

VIET NAM NATIONAL UNIVERSITY HO CHI MINH CITY  
UNIVERSITY OF INFORMATION TECHNOLOGY  
DEPARTMENT OF COMPUTER SCIENCE

THỊ GIÁC MÁY TÍNH NÂNG CAO - CS331.O21.KHCL



BÁO CÁO CUỐI KỲ  
ĐỒ ÁN: SKETCH BASED IMAGE RETRIEVAL

Instructor: PhD. Mai Tiến Dũng

Thành viên nhóm:

1. Nguyễn Như Hà 21522028
2. Hồ Thị Khánh Hiền 21522057

TPHCM, Ngày 25 tháng 6 năm 2024

## **Lời cảm ơn**

Nhóm chúng em xin gửi lời cảm ơn chân thành đến T.S Mai Tiến Dũng - Giảng viên Khoa Khoa học máy tính, Trường Đại học Công nghệ Thông tin, Đại học Quốc gia Thành phố Hồ Chí Minh. Thầy đã tận tình giảng dạy và hướng dẫn chúng em trong môn Thị giác máy tính nâng cao, thuộc lớp CS331.O21.KHCL, cung cấp cho chúng em nhiều kiến thức và kỹ năng hữu ích.

Trong quá trình thực hiện đồ án môn học, mặc dù nhóm đã nỗ lực hết sức nhưng vẫn không thể tránh khỏi một số sai sót. Qua buổi báo cáo trên lớp, nhóm đã nhận được những góp ý từ thầy và các câu hỏi từ các bạn. Nhờ đó, chúng em đã có cơ hội hoàn thiện trên bài báo cáo hơn. Những đóng góp này giúp nhóm tích lũy thêm nhiều kinh nghiệm và kiến thức. Qua môn học này, chúng em không chỉ củng cố kiến thức chuyên môn mà còn phát triển thêm các kỹ năng tìm kiếm và nghiên cứu về vấn đề. Chúng em xin chân thành cảm ơn!

Thành phố Hồ Chí Minh, tháng 6 năm 2024

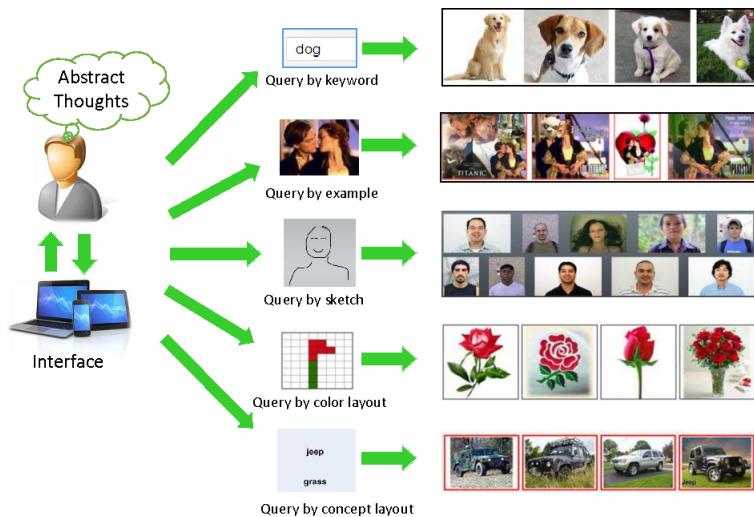
# Mục lục

<b>1 Mở đầu</b>	<b>3</b>
1.1 Giới thiệu về bài toán Image Retrieval . . . . .	3
1.2 Tổng quan về bài toán SBIR . . . . .	4
1.3 Lý do chọn đề tài . . . . .	5
1.4 Cài đặt . . . . .	5
<b>2 Nội dung</b>	<b>5</b>
2.1 Phát biểu Bài toán . . . . .	5
2.2 Phương pháp sử dụng . . . . .	6
2.2.1 Mạng nơ-ron tích chập . . . . .	6
2.2.2 Triplet Margin Loss . . . . .	8
2.2.3 Nearest Neighbor . . . . .	9
2.3 Dataset . . . . .	10
2.4 Thực nghiệm . . . . .	11
2.4.1 Tiền xử lý dữ liệu . . . . .	11
2.4.2 Tạo các triplets . . . . .	12
2.4.3 Quá trình huấn luyện . . . . .	13
2.5 Dánh giá . . . . .	15
2.5.1 Các độ đo sử dụng . . . . .	15
2.5.2 Kết quả thực nghiệm . . . . .	16
2.6 Demo . . . . .	17
<b>3 Kết luận</b>	<b>20</b>
3.1 Ưu điểm: . . . . .	20
3.2 Hạn chế: . . . . .	20
<b>4 Phân công công việc</b>	<b>22</b>

# 1 Mở đầu

## 1.1 Giới thiệu về bài toán Image Retrieval

Truy xuất ảnh là một lĩnh vực quan trọng trong xử lý ảnh và thị giác máy tính, nhằm mục đích tìm kiếm và truy vấn các hình ảnh từ một cơ sở dữ liệu lớn dựa trên các tiêu chí khác nhau. Các tiêu chí này có thể là văn bản, đặc trưng hình ảnh, hoặc nét vẽ, v.v. Việc truy xuất ảnh hiệu quả giúp người dùng dễ dàng tìm thấy các hình ảnh cần thiết trong kho dữ liệu khổng lồ, phục vụ cho nhiều mục đích khác nhau như tìm kiếm thông tin, phân tích dữ liệu, và hỗ trợ quyết định.



Hình 1: Một số phương pháp truy xuất hình ảnh.

Bài toán Image Retrieval có nhiều biến thể khác nhau, tùy thuộc vào loại dữ liệu đầu vào và cách thức truy vấn:

- Text-Based Image Retrieval (TBIR): Truy vấn hình ảnh dựa trên văn bản sử dụng các từ khóa hoặc mô tả văn bản để tìm kiếm hình ảnh.
- Content-Based Image Retrieval (CBIR): Truy vấn hình ảnh dựa trên nội dung sử dụng các đặc trưng của hình ảnh như màu sắc, kết cấu, hình dạng, và các đặc trưng khác. Các đặc trưng này được trích xuất từ hình ảnh và so sánh để tìm ra các hình ảnh tương tự trong cơ sở dữ liệu.
- Sketch-Based Image Retrieval (SBIR): Truy vấn hình ảnh dựa trên nét vẽ cho phép người dùng sử dụng một bút vẽ tay hoặc một nét vẽ đơn giản để tìm kiếm các hình ảnh tương tự.

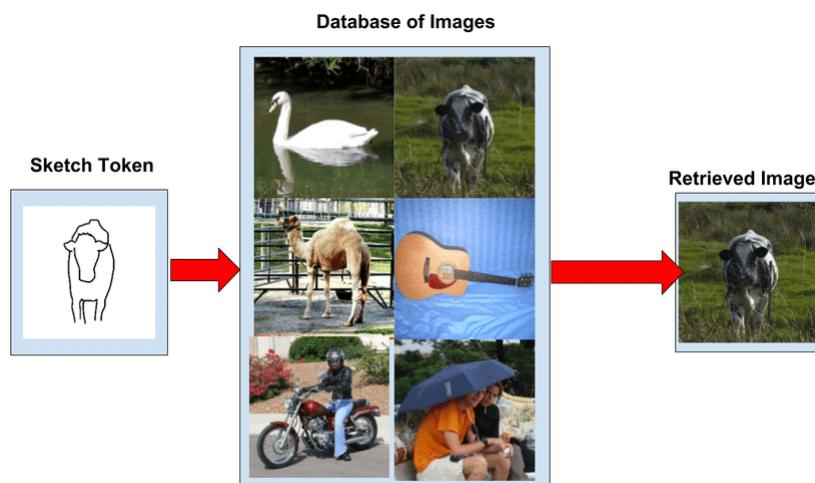
Truy xuất ảnh đóng vai trò quan trọng trong thế giới kỹ thuật số hiện đại, nơi mà hình ảnh và dữ liệu đa phương tiện ngày càng phổ biến và lớn mạnh. Việc có khả năng tìm kiếm và lấy lại các hình ảnh từ cơ sở dữ liệu khổng lồ không chỉ giúp tiết kiệm thời gian cho người dùng mà còn hỗ trợ cho nhiều ứng dụng thực tiễn. Trong lĩnh vực thương mại điện tử, truy

xuất ảnh giúp nâng cao trải nghiệm mua sắm bằng cách cung cấp các hình ảnh sản phẩm chất lượng và phù hợp. Trong giáo dục và nghiên cứu, nó hỗ trợ cho việc dạy và học bằng cách cung cấp nguồn tư liệu phong phú và đa dạng. Đặc biệt, trong y học và phân tích hình ảnh, truy xuất ảnh có thể giúp phát hiện bệnh lý và hỗ trợ quyết định điều trị. Sự phát triển của các thuật toán học sâu và công nghệ thị giác máy tính đã đóng góp tích cực vào việc cải thiện độ chính xác và hiệu quả của các hệ thống truy xuất ảnh, từ đó nâng cao khả năng ứng dụng và phát triển bền vững của các lĩnh vực này.

Kết luận lại bài toán Image Retrieval là một lĩnh vực phong phú và đa dạng trong xử lý ảnh và thị giác máy tính, với nhiều biến thể và kiểu truy vấn khác nhau. Sự phát triển của các phương pháp truy xuất ảnh hiệu quả không chỉ giúp người dùng dễ dàng tìm thấy các hình ảnh cần thiết mà còn mở ra nhiều ứng dụng mới trong các lĩnh vực khác nhau. Trong các nghiên cứu và ứng dụng hiện tại, các phương pháp học sâu đang ngày càng được sử dụng để cải thiện độ chính xác và hiệu quả của hệ thống truy xuất ảnh.

## 1.2 Tổng quan về bài toán SBIR

Một trong những phương pháp nổi bật trong lĩnh vực truy xuất ảnh là bài toán truy vấn hình ảnh dựa trên nét vẽ (Sketch-based Image Retrieval - SBIR). Phương pháp này cho phép người dùng sử dụng một bản vẽ tay hoặc một nét vẽ đơn giản để tìm kiếm các hình ảnh tương tự trong cơ sở dữ liệu. Thay vì dựa vào từ khóa hoặc mô tả văn bản, SBIR tận dụng khả năng trực quan của người dùng trong việc vẽ phác thảo để diễn đạt ý tưởng tìm kiếm của họ. Đây là một phương pháp độc đáo và hiệu quả, đặc biệt trong những tình huống mà mô tả văn bản khó diễn tả được hết các đặc điểm của đối tượng cần tìm.



Hình 2: Ví dụ về bài toán Sketch-based Image Retrieval.

Khi sử dụng SBIR, hệ thống sẽ phân tích bản phác thảo, trích xuất các đặc trưng hình học và so sánh chúng với các đặc trưng của hình ảnh trong cơ sở dữ liệu. Các hình ảnh có sự tương đồng với bản phác thảo đầu vào sẽ được hệ thống trả về, giúp người dùng nhanh chóng và dễ dàng tìm thấy những hình ảnh họ cần. Phương pháp này không chỉ mang lại cách tiếp cận trực quan và hiệu quả trong việc truy xuất hình ảnh, mà còn mở rộng khả năng

tìm kiếm hình ảnh trong nhiều lĩnh vực ứng dụng khác nhau như thiết kế đồ họa, nghiên cứu khoa học, và nhận diện đối tượng.

### 1.3 Lý do chọn đề tài

Mục tiêu của truy xuất hình ảnh dựa trên bản phác thảo là tạo ra một công cụ cho phép người dùng không cần phải là nghệ sĩ vẽ nội dung trực quan, nhưng vẫn có thể tìm kiếm và truy cập các hình ảnh phù hợp từ một bộ sưu tập lớn. Điều này đặc biệt hữu ích để:

- Tạo ra sự thuận tiện cho người dùng: Truy xuất hình ảnh dựa trên bản phác thảo giúp người dùng tiết kiệm thời gian và công sức bằng cách loại bỏ bước vẽ hoặc mô tả hình ảnh bằng ngôn ngữ. Thay vào đó, họ chỉ cần vẽ một phác thảo đơn giản để biểu diễn ý tưởng của họ.
- Khắc phục hạn chế ngôn ngữ: Trong một số trường hợp, việc mô tả hình ảnh bằng ngôn ngữ có thể gặp khó khăn, đặc biệt đối với những người không có kiến thức chuyên môn về ngôn ngữ hoặc về lĩnh vực đó. Truy xuất hình ảnh dựa trên phác thảo giúp vượt qua rào cản ngôn ngữ này.

### 1.4 Cài đặt

IDE: Kaggle (GPU T4 x2)

Ngôn ngữ lập trình: Python

## 2 Nội dung

### 2.1 Phát biểu Bài toán

Như đã được giới thiệu, mục tiêu chính của bài toán truy vấn hình ảnh dựa trên nét vẽ là phát triển một hệ thống cho phép người dùng tìm kiếm và truy vấn các hình ảnh trong một cơ sở dữ liệu lớn bằng cách sử dụng các bản vẽ tay đơn giản thay vì các từ khóa hoặc mô tả văn bản.

#### Input:

- 1 bản vẽ được người dùng cung cấp bằng cách tải lên hoặc vẽ bằng công cụ kỹ thuật số.
- Bộ dữ liệu bao gồm các ảnh thực và ảnh phác thảo tương ứng với hình ảnh đó.

#### Output:

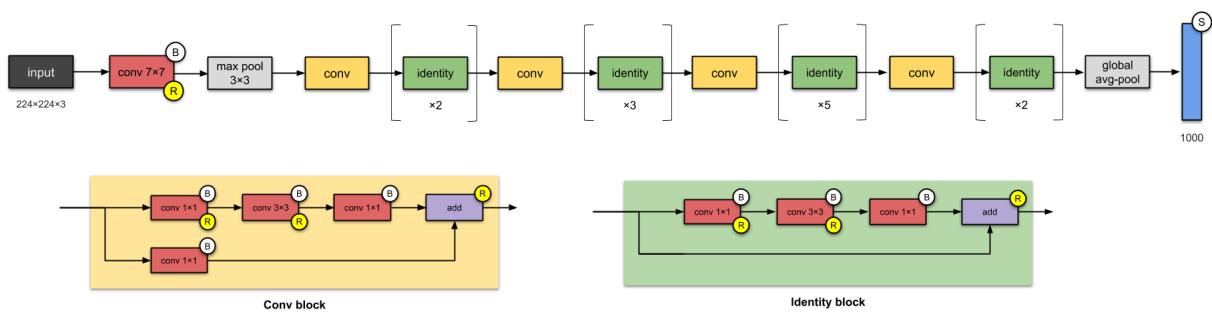
- Tập hợp các ảnh từ bộ dữ liệu mà được đánh giá là phù hợp nhất với ảnh phác thảo truy vấn. Các ảnh này được sắp xếp từ trên xuống dưới theo mức độ tương đồng với ảnh truy vấn.

## 2.2 Phương pháp sử dụng

### 2.2.1 Mạng nơ-ron tích chập

#### Resnet50

ResNet50 [1] là một mạng nơ-ron tích chập (CNN) có độ sâu 50 lớp, được phát triển bởi He et al. trong bài báo "Deep Residual Learning for Image Recognition" vào năm 2015. Đây là một phần của kiến trúc Residual Network (ResNet), được thiết kế để giải quyết các vấn đề về suy giảm độ chính xác khi độ sâu của mạng tăng lên. ResNet50 đã đạt được kết quả đột phá trên nhiều bài toán nhận dạng hình ảnh tiêu chuẩn, bao gồm ImageNet, COCO và Places.



Hình 3: Kiến trúc ResNet50.

ResNet50 [2] là một phiên bản cụ thể của kiến trúc Residual Network với tổng cộng 50 lớp (layers), bao gồm:

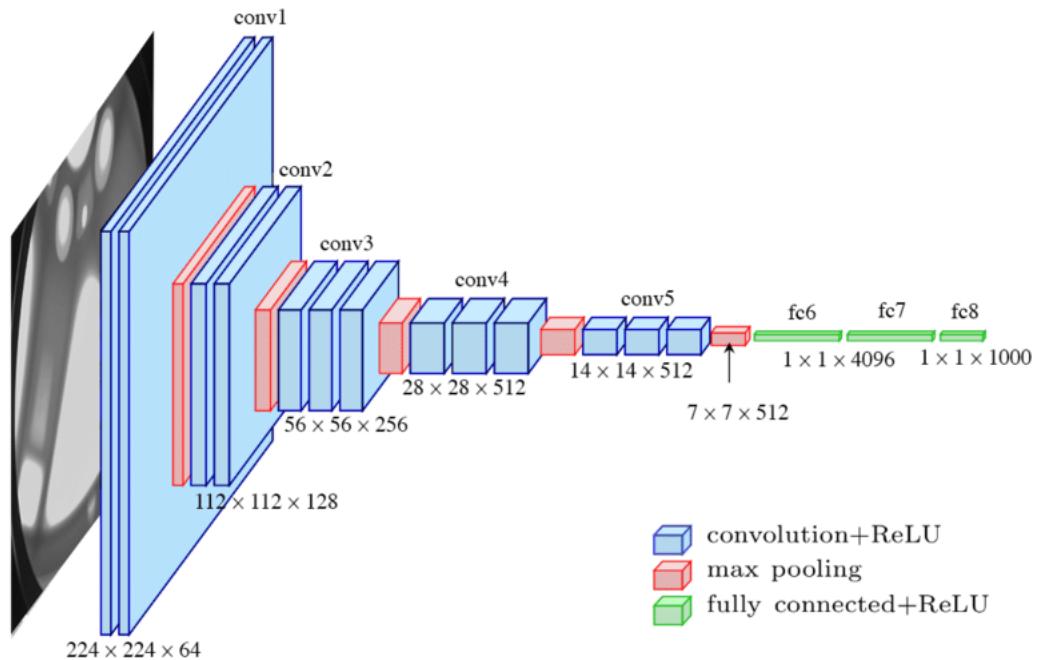
- 48 lớp tích chập (convolutional layers):
  - Các lớp tích chập trong ResNet-50 được sử dụng để trích xuất các đặc trưng (features) của hình ảnh đầu vào. Những lớp này sử dụng các bộ lọc (filters) với kích thước 1x1, 3x3, và 7x7 để quét qua hình ảnh, phát hiện các đặc trưng như cạnh, góc, và các họa tiết phức tạp hơn.
  - Các lớp tích chập được tổ chức thành các khối residual (residual blocks), mỗi khối gồm ba lớp tích chập. Các khối residual này giúp duy trì và truyền gradient hiệu quả, giảm thiểu vấn đề gradient biến mất, và cho phép mạng học được các mối quan hệ phức tạp giữa các đặc trưng hình ảnh.
- 1 lớp MaxPooling: được sử dụng sau lớp tích chập đầu tiên để giảm kích thước không gian (spatial size) của các đặc trưng, giảm số lượng tham số và tính toán, và làm nổi bật các đặc trưng quan trọng nhất. Lớp này có kích thước bộ lọc 3x3 và stride 2.
- 1 lớp AveragePooling được sử dụng ở cuối mạng trước khi lớp fully connected để giảm kích thước không gian cuối cùng của các đặc trưng. Lớp này giúp tổng hợp thông tin từ các đặc trưng cuối cùng và chuẩn bị dữ liệu cho lớp fully connected cuối cùng.

Cơ chế hoạt động:

- input đầu vào là một ảnh có kích thước  $3 \times 244 \times 244$ .
- Ảnh đi qua một lớp tích chập ban đầu  $7 \times 7$  để trích xuất các đặc trưng cơ bản, sau đó là lớp MaxPooling để giảm kích thước của ảnh.
- Các khối residual, gồm các lớp tích chập  $1 \times 1$ ,  $3 \times 3$  và  $1 \times 1$ , được lặp lại nhiều lần để học các đặc trưng phức tạp hơn và duy trì thông tin.
- Sau khi đi qua các khối residual, ảnh được gộp trung bình toàn cục thành một vector.
- Vector này được đưa vào các lớp fully connected để phân loại ảnh vào 1000 lớp khác nhau.
- Kết quả cuối cùng là một vector xác suất, biểu thị xác suất của mỗi lớp phân loại trong tập dữ liệu.

## VGG16

VGG16 là một mô hình mạng nơ-ron tích chập (Convolutional Neural Network - CNN) nổi tiếng, được phát triển bởi nhóm nghiên cứu tại Đại học Oxford, cụ thể là nhóm Visual Geometry Group (VGG). Mô hình này đã được giới thiệu lần đầu tại cuộc thi ImageNet Large Scale Visual Recognition Challenge (ILSVRC) năm 2014 và nhanh chóng trở thành một trong những mô hình phổ biến nhất trong lĩnh vực nhận diện và phân loại hình ảnh. [3]



Hình 4: Cấu trúc của mô hình VGG16.

VGG16 có tổng cộng 16 lớp (layers) có trọng số (weights), bao gồm:

- 13 lớp tích chập (convolutional layers): Các lớp này sử dụng bộ lọc (filter) có kích thước  $3 \times 3$  để quét qua hình ảnh đầu vào, giúp phát hiện các đặc trưng (features) khác nhau của hình ảnh như cạnh, góc và các họa tiết phức tạp hơn.
- 3 lớp kết nối đầy đủ (fully connected layers): Các lớp này nằm ở cuối mạng và chịu trách nhiệm đưa ra quyết định cuối cùng dựa trên các đặc trưng đã được trích xuất bởi các lớp tích chập.

Ngoài ra, mô hình còn sử dụng các lớp pooling (thường là MaxPooling) để giảm kích thước không gian của các đặc trưng, giúp giảm số lượng tham số và tính toán trong mô hình.

Cách thức hoạt động của VGG16:

- Đầu vào của mô hình là một ảnh RGB có kích thước  $3 \times 224 \times 224$
- Ảnh đầu vào được truyền qua các lớp tích chập (Convolutional layers) trong phần "features" của mô hình. Mỗi lớp tích chập thực hiện việc tích chập ảnh với các bộ lọc (kernel), sau đó kích hoạt bằng hàm ReLU.
- Sau mỗi lớp tích chập, ảnh được truyền qua lớp Max-Pooling để giảm kích thước của ảnh, giúp giảm số lượng tham số và tính toán.
- Sau khi ảnh đã đi qua các lớp tích chập và Max-Pooling, nó được làm phẳng (flattened) và đưa vào các lớp fully connected của mô hình.
- Kết quả cuối cùng sau khi ảnh đã đi qua các lớp phân loại đầy đủ là một vector xác suất, cho biết xác suất của mỗi lớp phân loại trong tập dữ liệu. Điều này thường được thực hiện bằng cách sử dụng hàm softmax để chuyển đổi các giá trị đầu ra thành phân phối xác suất.

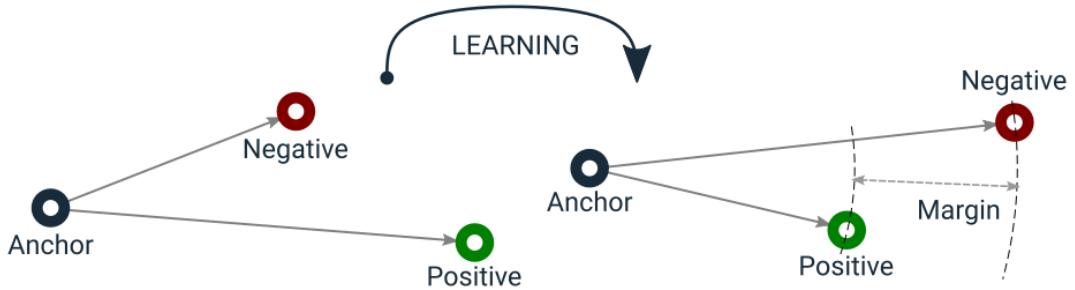
### 2.2.2 Triplet Margin Loss

Triplet Margin Loss là một hàm mất mát (loss function) được sử dụng phổ biến trong các bài toán học sâu liên quan đến nhận diện và xác thực danh tính, chẳng hạn như nhận diện khuôn mặt, nhận dạng hình ảnh và truy xuất hình ảnh dựa trên bản phác thảo (Sketch-Based Image Retrieval - SBIR). [4]

Triplet Margin Loss được sử dụng để học các biểu diễn đặc trưng (feature representations) sao cho các hình ảnh tương tự nhau được biểu diễn gần nhau trong không gian đặc trưng, còn các hình ảnh không tương tự nhau thì cách xa nhau.

Triplet Margin Loss sử dụng một bộ ba (triplet) gồm ba mẫu dữ liệu:

- Anchor (A): Hình ảnh hoặc đặc trưng chính mà các hình ảnh khác sẽ được so sánh với.
- Positive (P): Hình ảnh hoặc đặc trưng tương tự, thuộc cùng lớp với Anchor.
- Negative (N): Hình ảnh hoặc đặc trưng khác biệt, thuộc lớp khác với Anchor.

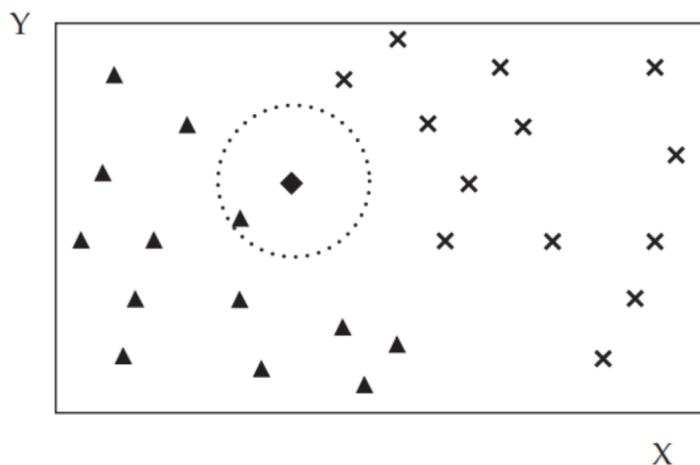


Hình 5: Mục tiêu học của Triplet Loss.

Mục tiêu chính của Triplet Margin Loss là tối ưu hóa hàm mất mát sao cho các biểu diễn đặc trưng của các cặp tương tự (Anchor và Positive) gần nhau trong không gian đặc trưng, trong khi đó các biểu diễn đặc trưng của các cặp không tương tự (Anchor và Negative) phải cách xa nhau ít nhất là một khoảng margin. Điều này giúp mô hình phân biệt hiệu quả hơn giữa các đối tượng tương tự và không tương tự.

### 2.2.3 Nearest Neighbor

Thuật toán Nearest Neighbor (NN) [5] là một trong những phương pháp học máy đơn giản và cơ bản nhất được sử dụng rộng rãi trong nhiều lĩnh vực, bao gồm phân loại, hồi quy và truy xuất thông tin. Nguyên tắc chính của thuật toán này là dựa vào khoảng cách giữa các điểm dữ liệu để tìm điểm dữ liệu gần nhất với một điểm cho trước.

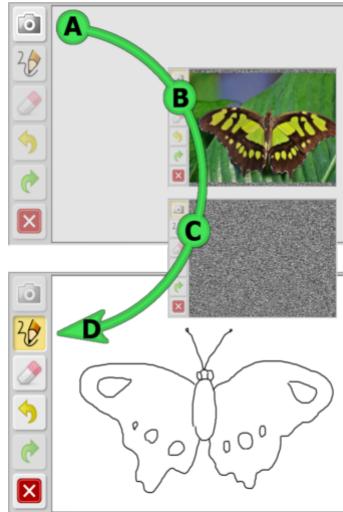


Hình 6: Tổng quan về Nearest Neighbor.

Nearest Neighbor hoạt động bằng cách tính toán khoảng cách giữa điểm cần dự đoán và tất cả các điểm trong tập dữ liệu huấn luyện. Khoảng cách thường được tính bằng các phương pháp như khoảng cách Euclidean, khoảng cách Manhattan hoặc khoảng cách cosine, tùy thuộc vào bài toán cụ thể. Sau khi tính toán khoảng cách, thuật toán sẽ chọn ra điểm dữ liệu gần nhất so với điểm cần dự đoán.

## 2.3 Dataset

The Sketchy Database [6] là một cơ sở dữ liệu lớn và phong phú dành cho các nghiên cứu liên quan đến nhận dạng hình ảnh và truy xuất hình ảnh dựa trên bản phác thảo (Sketch-Based Image Retrieval - SBIR). Cơ sở dữ liệu này được phát triển nhằm cung cấp một nguồn tài nguyên đa dạng và thực tế để thúc đẩy sự phát triển của các thuật toán và mô hình học máy trong lĩnh vực SBIR.

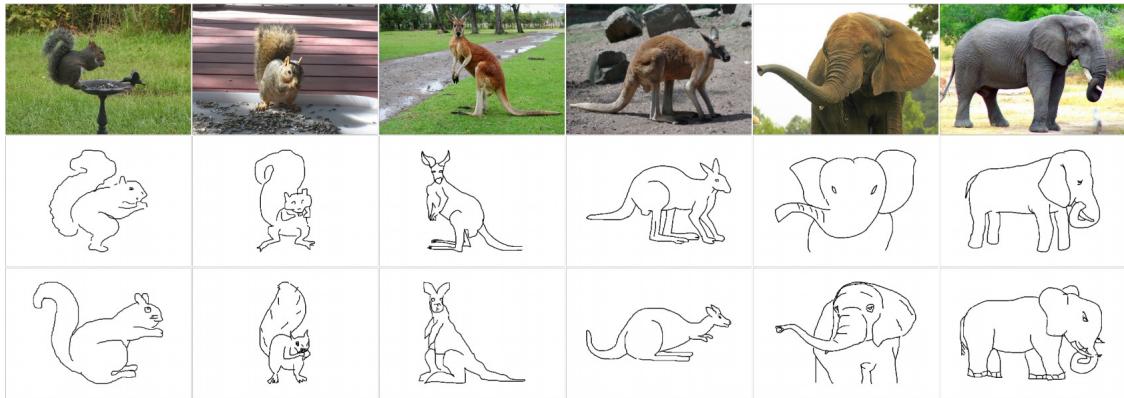


Hình 7: Giao diện thu thập các bản phác thảo.

Hình ảnh trên là giao diện thu thập các bản phác thảo của nhóm tác giả được thiết kế để thu thập các bản phác thảo từ người tham gia, giúp tạo ra một cơ sở dữ liệu phong phú và thực tế. Giao diện này đảm bảo rằng người tham gia có thể tạo ra các bản phác thảo một cách tự nhiên và thuận tiện, đồng thời đảm bảo tính chính xác và độ tin cậy của dữ liệu thu thập được.

Khi bắt đầu, người tham gia sẽ thấy một màn hình trắng (blank canvas). Đây là nơi mà người tham gia sẽ vẽ bản phác thảo của mình. Khi người tham gia nhấn một nút, một hình ảnh thực sẽ được hiển thị trong vòng 2 giây. Hình ảnh này là đối tượng mà người tham gia sẽ phác thảo.

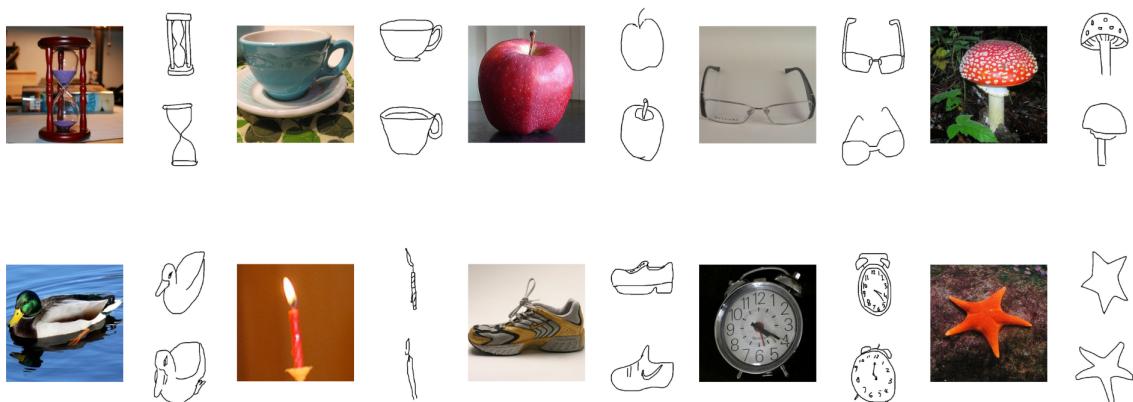
Sau khi hình ảnh thực được hiển thị, một noise mask sẽ xuất hiện trong vòng 1 giây. Noise mask này giúp làm mờ hình ảnh thực, tránh việc người tham gia sao chép trực tiếp hình ảnh mà phải dựa vào trí nhớ của mình để vẽ. Người tham gia sau đó sẽ sử dụng các công cụ như bút chì, tẩy và công cụ hoàn tác (undo) để tạo ra bản phác thảo của họ trên màn hình trắng ban đầu.



Hình 8: Các ví dụ minh họa về các cặp hình ảnh và bản phác thảo từ cơ sở dữ liệu Sketchy.

Nhóm tác giả đã thuê một nhóm người trong cộng đồng để vẽ phác thảo dựa trên 125 lớp gồm các ảnh vật thể khác nhau, thu thập được hơn 75 ngàn phác thảo của 12500 đối tượng.

Vì số lượng ảnh quá nhiều và để huấn luyện mô hình cho hơn 75,000 ảnh sẽ tốn rất nhiều thời gian, nhóm đã quyết định giảm xuống chỉ còn 10 lớp. Việc này sẽ giúp nhóm tối ưu hóa tài nguyên tính toán và tăng tốc quá trình huấn luyện.



Hình 9: Cặp hình ảnh và bản phác thảo minh họa cho 10 lớp.

Tỉ lệ cho tập train và test là 8:2, với mỗi lớp trong dataset có 100 ảnh thực, nhóm tiến hành lấy 80 ảnh cho tập train và 20 ảnh cho tập test. Với mỗi ảnh thực trong tập train và test thì sẽ lấy các ảnh sketch tương ứng với mỗi ảnh thực đó và nhóm sẽ huấn luyện trên số lượng ảnh sketch tương ứng, sẽ là 4584 ảnh sketch cho tập train.

## 2.4 Thực nghiệm

### 2.4.1 Tiền xử lý dữ liệu

Để chuẩn bị dữ liệu đầu vào cho các mô hình CNN như ResNet50 và VGG16 trong dự án nghiên cứu của nhóm, nhóm đã thực hiện một chuỗi các bước tiền xử lý ảnh. Đầu tiên, các hình ảnh được chuyển đổi sang không gian màu RGB và được thay đổi kích thước về 224x224 pixel. Việc này đảm bảo rằng tất cả các hình ảnh đều có kích thước chuẩn, phù hợp với yêu cầu của các mô hình CNN.

```

self.transform = transforms.Compose([
    transforms.Resize(224),
    transforms.CenterCrop(224),
    transforms.ToTensor(),
    transforms.Normalize(mean=[0.485, 0.456, 0.406],
                        std=[0.229, 0.224, 0.225])
])

```

Hình 10: Các bước tiền xử lý dữ liệu.

Sau đó, nhóm chuyển đổi các hình ảnh thành đối tượng tensor, định dạng dữ liệu mà các mô hình CNN của thư viện PyTorch có thể sử dụng để huấn luyện và dự đoán. Tiếp theo, nhóm tiến hành chuẩn hóa dữ liệu ảnh với các tham số mean và std được lựa chọn cẩn thận. Điều này giúp điều chỉnh giá trị của từng pixel trong ảnh để nằm trong phạm vi phù hợp, nhằm tối ưu hóa việc học tập và khả năng tổng quát hóa của mô hình trên dữ liệu mới.

Đặc biệt, nhóm đã cân nhắc các tham số mean=[0.485, 0.456, 0.406] và std=[0.229, 0.224, 0.225] cho việc chuẩn hóa dữ liệu. Đây là các giá trị thống kê tiêu chuẩn của tập dữ liệu ImageNet, được sử dụng rộng rãi để đảm bảo tính ổn định và hiệu quả của quá trình huấn luyện trên các mô hình CNN như ResNet50 và VGG16.

#### 2.4.2 Tạo các triplets



Hình 11: Minh họa về các triplets.

Trong quá trình huấn luyện với hàm mất mát Triplet Margin Loss, nhóm đã thực hiện việc lựa chọn các bộ ba dữ liệu (triplets) một cách tỉ mỉ để đảm bảo tính hiệu quả và đa dạng của mô hình. Mỗi bộ ba bao gồm ba điểm dữ liệu: anchor (điểm dữ liệu chính), positive (điểm dữ liệu cùng lớp với anchor), và negative (điểm dữ liệu khác lớp với anchor).

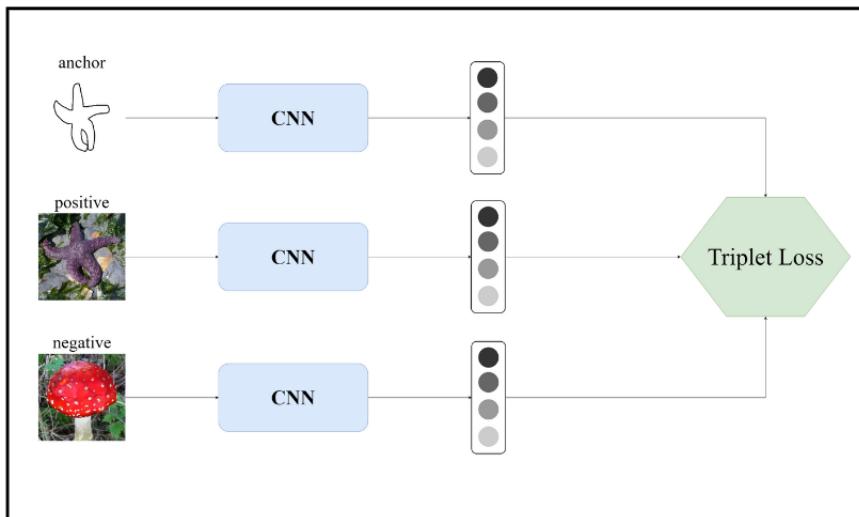
Dối với mỗi ảnh trong tập sketch, nhóm tiến hành lựa chọn một ảnh làm anchor, đại diện cho điểm dữ liệu chính của sketch đó. Tiếp theo, nhóm chọn ảnh thực tương ứng với ảnh sketch đó làm positive, đảm bảo rằng positive cùng lớp với anchor. Đồng thời, để tăng

tính đa dạng và độ phong phú của dữ liệu, nhóm cũng lựa chọn một ảnh thực khác, không thuộc cùng lớp với anchor, làm negative trong bộ ba dữ liệu.

Quá trình lựa chọn triplet này được thực hiện một cách tự động và ngẫu nhiên, nhằm đảm bảo rằng mô hình được huấn luyện trên các bộ dữ liệu đại diện và mang tính đại diện cao của từng lớp. Việc áp dụng hàm mất mát Triplet Margin Loss giúp mô hình học được sự tương đồng giữa các cặp sketch và ảnh thực cùng lớp, đồng thời tối thiểu hóa khoảng cách giữa các cặp sketch và ảnh thực khác lớp, từ đó nâng cao khả năng truy xuất hình ảnh dựa trên bản phác thảo (SBIR).

#### 2.4.3 Quá trình huấn luyện

Quá trình huấn luyện mô hình trong nghiên cứu của nhóm bao gồm việc áp dụng hàm mất mát Triplet Margin Loss để tối ưu hóa các biểu diễn đặc trưng của ảnh. Mỗi triplet gồm ba loại ảnh: anchor (ảnh phác thảo), positive (ảnh thực cùng lớp với anchor), và negative (ảnh thực khác lớp với anchor), được đưa vào mô hình để tính toán các biểu diễn đặc trưng tương ứng.



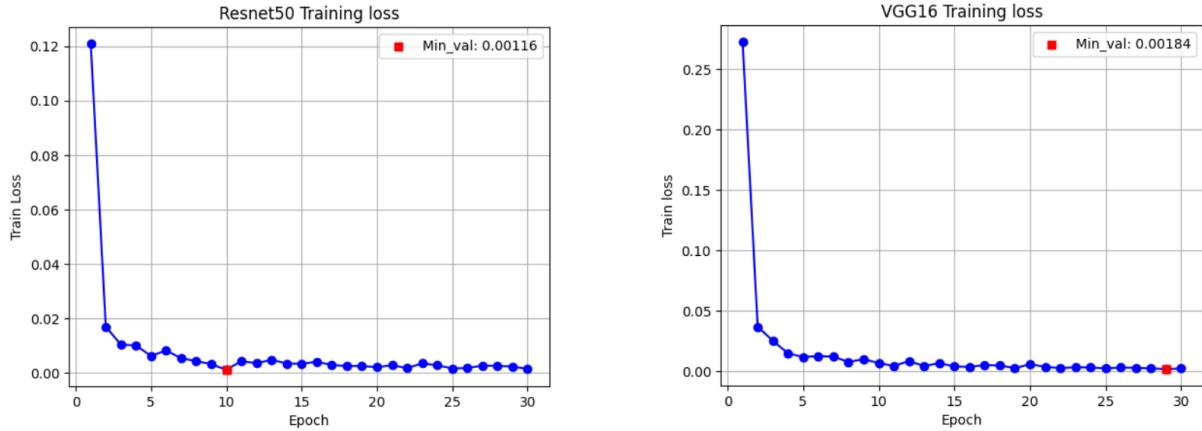
Hình 12: Các hoạt động chính trong quá trình huấn luyện mô hình.

Dối với mỗi triplet, nhóm sử dụng mô hình mạng nơ-ron tích chập (CNN) đã được huấn luyện trước trên ImageNet như ResNet50 hoặc VGG16 để trích xuất các đặc trưng từ ảnh. Các biểu diễn đặc trưng này được truyền vào hàm mất mát Triplet Margin Loss, nơi mà chúng được sử dụng để tính toán khoảng cách giữa các điểm trong không gian đặc trưng. Từ đó mô hình tiếp tục được huấn luyện trên bộ dữ liệu của nhóm bằng Triplet Margin Loss.

Trong quá trình huấn luyện, nhóm thiết lập các tham số của hàm mất mát như sau: **margin = 1** để định rõ ranh giới giữa các điểm dữ liệu, và **p = 2** để tính toán khoảng cách theo khoảng cách Euclid. Quá trình huấn luyện được thực hiện trong **30 epochs** với **learning rate = 1e-3** và **batch size = 32**.

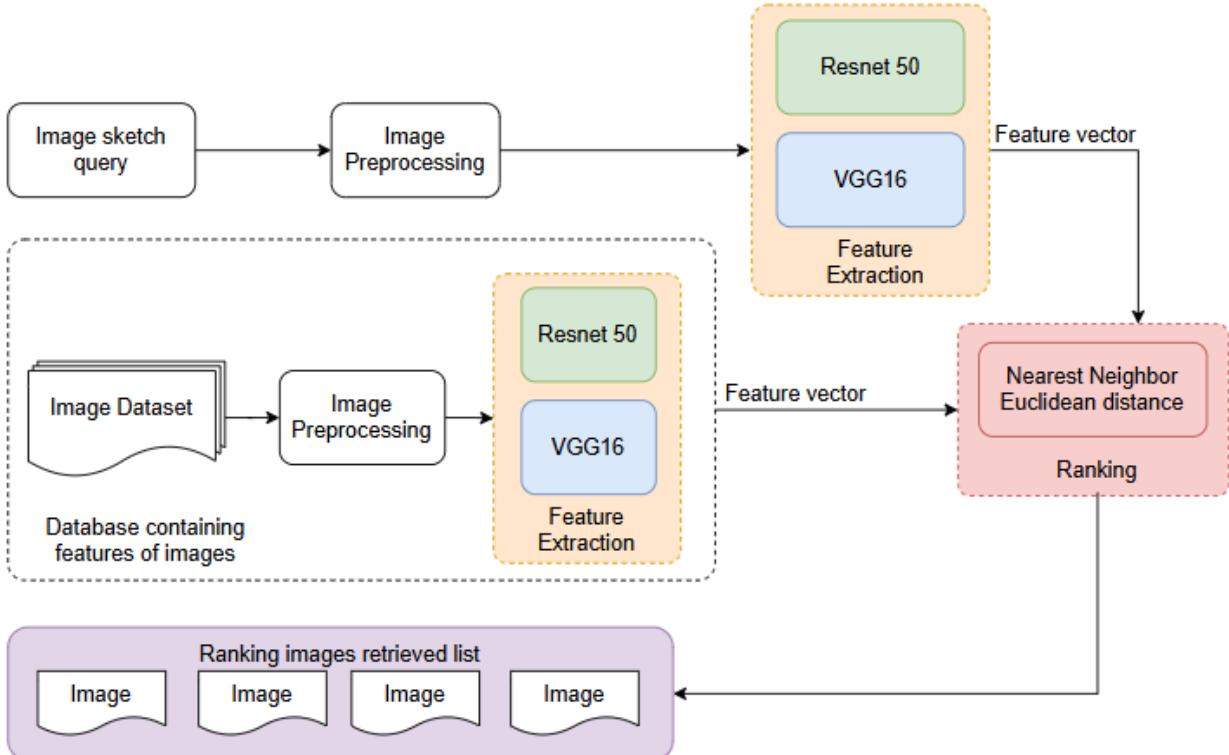
Việc lựa chọn các tham số này được thực hiện để đảm bảo rằng mô hình có thể học được các biểu diễn đặc trưng hiệu quả từ dữ liệu và cải thiện khả năng truy xuất hình ảnh dựa

trên bản phác thảo. Quá trình này cần đến sự cân nhắc kỹ lưỡng và thử nghiệm để đạt được hiệu quả cao nhất trong việc huấn luyện và đánh giá mô hình.



Hình 13: Training loss của Resnet50 và VGG16.

Khi một ảnh truy vấn được đưa vào hệ thống, quy trình xử lý diễn ra qua nhiều bước để đảm bảo rằng ảnh được so khớp chính xác với các ảnh trong cơ sở dữ liệu. Đầu tiên, ảnh truy vấn sẽ được tiền xử lý. Quá trình này bao gồm việc thay đổi kích thước ảnh về chuẩn 224x224 pixel, chuyển đổi ảnh sang không gian màu RGB nếu cần thiết, và sau đó chuyển đổi ảnh thành tensor. Các tensor này tiếp tục được chuẩn hóa với các tham số mean và std tương tự như quá trình huấn luyện để đảm bảo tính nhất quán và hiệu quả.



Hình 14: Workflow khi thực hiện truy vấn ảnh.

Sau khi ảnh truy vấn đã được chuẩn bị, nó được đưa vào mô hình mạng nơ-ron tích chập (CNN) đã được huấn luyện trước, ResNet50 hoặc VGG16, để trích xuất các đặc trưng. Các

mô hình này có khả năng trích xuất những đặc trưng quan trọng từ ảnh, chuyển chúng thành các biểu diễn đặc trưng dạng vector trong không gian đa chiều. Các vector này là một dạng biểu diễn nén của thông tin ảnh, giúp mô hình so sánh và phân loại ảnh một cách hiệu quả.

Tiếp theo, các đặc trưng của ảnh truy vấn được so sánh với các đặc trưng của tất cả các ảnh trong cơ sở dữ liệu. Quá trình so sánh này sử dụng khoảng cách Euclid để đo lường mức độ tương đồng giữa các vector đặc trưng. Khoảng cách càng nhỏ thì hai ảnh càng giống nhau. Do đó, hệ thống sẽ tính toán khoảng cách giữa vector đặc trưng của ảnh truy vấn và vector đặc trưng của từng ảnh trong cơ sở dữ liệu.

Dựa trên các khoảng cách đã tính, hệ thống sẽ sắp xếp các ảnh trong cơ sở dữ liệu theo thứ tự từ gần nhất đến xa nhất so với ảnh truy vấn. Các ảnh có khoảng cách nhỏ nhất, tức là các ảnh có đặc trưng gần giống nhất với ảnh truy vấn, sẽ được ưu tiên hiển thị trước. Danh sách các ảnh này, kèm theo thông tin liên quan như tên đường dẫn ảnh, được trả về cho người dùng.

Cuối cùng, hệ thống hiển thị các kết quả truy xuất lên giao diện người dùng. Người dùng có thể xem qua các ảnh được trả về để xác định ảnh nào phù hợp với mục đích tìm kiếm của mình. Quy trình này không chỉ đảm bảo tính chính xác trong việc so khớp ảnh mà còn tối ưu hóa tốc độ truy xuất, giúp người dùng nhanh chóng tìm thấy ảnh mong muốn trong một cơ sở dữ liệu lớn.

Qua quy trình chi tiết này, hệ thống có thể thực hiện việc truy xuất ảnh dựa trên bản phác thảo (Sketch-Based Image Retrieval) một cách hiệu quả, đảm bảo rằng người dùng nhận được các kết quả phù hợp và chính xác.

## 2.5 Đánh giá

### 2.5.1 Các độ đo sử dụng

Trong hệ thống truy xuất thông tin (IR), các phương pháp đánh giá đo lường đánh giá mức độ hiệu quả của hệ thống từ việc trả về kết quả từ một bộ tài nguyên dữ liệu để đáp ứng yêu cầu của người dùng. Các kết quả hiệu suất này là nền tảng cho sự thành công hay không của hệ thống truy xuất thông tin. Yếu tố quan trọng nhất trong việc xác định tính hiệu quả của hệ thống đối với người dùng là mức độ liên quan tổng thể của các kết quả được truy xuất.

Trong bài toán truy vấn hình ảnh dựa trên nét vẽ (SBIR), nhóm sử dụng một độ đo đánh giá phổ biến trong truy xuất thông tin, gọi là Mean Average Precision (mAP)[7]. mAP đánh giá chất lượng của các kết quả truy xuất được xếp hạng bằng cách đánh giá cả precision và recall của dự đoán của hệ thống. Với mỗi truy vấn hình ảnh, hệ thống trả về 30 hình ảnh.

Mean Average Precision (MAP) [8] cho một tập các truy vấn là trung bình của các điểm Average Precision (AP) cho mỗi truy vấn. AP tính toán trung bình các giá trị precision tại mỗi vị trí liên quan trong danh sách xếp hạng các mục đã được truy xuất.

Precision [8] là tỷ lệ giữa số lượng hình ảnh liên quan được hệ thống truy xuất đúng so

với tổng số hình ảnh được truy xuất. Có công thức:

$$Precision = \frac{NumberOfRelevantItemsRetrieved}{TotalNumberOfItemsRetrieved}$$

Average Precision [8] là giá trị trung bình của precision tại các vị trí trong danh sách kết quả nơi có hình ảnh liên quan. Có công thức:

$$AveP = \frac{\sum_{k=1}^n P(k) * rel(k)}{TotalNumberOfRelevantItems}$$

,trong đó:

- k duyệt qua các vị trí mà các mục liên quan được truy xuất.
- n là tổng số lượng các mục liên quan.
- rel(k) là một hàm chỉ thị bằng 1 nếu mục tại vị trí k là một hình ảnh (item) liên quan và bằng 0 nếu không.
- P(k) là độ chính xác khi chỉ xem xét các mục được truy xuất đầu tiên k.

Lưu ý rằng điểm trung bình sẽ được tính trên các truy vấn liên quan trong top-k và các ảnh (item) liên quan không được truy xuất sẽ có điểm precision bằng không.

Công thức tính Mean Average Precision (mAP):

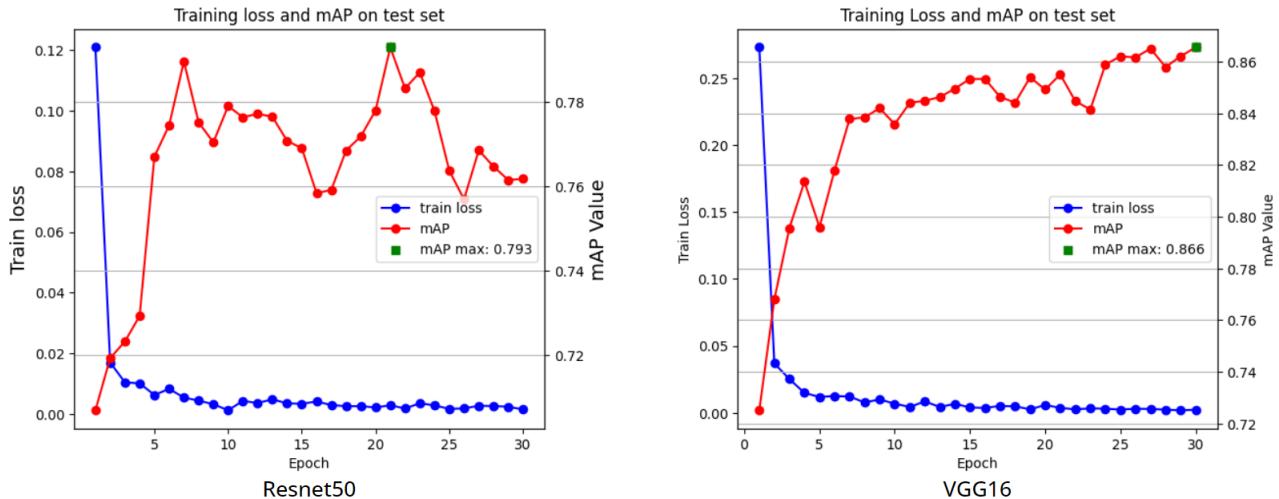
$$mAP = \frac{\sum_{q=1}^Q AveP(q)}{Q}$$

,trong đó Q là số lượng truy vấn.

Điểm mAP càng cao, hiệu suất của hệ thống SBIR càng tốt. MAP là một độ đo được sử dụng rộng rãi để đánh giá các hệ thống SBIR, vì nó tính đến sự liên quan của mỗi hình ảnh được truy xuất và cung cấp một đánh giá toàn diện về hiệu suất của hệ thống qua nhiều truy vấn.

### 2.5.2 Kết quả thực nghiệm

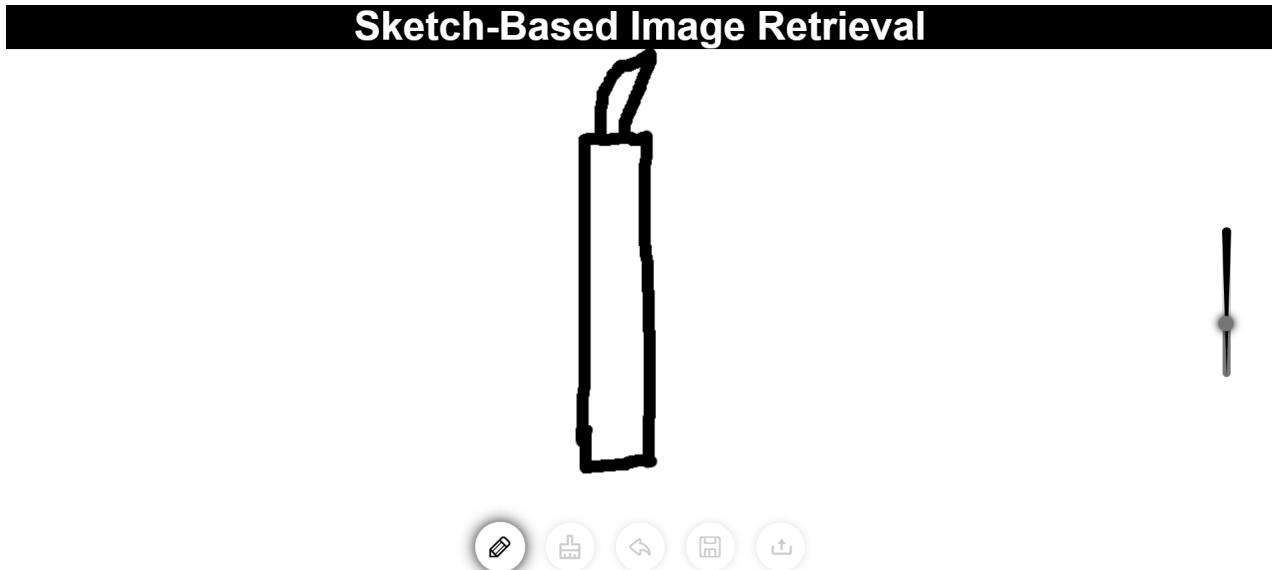
Dưới đây là hai đồ thị biểu diễn điểm mAP và độ mất mát (loss) qua từng epoch của hai phương pháp ResNet-50 và VGG16. Đường màu xanh biểu thị giá trị loss trung bình của mỗi epoch, trong khi đường màu đỏ thể hiện giá trị mAP qua từng epoch, với số lượng truy vấn top k = 30. Điểm mAP cao nhất của ResNet-50 đạt 0.793 vào epoch thứ 21; trong khi đó, VGG16 đạt điểm mAP là 0.866 vào epoch 30. Mặc dù VGG16 có cấu trúc đơn giản hơn so với ResNet-50, nhưng kết quả thử nghiệm cho thấy chỉ số mAP của VGG16 cao hơn so với ResNet-50. Điều này cho thấy VGG16 có khả năng truy xuất hình ảnh phù hợp với ảnh phác thảo của người dùng tốt hơn trong trường hợp cụ thể này.



Hình 15: Điểm mAP của Resnet50 và VGG16.

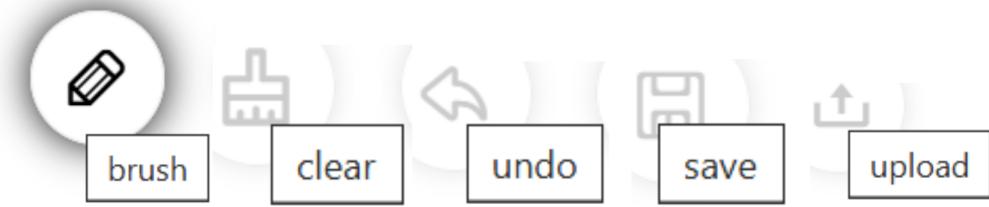
## 2.6 Demo

Khi người dùng mở giao diện website của hệ thống truy xuất ảnh dựa trên phác thảo (Sketch-Based Image Retrieval), họ sẽ thấy một giao diện chính được thiết kế để tạo sự thuận tiện và linh hoạt trong quá trình sử dụng. Trên giao diện chính, một canvas trắng lớn chiếm phần lớn diện tích màn hình, cho phép người dùng dễ dàng vẽ các bản phác thảo. Đây là nơi người dùng có thể sáng tạo và vẽ bất kỳ hình ảnh nào mà họ muốn sử dụng làm truy vấn tìm kiếm.



Hình 16: Giao diện của bài toán Sketch-Based Image Retrieval.

Một thanh công cụ được đặt dọc theo cạnh phải của màn hình, cho phép người dùng điều chỉnh độ dày của bút vẽ. Thanh này giúp người dùng có thể dễ dàng thay đổi độ dày hoặc mỏng của các đường vẽ, tùy thuộc vào chi tiết và phong cách của phác thảo mà họ muốn tạo ra. Việc này cung cấp thêm một mức độ tùy chỉnh, giúp người dùng có thể vẽ chính xác hơn và dễ dàng tạo ra các bản phác thảo phù hợp với yêu cầu tìm kiếm.



Hình 17: Thanh bottom bar của giao diện.

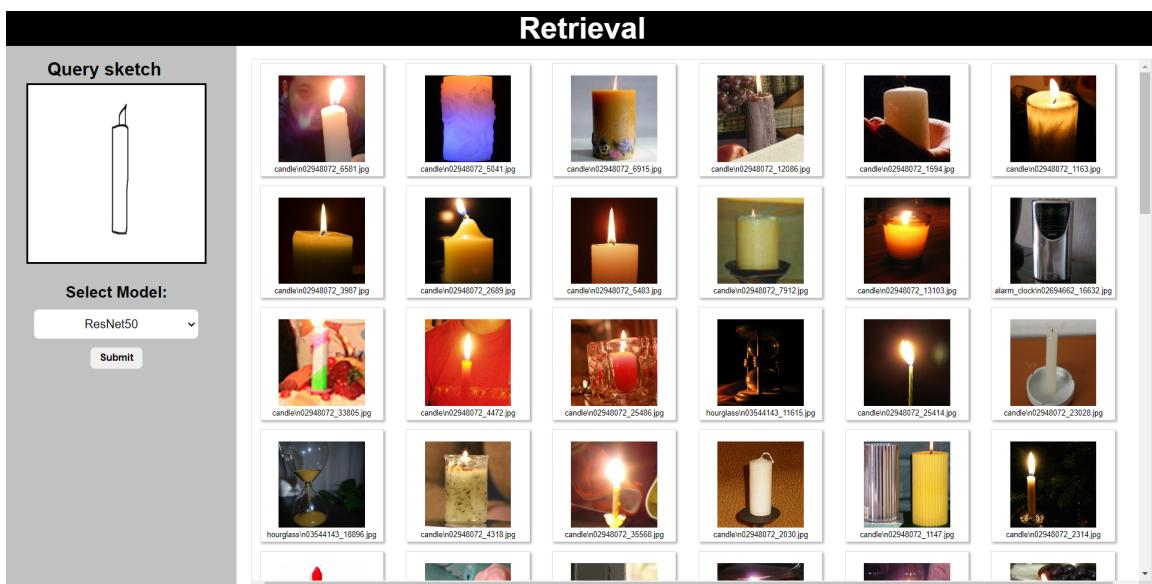
Bên cạnh đó, phía dưới canvas là một bottom bar chứa các công cụ hữu ích hỗ trợ việc vẽ, bao gồm các nút chức năng như:

- bút vẽ để tạo các đường nét
- tẩy để xóa các phần không mong muốn của bản vẽ
- nút undo để quay lại bước vẽ trước đó
- nút upload cho phép người dùng tải lên một tấm ảnh có sẵn từ máy tính của họ để tìm kiếm thay vì phải vẽ phác thảo

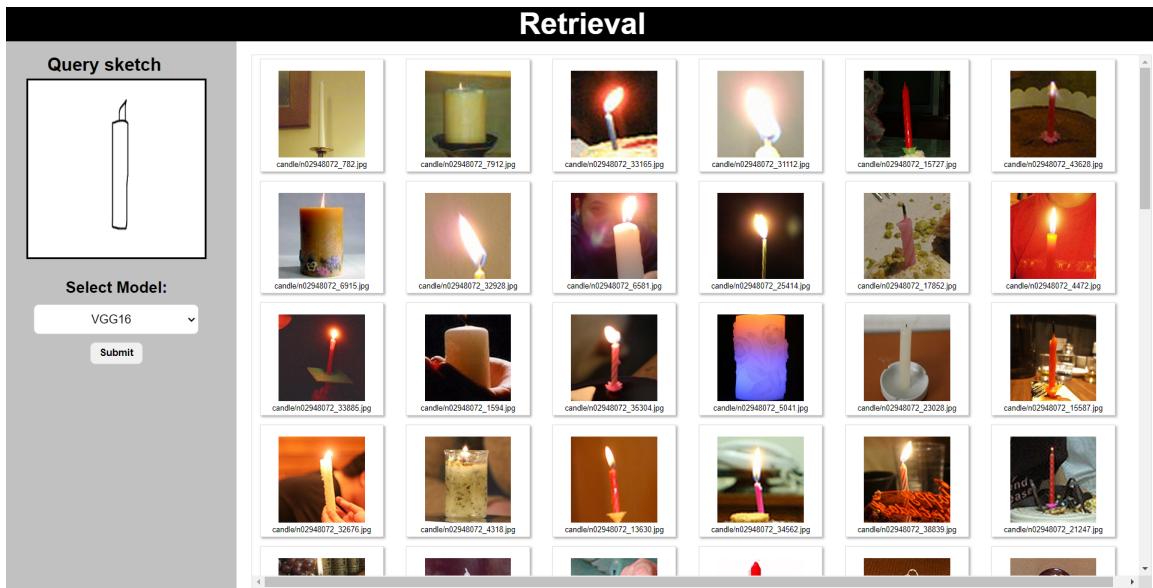
=> Điều này mang lại sự linh hoạt cho người dùng, cho phép họ sử dụng cả bản vẽ tay và hình ảnh có sẵn để truy vấn.

Sau khi hoàn tất quá trình vẽ hoặc tải lên ảnh, người dùng có thể nhấn nút lưu. Khi nút này được nhấn, hệ thống sẽ chuyển hướng người dùng sang một trang mới, nơi các kết quả tìm kiếm tương ứng với ảnh truy vấn sẽ được hiển thị.

Trên trang kết quả này, người dùng có tùy chọn để chọn giữa hai mô hình trích xuất đặc trưng ảnh truy vấn là ResNet50 hoặc VGG16. Mỗi mô hình có các đặc điểm và ưu điểm riêng, và việc cho phép người dùng chọn mô hình phù hợp giúp tăng tính linh hoạt và hiệu quả của hệ thống.



Hình 18: Giao diện kết quả truy vấn SBIR với phương pháp Resnet50.



Hình 19: Giao diện kết quả truy vấn SBIR với phương pháp VGG16.

### 3 Kết luận

#### 3.1 Ưu điểm:

Bài toán truy xuất ảnh dựa trên nét vẽ mà nhóm đã thực hiện có những ưu điểm nổi bật sau:

- Hệ thống SBIR hiệu quả: Xây dựng được hệ thống SBIR giúp người dùng tìm kiếm ảnh dựa trên hình dáng và cấu trúc của đối tượng, không phụ thuộc vào màu sắc hoặc các đặc điểm khác. Điều này đặc biệt hữu ích trong các trường hợp mà màu sắc hoặc chi tiết bề mặt không quan trọng.
- Khơi nguồn sáng tạo: Việc tìm kiếm dựa trên nét vẽ đơn giản giúp người dùng sáng tạo hơn khi tìm kiếm các ý tưởng thiết kế hoặc các đối tượng tương tự.
- Hiệu suất ổn định và truy xuất nhanh: Hiệu suất của hệ thống đạt được tương đối ổn định trên các lớp, và thời gian truy xuất ảnh của mô hình ResNet tương đối nhanh, đảm bảo trải nghiệm người dùng mượt mà và hiệu quả.

#### 3.2 Hạn chế:

Tuy nhiên, trong đồ án này vẫn còn một số hạn chế:

- Yêu cầu dữ liệu lớn: Để huấn luyện các mô hình SBIR hiệu quả, cần có một lượng lớn dữ liệu phác thảo và hình ảnh tương ứng, dẫn đến tiêu tốn dung lượng khá lớn.
- Hiệu suất và độ chính xác: Hiệu suất và độ chính xác của các hệ thống SBIR hiện tại vẫn chưa đạt được mức độ cao như các phương pháp tìm kiếm dựa trên văn bản hoặc hình ảnh đầy đủ chi tiết. Đặc biệt, trong trường hợp ResNet50, do tài nguyên không đủ nên chưa thể tiếp tục huấn luyện cho đến khi mô hình hội tụ.
- Thời gian truy xuất: Thời gian truy xuất của VGG16 khá lâu so với ResNet50.
- Đa dạng phong cách vẽ: Phong cách vẽ của mỗi người có thể khác nhau rất nhiều, tạo ra thách thức lớn trong việc phát triển các mô hình SBIR có thể nhận dạng chính xác các phác thảo từ nhiều người dùng khác nhau.

**Kết Luận:** Báo cáo này đã trình bày một khám phá toàn diện về bài toán truy xuất hình ảnh dựa trên nét vẽ (Sketch-Based Image Retrieval - SBIR) sử dụng bộ dữ liệu Sketchy. Quá trình đánh giá bao gồm việc áp dụng hai phương pháp học sâu tiên tiến là ResNet50 và VGG16 để trích xuất đặc trưng từ hình ảnh và nét vẽ. Các mô hình này được huấn luyện sử dụng phương pháp triplet margin để tối ưu hóa việc học các đặc trưng.

Sau khi các đặc trưng được trích xuất, chúng được sử dụng để tìm các hình ảnh tương đồng bằng thuật toán nearest neighbor. Phân tích và so sánh hai mô hình đã làm sáng tỏ hiệu suất của chúng trong bối cảnh truy xuất hình ảnh dựa trên nét vẽ. Đặc biệt, VGG16

cho thấy khả năng truy xuất hình ảnh phù hợp với ảnh phác thảo của người dùng tốt hơn so với ResNet50 trong trường hợp cụ thể này.

Hệ thống đã được triển khai vào một giao diện web thân thiện với người dùng, cho phép người dùng thử nghiệm truy xuất hình ảnh bằng nét vẽ một cách trực quan và tương tác. Điều này giúp người dùng dễ dàng tìm kiếm và truy xuất các hình ảnh liên quan từ bộ dữ liệu Sketchy.

Trong tương lai, nhóm sẽ phát triển nhiều phương pháp hơn để tăng độ chính xác và tốc độ xử lý. Điều này sẽ được thực hiện bằng cách nghiên cứu và sử dụng các mạng học sâu hiện đại hơn để trích xuất các đặc trưng cao cấp, kết hợp với các biện pháp phức tạp hơn để so sánh các vector đặc trưng, và sử dụng thêm các kỹ thuật tìm kiếm tiên tiến trên các cơ sở dữ liệu hình ảnh lớn.

## 4 Phân công công việc

	<b>Code</b>	<b>Slide</b>	<b>Report</b>	<b>Hoàn thành</b>
Hồ Thị Khánh Hiền	VGG16	II. Phương pháp III. Dataset IV. Thực nghiệm	2.2 Phương pháp sử dụng 2.3 Dataset 2.4 Thực nghiệm 2.6 Demo	100 %
Nguyễn Như Hà	ResNet50	I. Tổng quan II. Phương pháp V. Đánh giá	1. Mở đầu 2.1 Phát biểu bài toán 2.2 Phương pháp sử dụng 2.5 Đánh giá 2.6 Demo	100 %

Bảng 1: Bảng phân công công việc.

## Tài liệu

- [1] Resnet-50: The basics and a quick tutorial. <https://datagen.tech/guides/computer-vision/resnet-50/>.
- [2] Srinivas Rahul Sapireddy. Resnet-50: Introduction. <https://srsapireddy.medium.com/resnet-50-introduction-b5435fdb66f>, 2023.
- [3] pawangfg. Vgg-16 | cnn model. <https://www.geeksforgeeks.org/vgg-16-cnn-model/>.
- [4] Deval Shah. Triplet loss: Intro, implementation, use cases. <https://www.v7labs.com/blog/triplet-loss>, 2016.
- [5] Scikit learn Organization. Nearest neighbors (nn). <https://scikit-learn.org/stable/modules/neighbors.html>, 2024.
- [6] Cusuh Ham James Hays Patsorn Sangkloy, Nathan Burnell. The sketchy database: Learning to retrieve badly drawn bunnies. <https://sketchy.eye.gatech.edu/>.
- [7] Fan Yang, Nor Azman Ismail, Yee Yong Pang, Victor R. Kebande, Arafat Al-Dhaqm, and Tieng Wei Koh. A systematic literature review of deep learning approaches for sketch-based image retrieval: Datasets, metrics, and future directions. *IEEE Access*, 2024.
- [8] Wikipedia. Evaluation measures (information retrieval). [https://en.wikipedia.org/wiki/Evaluation\\_measures\\_\(information\\_retrieval\)](https://en.wikipedia.org/wiki/Evaluation_measures_(information_retrieval)), 2023.