

**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**  
**TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN**  
**KHOA CÔNG NGHỆ THÔNG TIN**



**TRUY VĂN THÔNG TIN THỊ GIÁC**

---

**TRUY VĂN ẢNH THỜI TRANG**

---

**Giảng viên hướng dẫn**

Thầy Võ Hoài Việt

Thầy Nguyễn Trọng Việt

Thầy Phạm Minh Hoàng

**Sinh viên thực hiện**

Võ Nguyễn Hoàng Kim - 21127090

Trần Thanh Ngân - 21127115

Lâm Thanh Ngọc - 21127118

# NỘI DUNG

<b>A. THÔNG TIN SINH VIÊN</b>	3
<b>B. NỘI DUNG BÁO CÁO</b>	3
1. Giới thiệu.....	3
2. Nghiên cứu liên quan.....	4
a. Which is Plagiarism: Fashion Image Retrieval based on Regional Representation for Design Protection [5].....	4
b. Fashion Image Retrieval with Capsule Networks [3].....	5
c. FashionVLP: Vision Language Transformer for Fashion Retrieval with Feedback [2].....	5
d. Fashion Retrieval via Graph Reasoning Networks on a Similarity Pyramid [4].....	6
e. Leveraging Weakly Annotated Data for Fashion Image Retrieval and Label Prediction [1].....	7
f. Study on Fashion Image Retrieval Methods for Efficient Fashion Visual Search [6].....	8
3. Tập dữ liệu.....	9
a. DeepFashion [7].....	9
b. FashionIQ [8].....	9
4. Phát biểu bài toán.....	11
a. Phát biểu bài toán.....	11
b. Giới hạn bài toán.....	11
5. Phương pháp.....	11
a. Chuẩn bị cơ sở dữ liệu.....	12
b. Truy vấn.....	13
6. Thực nghiệm.....	14
a. Chuẩn bị dữ liệu.....	15
b. Huấn luyện mô hình.....	15
c. Xây dựng cơ sở dữ liệu.....	17
d. Truy vấn.....	18
e. Kết quả.....	18
7. Tái thực nghiệm.....	19
a. Thực nghiệm trên tập dữ liệu DeepFashion.....	19
b. Ứng dụng các tính năng từ mô hình OpenAI CLIP.....	22

<b>8.</b>	<b>Cải tiến</b>	25
a.	Vân đề hiện tại	25
b.	Giải pháp cải tiến	25
c.	Kết quả	26
<b>9.</b>	<b>Deploy sản phẩm</b>	27
<b>10.</b>	<b>Kết luận</b>	28
<b>C.</b>	<b>TÀI LIỆU THAM KHẢO</b>	29

## A. THÔNG TIN SINH VIÊN

Họ và tên	Mã số sinh viên	Mức độ hoàn thành
Võ Nguyễn Hoàng Kim	21127090	100%
Trần Thanh Ngân	21127115	100%
Lâm Thanh Ngọc	21127118	100%

## B. NỘI DUNG BÁO CÁO

### 1. Giới thiệu

- Truy vấn trang phục thời trang (Fashion Image Retrieval) là một tác vụ phổ biến trong lĩnh vực thị giác máy tính, nhằm tìm kiếm và nhận diện các hình ảnh thời trang tương tự nhau trong một tập hợp lớn các hình ảnh (bộ dữ liệu). Mục tiêu chính của tác vụ này là xây dựng một hệ thống có khả năng xác định và truy vấn các sản phẩm thời trang dựa trên các đặc điểm hình ảnh, như kiểu dáng, màu sắc,... Thông qua hệ thống, người dùng có thể tìm thấy các sản phẩm có đặc tính tương tự một cách nhanh chóng và chính xác.
- Việc thực hiện truy vấn trang phục đang dành được sự quan tâm rất lớn từ cộng đồng, đặc biệt trong bối cảnh phát triển mạnh mẽ của các sàn thương mại điện tử, cùng với xu hướng mua sắm trực tuyến đang ngày một gia tăng. Nhờ vào việc truy vấn trang phục thời trang, các doanh nghiệp có thể tiết kiệm chi phí nhưng vẫn có thể quảng bá sản phẩm của mình đến người dùng, góp phần nâng cao lợi nhuận.

Đồng thời, đối với người tiêu dùng, nhờ vào hệ thống truy vấn này, họ dễ dàng tiếp cận các sản phẩm tương tự mà không phải tốn quá nhiều thời gian cho việc tìm kiếm từ đầu.

## 2. Nghiên cứu liên quan

- Việc xây dựng hệ thống truy vấn trang phục đã được khai thác, nghiên cứu và trình bày rất nhiều trong cộng đồng thị giác máy tính. Dưới đây là những công trình nghiên cứu liên quan mà nhóm đã tìm hiểu:
  - a. Which is Plagiarism: Fashion Image Retrieval based on Regional Representation for Design Protection [5]
    - Với sự phát triển nhanh chóng của thương mại điện tử và mua sắm trực tuyến, truy vấn trang phục đã nhận được nhiều sự quan tâm. Khác với các nghiên cứu hiện có tập trung vào truy xuất mặt hàng thời trang giống hệt hoặc tương tự, bài báo này nghiên cứu việc truy vấn trang phục đạo nhái, một lĩnh vực bị phớt lờ trong cộng đồng học thuật. Trang phục đạo nhái thường được chỉnh sửa ở một vùng nhất định để tránh giám sát bởi các phương pháp truy xuất truyền thống. Bài báo đề xuất mô hình Plagiarized-Search-Net (PS-Net) dựa trên đại diện khu vực, sử dụng các mốc để hướng dẫn việc học và so sánh các mặt hàng thời trang theo khu vực. Đồng thời, các tác giả cũng giới thiệu bộ dữ liệu Plagiarized Fashion (Thời trang đạo nhái) để truy xuất quần áo đạo nhái, cung cấp bổ sung quan trọng cho lĩnh vực truy vấn thời trang. Các thử nghiệm trên bộ dữ liệu này cho thấy phương pháp của họ vượt trội hơn các công trình khác trong truy vấn y phục đạo nhái, bảo vệ thiết kế ban đầu. PS-Net cũng có thể điều chỉnh cho các nhiệm vụ truy vấn thời trang truyền thống và ước tính điểm mốc, đạt hiệu suất cao trên tập dữ liệu DeepFashion và DeepFashion2.
    - Các phương pháp truy xuất quần áo thường học sự tương đồng trên toàn bộ phiên bản của quần áo mà không tập trung vào các điểm cụ thể, dẫn đến dễ bị ảnh hưởng bởi các đặc điểm không liên quan. Mặc dù gần đây, có sự cải thiện hiệu suất khi sử dụng biểu diễn thuộc tính để hướng dẫn truy xuất, việc này không khả thi trong vấn đề truy xuất trang phục đạo nhái. Bài báo đề xuất PS-Net, một mô

hình dựa trên đại diện khu vực để so sánh và thực hiện truy xuất. Thay vì sử dụng các thuộc tính quần áo, PS-Net dựa vào các đặc tính hình học, giúp duy trì sự ổn định đối với các mẫu bị biến dạng và che khuất. Các mốc quần áo được sử dụng để hướng dẫn cách thể hiện và so sánh giữa các khu vực.

- Do thiếu tập dữ liệu phù hợp, các tác giả đã tạo bộ dữ liệu mới tên là Plagiarized Fashion, thách thức trong việc xây dựng tập dữ liệu này là phân biệt quần áo có khác biệt nhỏ về kiểu dáng từ số lượng lớn trang phục giống hệt nhau.

### b. Fashion Image Retrieval with Capsule Networks [3]

- Trong nghiên cứu này, họ điều tra hiệu suất truy vấn quần áo của mạng Capsule với tính năng định tuyến động. Bài báo đề xuất kiến trúc mạng Capsule dựa trên Triplet với hai phương pháp trích xuất đặc trưng khác nhau, sử dụng các khối Stacked-convolution (SC) và Residual-connected (RC) làm đầu vào của các lớp Capsule. Kết quả thực nghiệm cho thấy cả hai thiết kế đều vượt trội hơn các nghiên cứu cơ sở như FashionNet mà không cần thông tin mang tính bước ngoặt. So với kiến trúc SOTA trong truy vấn y phục, mạng Triplet Capsule của chúng tôi đạt tỷ lệ thu hồi tương đương với chỉ một nửa số tham số.
- Việc truy xuất hình ảnh quần áo mong muốn từ một bộ sưu tập là nhiệm vụ khó khăn trong lĩnh vực thời trang, thường được giải quyết bằng cơ chế học cách nắm bắt các khái niệm về sự tương đồng giữa các hình ảnh trong một không gian con chung. Nhiều nghiên cứu sử dụng mô hình mạng nơ-ron tích chập (CNNs), nhưng CNNs có hạn chế về mất thông tin không gian phân cấp và không bền vững trước các phép biến đổi Affine. Bài báo này đề xuất sử dụng mô hình mạng Capsule và thuật toán định tuyến động để cải thiện truy xuất hình ảnh quần áo. Thuật toán định tuyến bằng thỏa thuận giúp học nhiều thông tin mô tả về đối tượng mà không mất mối quan hệ không gian nội tại giữa đối tượng và các phần của nó. Mạng Capsule có khả năng nhận dạng hình ảnh bất kể góc nhìn và không cần các phép biến đổi, nhờ khả năng tự học cấu hình tư thế chiều cao hơn của các hình ảnh.

### c. FashionVLP: Vision Language Transformer for Fashion Retrieval with Feedback [2]

- Truy vấn ảnh thời trang (FIR) dựa trên cặp truy vấn gồm hình ảnh tham chiếu và phản hồi ngôn ngữ tự nhiên là một nhiệm vụ đầy thách thức, yêu cầu mô hình phải đánh giá đồng thời thông tin từ cả thị giác và văn bản. Bài báo đề xuất mô hình FashionVLP, dựa trên biến đổi thị giác-ngôn ngữ, đưa kiến thức từ các tập lớn hình ảnh-văn bản vào lĩnh vực FIR, và kết hợp thông tin thị giác từ nhiều cấp độ ngữ cảnh để nắm bắt thông tin thời trang hiệu quả. Truy vấn được mã hóa qua các lớp biến đổi (transformer), và thiết kế bất đối xứng của mô hình áp dụng cách tiếp cận mới dựa trên sự chú ý để hợp nhất các đặc trưng hình ảnh mục tiêu mà không liên quan đến văn bản hoặc các lớp biến đổi trong quá trình này. Kết quả cho thấy FashionVLP đạt hiệu suất vượt trội trên các tập dữ liệu chuẩn, với mức cải thiện 23% trên tập dữ liệu FashionIQ đầy thách thức, chừa phản hồi ngôn ngữ tự nhiên phức tạp.
- Truy vấn ảnh thời trang dựa trên cặp truy vấn gồm hình ảnh tham chiếu và phản hồi ngôn ngữ tự nhiên là nhiệm vụ khó khăn, do cần kết hợp thông tin từ cả hình ảnh và văn bản. Điều này đòi hỏi mô hình phải hiểu sâu sắc cả hai phương thức để nắm bắt ngữ cảnh và yêu cầu của người dùng. Phản hồi ngôn ngữ tự nhiên thường rất đa dạng và phức tạp, tăng thêm độ khó của nhiệm vụ. Thông tin thị giác trong thời trang không chỉ ở cấp độ bề mặt mà còn phải xem xét nhiều cấp độ ngữ cảnh khác nhau.
- Mô hình FashionVLP giải quyết những thách thức này bằng cách sử dụng bộ biến đổi thị giác-ngôn ngữ, giúp kết hợp hiệu quả thông tin từ hình ảnh và văn bản. Thiết kế bất đối xứng với phương pháp dựa trên sự chú ý mới cho phép kết hợp các đặc trưng của hình ảnh mục tiêu mà không cần liên quan đến văn bản hoặc các lớp biến đổi, nâng cao hiệu suất và độ chính xác của mô hình.

### d. Fashion Retrieval via Graph Reasoning Networks on a Similarity Pyramid [4]

- Việc kết hợp hình ảnh quần áo từ khách hàng và các cửa hàng mua sắm trực tuyến có ứng dụng rộng rãi trong thương mại điện tử. Các thuật toán hiện tại mã hóa

hình ảnh thành vectơ đặc trưng toàn cục và thực hiện truy vấn dựa trên biểu diễn này, nhưng thông tin cục bộ quan trọng về quần áo bị mất đi, dẫn đến hiệu suất không tối ưu. Để giải quyết vấn đề này, bài báo đề xuất Mạng Lý Luận Đồ thị (GRNet) trên một Kim tự tháp tương đồng. GRNet tìm hiểu sự tương đồng giữa truy vấn và bộ sưu tập quần áo bằng cách sử dụng cả biểu diễn toàn cục và cục bộ ở nhiều tỷ lệ. Kim tự tháp tương đồng được biểu diễn bằng một Đồ thị tương tự, trong đó các nút đại diện cho sự tương đồng giữa các thành phần quần áo ở các tỷ lệ khác nhau, và điểm phù hợp cuối cùng được xác định bằng cách truyền thông điệp dọc theo các cạnh. GRNet sử dụng mạng tích chập đồ thị để căn chỉnh các thành phần quần áo nổi bật hơn, cải thiện việc truy vấn quần áo. Để hỗ trợ nghiên cứu tương lai, bài báo giới thiệu bộ kiểm tra mới FindFashion, chứa các chủ thích về hộp giới hạn, góc nhìn, che khuất và cắt xén. Các thử nghiệm cho thấy GRNet đạt kết quả vượt trội trên hai bộ kiểm tra, cải thiện độ chính xác top-1, top-20 và top-50 trên DeepFashion lần lượt là 26%, 64% và 75%, và đạt cải tiến đáng kể trên FindFashion.

- Các phương pháp hiện có thường mã hóa hình ảnh thành biểu diễn toàn cục và thực hiện truy vấn dựa trên biểu diễn này, nhưng thông tin chi tiết về các phần của quần áo thường bị mất, dẫn đến hiệu suất truy vấn không tối ưu. Để giải quyết vấn đề này, bài báo đề xuất Mạng Lý Luận Đồ thị (GRNet) dựa trên Kim tự tháp tương đồng. GRNet học sự tương đồng giữa các truy vấn và hình ảnh quần áo trong bộ sưu tập bằng cách sử dụng cả biểu diễn toàn cục và cục bộ ở nhiều tỷ lệ khác nhau, giúp tái tạo thông tin chi tiết về các phần của quần áo và cải thiện hiệu suất truy vấn hình ảnh quần áo.

**e. Leveraging Weakly Annotated Data for Fashion Image Retrieval and Label Prediction [1]**

- Trong bài báo này, họ giới thiệu một phương pháp huấn luyện trình bày trực quan cho sản phẩm thương mại điện tử. Dựa trên phương pháp học ít giám sát, mô hình của họ học từ các tập dữ liệu nhiều thu thập từ các trang web thương mại điện tử mà không cần đánh nhãn thủ công. Chúng tôi cho thấy phương pháp này có thể được áp dụng cho phân loại quần áo với các mức độ chi tiết khác nhau và đã được

huấn luyện để phù hợp với truy vấn hình ảnh. Kết quả cho thấy rằng công trình đã đạt được độ chính xác cao trong việc truy vấn ảnh quần áo tại cửa hàng DeepFashion và dự đoán thuộc tính danh mục mà không cần sử dụng bộ huấn luyện cung cấp.

- Mục tiêu bài báo là phát triển một công cụ trích xuất đặc trưng hình ảnh cho lĩnh vực thương mại điện tử, thông qua việc phân tích chuyên sâu các ứng dụng nhận dạng hình ảnh thời trang như truy xuất hình ảnh và gắn thẻ thuộc tính, đồng thời phát triển kiến trúc CNN nâng cao và xử lý nhiều ngôn ngữ. Để đạt được mục tiêu này, các tác giả đã huấn luyện quy trình trích xuất đặc trưng hình ảnh trên một tập dữ liệu lớn các hình ảnh có chủ thích từ Internet, với các chủ thích tương ứng là mô tả văn bản liên quan. Mô hình được đào tạo trên dữ liệu với chi phí ghi nhãn bằng 0, sử dụng điểm dữ liệu trích xuất từ các trang web thương mại điện tử.

#### **f. Study on Fashion Image Retrieval Methods for Efficient Fashion Visual Search [6]**

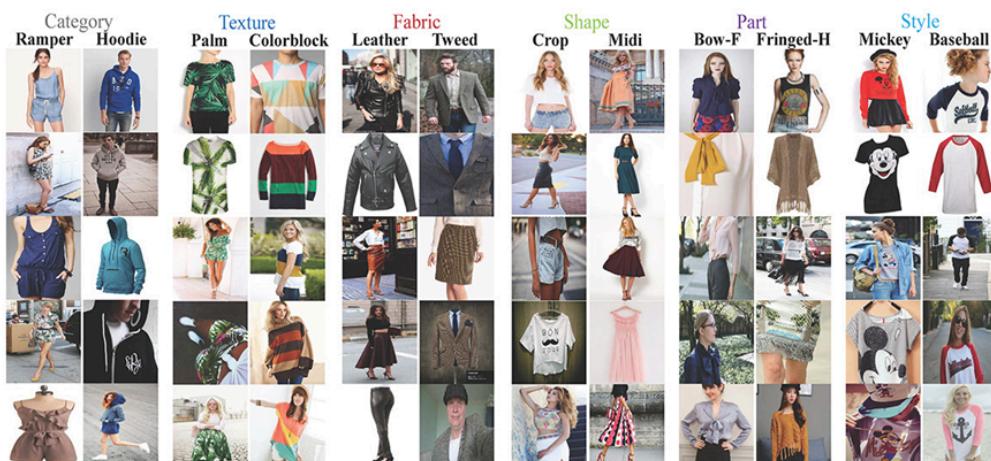
- Truy vấn ảnh thời trang (FIR) là nhiệm vụ khó khăn, yêu cầu tìm kiếm chính xác các mặt hàng từ bộ sưu tập lớn dựa trên hình ảnh truy vấn. Dù có những tiến bộ gần đây, tuy nhiên, FIR vẫn gặp hạn chế trong việc áp dụng vào tìm kiếm trực quan thực tế, do sự cân bằng giữa độ phức tạp của mô hình và hiệu suất, cũng như sự biến đổi của các hình ảnh thời trang được chụp trong các hoàn cảnh khác nhau. Đặc biệt, hình ảnh thời trang dễ bị biến dạng và không nhất quán giữa ảnh truy vấn và ảnh sản phẩm. Bài báo này đề xuất một phương pháp FIR tối ưu hóa cho lĩnh vực thời trang, nghiên cứu các chiến lược huấn luyện và mô hình học sâu để cải thiện hiệu suất truy vấn. Các kết quả thử nghiệm trên ba điểm chuẩn từ bộ dữ liệu DeepFashion cho thấy phương pháp này đạt được cải tiến đáng kể so với các phương pháp FIR trước đó.
- Bài báo giải quyết các vấn đề trong việc truy vấn ảnh thời trang, chẳng hạn như việc các hình ảnh thời trang thường có nhiều món đồ và sự khác biệt lớn về quan điểm và phong cách, cùng với sự biến dạng và che khuất tùy thuộc vào môi trường chụp ảnh. Mục tiêu của bài báo là nghiên cứu một cách tiếp cận hiệu quả để huấn luyện mô hình FIR, bằng cách xem xét chiến lược huấn luyện và hàm mất

mát, kiểm tra các cải tiến cấu trúc để đạt được một phương pháp FIR hiệu quả. Các thí nghiệm cho thấy việc lựa chọn chiến lược huấn luyện và hàm mất mát phù hợp có thể cải thiện đáng kể độ chính xác.

### 3. Tập dữ liệu

#### a. DeepFashion [7]

- DeepFashion là một tập dữ liệu lớn được sử dụng phổ biến trong các nghiên cứu về nhận diện và tìm kiếm thời trang. Tập dữ liệu này được giới thiệu với hơn 800.000 ảnh thời trang đa dạng, từ những hình ảnh chụp trong cửa hàng đến những bức ảnh chụp bởi người tiêu dùng, tạo nên một cơ sở dữ liệu phân tích thời trang lớn nhất. Hơn nữa, DeepFashion được chú thích với thông tin phong phú về các mặt hàng quần áo. Mỗi hình ảnh trong bộ dữ liệu này được gán nhãn với 50 loại, 1.000 thuộc tính mô tả, hộp giới hạn và điểm mốc của quần áo. Bên cạnh đó, DeepFashion còn chứa hơn 300.000 cặp hình ảnh có sự thay đổi về tư thế/ miền.
- Bốn tiêu chuẩn được phát triển bằng cách sử dụng cơ sở dữ liệu DeepFashion, bao gồm Dự đoán Thuộc tính (Attribute Prediction), Truy xuất Quần áo Từ Người Tiêu Dùng đến Cửa Hàng (Consumer-to-shop Clothes Retrieval), Truy xuất Quần Áo Trong Cửa Hàng (In-shop Clothes Retrieval), và Phát Hiện Điểm Mốc (Landmark Detection). Dữ liệu và chú thích của các tiêu chuẩn này cũng có thể được sử dụng làm bộ đào tạo và kiểm tra cho các tác vụ thị giác máy tính sau, như Phát hiện quần áo, Nhận dạng quần áo, và Truy xuất hình ảnh.

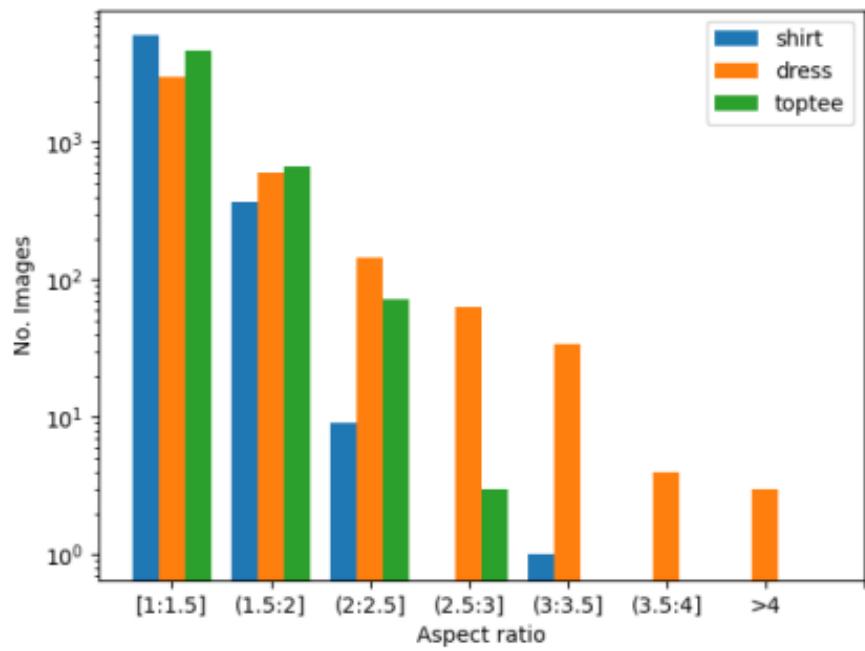


## b. FashionIQ [8]

- FashionIQ là một bộ dữ liệu truy vấn thời trang có chứa các chủ thích ngôn ngữ tự nhiên có tính tương tác. Các sản phẩm thời trang trong bộ dữ liệu sẽ thuộc về ba loại chính là Đầm, Top&Tees và Áo sơ mi.



- Bộ dữ liệu này có tổng cộng 77.000 hình ảnh, trong đó, 46.000 hình ảnh được sử dụng cho việc huấn luyện và có sẵn 18.000 cặp hình ảnh. Mỗi cặp có hai chủ thích được thu thập từ cộng đồng, mô tả các thay đổi từ hình ảnh tham chiếu đến mục tiêu. Phản hồi là phức tạp và đôi khi một câu bao gồm nhiều khái niệm cần thay đổi, ví dụ: "có hoa văn và đường viền cổ áo dây buộc", "màu đen với họa tiết hoa",.... Tính phức tạp nhưng thực tế của phản hồi trong bộ dữ liệu này làm cho nó trở thành một thách thức đặc biệt cho nhiệm vụ truy vấn.



**(b) FashionIQ categories**

*Histogram tý lệ khung hình ảnh và ba loại trang phục trong tập dữ liệu FashionIQ. Trục x biểu thị tý lệ khung hình được xác định là  $\max(\text{width}, \text{height}) / \min(\text{width}/\text{height})$  trong khi trục y biểu thị số lượng hình ảnh (theo thang logarit). Chiều rộng của mỗi bin là 0,5 và bin đầu tiên bắt đầu từ 1. Hơn một nửa số hình ảnh trong tập dữ liệu bị lệch và có tý lệ khung hình ít nhất là 1,5. Trong tập dữ liệu FashionIQ, vấn đề này thể hiện rõ trong phân loại Váy.*

## 4. Phát biểu bài toán

### a. Phát biểu bài toán

- Đầu vào (Input): Hệ thống nhận vào là một ảnh về thời trang mà người dùng muốn truy vấn.
- Đầu ra (Output): Hệ thống sẽ trả ra các ảnh kết quả ảnh (trong tập cơ sở dữ liệu) có độ tương tự cao nhất với ảnh được truy vấn.

### b. Giới hạn bài toán

- Để hệ thống truy vấn đạt được độ chính xác cao, một số giới hạn được đặt ra để giảm bớt độ phức tạp:

- + Loại thời trang trong tập cơ sở dữ liệu chỉ có 3 loại: đầm (dress), áo sơ-mi (shirt) và áo thun (top-tee). Do đó, hệ thống chỉ có thể truy vấn với 3 loại thời trang đó.
- + Với ảnh truy vấn, hệ thống chỉ có thể thực hiện truy vấn tốt khi nó được đưa vào là một loại thời trang cụ thể (nếu truy vấn áo sơ-mi thì ảnh chỉ nên tập trung vào chỗ có áo sơ mi). Nếu cung cấp ảnh đầu vào có quá nhiều yếu tố ngoại cảnh (như người, vật,...) điều này sẽ làm giảm độ chính xác của hệ thống khi thực hiện rút trích đặc trưng và so khớp. Điều này có thể gây ảnh hưởng tiêu cực đến kết quả cuối cùng.

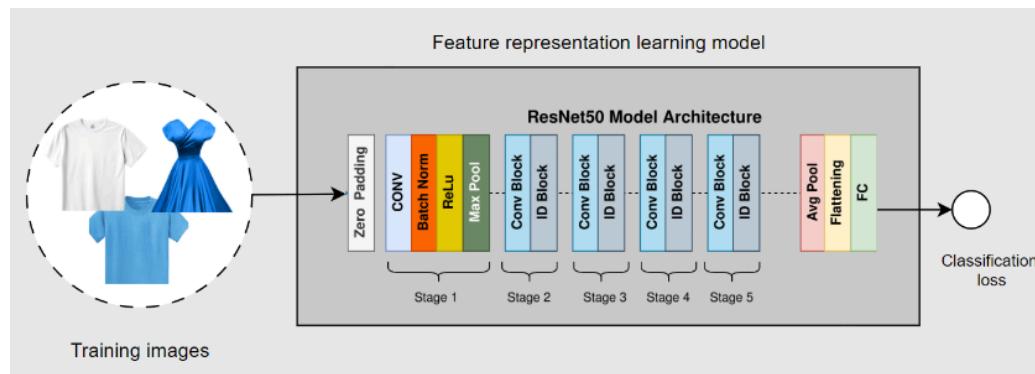
## 5. Phương pháp

- Việc xây dựng một hệ thống truy vấn thường được chia thành hai bước chính: thứ nhất là chuẩn bị cơ sở dữ liệu và thứ hai là thực hiện truy vấn. Điều này có thể phát biểu tổng quan như sau:
  - + Hệ thống sẽ sử dụng ảnh trong tập dữ liệu FashionIQ đi qua một mô hình mạng nơ-ron để thực hiện trích xuất đặc trưng của ảnh. Các đặc trưng này sau khi trích xuất sẽ được lưu lại làm cơ sở dữ liệu cho hệ thống truy vấn.
  - + Khi ảnh truy vấn được đưa vào, sử dụng mô hình tương tự ở bước chuẩn bị cơ sở dữ liệu để tiến hành rút trích đặc trưng của ảnh truy vấn, đặc trưng này sẽ được so khớp với các đặc trưng trong cơ sở dữ liệu để có thể tìm ra những ảnh có độ tương đồng cao và cuối cùng trả ra kết quả truy vấn.

### a. Chuẩn bị cơ sở dữ liệu

- Việc lựa chọn mô hình để tiến hành rút trích đặc trưng cho ảnh là một vấn đề quan trọng, cần được lựa chọn kỹ lưỡng. Các bộ khung (backbone) của mô hình CNN đã được cân nhắc cho tác vụ này như ResNet18, ResNet50, DenseNet121. Tuy nhiên, do một số vấn đề như giới hạn tài nguyên, chi phí thời gian, độ phức tạp và giới hạn kiến thức của các thành viên trong nhóm, nên việc lựa chọn được thực hiện dựa trên các tiêu chí như mức độ phổ biến, dễ dàng tiếp cận, tiết kiệm tài nguyên và chi phí nhưng vẫn đem lại hiệu quả ổn định. Do đó, nhóm đã lựa chọn ResNet50 để phục vụ cho tác vụ này.

- Mô hình cơ bản của ResNet50 được sử dụng lại, tuy nhiên được chỉnh sửa và thay đổi lớp liên kết đầy đủ (Fully Connected) cuối cùng để phù hợp với số lượng lớp của tập dữ liệu trong bài toán. Trong mô hình này, đường dẫn mất mát phân loại (Classification Loss) được sử dụng để đào tạo mô hình, khuyến khích các đặc trưng học được có khả năng phân biệt giữa các loại thời trang với nhau. Hàm mất mát Cross-Entropy được sử dụng cho đường dẫn mất mát phân loại.
- Hình dưới đây mô tả tổng quan về cấu trúc mô hình được sử dụng để huấn luyện trích xuất đặc trưng của tập dữ liệu ảnh thời trang.



- Sau huấn luyện mô hình hoàn tất và đạt được một độ chính xác ổn định, nhóm sẽ sử dụng mô hình đó để tiến hành rút trích đặc trưng cho toàn bộ ảnh trong tập dữ liệu.

### b. Truy vấn

- Sau khi hoàn tất việc chuẩn bị cơ sở dữ liệu cho hệ thống truy vấn, việc thực hiện truy vấn sẽ là bước cuối cùng cho tác vụ này. Người dùng sẽ đưa ảnh cần truy vấn vào, hệ thống sẽ sử dụng mô hình ResNet50 đã được huấn luyện trước đó để tiến hành trích xuất đặc trưng cho ảnh truy vấn. Đặc trưng của ảnh truy vấn sau đó được đếm đi so khớp với các đặc trưng có trong cơ sở dữ liệu và tìm ra những đặc trưng có độ tương đồng cao nhất với ảnh truy vấn và trả ra kết quả.
- Việc so khớp được dựa vào các hàm tính toán độ tương đồng / dị biệt như Cosine hay Euclidean để tính toán kết quả giữa đặc trưng của ảnh truy vấn và đặc trưng trong cơ sở dữ liệu.
  - + Cosine là một độ đo tương đồng, đo lường góc giữa hai vector thay vì khoảng cách thực sự giữa chúng. Nó tập trung vào hướng của các vector

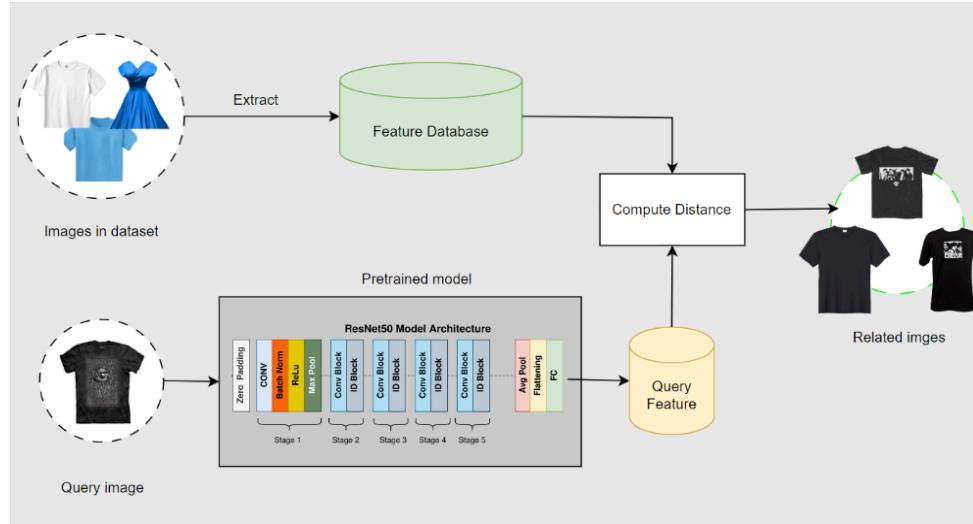
hơn là độ lớn của chúng. Kết quả mà Cosine cho ra nằm trong khoảng từ -1 đến 1, với 1 biểu thị cho sự tương đồng hoàn hảo, 0 là không tương đồng và -1 là dị biệt. Cosine Similarity hiệu quả khi chỉ cần quan tâm đến hướng của các vector, bất kể độ lớn của chúng, đặc biệt khi làm việc với dữ liệu thưa hoặc các vector có độ lớn khác nhau. Nếu sử dụng Cosine, hệ thống sẽ đo độ tương đồng của đặc trưng ảnh truy vấn với toàn bộ cơ sở dữ liệu đặc trưng. Và kết quả trả ra sẽ là những ảnh có độ tương đồng cao nhất so với ảnh truy vấn:

$$\text{Cosine Similarity} = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \cdot \|\mathbf{y}\|}$$

- + Euclidean là độ đo về sự dị biệt, nó đo khoảng cách trực tiếp giữa hai điểm trong không gian nhiều chiều. Khoảng cách Euclid càng nhỏ, các vector càng gần nhau, và ngược lại. Euclidean tập trung vào độ lớn của các vector và khoảng cách giữa chúng. Nếu sử dụng Euclidean, hệ thống sẽ tính toán độ dị biệt giữa đặc trưng ảnh truy vấn với toàn bộ cơ sở dữ liệu đặc trưng. Và kết quả trả ra sẽ là những ảnh có độ dị biệt thấp nhất so với ảnh truy vấn:

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

- Ảnh dưới đây minh họa ngắn gọn mà hệ thống truy vấn sẽ thực hiện khi người dùng cung cấp ảnh đầu vào cho hệ thống truy vấn.



## 6. Thực nghiệm

- Tập dữ liệu Fashion IQ được sử dụng trong bước thực nghiệm với 3 phân loại trang phục là Đầm (Dress), Áo sơ-mi (Shirt) và Áo thun (Top-Tee)
- Thuật toán được triển khai trên nền tảng Kaggle với trình tăng tốc phần cứng sử dụng T4 GPU và ngôn ngữ Python.

### a. Chuẩn bị dữ liệu

- Trong tập FashionIQ, các tác giả đã hỗ trợ chúng ta trong việc phân chia tập train - valid và test đối với từng loại trang phục, điều này được thực hiện dưới các file .json trong thư mục images\_splits. Như vậy, ta có thể nhờ vào đây để phân chia các ảnh thành những tập dữ liệu: như tập huấn luyện (train), kiểm tra (test) và đánh giá (valid) bằng cách gộp các tập tương ứng của cả ba loại thời trang với nhau. Còn về phía nhãn (label) của chúng sẽ được chuyển về dạng số để dễ lưu trữ và truy cập, thao tác nhanh chóng, với 0 đại diện cho “dress”, 1 đại diện cho “shirt” và 2 đại diện cho “top-tee”.
- Hiệu suất của mô hình huấn luyện phụ thuộc rất nhiều vào tập dữ liệu sử dụng, để có thể đạt được kết quả tốt ở bước tiếp theo, dữ liệu ảnh cần được chuẩn bị và tiền xử lý trước khi đưa vào mô hình. Trước hết, khởi tạo các tập dữ liệu huấn luyện (train), kiểm tra (test) và đánh giá (valid) thông qua lớp FashionIQDataset (đây là một lớp được tạo và kế thừa lớp Dataset trong thư viện Pytorch):
  - + Để mô hình hoạt động chính xác, dữ liệu ảnh cần được tiền xử lý nhằm đảm bảo tính nhất quán về kích thước và giá trị pixel. Các ảnh được thay đổi kích thước về chuẩn 224x224 pixel, phù hợp với yêu cầu đầu vào của ResNet-50. Sau đó, ảnh được chuyển đổi sang tensor, một định dạng dữ liệu đặc trưng của PyTorch, giúp tăng cường khả năng xử lý trên GPU.

Cuối cùng, ảnh được chuẩn hóa bằng cách sử dụng giá trị trung bình và độ lệch chuẩn tương ứng với tập dữ liệu ImageNet, giúp cân bằng độ sáng tối và giảm sự chênh lệch giữa các ảnh.

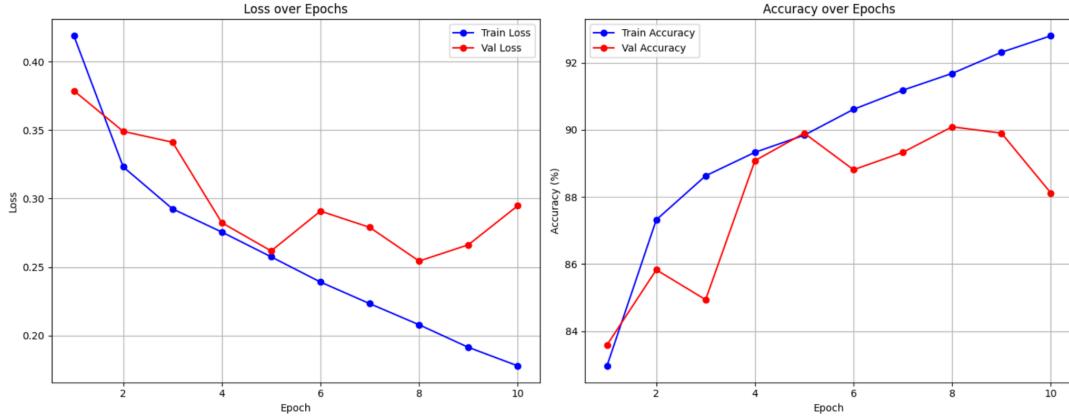
- + Lớp FashionIQDataset được định nghĩa để quản lý việc nạp ảnh và nhãn tương ứng từ các tập dữ liệu. Trong đó, mỗi ảnh sẽ được đọc và chuyển đổi thành định dạng RGB, sau đó áp dụng các bước tiền xử lý (đã được định nghĩa ở trên). Mỗi lần truy xuất dữ liệu, lớp này sẽ trả về một cặp gồm ảnh và nhãn tương ứng, đảm bảo dữ liệu được chuẩn bị sẵn sàng trước khi đưa vào mô hình huấn luyện.
- Đồng thời sử dụng DataLoader để chia nhỏ dữ liệu (với từng tập dữ liệu) thành các batch có kích thước cố định (ở đây, batch\_size được sử dụng là 32). Trong tập huấn luyện (train), dữ liệu được xáo trộn ( thông qua tham số “shuffle”), còn tập đánh giá (valid) và kiểm tra (test) thì không. Việc chia dữ liệu thành batch giúp mô hình xử lý dữ liệu lớn hiệu quả hơn và hỗ trợ quá trình tối ưu hóa thông qua tính toán song song.

## b. Huấn luyện mô hình

- Xây dựng mô hình ResNet50:
  - + Trong phần thực nghiệm này, ta sẽ sử dụng mô hình ResNet50 đã được huấn luyện sẵn trên tập dữ liệu ImageNet, điều này giúp mô hình có được các đặc trưng cơ bản từ tập dữ liệu lớn, giảm thiểu thời gian huấn luyện và cải thiện hiệu suất. Thay đổi lớp cuối cùng của mô hình để phù hợp với số lượng lớp trong bài toán hiện tại. Vì ResNet-50 gốc có lớp phân loại cuối cùng là một lớp tuyến tính (Linear) với số lượng đầu vào bằng số đặc trưng đầu ra của các lớp trước đó, nên vì thế, khi thực hiện xây dựng mô hình này, nó cần được thay thế bằng một lớp tuyến tính (Linear) mới với số lượng đầu ra tương ứng với số lớp phân loại trong bài toán (ở đây là 3).
  - + Việc thực hiện huấn luyện được thực hiện và tận dụng trên GPU để tăng tốc quá trình huấn luyện. Vì vậy, sau khi khởi tạo mô hình xong, ta sẽ di chuyển nó lên thiết bị tính toán là GPU.
- Hàm mất mát và thuật toán tối ưu:
  - + Hàm mất mát (Criterion) là một hàm dùng để đo lường mức độ sai lệch giữa dự đoán của mô hình và giá trị thực tế. Nó giúp đánh giá hiệu suất của mô hình trong quá trình huấn luyện. Mục tiêu của huấn luyện là tối thiểu hóa giá trị của hàm mất mát. Trong phần này, nhóm chỉ định CrossEntropyLoss là hàm mất mát vì nó phù hợp cho các bài toán phân loại nhiều lớp. Hàm mất mát này đo lường sự khác biệt giữa phân phối dự đoán của mô hình và nhãn thực tế.
  - + Thuật toán tối ưu (Optimizer) là thuật toán dùng để điều chỉnh trọng số của mô hình dựa trên gradient của hàm mất mát. Nó cập nhật các tham số của

mô hình (như trọng số và độ lệch) nhằm giảm giá trị của hàm mất mát qua các bước huấn luyện. Optimizer thực hiện các bước cập nhật trọng số để cải thiện hiệu suất của mô hình qua từng lần lặp huấn luyện. Ở đây, Adam là thuật toán được sử dụng, một trong những thuật toán tối ưu hóa phổ biến và hiệu quả cho huấn luyện mô hình. Hệ số học (learning rate) được đặt là 0.001, quyết định kích thước bước cập nhật trọng số trong quá trình huấn luyện.

- Huấn luyện mô hình
  - + Quá trình huấn luyện trong phần thực nghiệm này được thực hiện qua 10 epoch, trong đó mỗi epoch đại diện cho một lượt huấn luyện toàn bộ tập dữ liệu. Trong mỗi epoch, mô hình được đặt vào chế độ huấn luyện và mất mát tổng hợp của epoch được theo dõi.
  - + Dữ liệu từ train\_loader được xử lý theo từng batch, với các gradient của bộ tối ưu hóa được xóa trước khi dự đoán được thực hiện. Mất mát giữa dự đoán của mô hình và nhãn thực tế được tính toán và gradient được tính toán thông qua lan truyền ngược. Các trọng số của mô hình sau đó được cập nhật dựa trên các gradient này, và mất mát của từng batch được cộng dồn để tính toán mất mát trung bình của toàn bộ epoch.
  - + Sau khi hoàn tất một epoch huấn luyện, mô hình được đánh giá trên tập validation để kiểm tra độ chính xác. Hàm evaluate được sử dụng để tính toán độ chính xác của mô hình trên tập validation, giúp theo dõi sự cải thiện của mô hình qua các epoch. Nếu độ chính xác hiện tại của mô hình trên tập validation tốt hơn các mức trước đó, trọng số của mô hình sẽ được lưu lại, ghi nhận mô hình tốt nhất đạt được cho đến thời điểm đó.
  - + Hàm evaluate được sử dụng để thực hiện đánh giá mô hình, giúp tính toán độ chính xác của mô hình trên tập dữ liệu đánh giá mà không cần tính toán gradient. Mô hình được đặt vào chế độ đánh giá để đảm bảo rằng các tính năng huấn luyện như dropout không ảnh hưởng đến kết quả dự đoán. Độ chính xác được tính toán bằng cách so sánh dự đoán của mô hình với nhãn thực tế, cung cấp cái nhìn về hiệu suất của mô hình trên dữ liệu chưa thấy trước.
- Sau bước huấn luyện mô hình, chương trình đã có được trọng số của mô hình, là phiên bản tốt nhất trong quá trình huấn luyện, được lưu dưới tên **best\_model.pth**. Kết quả của quá trình huấn luyện được ghi lại và biểu diễn như biểu đồ sau:



- Để đảm bảo rằng mô hình đã thực sự tốt, ta sẽ sử dụng bộ dữ liệu kiểm tra (test) để kiểm thử độ chính xác của nó. Kết quả độ chính xác mà mô hình đạt được là 0.91. Điều này cho thấy rằng mô hình đã ở mức ổn định và có thể đem đi sử dụng cho các tác vụ sau.

#### c. Xây dựng cơ sở dữ liệu

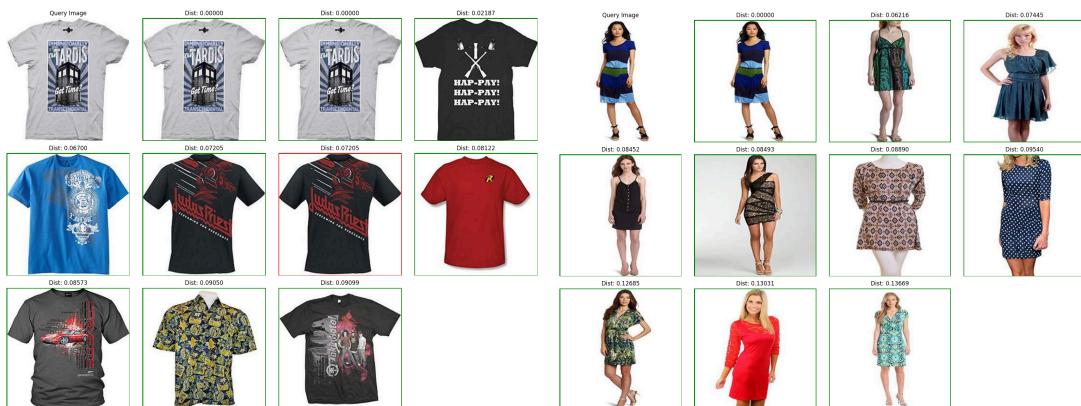
- Tất cả các ảnh có trong tập dữ liệu FashionIQ được sử dụng để làm cơ sở dữ liệu cho hệ thống truy vấn. Ta sử dụng mô hình ResNet50 cùng trọng số **best\_model.pth** để tiến hành rút trích đặc trưng của toàn bộ ảnh trong tập dữ liệu. Tất cả các đặc trưng được trích xuất và lưu trữ dưới dạng numpy với tên **all\_features.npy**, đồng thời ta cũng sẽ lưu trữ luôn cả nhãn (label) tương ứng của chúng với tên **all\_labels.npy**.
- Hai tập tin numpy trên chính là cơ sở dữ liệu của hệ thống truy vấn ảnh.

#### d. Truy vấn

- Như đã trình bày, hệ thống sẽ nhận vào một ảnh đầu vào là ảnh truy vấn cùng với nhãn tương ứng của nó. Việc nhận nhãn đầu vào không được sử dụng cho bước truy vấn mà dùng để thực hiện kiểm tra xem mức độ chính xác mà hệ thống có thể trả ra cho ảnh truy vấn là như thế nào.
- Với ảnh đầu vào đó, hệ thống sẽ tiến hành sử dụng mô hình ResNet50 cùng trọng số **best\_model.pth** để tiến hành rút trích đặc trưng của ảnh (tạm gọi là đặc trưng truy vấn). Với đặc trưng truy vấn này, hệ thống sẽ đếm so khớp, thực hiện tính toán với các đặc trưng có trong cơ sở dữ liệu. Việc tính toán này sử dụng các độ tính toán như Euclidean hoặc Cosine tùy theo lựa chọn của người dùng (trong chương trình được đặt mặc định là Cosine).
- Sau khi tính toán xong, hệ thống sẽ sắp xếp các ảnh có đặc trưng tương đồng cao nhất so với ảnh truy vấn và lấy ra top-k ảnh để biểu diễn kết quả.

### e. Kết quả

- Kết quả sau khi truy vấn sẽ được biểu diễn trên hệ thống (tạm biểu diễn trên màn hình Console). Để có cơ sở đánh giá trực quan, ta sẽ nhờ vào nhãn của tập dữ liệu để xem rằng khi thực hiện truy vấn một ảnh, hệ thống có thể trả ra các ảnh cùng nằm trong loại thời trang đó không. Các ảnh có cùng nhãn với ảnh truy vấn sẽ được biểu diễn trong khung ảnh màu xanh lá (đại diện cho ảnh truy vấn đúng), ngược lại đối với ảnh truy vấn sai sẽ nằm trong khung ảnh màu đỏ.
- Dưới đây là một số kết quả:



(1)



(2)



(3)



(4)

- Có thể thấy, hệ thống thực hiện truy vấn tốt, các ảnh kết quả được trả ra hầu như đều có độ tương tự với ảnh truy vấn, đặc biệt chúng có cùng một nhãn.
- Tuy nhiên, với một số trường hợp (như kết quả tại 3), mặc dù hệ thống truy vấn các ảnh áo có cùng độ tương tự với ảnh cần truy vấn, nhưng chúng lại không được đánh dấu bằng khung ảnh đỏ, tức là không cùng nhãn. Việc này không thể chứng minh rằng hệ thống có sự sai sót, vì nếu không đánh dấu bằng các khung màu, bằng mắt thường và kinh nghiệm của bản thân, ta có thể thấy rằng giữa ảnh kết quả và ảnh truy vấn vẫn có độ tương đồng nhau. Vì thế mà nhãn chỉ là một yếu tố nhỏ để ta kiểm thử mô hình chứ không thể dựa vào đó để đánh giá rằng mô hình đang hoạt động đúng hay sai.

- Các kết quả truy vấn được chỉ nằm ở mức tương đối (ở cùng loại, hoặc màu sắc tương tự nhau, ...) vì chúng gặp phải vấn đề về ngữ nghĩa (semantic gap) cũng như giới hạn về ảnh trong bộ dữ liệu.

## 7. Tái thực nghiệm

### a. Thực nghiệm trên tập dữ liệu DeepFashion

- Trong bước tái thực nghiệm này, nhóm sử dụng ba điểm chuẩn (benchmark) của tập dữ liệu DeepFashion là Category Prediction, InShop Clothes Retrieval và Consumer-to-Shop Clothes Retrieval.
- Thuật toán được triển khai và chạy bước đầu trên nền tảng Google Colab với loại thời gian chạy sử dụng Python 3, trình tăng tốc phần cứng sử dụng T4 GPU và được kết nối đến Google Drive để tải dữ liệu và lưu trữ. Tuy nhiên, vì lý do giới hạn về tài nguyên chạy, nhóm đã đổi sang nền tảng Kaggle với tập dữ liệu tương tự và được chuẩn bị trong thư mục /kaggle/working với cấu trúc các file như sau:
  - + Thư mục img thuộc tập dữ liệu fir-dataset được chứa trong thư mục Category and Attribute Prediction Benchmark
    - ▼ □ /kaggle/working
    - ▼ □ deep-fashion-retrieval
    - ▼ □ Category and Attribute Prediction Benchmark
      - ▶ □ img
  - + Hai thư mục MEN và WOMEN thuộc tập dữ liệu deepfashion-inshop-clothes-retrieval được chứa trong thư mục img thuộc đường dẫn /kaggle/working/deep-fashion-retrieval/Category and Attribute Prediction Benchmark/in\_shop
    - ▼ □ /kaggle/working
    - ▼ □ deep-fashion-retrieval
    - ▼ □ Category and Attribute Prediction Benchmark
      - ▶ □ img
      - ▶ □ Eval
      - ▶ □ models
    - ▼ □ in\_shop
      - ▼ □ img
        - ▶ □ WOMEN
        - ▶ □ MEN

- Sau khi thiết lập cấu trúc thư mục như trên, tiến hành thay đổi đường dẫn đến thư mục chứa tập dữ liệu tương ứng trong tập tin config.py: /kaggle/working/deep-fashion-retrieval/Category and Attribute Prediction Benchmark và gán cho DATABASE\_BASE
- Gọi tập tin train.py để tiến hàng huấn luyện mô hình sử dụng mạng ResNet50. Quá trình huấn luyện diễn ra với 10 vòng lặp với các models được lưu tại thư mục models.

```
%cd /kaggle/working/deep-fashion-retrieval
!python train.py
```

```
Test set: Average loss: 1.6263, Accuracy: 510/960 (53%)

Train Epoch: 10 [99200/99577 (100%)]    All Loss: 3.8972      Triple Loss(0): 1.1666 C1
assification Loss: 1.5640
Train Epoch: 10 [99520/99577 (100%)]    All Loss: 1.7036      Triple Loss(1): 0.0663 C1
assification Loss: 1.5710
Model saved to /kaggle/working/deep-fashion-retrieval/Category and Attribute Prediction Be
nchmark/models/model_10_final.pth.tar
```

- Sau quá trình huấn luyện, model mới nhất sẽ được gọi bằng cách gán tên model cho DUMPED\_MODEL trong tập tin config.py. (Ví dụ nếu dùng quá trình huấn luyện sau khi hoàn tất vòng lặp thứ 2 thì sẽ gán DUMPED\_MODEL = “model\_2\_final.pth.tar”).
- Khởi tạo tập dữ liệu đặc trưng (feature database) bằng cách chạy tập tin feature\_extractor.py. Các đặc trưng bao gồm đặc trưng màu sắc (color\_feat) và tất cả các đặc trưng (all\_feat) sau khi được trích xuất sẽ được lưu dưới dạng các tập tin: all\_feat.npy, all\_color\_feat.npy và all\_feat.list.
- Tăng tốc độ truy vấn bằng cách phân cụm sử dụng thuật toán kmeans với tập tin kmeans.py với số lượng cụm mặc định là 50 cụm. Kết quả phân cụm sẽ được lưu dưới dạng tập tin: kmeans.m trong thư mục models.
- Tiến hành truy vấn ảnh với câu lệnh retrieval.py <img\_path>, trong đó <img\_path> là hình ảnh người dùng muốn truy vấn. Có thể thay đổi các độ đo (metrics) khác nhau bằng việc chỉnh sửa DISTANCE\_METRIC trong tập tin config.py như cosine, euclidean,...

```
!python retrieval.py img/Sheer_Pleated-Front_Blouse/img_00000005.jpg
```

- + Khi thực hiện truy vấn, hệ thống sẽ trả ra kết quả là tập ảnh truy vấn được cùng với độ giá trị cùng độ tương đồng giữa ảnh truy vấn và ảnh kết quả.
- + Chương trình cho phép vừa hiển thị kết quả (là tập ảnh) vừa hiển thị các giá trị kết quả trên màn hình Console.

```

Loading model Done. Time: 0.962 sec
Loading feature database...
Loading feature database Done. Time: 0.196 sec
Extracting image feature...
Extracting image feature Done. Time: 0.409 sec
Loading feature K-means model...
Loading feature K-means model Done. Time: 0.468 sec
Doing naive query...
Doing naive query Done. Time: 0.290 sec
Doing query with k-Means...
Doing query with k-Means Done. Time: 0.228 sec
Naive query result: [('img/Sheer_Pleated-Front_Blouse/img_00000005.jpg', -2.7492896591686836e-06), ('img/Classic_Collared_Button-Down/img_0000054.jpg', -1.8410608924718095), ('img/WOMEN/Sweaters/id_000006070/02_2_side.jpg', -1.8460663112106792), ('img/WOMEN/Blouses_Shirts/id_0007627/05_2_side.jpg', -1.9297835175668132), ('img/Cities_Graphic_Trapeze_Top/img_00000005.jpg', -1.9596936584903952)]
K-Means query result: [('img/Sheer_Pleated-Front_Blouse/img_00000005.jpg', -2.7492896591686836e-06), ('img/Classic_Collared_Button-Down/img_0000054.jpg', -1.8410608924718095), ('img/WOMEN/Sweaters/id_000006070/02_2_side.jpg', -1.8460663112106792), ('img/WOMEN/Blouses_Shirts/id_0007627/05_2_side.jpg', -1.9297835175668132), ('img/Cities_Graphic_Trapeze_Top/img_000000031.jpg', -1.9596936584903952)]

```

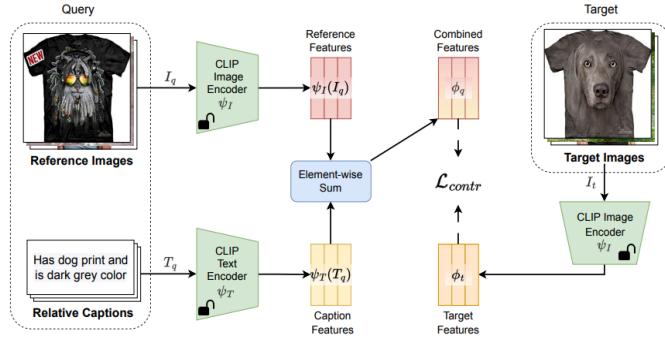
- + Kết quả trả ra là top-5 ảnh có độ tương đồng cao nhất đối với ảnh truy vấn so với tập cơ sở dữ liệu.

## b. Ứng dụng các tính năng từ mô hình OpenAI CLIP

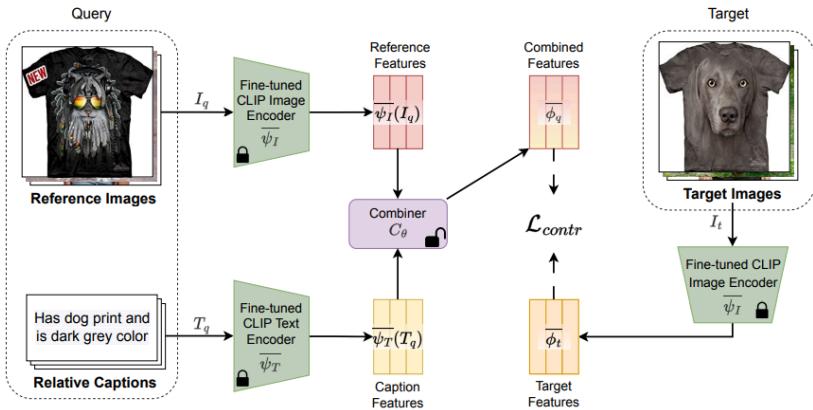
- Với một truy vấn bao gồm một hình ảnh tham chiếu và một chú thích tương đối, mô hình sẽ truy xuất các hình ảnh tương tự về mặt hình ảnh đi kèm với các sửa đổi được thể hiện bởi chú thích.

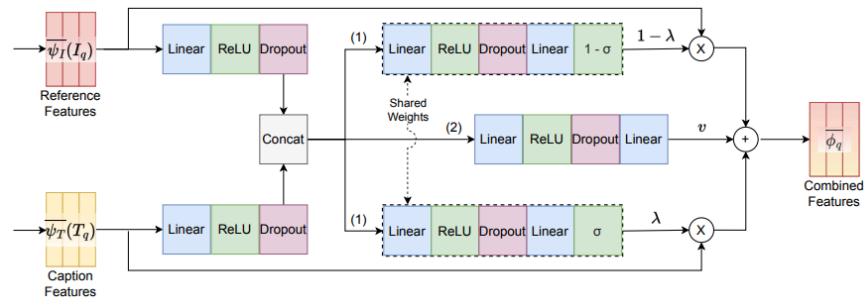


- **Giai đoạn đầu tiên của quá trình huấn luyện.** Trong giai đoạn này, chúng tôi thực hiện task-oriented fine-tuning của bộ mã hóa CLIP nhằm giảm sự không khớp giữa quá trình tiền huấn luyện quy mô lớn (large-scale pre-training) và nhiệm vụ phía sau (downstream task). Chúng tôi bắt đầu bằng cách trích xuất các đặc trưng truy vấn hình ảnh-văn bản và kết hợp chúng thông qua element-wise sum. Sau đó, chúng tôi sử dụng một hàm mất mát tương phản (contrastive loss) để tối thiểu hóa khoảng cách giữa các đặc trưng đã kết hợp và các đặc trưng ảnh mục tiêu trong cùng một triplet và tối đa hóa khoảng cách giữa các ảnh khác trong batch. Cuối cùng chúng tôi cập nhật trọng số của cả hai bộ mã hóa CLIP.



- **Giai đoạn huấn luyện thứ hai.** Trong giai đoạn này, chúng tôi huấn luyện một mạng Combiner học cách hợp nhất các đặc trưng đa phương thức (multimodal) được trích xuất bằng bộ mã hóa CLIP. Chúng tôi bắt đầu bằng cách trích xuất các đặc trưng truy vấn hình ảnh-văn bản bằng các bộ mã hóa được tinh chỉnh, và kết hợp chúng bằng mạng Combiner. Sau đó, chúng tôi sử dụng một hàm mất mát tương phản để tối thiểu hóa khoảng cách giữa các đặc trưng đã kết hợp và các đặc trưng ảnh mục tiêu trong cùng một triplet và tối đa hóa khoảng cách giữa các ảnh khác trong batch. Chúng tôi giữ cả hai bộ mã hóa CLIP ở trạng thái đóng băng trong khi chỉ cập nhật trọng số của mạng Combiner. Tại thời điểm suy luận, các bộ mã hóa được tinh chỉnh và Combiner đã huấn luyện sẽ được sử dụng để tạo ra biểu diễn hiệu quả được dùng để truy vấn cơ sở dữ liệu. Hình dưới đây mô tả kiến trúc của mạng Combiner (lấy các đặc trưng truy vấn đa phương thức làm đầu vào và cho đầu ra là một biểu diễn thống nhất).





- Kết quả:



CF	IFT	TFT	Shirt		Dress		Toptee		Average	
			R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50
Sum	X	X	19.53	35.57	17.70	36.29	21.88	42.93	19.70	38.26
	✓	X	30.08	52.94	29.10	52.01	34.42	57.62	31.20	54.19
	X	✓	32.29	53.73	27.76	52.31	35.14	60.12	31.73	55.39
	✓	✓	38.67	59.42	35.99	62.22	43.35	67.52	39.34	63.05
Combiner	X	X	31.85	52.50	27.22	50.62	33.81	57.57	30.96	53.56
	✓	X	34.30	55.79	32.47	55.18	38.45	62.36	35.07	57.78
	X	✓	35.87	57.21	31.43	54.98	38.20	63.22	35.16	58.47
	✓	✓	39.87	60.84	37.67	63.16	44.88	68.59	40.80	64.20

- + Recall@K trên tập dữ liệu validation FashionIQ khi thay đổi hàm kết hợp và phương thức tinh chỉnh CLIP. Trong đó IFT (image encoder fine-tuning) và TFT (text encoder fine-tuning) biểu thị bộ mã hóa hình ảnh hay văn bản được tinh chỉnh ở giai đoạn đầu tiên. CF (combining function) đại diện cho hàm được sử dụng để kết hợp các đặc trưng truy vấn. Điểm số tốt nhất được **in đậm** và điểm số tốt thứ hai được gạch chân.

Method	Encoder		Shirt		Dress		Toptee		Average	
	Visual	Textual	R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50
TRACE [27]	RN-50	BERT [15]	20.80	40.80	22.70	44.91	24.22	49.80	22.57	46.19
VAL [8]	RN-50	LSTM(GloVe) [23]	22.38	44.15	22.53	44.00	27.53	51.68	24.15	46.61
CurlingNet [52]	RN-152	biGRU [10]	21.45	44.56	26.15	53.24	30.12	55.23	25.90	51.01
RTIC-GCN [45]	RN-50	LSTM(GloVe)	23.79	47.25	29.15	54.04	31.61	57.98	28.18	53.09
CoSMo [31]	RN-50	LSTM	24.90	49.18	25.64	50.30	29.21	57.46	26.58	52.31
DCNet [29]	RN-50	Conv1D(GloVe)	23.95	47.30	28.95	56.07	30.44	58.29	27.78	53.89
CLVC-Net [50]	RN-50	LSTM	28.75	54.76	29.85	56.47	33.50	64.00	30.70	58.41
CIRPLANT [37]	RN-152	OSCAR [34]	17.53	38.81	17.45	40.41	21.64	45.38	18.87	41.53
MAAF [16]	RN-50	BERT	18.55	37.63	18.59	39.66	23.05	45.95	20.06	41.08
SAC [26]	RN-50	BERT	28.02	51.86	26.52	51.01	32.70	61.23	29.08	54.70
FashionViL [21]	RN-50	BERT	25.17	50.39	33.47	59.94	34.98	60.79	31.20	57.04
Ours	RN-50	Transformer	39.87	60.84	37.67	63.16	44.88	68.59	40.80	64.20
Ours	RN-50x4	Transformer	44.41	65.26	39.46	64.55	47.48	70.98	43.78	66.93

- + So sánh với các mô hình SOTA trên bộ FashionIQ validation. Điểm số tốt nhất được **in đậm** và điểm số tốt thứ hai được gạch chân. "RN" viết tắt cho ResNet.

## 8. Cải tiến

### a. Vấn đề hiện tại

- Với phương pháp hiện tại, ở bước truy vấn, việc so khớp đặc trưng truy vấn với toàn bộ đặc trưng có trong cơ sở dữ liệu tiềm ẩn nhiều vấn đề có thể gây ra cho hệ thống như tiêu tốn thời gian, tài nguyên và chi phí tính toán

### b. Giải pháp cải tiến

- Với lợi thế của bộ dữ liệu FashionIQ, các tác giả đã cung cấp sẵn nhãn tương ứng cho từng tấm ảnh, ta có thể tận dụng nó để làm ưu thế cho việc truy vấn bằng cách chỉ thực hiện truy vấn ảnh đầu vào với loại thời trang tương ứng của nó. Với giải pháp này, hệ thống sẽ có thể tiết kiệm thời gian truy vấn, các chi phí tính toán cũng được giảm dần do việc so khớp các đặc trưng đã được thu hẹp lại.
- Để thực hiện điều này, hệ thống cần bổ sung một lớp SVM ở bước truy vấn, giúp ta phân loại ảnh truy vấn thuộc về loại thời trang nào trước khi đếm so khớp.
  - + Support Vector Machine (SVM) là một mô hình học máy được sử dụng phổ biến trong các bài toán phân lớp. Mục tiêu của SVM là tìm ra một siêu phẳng (hyperplane) phân tách dữ liệu thành các lớp khác nhau với khoảng cách lớn nhất có thể giữa các điểm gần nhất của hai lớp (gọi là support vectors).
  - + Để có thể sử dụng mô hình SVM hiệu quả cho bài toán của mình, ta cần huấn luyện nó trên tập dữ liệu FashionIQ. Mô hình được thử nghiệm với nhiều bộ siêu tham số khác nhau để tìm ra bộ tốt nhất cho mô hình. Cuối cùng, bộ tham số tìm được gồm có  $C = 1$ , Kernel = “rbf” và Gamma = “auto”, kết quả huấn luyện cho ra đạt được độ chính xác 0.91 cho tập đánh giá (valid) và 0.898 cho tập kiểm thử (test).
  - + Sau khi huấn luyện được mô hình đạt mức ổn định, ta sẽ lưu lại trọng số của nó để phục vụ cho các bước sau, trọng số này được lưu dưới tên **svm\_model.joblib**.
- Lúc này, khi nhận ảnh đầu vào, hệ thống vẫn sử dụng mô hình ResNet50 để trích xuất đặc trưng của ảnh truy vấn (là đặc trưng truy vấn). Nhưng trước khi đếm so khớp với cơ sở dữ liệu, đặc trưng truy vấn này sẽ đi qua mô hình SVM (với trọng số **svm\_model**) để thực hiện phân loại nhãn cho nó. Với nhãn được dự đoán, hệ thống sẽ dựa vào đó để so khớp đặc trưng truy vấn với các đặc trưng có cùng nhãn với nó, nghĩa là nó sẽ bỏ qua việc so khớp với các ảnh không cùng nhãn với ảnh truy vấn trong cơ sở dữ liệu. Sau khi so khớp và có được kết quả về độ tương đồng, hệ thống sẽ sắp xếp và trả ra các ảnh có độ tương đồng cao nhất (hoặc dị biệt thấp nhất) so với ảnh truy vấn. Các bước ở sau vẫn được giữ nguyên.

### c. Kết quả

- Để có thể đánh giá tổng quan về mức độ cải thiện của phương pháp cải tiến và phương pháp ban đầu, ta sử dụng cả hai phương pháp trên tập kiểm thử để trả ra kết quả top-k ảnh với 4 giá trị k lần lượt là 10, 100, 1000 và 10000. Kết quả được biểu diễn bằng ảnh sau: (1) đại diện cho kết quả phương pháp ban đầu; (2) đại diện cho kết quả phương pháp cải tiến

```

Top-10:
Mean mAP: 0.933
Mean Accuracy: 0.875
Mean Query Time: 0.009 seconds

Top-100:
Mean mAP: 0.885
Mean Accuracy: 0.866
Mean Query Time: 0.009 seconds

Top-1000:
Mean mAP: 0.870
Mean Accuracy: 0.866
Mean Query Time: 0.009 seconds

Top-10000:
Mean mAP: 0.865
Mean Accuracy: 0.858
Mean Query Time: 0.009 seconds

```

(1)

```

Top-10:
mAP: 0.898
Mean Accuracy: 0.898
Mean Query Time: 0.006 seconds

Top-100:
mAP: 0.898
Mean Accuracy: 0.898
Mean Query Time: 0.006 seconds

Top-1000:
mAP: 0.898
Mean Accuracy: 0.898
Mean Query Time: 0.006 seconds

Top-10000:
mAP: 0.898
Mean Accuracy: 0.898
Mean Query Time: 0.006 seconds

```

(2)

- **Thời gian:** Có thể thấy, khi kết hợp mô hình SVM, tốc độ truy vấn của mô hình đã được cải thiện, trung bình thời gian từ 0.09 ban đầu giờ chỉ còn 0.06 giây.
- **Độ chính xác:** Với phương pháp ban đầu, độ chính xác của nó với giá trị k nhỏ (như k=10) có phần tốt hơn so với phương pháp cải tiến, tuy nhiên, khi các giá trị k tăng dần thì độ chính xác lại càng giảm. Trong khi với phương pháp cải tiến, các giá trị được ổn định và không thay đổi khi giá trị k ngày càng tăng. Điều này có thể lý giải do độ chính xác của phương pháp cải tiến được quyết định ở bước phân lớp. Chính vì thế, nếu mô hình SVM phân lớp ảnh truy vấn đúng thì toàn bộ các ảnh được truy vấn sẽ đúng, ngược lại sẽ sai hoàn toàn. Đây là một bước đánh đổi của phương pháp cải tiến này.

## 9. Deploy sản phẩm

- Đầu tiên, chạy cell code bên dưới để run webapp sử dụng streamlit, output cell sẽ hiển thị 2 dòng thông tin quan trọng như ở hình bên dưới. Dòng đầu tiên là tunnel password (sẽ sử dụng cho bước tiếp theo), và dòng thứ hai là địa chỉ url của webapp.

```

!wget -q -O - https://loca.lt/mytunelpassword
!streamlit run src/webapp.py & npx localtunnel --port 8501

→ 34.143.134.81
Collecting usage statistics. To deactivate, set browser.gatherUsageStats to false.

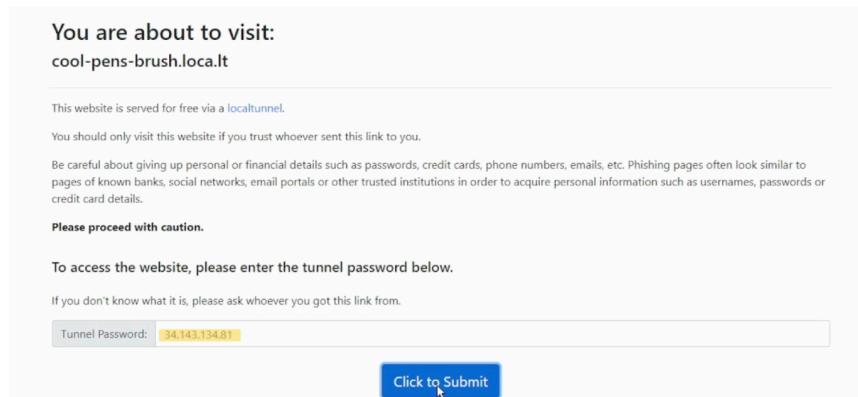
You can now view your Streamlit app in your browser.

Local URL: http://localhost:8501
Network URL: http://172.28.0.12:8501
External URL: http://34.143.134.81:8501

your url is: https://evil-tires-pick.loca.lt

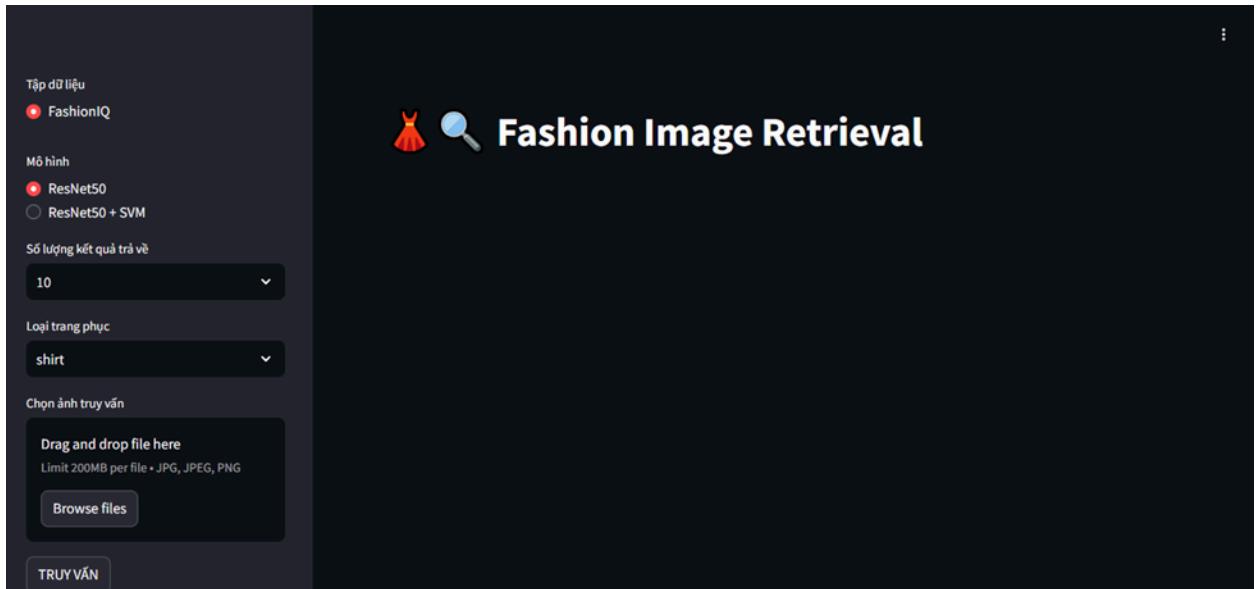
```

- Sau khi truy cập địa chỉ url trên, màn hình sẽ yêu cầu nhập password, thực hiện xong thì nhấn chọn vào “Click to Submit”.

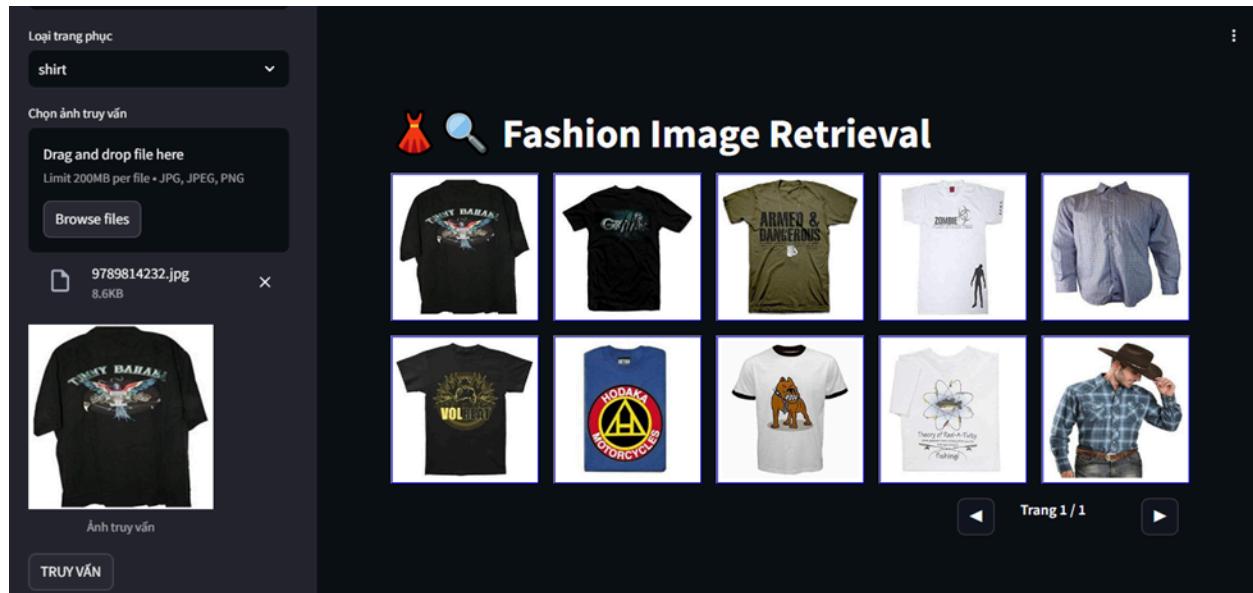


- Dẫn đến một trang chủ chính của streamlit webapp, tại đây ta thực hiện nhập các đầu vào cần thiết như:
  - Mô hình: ResNet50 (mô hình ban đầu, trước khi cài tiến) hay ResNet50+SVM (mô hình sau khi cài tiến)
  - Số lượng kết quả trả về (10, 100, 1000, 10000)
  - Loại trang phục (dress, shirt, toptee)
  - Ảnh truy vấn

Sau đó nhấn “TRUY VẤN”



- Kết quả sau khi truy vấn là top ảnh được sắp xếp theo thứ tự giảm dần mức độ phù hợp.



## 10. Kết luận

- Việc khai thác và sử dụng mô hình CNN (cụ thể là backbone ResNet50) để phục vụ cho tác vụ truy vấn ảnh thời trang đem lại một kết quả ổn định và khá tốt mà không phức tạp. Tuy nhiên, vẫn còn những thách thức cần phải đổi mới để nâng cao hiệu quả của hệ thống.
- Một trong những thách thức chính là việc xử lý và quản lý khối lượng dữ liệu lớn. Cần có các phương pháp tối ưu hóa và làm giàu dữ liệu để hệ thống có thể nhận diện và truy vấn ảnh một cách chính xác hơn. Ngoài ra, việc cải thiện khả năng xử lý các biến đổi về ánh sáng, góc nhìn, và bối cảnh cũng là một yếu tố quan trọng để tăng cường độ chính xác của hệ thống.
- Một thách thức khác là việc tối ưu hóa thời gian xử lý để đáp ứng nhu cầu truy vấn thời gian thực. Sự phát triển của các phương pháp giảm thiểu độ phức tạp tính toán, như kỹ thuật nén mô hình hoặc tối ưu hóa phần cứng, sẽ đóng vai trò quan trọng trong việc cải thiện tốc độ và hiệu suất của hệ thống.
- Tóm lại, mặc dù đã đạt được những kết quả đáng khích lệ với mô hình CNN và ResNet50 trong việc truy vấn ảnh thời trang, vẫn còn nhiều cơ hội để cải thiện và phát triển hơn nữa. Việc tiếp tục nghiên cứu và tối ưu hóa các mô hình cũng như áp dụng những công nghệ tiên tiến sẽ giúp nâng cao chất lượng và hiệu quả của hệ thống, đáp ứng ngày càng tốt hơn nhu cầu của người dùng và doanh nghiệp.

## C. TÀI LIỆU THAM KHẢO

- [1] Charles Corbière et al. “Leveraging Weakly Annotated Data for Fashion Image Retrieval and Label Prediction”. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 2017, pp. 14105– 14115.
- [2] Sonam Goenka et al. “FashionVLP: Vision Language Transformer for Fashion Retrieval with Feedback”. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022, pp. 14105– 14115.
- [3] Furkan Kınlı, Barış Ozcan, and Furkan Kırıcı. “Fashion Image Retrieval with Capsule Networks”. In: Proceedings of the International Conference on Computer Vision (ICCV) Workshops. 2019.
- [4] Zhanghui Kuang et al. “Fashion Retrieval via Graph Reasoning Networks on a Similarity Pyramid”. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 2019, pp. 14105–14115.
- [5] Yining Lang et al. “Which is Plagiarism: Fashion Image Retrieval based on Regional Representation for Design Protection”. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020, pp. 1–9.
- [6] Sanghyuk Park et al. “Study on Fashion Image Retrieval Methods for Efficient Fashion Visual Search”. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. 2019, pp. 14105–14115.
- [7] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, Xiaoou Tang “DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations”. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1096-1104.

[8] Y. Zhang, J. Chen, J. Wu, and S. Wang "Fashion IQ: A New Dataset Towards Retrieving Images by Natural Language Feedback". In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)\*, 2020, pp. 601-610.

[9] Baldrati, Alberto and Bertini, Marco and Uricchio, Tiberio and Bimbo, Alberto Del "Composed Image Retrieval using Contrastive Learning and Task-oriented CLIP-based Features". In: ACM Transactions on Multimedia Computing, Communications and Applications.

Mã nguồn tham khảo

<https://github.com/ihciah/deep-fashion-retrieval?tab=readme-ov-file>

<https://github.com/ABaldrati/CLIP4Cir>