TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
KHOA KHOA HỌC MÁY TÍNH

BÁO CÁO ĐỒ ÁN
XỬ LÝ ẢNH VÀ ỨNG DỤNG

ĐỀ TÀI
TRUY VẤN ẢNH

GV hướng dẫn: Mai Tiến Dũng
Lớp: CS406.O11.KHCL
Nhóm thực hiện:

1. Lê Trần Bảo Lợi - 21522295
2. Bùi Đình Quân - 21522487
3. Ngô Đức Hoàng Hiệp - 21520846

TP.HCM, tháng 1 năm 2024

MỤC LỤC

1 GIỚI THIỆU BÀI TOÁN

1.1 Giới thiệu bài toán

- Image retrieval là hệ thống tìm kiếm ảnh dựa trên ảnh tham khảo
- Tính hiệu quả của hệ thống
  - Tối ưu về không gian lưu trữ
  - Tốc độ tìm kiếm nhanh
  - Kết quả tìm kiếm phù hợp
  - Tiện dụng

1.2 Động lực thực hiện bài toán

- Tìm kiếm luôn là nhu cầu cần thiết của con người.
- Khối lượng dữ liệu ngày càng lớn, phức tạp và đa dạng.
- Ảnh là kiểu dữ liệu phổ biến có nhu cầu tìm kiếm cao.

1.3 Mô tả bài toán

Input:

- Một hình ảnh truy vấn.
- Tập dữ liệu hình ảnh mà hệ thống truy vấn sẽ tìm kiếm.

Output:

- Danh sách xếp hạng các hình ảnh tương đồng hoặc liên quan đến hình ảnh truy vấn trong tập dữ liệu.

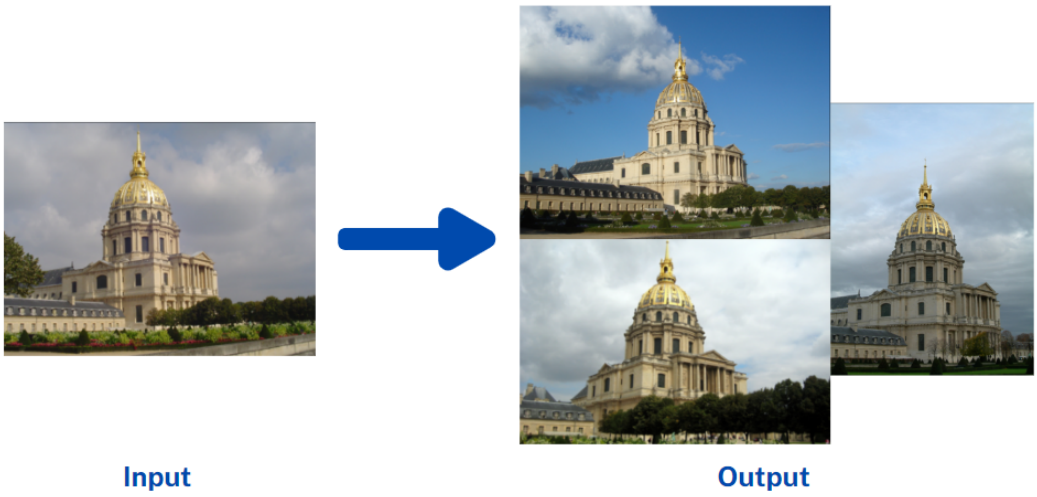

Figure 1: Ví dụ mô tả input/ output của bài toán

2 Dataset

2.1 revisited Paris dataset

- Tập dữ liệu "Revisited Paris" : benchmark dataset cho việc đánh giá content-based image retrieval (CBIR) systems.
- Ảnh được thu thập từ Flickr của 11 địa danh của Pháp được đăng tải trên mạng xã hội.
- Được gán nhãn lại trên tập Paris giải quyết các trường hợp bị sai ở tập ground truth và mở rộng tập query.

2.2 Cấu trúc dataset

- Dataset gồm: 6412 ảnh.
- Số lượng query: 70 ảnh query.
- Mỗi ảnh query được con người gán thủ công ground truth tương ứng.

Figure 2: Một số hình ảnh trong tập dataset

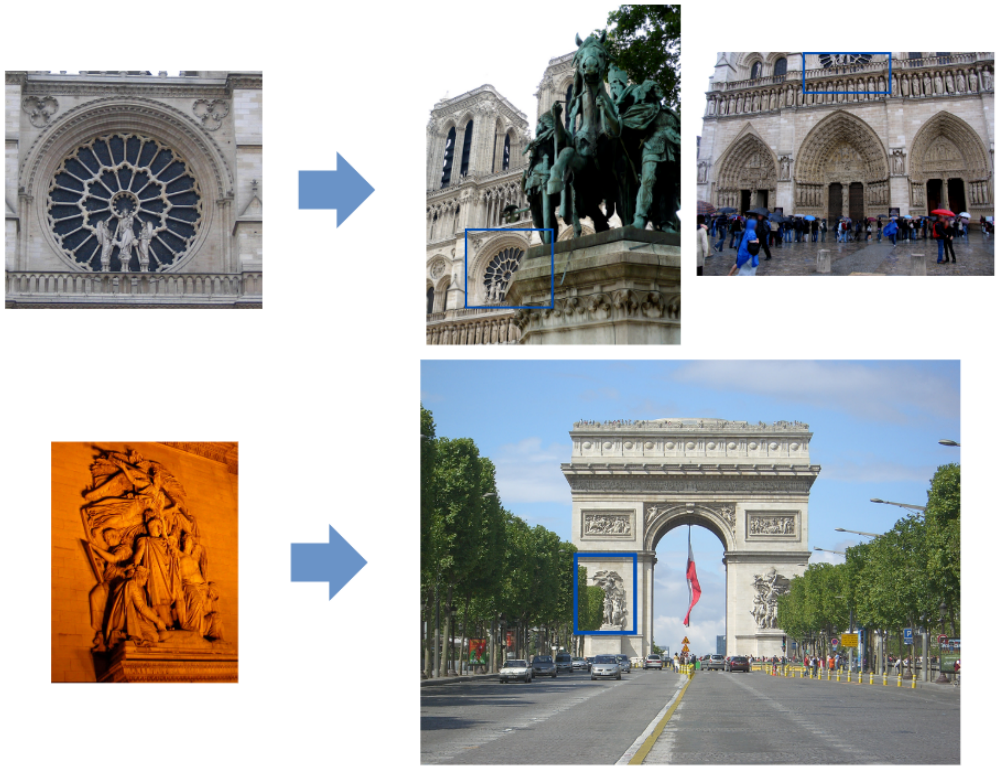

Figure 3: Một số trường hợp phức tạp

3 Phương pháp thực hiện

3.1 Tổng quan

Với bài toán đã được định nghĩa, nhóm lựa chọn hướng tiếp cận của bài toán truy vấn ảnh với đầu vào là ảnh. Với mong đợi là có thể tìm kiếm các ảnh tương đồng(với ảnh truy vấn) về mặt nội dung thị giác chứa trong ảnh.

Những ý tưởng đầu tiên về việc sử dụng máy tính cho bài toán truy vấn, tìm kiếm lần đầu tiên được nhắc đến ở những năm 1940 [10]. Và được nghiên cứu sâu rộng cho đến ngày nay, do khối lượng dữ liệu lớn và tăng nhanh, dẫn đến gia tăng nhu cầu tìm kiếm một cách hiệu quả.

Để giải quyết bài toán truy vấn thông tin đã có rất nhiều mô hình được đề xuất như: mô hình boolean, mô hình xác xuất(probabilistic model) hay mô hình không gian vector(vector space model).

Do thời gian thực hiện đồ án có hạn nhóm đã lựa chọn mô hình không gian vector làm thành phần chính cho hệ thông của nhóm(bỏ qua các phương pháp khác). Bên cạch đó, mô hình có cấu trúc đơn giản mà mang lại hiệu quả cao, có tính mở rộng và được ứng dụng rộng rãi cho tới tận ngày nay trong các hiện thống truy vấn thông tin hiện đại. Với sự phổ biến của mô hình này, đồng nghĩa với việc nhóm có nhiều cơ hội tiếp xúc cũng như tìm hiểu nhiều hơn so với các mô hình khác, dẫn đến việc nhóm sẽ tự tin hơn khi trình bày về nội dung của phương pháp này.

3.2 Mô hình không gian vector(vector space model)

Ý tưởng chính của mô hình không gian vector là việc biểu diễn các đối tượng tìm kiếm(các đặc trưng) dưới dạng **vector**. Khoảng cách giữa hai vector trong không gian vector thể hiện cho **độ phù hợp** giữa hai đối tượng ứng với hai vector đang xét. Kết quả truy vấn là danh sách đối tượng có vector biểu diễn gần nhất với vector biểu diễn cho đối tượng truy vấn(query).

Phương pháp tương đối đơn giản chỉ bao gồm hai thao tác chính là "rút trích đặc trưng"(feature extraction hay feature representation) và "so sánh đặc trưng"(feature comparison). Phương pháp khá linh hoạt có thể giải quyết cho nhiều bài toán khác nhau, trên các đối tượng tìm kiếm khác nhau như văn bản, hình ảnh, video, âm thanh,...

Ở phần tiếp theo nhóm sẽ diễn giải chi tiết hơn về quá trình ứng dụng mô hình không gian vector cho bài toán truy vấn ảnh.

3.3 Mô hình không gian vector cho bài toán truy vấn ảnh

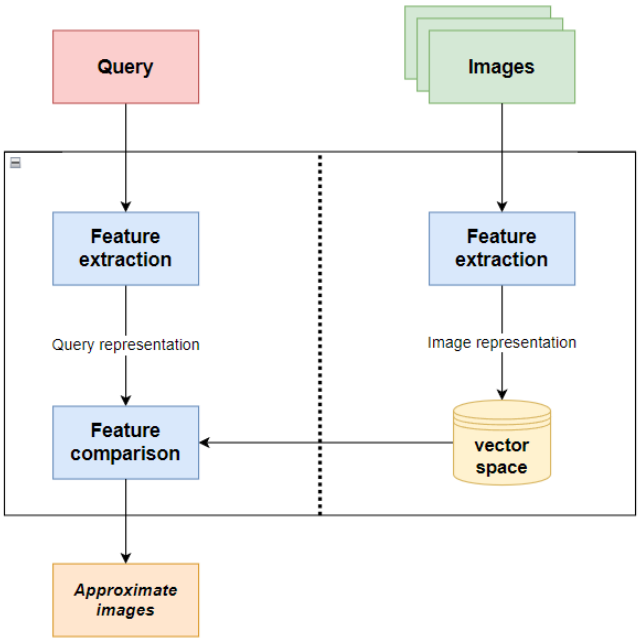

Figure 4: Vector space model cho bài toán truy vấn ảnh

Như đã đề cập ở phần trước, mô hình không gian vector đơn giản, dễ dàng mở rộng cho nhiều bài toán. Hiệu quả của phương pháp phần lớn được quyết định bởi hai giai đoạn: **rút trích đặc trưng** và **so sánh đặc trưng**.

3.3.1 Rút trích đặc trưng

Biểu diễn đặc trưng(hay rút trích đặc trưng) hiệu quả sẽ giúp tăng độ chính xác của hệ thống truy vấn. Một phương pháp biểu diễn đặc trưng tốt sẽ thể hiện được sự phân biệt giữa các đối tượng, đồng nghĩa rằng các ảnh tương đồng về nội dung sẽ có đặc trưng giống hoặc gần giống nhau, ngược lại các ảnh có sự khác biệt về nội dung sẽ có đặc trưng khác nhau.

Bên cạnh đó, đối với bài toán truy vấn ảnh, đặc biệt đối với ảnh đời thực, những ảnh cùng nội dung có thể sẽ chịu sự biến động, thay đổi của các yếu tố môi trường như: độ sáng, góc nhìn,...gây khó khăn khi biểu diễn và rút trích đặc trưng.

Yêu cầu đặt ra, sử dụng phương pháp rút trích đặc trưng, có khả năng phân biệt giữa các ảnh(theo nội dung thị giác) và bất biến với các yếu tố của môi trường.

Trước những yêu cầu đặt ra, nhóm quyết định lựa chọn các phương pháp học sâu để rút trích đặc trưng ảnh, bởi vì theo nhiều kết quả thực nghiệm cho thấy các đặc trưng

học sâu cho kết quả vượt rội hơn so với các đặc trưng thủ công(handcrafted feature) về khả năng biểu diễn đặc trưng thị giác.

Trong phần thực nghiệm 5 nhóm lựa chọn mô hình học sâu Resnet [11] để rút trích đặc trưng, vì đây là mô hình nổi tiếng với cơ chế skip connection mang lại kết quả vượt trội trong bài toán phân loại ảnh với độ lỗi top 5 chỉ 3.57% trên tập ImageNet.

Tuy nhiên để minh chứng cho khả năng vượt trội của Resnet trước biến động của yếu tố môi trường và so sánh với các đặc trưng thủ công thì nhóm thực nghiệm thêm trên đặc trưng HOG cùng với các thao tác biến đổi ảnh cơ bản.

3.3.2 So sánh đặc trưng

Danh sách kết quả trả về là các ảnh có vector đặc trưng có độ tương đồng cao(similarity) trong không gian vector theo một độ đo tương đồng nào đó. Nhóm lựa chọn cosine là độ đo tương đồng.

Bên cạnh đó việc tổ chức và lưu trữ không gian vector cũng là một thách thức, đặc biệt đối với hệ thống hoạt động trên tập dữ liệu lớn. Từ đó nhóm đã lựa chọn một phương pháp đã được phát triển từ trước của nhóm nghiên cứu thuộc Meta có tên là FAISS, phương pháp này đã được triển khai thành ứng dụng, hiệu quả cao cho bài toán truy vấn theo mô hình không gian vector.

3.4 Phương pháp biến đổi ảnh truy vấn

Với một ảnh truy vấn đầu vào nhóm thực hiện một số phép biến đổi trên ảnh truy vấn và thực hiện truy vấn(theo tuần tự) bằng cách ảnh đã biến đổi sau đó thực hiện tổng hợp kết quả. Với mong đợi rằng việc đa dạng hóa ảnh truy vấn giúp hệ thống tìm ra được thêm đáp án đúng.

Sau đây là một số phép biến đổi mà nhóm sử dụng để thực nghiệm:

- Phép đóng/mở với kích thước bộ lọc là 5x5.
- Phép xoay với góc ngẫu nhiên từ -30 độ đến 30 độ.
- Phép cắt ảnh vị trí ngẫu nhiên, với kích thước ảnh cắt bằng 1/2 ảnh gốc.
- Phép lật ảnh theo chiều dọc.
- Phép làm sắc, làm mượt ảnh.

Sau khi thực hiện biến đổi, ta có tổng cộng là 9 ảnh truy vấn(tính cả ảnh gốc). Ứng với mỗi ảnh nhóm thực hiện 9 câu truy vấn độc lập. Danh sách kết quả vẫn đảm bảo chỉ là top K ảnh, việc tổng hợp diễn ra như sau:

- Khởi tạo danh sách kết quả ban đầu bằng với danh sách kết quả khi truy vấn bằng ảnh gốc(danh sách đã được sắp xếp)
- Thực hiện chèn đáp án từ các câu truy vấn còn lại vào danh sách kết quả theo độ tương đồng(ảnh đáp án so với ảnh biến đổi, không phải ảnh gốc), sao cho đảm bảo thứ tự.
- Sau khi chèn đáp án dựa trên 8 kết quả truy vấn còn lại, danh sách được chèn là kết quả cuối cùng.

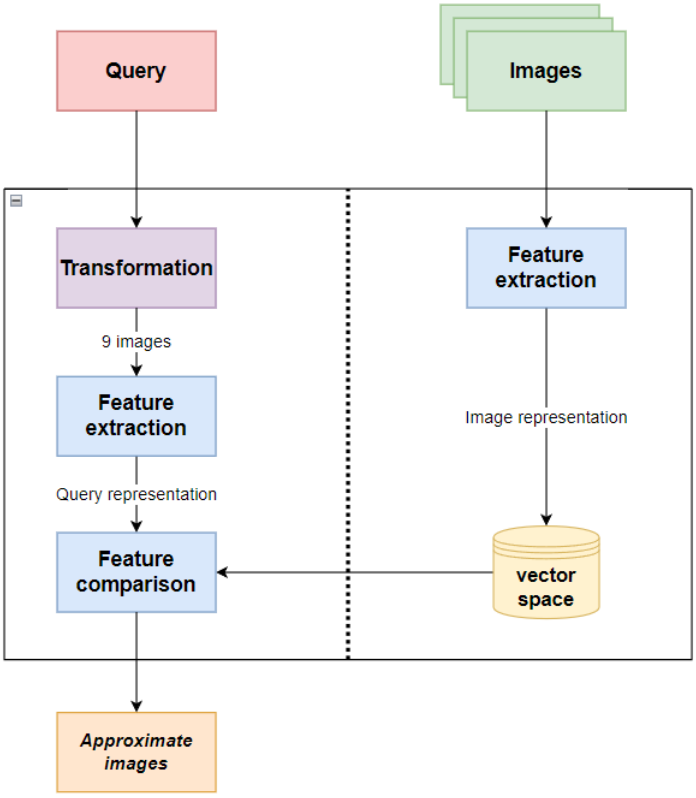

Figure 5: Hệ thống với phép biến đổi ảnh truy vấn.

Kết quả thực nghiệm được thể hiện ở phần 5

3.5 Tăng cường kho dữ liệu

Sau khi thực hiện phương pháp biến đổi ảnh truy vấn, nhóm nhận thấy kết quả dường như không cải thiện đối với đặc trưng học sâu Resnet. Kết quả thực nghiệm và đánh

giá cho thấy, các trường hợp đạt kết quả thấp là do vùng cần tìm kiếm chiếm một phần tương đối nhỏ trên ảnh mục tiêu, dẫn đến việc đặc trưng được rút trích từ ảnh bị tác động bởi các thông tin nhiễm(không phải vùng ảnh cần tìm kiếm), gây khó cho hệ thống đối với các trường hợp ấy.

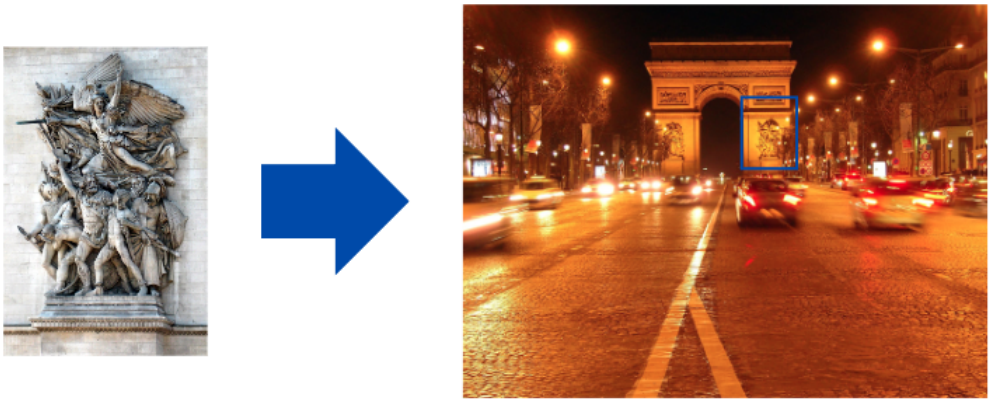

Figure 6: Ảnh truy vấn bên trái, Ảnh đáp án bên phải, với vùng cần tìm được đánh dấu

Trước quan sát đó, nhóm nhận thấy rằng nếu ta có thể giới hạn vùng cần tìm trên ảnh đáp án, có thể sẽ cải thiện được độ chính xác của hệ thống truy vấn.

Trong phạm vi, đồ án môn học xử lý ảnh, nhóm lựa chọn một hướng tiếp cận đơn giản, với ý tưởng chính là: chia ảnh thành 4 phần tư bằng nhau, và thực hiện đánh chỉ mục cho 4 ảnh đó cùng với ảnh gốc(5 ảnh cùng một id).

Ý tưởng hoạt động trên giả định rằng: các vùng cần tìm nằm ở các góc sẽ khó tìm hơn các, khi nó nằm chính giữa ảnh. Việc chia ảnh thành 4 vùng có thể mang lại hiệu quả nếu vùng cần tìm nằm trọn trong một phần tư, lúc đó sẽ loại bỏ được các đặc trưng dư thừa(gây nhiễu).

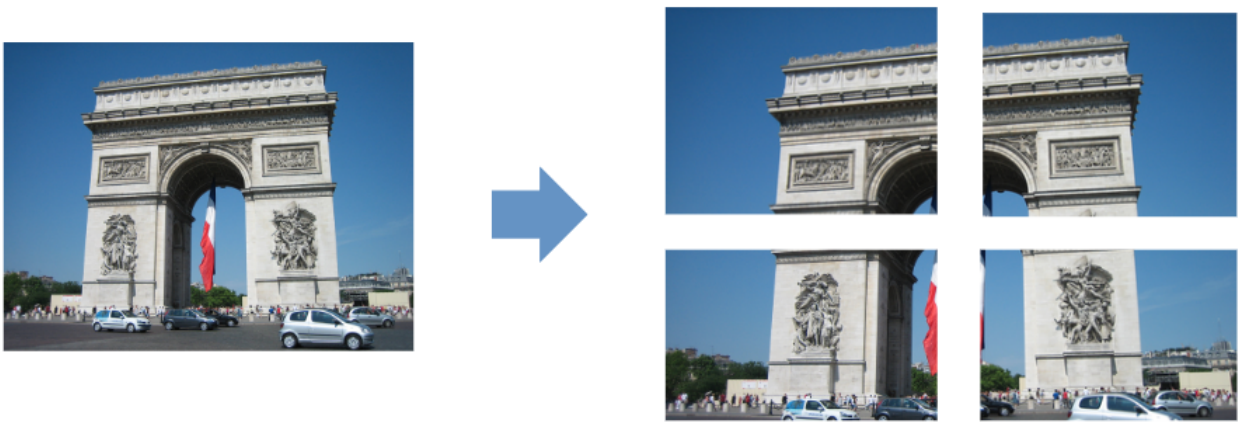

Figure 7: Chia ảnh thành 4 phần tư trước khi đánh chỉ mục.

Kết quả thực nghiệm ở phần 5 cho thấy, phương pháp tăng cường dữ liệu cho kết quả tốt hơn so với mô hình gốc, và mô hình áp dụng biến đổi ảnh truy vấn. Kết quả cải tiện từ 1-4% tuy vào từng đánh giá.

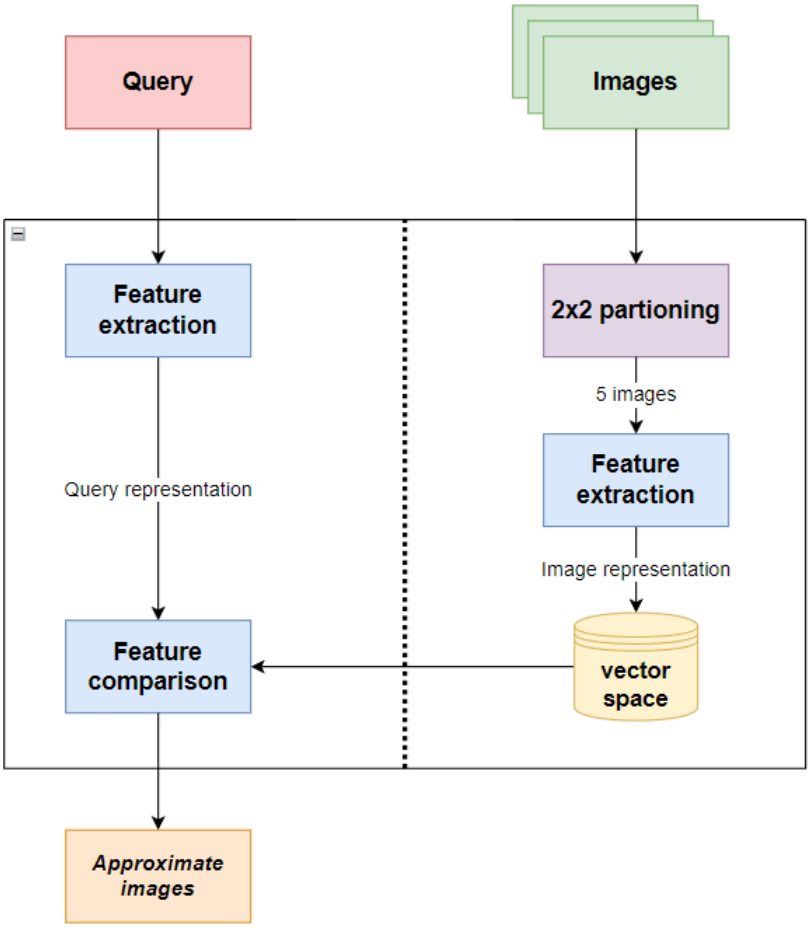

Figure 8: Hệ thống sau khi kết phương pháp tăng cường dữ liệu

4 Môi trường thực nghiệm

4.1 Môi trường thực nghiệm

Google cung cấp một môi trường thực nghiệm trực tuyến miễn phí đó là Google Colab.

Colab cung cấp GPU miễn phí thuận tiện cho quá trình huấn luyện các model deep learning.

Colab giúp làm việc nhóm hiệu quả và dễ sử dụng.

4.2 Độ đo đánh giá

mAP được tính bằng cách lấy trung bình của các giá trị AP qua tất cả các lớp.Cung cấp một giá trị số duy nhất biểu thị hiệu suất tổng thể của mô hình phát hiện đối tượng trên các loại đối tượng khác nhau.

5 Kết quả thực nghiệm

5.1 Một số thực nghiệm

Đối với phương pháp biến đổi ảnh truy vấn, nhóm lựa chọn đánh giá bằng mAP và so sánh với mô hình gốc không biến đổi ảnh truy vấn. Riêng với phương pháp tăng cường dữ liệu, nhóm mở rộng đánh giá với recall, precision và MRR, nhóm cũng thực hiện so sánh với kết quả của mô hình gốc.

5.1.1 Biến đổi ảnh truy vấn + đặc trưng HOG

	mAP@5	mAP@20	mAP@50
ORIGINAL	**0.5779**	**0.4884**	**0.4117**
CROP	0.443	0.3685	0.2934
RANDOM CROP	0.5378	0.4242	0.3489
RANDOM ROTATE	0.4740	0.4137	0.3487
HORIZONTAL FLIP	0.5708	0.4870	0.4057
BLUR	0.5720	0.4849	0.4079
SHARPEN	0.4593	0.3725	0.3395
SMOOTH	0.5748	0.4880	0.4116

5.1.2 Biến đổi ảnh truy vấn + đặc trưng Resnet50

	mAP@5	mAP@20	mAP@50
ORIGINAL	0.9519	0.9213	0.8914
CROP	**0.9636**	0.9294	0.9056
RANDOM CROP	**0.9627**	0.9300	**0.9087**
RANDOM ROTATE	**0.9627**	**0.9301**	0.9086
HORIZONTAL FLIP	**0.9627**	**0.9301**	0.9044
BLUR	**0.9627**	**0.9301**	0.9044
SHARPEN	**0.9627**	**0.9301**	0.9044
SMOOTH	**0.9627**	**0.9301**	0.9044

5.1.3 Tăng cường dữ liệu + đặc trưng Resnet50

	ORIGINAL DATABASE	AUGMENTATED DATABASE
mAP@5	0.9543	**0.9613**
mAP@20	0.9297	**0.9328**
mAP@50	0.9058	**0.9062**
mean Precision@50	0.8274	**0.8346**
Recall@Overall	0.4855	**0.5232**
Recall@Hard	0.2954	**0.3206**
MRR@Overall	**0.9659**	**0.9659**

6 Tài liệu tham khảo

[1] Computer Vision - Wikipedia

[2] Logistic Regression 1

[3] Logistic Regression 2

[4] k-Nearest Neighbors

[5] Support Vector Machine 1

[6] Support Vector Machine 2

[7] V.Vapnik - The Nature Of Statistical Learning

[8] Histogram of Oriented Gradients

[9] DATASET

[10] Information retrieval - Wikipedia

[11] Deep Residual Learning for Image Recognition

[12] Faiss - Billion-scale similarity search with GPUs