. Phương pháp nghiên cứu

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. Trụ cột thứ hai xoay quanh công tác phát triển mô hình và thực hiện kiểm chứng. 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.

6. Đóng góp của đề tài

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.

Đề tài cũng đề cập đến đóng góp trong việc cải thiện tốc độ truy vấn thông qua cơ chế phân cụm vector nhúng.

7. Cấu trúc luận văn

Nội dung nghiên cứu 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. Nội dung kế tiếp sẽ đi sâu vào mô tả tường tận giải pháp kỹ thuật được xây dựng trong khuôn khổ đề tài. 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. 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.

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.

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 lớn 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.

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.

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 lân cận xấp xỉ hoặc triển khai hệ thống phân tán.

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.

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 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.

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.

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. 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. 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 bài toán tra cứu hình ảnh, các phương pháp truyền thống chủ yếu dựa trên việc trích chọn đặc trưng thị giác và so khớp trong không gian đặc trưng. Sự phát triển của học sâu, đặc biệt là các mạng CNN [3] và gần đây là Vision Transformer [4, 5], đã giúp cải thiện đáng kể chất lượng biểu diễn nhờ khả năng khai thác ngữ nghĩa và ngữ cảnh toàn cục của ảnh. Tuy nhiên, các mô hình này thường đi kèm chi phí tính toán cao và yêu cầu dữ liệu huấn luyện lớn, gây khó khăn cho các hệ thống truy xuất quy mô lớn.

Để giải quyết vấn đề hiệu suất truy vấn, các phương pháp hashing đã được đề xuất nhằm ánh xạ đặc trưng liên tục sang mã nhị phân, giúp tăng tốc tìm kiếm và giảm chi phí lưu trữ. Mặc dù hiệu quả về mặt tính toán, hashing có thể làm suy giảm thông tin ngữ nghĩa do quá trình lượng tử hóa, đặc biệt trong các kịch bản đòi hỏi phân biệt tinh tế giữa các lớp tương đồng.

Nhìn chung, các nghiên cứu hiện nay tập trung vào hai hướng chính: nâng cao chất lượng biểu diễn đặc trưng bằng các mô hình học sâu và cải thiện hiệu suất truy vấn thông qua hashing. Tuy nhiên, việc dung hòa hiệu quả giữa độ chính xác truy xuất và khả năng mở rộng hệ thống vẫn là một thách thức mở, thúc đẩy nhu cầu nghiên cứu các phương pháp mới cho bài toán tra cứu hình ảnh.

5. Convolutional Neural Network (CNN)

Điểm ưu việt của CNN [3] 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. 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 để hợp nhất với đầu ra của chuỗi xử lý phi tuyến.

Để nâng cao khả năng học đặc trưng và khắc phục các hạn chế hiện có của các mạng nơ-ron rất sâu, [7]−một biến thể nâng cao của ResNet – đã được phát triển. Bằng cách sử dụng các kết nối tắt có đệm số 0 (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.

6. Các mô hình 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.

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.

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.

Mỗi lớp của GCN tương ứng với một mức độ lan truyền thông tin khác nhau trong đồ thị. 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.

6.2. Graph Attention Network (GAT)

Graph Attention Network (GAT) [9] 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. GAT cho phép mô hình tự động học được mức độ ảnh hưởng của từng đỉnh lân cận đối với nút trung tâm.

Một trong những điểm nổi bật của GAT là việc ứng dụng multi-head attention. 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 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.

6.3. Graphormer

Graphormer [10] 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.

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. 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 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.

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 [11] nhằm học các biểu diễn nút trong đồ thị.

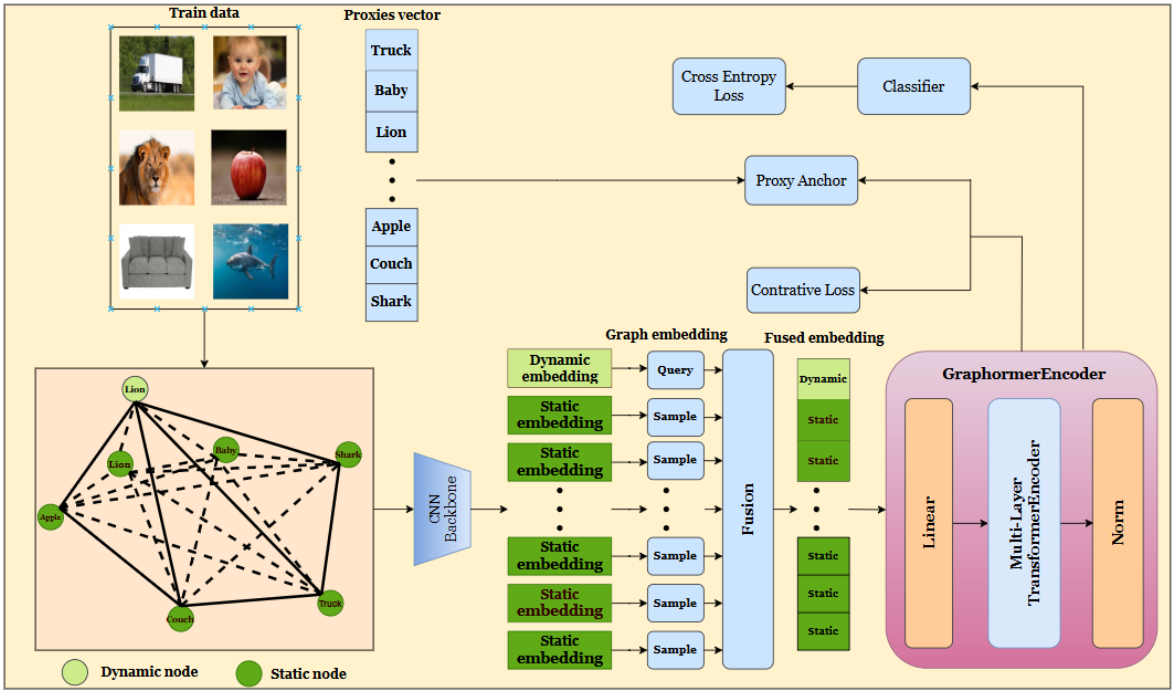
7. Tiểu kết chương

Graphormer 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.

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.

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.

CHƯƠNG II: PHƯƠNG PHÁP NGHIÊN CỨU ĐỀ XUẤT



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

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

Chi tiết hơn, quy trình xử lý được bắt đầu bằng việc đưa từng hình ảnh qua mạng CNN tiền huấn luyện nhằm trích xuất vector đặc trưng tại lớp liền kề lớp cuối cùng (penultimate layer). Kế đến, kỹ thuật chuẩn hóa L2 (L2 normalization) được áp dụng.

2. Biểu diễn đồ thị

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: $N_s = \{N_1, N_2, \dots, N_C\}$.

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.

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.

3. Học biểu diễn bằng Graphormer

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:

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

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 [11], 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:

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

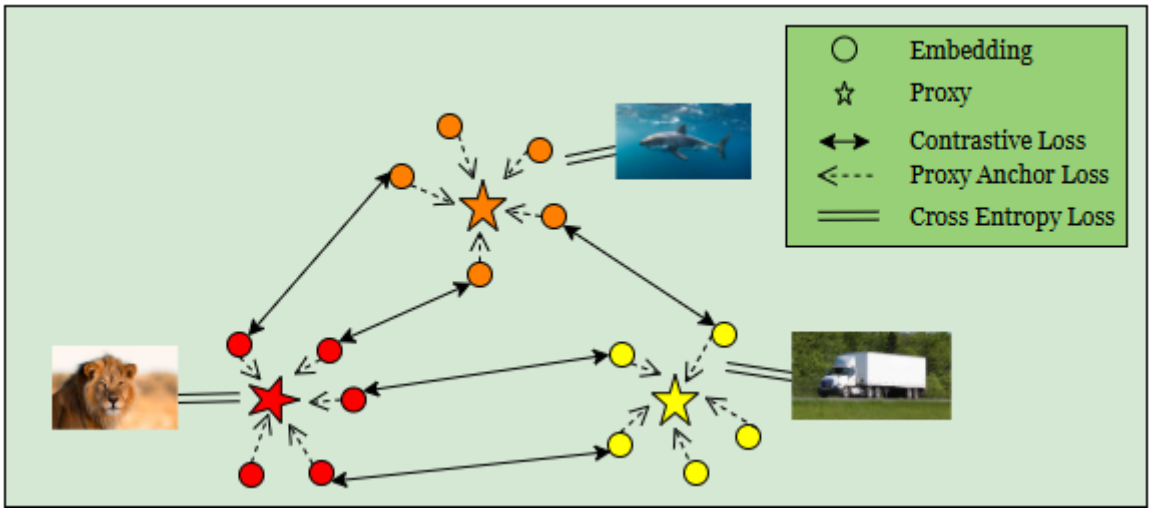
Sau khi đi qua nhiều lớp Transformer Encoder, đặc trưng của mỗi node trong đồ thị bao gồm node động N_0và các node tĩnh $N_1, \dots, N_C$ được cập nhật để phản ánh ngữ cảnh từ tất cả các node còn lại. Tiếp theo, các vector đặc trưng của từng node được chuẩn hóa L2, hoàn tất quá trình mã hóa của Graphormer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

4. Huấn luyện và tối ưu hóa mô hình

Mô hình được huấn luyện đa nhiệm dựa trên tổ hợp ba hàm mất mát: 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.



Hình 2.2: Hoạt động của các hàm mất mát.

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.

Proxy-anchor loss [12] 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.

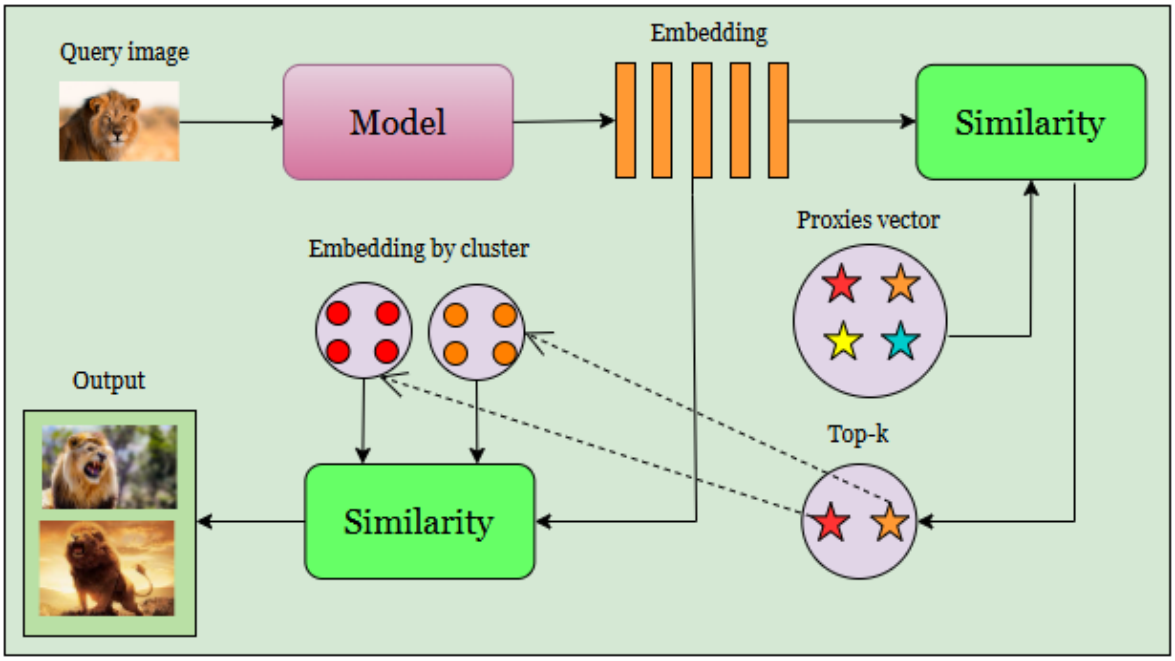
Ba thành phần trên được kết hợp tuyến tính để hình thành hàm mất mát tổng hợp:

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

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.

5. Truy xuất nhanh dựa trên cụm



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

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.

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 $\mathcal{C} = \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k\}$.

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 $\mathcal{C}$. 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; Độ chính xác cao hơn; Khả năng mở rộng tốt.

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; 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.

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 thành phần (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 huấn luyện sử dụng tổng hợp nhiều hàm mất mát. 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.

CHƯƠNG III: 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.

1. Mục tiêu thực nghiệm

Bước đầu tiên là so sánh với những kỹ thuật tiên tiến, chiến lược tìm kiếm tốc độ cao thông qua phân nhóm cũng được xem xét kỹ lưỡng. 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. Tiếp đến là đưa mô hình vào truy xuất ra các ảnh thực tế. Cuối cùng, phương pháp phân tích thành phần (ablation study) được áp dụng.

2. Chi tiết cài đặt

Việc huấn luyện diễn ra trong 100 epoch sử dụng thuật toán AdamW [13] 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.

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.

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].

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.

5. Phân tích kết quả

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 cho thấy mô hình đạt hiệu suất cạnh tranh trên CIFAR-100 so với các phương pháp hiện có. 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).

Tiếp theo, một thử nghiệm benchmark được thực hiện trên CUB-200-2011. Đối với backbone, mô hình sử dụng mô hình BiT-L [14] (một ResNetV2-152×2 [15]) được pre-train trên ImageNet-21k, với batch size 64 và embedding dimension là 512. Trên benchmark CUB-200-2011, mô hình tiếp tục cho thấy hiệu suất cạnh tranh cao. 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.

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.

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.

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.

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).

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.
- mAP@K: Tại các ngưỡng K lớn hơn 1, mAP@K của k = 3 thậm chí còn cao hơn so với $k = C$.
- 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.

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.

Hình ảnh cho thấy một cách trực quan chất lượng của không gian embedding do mô hình tạo ra. 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:

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

Các kết quả 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.

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). Việc chuyển đổi sang kiến trúc $Graph \rightarrow FC$ cũng không tạo ra sự đột phá.
- 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.
- Một điểm nhấn khác là chỉ số Precision@R đạt 64.94, bỏ xa mức ~44.35 của các biến thể nền tảng.

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 [12]. Cả 2 đều được thực hiện truy xuất trên 3 cụm, các số liệu thống kê 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.

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.

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 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).

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.

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.

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.

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.

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.
- 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.

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. 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) [16] nhằm học không gian biểu diễn chung cho cả ảnh và ngôn ngữ.

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 tài:

- 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] 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.

[12] 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.

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

[14] 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.

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

[16] 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.