Fashion Image Search Engine

Nguyễn Công Nguyên, Trần Lê Bảo Trung
Khoa Khoa học máy tính, Trường Đại học Công nghệ Thông tin

Tóm tắt nội dung
Báo cáo này trình bày quy trình xây dựng công cụ truy vấn hình ảnh trong lĩnh vực thời trang, sử dụng đặc trưng được trích xuất từ ba mô hình đã được fine-tuning: VGG16, ResNet50 và InceptionV3. Các mô hình này được pretrained trên ImageNet và sau đó được fine-tuning trên tập dữ liệu Farfetch-listings. Kết quả đánh giá qua độ đo mAP@10 cho thấy mô hình ResNet50 đạt hiệu suất tốt nhất với mAP@10 là 0.8433, tiếp theo là InceptionV3 (0.8251) và VGG16 (0.8241). Các kết quả thực nghiệm cung cấp những nhận xét quan trọng về hiệu suất của các mô hình học sâu trong bài toán truy vấn hình ảnh thời trang.

Mục lục

1 Tổng quan đề tài

1.1 Bối cảnh

Trong thời đại ngày nay, việc mua sắm trực tuyến đang trở nên ngày càng phổ biến, đặc biệt là trong lĩnh vực thời trang. Nhu cầu của người dùng tăng cao và cũng ngày một trở nên phức tạp. Việc tìm kiếm những sản phẩm mong muốn có thể trở nên khá thách thức khi người tiêu dùng muốn định vị những mẫu sản phẩm tương tự với hình ảnh họ thấy trên mạng Internet, hay đơn giản là giống với một sản phẩm mà họ vô tình nhìn thấy và chụp lại. Nhận thức được điều đó, nhóm đã quyết định thực hiện phát triển một hệ thống truy vấn hình ảnh trong lĩnh vực thời trang, với mục tiêu giúp người dùng có thể dễ dàng tìm kiếm và xác định những sản phẩm thông qua hình ảnh cụ thể.

1.2 Lý do lựa chọn đề tài

Hệ thống truy vấn hình ảnh giúp người dùng dễ dàng tìm thấy những sản phẩm mà họ yêu thích chỉ bằng cách cung cấp một hình ảnh, tiết kiệm thời gian và công sức so với việc phải tìm kiếm thủ công. Các cửa hàng thời trang trực tuyến có thể tận dụng công nghệ này để cải thiện khả năng gợi ý sản phẩm, từ đó tăng tỷ lệ chuyển đổi và doanh thu. Ngoài ra, công nghệ này còn giúp đáp ứng tốt hơn nhu cầu cá nhân của người dùng, tạo ra những gợi ý phù hợp với sở thích và phong cách cá nhân, từ đó nâng cao sự hài lòng và trung thành của khách hàng.

1.3 Ứng dụng

1. Các nền tảng mua sắm trực tuyến: Ứng dụng truy vấn hình ảnh có thể tích hợp vào các trang web và ứng dụng mua sắm, giúp người dùng dễ dàng tìm kiếm sản phẩm theo hình ảnh.
2. Công cụ gợi ý thời trang: Các ứng dụng gợi ý thời trang có thể sử dụng công nghệ này để đưa ra những gợi ý phù hợp với phong cách và sở thích của người dùng, từ đó giúp họ dễ dàng tạo nên các bộ trang phục hoàn hảo.
3. Quản lý và kiểm kê kho hàng: Hệ thống quản lý kho có thể sử dụng công nghệ này để tự động nhận diện và phân loại sản phẩm, giúp việc kiểm kê và quản lý kho trở nên hiệu quả hơn.
4. Mạng xã hội và cộng đồng thời trang: Các ứng dụng mạng xã hội về thời trang có thể sử dụng

tính năng truy vấn hình ảnh để kết nối người dùng với những sản phẩm và xu hướng mới nhất, từ đó tạo ra một cộng đồng chia sẻ và trao đổi thông tin phong phú hơn.

5. Ngành công nghiệp bán lẻ: Các cửa hàng bán lẻ có thể sử dụng công nghệ này để cung cấp dịch vụ tìm kiếm sản phẩm cho khách hàng tại chỗ.

2 Giới thiệu bài toán

2.1 Tổng quát bài toán

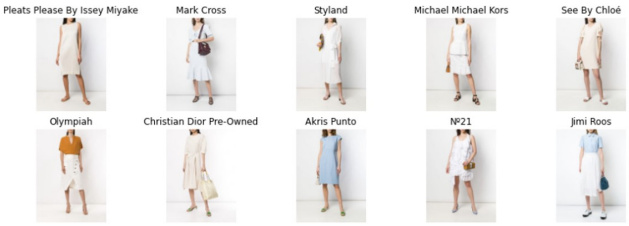
Bài toán được trình bày trong báo cáo này liên quan đến việc truy vấn hình ảnh trong lĩnh vực thời trang. Người dùng cung cấp một hình ảnh sản phẩm như quần áo, giày dép, hoặc ví. Hệ thống sẽ trả về danh sách các sản phẩm tương tự với hình ảnh đầu vào, được xếp hạng theo mức độ tương đồng. Hình 1 minh họa một ví dụ về hình ảnh đầu vào.

2.2 Mô tả bài toán

Input (Đầu vào): Một hình ảnh truy vấn (query image), ví dụ như hình minh họa ở Hình 1, thể hiện sản phẩm thời trang mà người dùng quan tâm.

Output (Đầu ra): Một danh sách xếp hạng các hình ảnh tương tự nhất từ cơ sở dữ liệu, với hình ảnh giống nhất đứng đầu danh sách. Hình 2 minh họa kết quả đầu ra, bao gồm các sản phẩm tương tự với sản phẩm truy vấn.

Hình 1: Ví dụ hình ảnh đầu vào (Input)

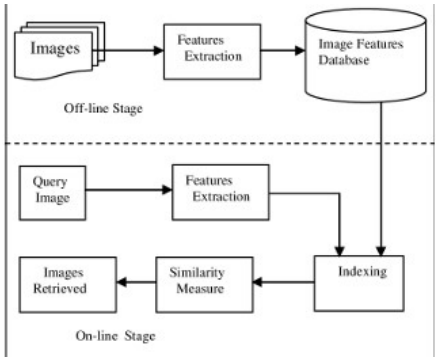


Hình 2: Ví dụ hình ảnh đầu ra (Output)

3 Cơ sở lý thuyết

3.1 Hệ thống truy vấn hình ảnh

Hệ thống truy vấn hình ảnh cho phép tìm kiếm các hình ảnh tương tự từ cơ sở dữ liệu. Quá trình gồm hai bước: đầu tiên, trích xuất đặc trưng từ tất cả hình ảnh trong cơ sở dữ liệu; tiếp theo, hệ thống so sánh độ tương đồng giữa hình ảnh truy vấn và các hình ảnh trong cơ sở dữ liệu. Hình minh họa 3 cho thấy sơ đồ hoạt động của hệ thống truy vấn hình ảnh.



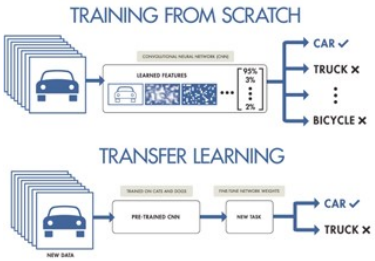
Hình 3: Hệ thống truy vấn hình ảnh

3.2 CNN và ImageNet

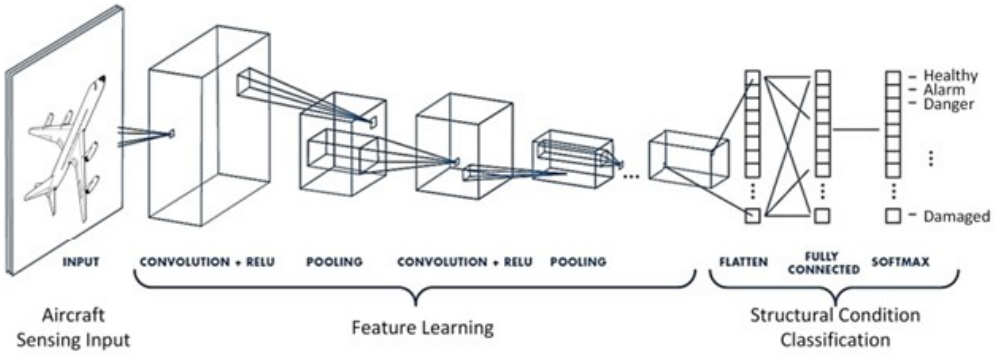
Convolutional Neural Network (CNN) là mô hình học sâu mạnh mẽ để xử lý và phân tích hình ảnh. CNN bao gồm các lớp như Convolutional, Pooling, và Fully Connected, giúp mô hình học được các đặc trưng từ hình ảnh một cách tự động. Hình minh họa5 cho thấy kiến trúc cơ bản của một mạng CNN. ImageNet [2]là một tập dữ liệu lớn với hàng triệu hình ảnh được phân loại vào nhiều danh mục. Tập dữ liệu này thường được dùng để huấn luyện các mô hình học sâu như CNN, tạo điều kiện thuận lợi cho việc phát triển và kiểm tra các mô hình xử lý hình ảnh. Mặc dù không có hình minh họa trực tiếp, ImageNet đóng vai trò quan trọng trong nghiên cứu thị giác máy tính và đã thúc đẩy sự phát triển của nhiều mô hình hiệu quả.

3.3 Học chuyển giao

Học chuyển giao (Transfer Learning) là một kỹ thuật trong học máy, trong đó một mô hình đã được huấn luyện trước trên một tập dữ liệu lớn sẽ được sử dụng làm nền tảng để huấn luyện một mô hình mới trên một tập dữ liệu nhỏ hơn hoặc liên quan đến một tác vụ mới. Ý tưởng chính của học chuyển giao là tận dụng kiến thức đã học từ một tác vụ trước đó để cải thiện hiệu suất của mô hình trên một tác vụ mới, thường có ít dữ liệu huấn luyện. Thay vì xây dựng mô hình từ đầu, chúng ta sử dụng một mô hình được huấn luyện trước (pre-trained model) và điều chỉnh nó cho phù hợp với tác vụ mới. Học chuyển giao mang lại nhiều lợi ích như rút ngắn thời gian huấn luyện và đạt hiệu suất tốt hơn khi dữ liệu hạn chế. Dựa trên ý tưởng này, đồ án sử dụng một số kiến trúc CNN đã được huấn luyện trước, vốn phổ biến trong các tác vụ thị giác máy tính, để trích xuất đặc trưng từ hình ảnh. Hình minh họa 4 thể hiện quy trình cơ bản của học chuyển giao.



Hình 4: Quy trình học chuyển giao



Hình 5: Kiến trúc CNN cơ bản

4 Phương pháp cụ thể

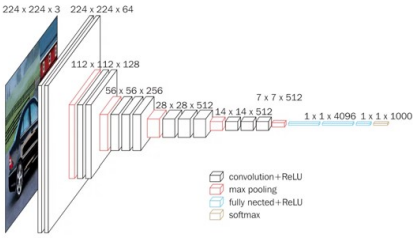
4.1 Trích xuất đặc trưng cho Fashion Image Retrieval

Trong bài toán truy vấn hình ảnh thời trang (Fashion Image Retrieval), mục tiêu là tìm kiếm các hình ảnh tương tự trong cơ sở dữ liệu dựa trên một hình ảnh truy vấn từ người dùng. Để thực hiện điều này, việc trích xuất đặc trưng hình ảnh (feature extraction) là bước quan trọng, giúp hệ thống nhận diện và so sánh các yếu tố quan trọng của hình ảnh. Các mô hình học sâu như VGG16 [3], ResNet50 [1] và InceptionV3 [4] thường được sử dụng để trích xuất đặc trưng từ các hình ảnh thời trang. Các mô hình này học được các đặc trưng sâu sắc từ các lớp convolution và pooling, giúp nhận diện các yếu tố như hình dáng, màu sắc, chất liệu, và kiểu dáng của sản phẩm.

4.2 VGG16

VGG16 là một trong những kiến trúc CNN cơ bản nhưng mạnh mẽ cho việc trích xuất đặc trưng hình ảnh. Mô hình này, được giới thiệu bởi Simonyan và

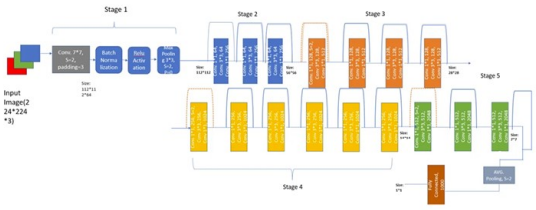
Zisserman[3], sử dụng một loạt các lớp convolution và sau đó sử dụng một số lớp fully connected trước khi phân loại để trích xuất các đặc trưng quan trọng từ hình ảnh. Các lớp fully connected cuối cùng của VGG16 (cụ thể là lớp trước lớp phân loại) giúp mô hình tạo ra một đại diện đặc trưng có thể sử dụng để so sánh hình ảnh. Trong bài toán fashion image retrieval, các đặc trưng này phản ánh thông tin về hình dáng và kiểu dáng sản phẩm thời trang, giúp so sánh hình ảnh truy vấn với cơ sở dữ liệu. Hình 6 minh họa cấu trúc của VGG16.



Hình 6: Kiến trúc VGG16

4.3 ResNet50

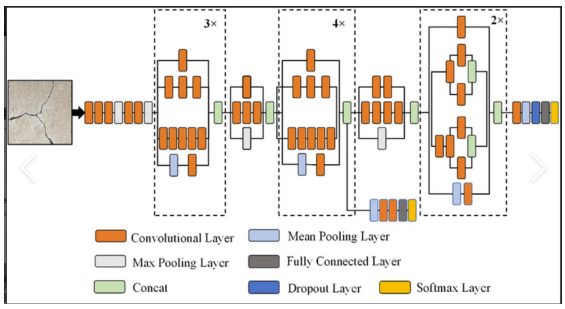
ResNet50 cải tiến so với VGG16 nhờ vào việc sử dụng các kết nối dư (residual connections), giúp cải thiện khả năng học của mô hình khi xử lý các hình ảnh phức tạp. Được giới thiệu bởi He et al.[1], ResNet50 sử dụng lớp Average Pooling (avgpool) thay vì lớp fully connected như VGG16 để trích xuất đặc trưng. Lớp avgpool giúp giảm chiều của các đặc trưng đầu ra, tạo ra một đại diện tổng quát hơn của hình ảnh mà không cần phải sử dụng nhiều lớp fully connected. Các đặc trưng học được từ ResNet50 có thể giúp hệ thống phân loại các sản phẩm thời trang với độ chính xác cao hơn và là một lựa chọn hiệu quả cho việc trích xuất đặc trưng trong bài toán fashion image retrieval. Hình 7 minh họa cấu trúc của ResNet50.



Hình 7: Kiến trúc ResNet50

4.4 InceptionV3

InceptionV3, được giới thiệu bởi Szegedy et al.[4], là một kiến trúc tiên tiến trong học sâu, nổi bật với việc sử dụng các khối inception, giúp xử lý hiệu quả thông tin ở nhiều kích thước kernel khác nhau trong cùng một lớp. Thay vì sử dụng các lớp fully connected, InceptionV3 sử dụng lớp Average Pooling (avgpool) để trích xuất đặc trưng từ hình ảnh. Điều này giúp mô hình học các đặc trưng đa dạng và phức tạp từ hình ảnh, như kết cấu vải, kiểu dáng sản phẩm, hoặc các yếu tố nhỏ trong hình ảnh mà các lớp convolution không thể nắm bắt được. Trong bài toán fashion image retrieval, InceptionV3 cung cấp độ chính xác cao trong việc tìm kiếm các hình ảnh tương tự, đặc biệt trong các trường hợp có sự thay đổi về góc nhìn hay độ sáng. Hình 8 minh họa cấu trúc của InceptionV3.



Hình 8: Kiến trúc InceptionV3

5 Thực nghiệm

5.1 Bộ dữ liệu

Bộ dữ liệu sử dụng trong bài toán này là *Farfetch-Listing*[5], một bộ dữ liệu chứa 187,620 hình ảnh thời trang được thu thập từ trang web Farfetch. Các hình ảnh này đều được lưu ở định dạng .jpg và bao gồm các sản phẩm thời trang với các mô tả chi tiết về tên, giá trị, chất liệu, kiểu dáng, màu sắc, và các thuộc tính khác của từng sản phẩm. Bộ dữ liệu này được gắn nhãn với các đặc điểm của sản phẩm, giúp cho bài toán truy vấn hình ảnh thời trang trở nên thực tế và đa dạng.

5.2 Độ đo

Trong bài toán *Fashion Image Retrieval*, độ đo quan trọng được sử dụng để đánh giá hiệu quả của các mô hình là *Mean Average Precision (mAP)*. Độ đo mAP tính toán độ chính xác trung bình của các kết quả trả về từ mô hình ở các vị trí top-k, với k là các giá trị 1, 5, và 10 trong trường hợp này.

Công thức tính mAP cho một truy vấn là:

$$\text{AP}(k) = \frac{1}{|R|} \sum_{r \in R} P(k_r)$$

Trong đó: - R là tập hợp các hình ảnh phù hợp (relevant images). - $P(k_r)$ là độ chính xác tại vị trí k_r của hình ảnh thứ r trong kết quả truy vấn.

Sau đó, mAP được tính bằng trung bình của độ chính xác trung bình tại các top-k khác nhau. Độ đo này cho thấy khả năng của mô hình trong việc tìm kiếm các hình ảnh tương tự với độ chính xác cao, đặc biệt khi cần tìm kiếm các sản phẩm tương tự trong các tập dữ liệu lớn.

Top-k	**InceptionV3**	**ResNet50**	**VGG16**
Top 1	0.7938	0.8122	0.7885
Top 5	0.8355	0.8550	0.8329
Top 10	0.8241	0.8433	0.8251

Bảng 1: Kết quả thực nghiệm mAP của các mô hình trên bộ dữ liệu Farfetch-Listing

5.3 Kết quả thực nghiệm

Kết quả thực nghiệm được thực hiện trên bộ dữ liệu *Farfetch-Listing* [5] sử dụng ba mô hình phổ biến trong việc trích xuất đặc trưng hình ảnh: *InceptionV3*, *ResNet50* và *VGG16*. Các mô hình này đã được huấn luyện và áp dụng cho bài toán Fashion Image Retrieval, với các kết quả như sau:

Từ bảng trên, có thể thấy rằng ResNet50 đạt điểm mAP cao nhất ở các top-k so với InceptionV3 và VGG16. Cụ thể: - Ở top 1, ResNet50 đạt độ chính xác 0.8122, cao hơn một chút so với InceptionV3 (0.7938) và VGG16 (0.7885). - Ở top 5, ResNet50 cũng đạt điểm cao nhất với mAP là 0.8550, vượt trội hơn so với InceptionV3 (0.8355) và VGG16 (0.8329). - Ở top 10, ResNet50 vẫn giữ vị trí dẫn đầu với mAP là 0.8433, trong khi InceptionV3 (0.8241) và VGG16 (0.8251) gần như ngang bằng nhau.

Các kết quả này cho thấy rằng ResNet50 với kiến trúc residual connections giúp mô hình học được các đặc trưng phức tạp hơn, cải thiện độ chính xác trong việc tìm kiếm các hình ảnh tương tự. Trong khi đó, InceptionV3 và VGG16 cũng thể hiện hiệu quả tốt, với các độ chính xác mAP khá gần nhau, nhưng không thể đạt được mức độ chính xác cao như ResNet50.

Nhìn chung, các mô hình học sâu như InceptionV3, ResNet50 và VGG16 đều có khả năng hiệu quả trong việc trích xuất đặc trưng và tìm kiếm hình ảnh thời trang, nhưng ResNet50 cho thấy ưu thế vượt trội trong việc tối ưu hóa độ chính xác ở các mức độ tìm kiếm top-k.

6 Tổng kết

6.1 Kết luận

Mô hình CNN đã thể hiện khả năng xuất sắc trong việc tìm kiếm hình ảnh tương tự với hình ảnh đầu vào. Điều này chứng tỏ hiệu quả và hiệu suất cao của các mô hình CNN trong ngữ cảnh bài toán truy vấn hình ảnh. Nhóm nghiên cứu đã thành công trong việc giải quyết bài toán này, đặc biệt khi sử dụng các mô hình pre-trained như ResNet50 ,VGG16 và InceptionV3 mang lại hiệu suất cao và giảm đáng kể thời gian cũng như tài nguyên tính toán. Việc áp dụng học chuyển giao trong truy vấn hình ảnh thời trang giúp cải thiện khả năng trích xuất đặc trưng, tăng cường tổng quát hóa và giảm nguy cơ overfitting, làm cho mô hình trở nên ổn định và đáng tin cậy.

Khoảng cách Euclidean, mặc dù đơn giản và dễ hiểu, đã chứng minh là một phương pháp tốt để phản ánh độ tương đồng, đặc biệt khi áp dụng trong không gian đặc trưng đã được trích xuất từ các mô hình pre-trained. Về phần Indexing, mặc dù đoạn mã hiện tại chưa tạo ra chỉ mục hoàn chỉnh, nó đã đặt nền tảng cho việc indexing bằng cách lưu trữ các đặc trưng cùng với chỉ mục hình ảnh trong từ điển. Việc này là quan trọng để tối ưu hóa quá trình truy vấn và tăng cường hiệu suất của hệ thống.

6.2 Hướng phát triển

Trong quá trình tạo chỉ mục bằng cách lưu trữ đặc trưng cùng với chỉ mục hình ảnh, để tối ưu hóa hiệu suất và tốc độ truy vấn, có thể áp dụng các thư viện chuyên dụng như Faiss, Annoy, hoặc Elasticsearch. Điều này không chỉ giúp nâng cao hiệu suất của truy vấn mà còn tăng cường khả năng mở rộng của hệ thống. Nhóm nghiên cứu sẽ mở rộng và nâng cao khả năng học của hệ thống truy vấn hình ảnh thời trang bằng cách thử nghiệm và đánh giá hiệu suất của các mô hình pre-trained khác nhau như EfficientNet, MobileNet, và các mô hình mới hơn. Mục tiêu là tìm ra mô hình phù hợp nhất cho bài toán cụ thể của nhóm, giúp cải thiện kết quả và trải nghiệm người

dùng trong quá trình tìm kiếm sản phẩm thời trang.

Tích hợp các tính năng tương tác người dùng có thể làm tăng trải nghiệm sử dụng. Các tính năng như tìm kiếm theo màu sắc, xu hướng thời trang mới, hoặc gợi ý sản phẩm có thể cải thiện sự tương tác và tìm kiếm của người dùng. Nhóm cũng sẽ nghiên cứu và so sánh các phương pháp đo lường độ tương đồng khác như Cosine Similarity, Jaccard Similarity để xem liệu chúng có thể cung cấp kết quả tốt hơn khoảng cách Euclidean trong ngữ cảnh bài toán cụ thể này.

Với mAP đạt được từ các mô hình ResNet50 và VGG16, kết quả thực nghiệm cho thấy sự cải thiện đáng kể trong hiệu suất truy vấn hình ảnh. Nhóm sẽ tiếp tục tối ưu hóa các mô hình và nâng cao khả năng ứng dụng của hệ thống trong các trường hợp thực tế.

Tài liệu

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[2] Olga Russakovsky, Jia Deng, Hongwei Su, Jonathan Krause, Sanjeev Satheesh, Shuohang Ma, Ziwei Huang, Andrej Karpathy, Abhinav Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[3] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2015.

[4] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[5] Liling Tan. Farfetch listings dataset. https://www.kaggle.com/datasets/alvations/farfetch-listings, 2019. Accessed: 2024-05-20.