TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN – ĐHQG TPHCM

KHOA CÔNG NGHỆ THÔNG TIN (CLC)

IMAGE RETRIEVAL

Chuyên đề Thị giác máy tính

GVHD: Thầy Võ Hoài Việt

Sinh viên thực hiện: Nguyễn Hoàng Sơn

Hoàng Thị Quỳnh Liên

Huỳnh Thị Mỹ Thanh

Nhóm 6

MỤC LỤC:

I. Giới thiệu chung về bài toán Image Retrieval:

1. Tổng quan

Image retrieval là một lĩnh vực khá nổi bật trong thị giác máy tính, liên quan đến việc tìm kiếm và truy xuất các hình ảnh dựa trên các đặc trưng của chúng. Nó đang trở thành một chủ đề quan tâm đến trong thị giác máy tính vì sự phát triển của công nghệ số. Image retrieval có thể được sử dụng trong nhiều lĩnh vực, bao gồm tìm kiếm ảnh trực tuyến, quản lý ảnh trong hệ thống tập tin và phân tích hình ảnh trong các ứng dụng y tế.

Với sự phát triển của công nghệ số hóa, số lượng hình ảnh trên Internet và các nền tảng truyền thông xã hội ngày càng tăng, làm cho việc tìm kiếm và truy xuất hình ảnh trở nên phức tạp và tốn nhiều thời gian.

Qua đó, Image retrieval giúp cho việc tìm kiếm và truy xuất hình ảnh trở nên dễ dàng và nhanh chóng hơn. Các hệ thống image retrieval thường sử dụng các kỹ thuật phân tích và đặc trưng hóa hình ảnh để mô tả và tìm kiếm các hình ảnh tương tự.

Image retrieval còn có thể sử dụng các phương pháp học máy để tìm kiếm và truy xuất hình ảnh. Các thuật toán học máy được sử dụng để tìm hiểu các đặc trưng của hình ảnh và tìm kiếm các hình ảnh có đặc trưng tương tự. Các thuật toán và kỹ thuật image retrieval phải đảm bảo rằng các hình ảnh được xử lý nhanh chóng và chính xác để đảm bảo hiệu quả của quá trình tìm kiếm.

2. Ý nghĩa khoa học:

Việc nghiên cứu image retrieval là một lĩnh vực nghiên cứu quan trọng trong khoa học máy tính và trí tuệ nhân tạo. Nó liên quan đến việc xử lý dữ liệu ảnh và tìm kiếm thông tin trong ảnh, đó là những vấn đề rất thú vị và phức tạp. Các nhà nghiên cứu trong lĩnh vực này đang phát triển các phương pháp mới để cải thiện hiệu quả và độ chính xác của hệ thống tìm kiếm ảnh. Họ cũng đang tìm hiểu cách để ứng dụng các công nghệ mới nhất như học sâu (deep learning) vào nghiên cứu của họ.

3. Ứng dụng thực tiễn

Với sự phát triển của công nghệ và internet, hình ảnh đã trở thành một phương tiện truyền thông phổ biến và quan trọng trong cuộc sống hàng ngày. Việc phát triển các công nghệ tìm kiếm ảnh sẽ giúp chúng ta tìm kiếm các hình ảnh cần thiết nhanh chóng và chính xác hơn. Điều này rất hữu ích cho các ứng dụng thực tế như: tìm kiếm sản phẩm trên các trang web thương mại điện tử, tìm kiếm thông

tin y học từ các hình ảnh y khoa, tìm kiếm thông tin về môi trường từ các hình ảnh địa lý v.v.

Ví dụ, trong lĩnh vực y tế, image retrieval có thể được sử dụng để tìm kiếm các hình ảnh y học để hỗ trợ chẩn đoán bệnh, giúp cho các chuyên gia y tế và bác sĩ có thể tiếp cận với các thông tin y học một cách nhanh chóng và hiệu quả.

Ngoài ra, image retrieval còn có thể được áp dụng trong lĩnh vực giáo dục, khi giáo viên và học sinh cần tìm kiếm và sử dụng các hình ảnh liên quan đến nội dung học tập. Trong lĩnh vực marketing, image retrieval cũng có thể được sử dụng để tìm kiếm các hình ảnh để sử dụng trong quảng cáo và truyền thông.

4. Thách thức

- Thách thức về dataset

Bài toán Image retrieval phải đối mặt với nhiều thách thức. Ví dụ, hình ảnh có thể đến từ nhiều nguồn khác nhau và có thể được chụp từ nhiều góc độ khác nhau. Hình ảnh cũng có thể bị nhiễu và mất chất lượng, gây ảnh hưởng đến kết quả truy xuất. Ngoài ra, bài toán image retrieval còn gặp khó khăn với vấn đề xử lý ảnh số lượng lớn và đa dạng.

Điều này đã đưa ra động lực cho nhiều nhà nghiên cứu trên thế giới để tìm kiếm các phương pháp và thuật toán mới để giải quyết các thách thức của bài toán image retrieval. Các kỹ thuật và thuật toán image retrieval được phát triển để giúp cho quá trình tìm kiếm và truy xuất hình ảnh trở nên nhanh chóng và hiệu quả hơn. Ngoài ra, việc phát triển các ứng dụng mới của image retrieval cũng đóng vai trò quan trọng trong động lực nghiên cứu của bài toán này.

- Cải thiện khả năng mở rộng truy xuất

Các hệ thống truy vấn ảnh trực tuyến ngày càng phát triển với số lượng lớn và tính đa dạng của bộ dữ liệu nên các hệ thống truy vấn hiện tại có thể không phù hợp. Các mạng học sâu hiện có được đào tạo cho việc phân loại hình ảnh nên việc trích xuất đặc trưng trở thành một thách thức trong bài toán truy vấn. Với các thách thức như vậy nên khả năng mở rộng truy xuất đối tượng trên các bộ dữ liệu tương đối kém.

- Khoảng cách ngữ nghĩa

Các phương pháp tìm kiếm phổ biến hiện nay gồm: tìm theo từ khóa (Text-based Image Retrieval), tìm theo nội dung (Content-based Image Retrieval) và tìm theo ngữ nghĩa (Semantic-based Image Retrieval). Trong đó, tìm kiếm bằng từ khóa (TBIR) là hệ thống truy xuất hình ảnh bằng các câu truy vấn từ các mô tả, chú thích hình ảnh được cung cấp bởi người dùng, đây cũng là hạn

chế của phương pháp TBIR. Tiếp theo là phương pháp tìm kiếm bằng nội dung (CBIR), đây là phương pháp truy xuất ảnh bằng cách trích xuất và so sánh các đặc trưng từ hình ảnh, phương pháp này được pháp triển để giải quyết hạn chế của phương pháp TBIR,nhưng phương pháp CBIR cũng có hạn chế là bị phụ thuộc vào kỹ thuật trích xuất đặc trưng. Để khắc phục hạn chế của phương pháp CBIR, phương pháp tìm kiếm bằng ngữ nghĩa (SBIR) được đề xuất, phương pháp SBIR truy xuất ảnh dựa trên ngữ nghĩa. Tuy nhiên, việc tạo ra ngữ nghĩa cao cho ảnh và giảm 'khoảng cách ngữ nghĩa' giữa đặc trưng cấp thấp và ngữ nghĩa cấp cao vẫn là một thách thức lớn trong cộng đồng truy vấn ảnh.

II. Phát biểu bài toán

Image Retrieval là một bài toán mà trong đó chúng ta sẽ thực hiện truyền, hoặc một hình ảnh query, hoặc văn bản mô tả hình ảnh, hoặc cả hai và sau quá trình xử lý, hệ thống sẽ trả kết quả là một tập các hình ảnh tương đồng với hình ảnh query theo thứ tự. Đây là một bài toán quan trọng trong lĩnh vực xử lý ảnh và máy học, có mục đích tìm kiếm các hình ảnh tương tự trong một cơ sở dữ liệu hình ảnh dựa trên một hình ảnh truy vấn.

Bài toán Image Retrieval có ứng dụng rộng rãi trong các lĩnh vực như tìm kiếm ảnh trực tuyến, quản lý thư viện ảnh, phân tích hình ảnh và video, nhận diện ảnh và video, … Một số ví dụ việc ứng dụng Image Retrieval phổ biến ta thường thấy là các công cụ tìm kiếm như Google Lens, Google hình ảnh, hệ thống đề xuất sản phẩm trong các ứng dụng mua bán, giao hàng như Gojek, Grab, Shopee, đặc biệt hơn Image Retrieval còn được áp dụng trong các hệ thống phát hiện, chẩn đoán bệnh, xác định các loại khối u, ...

1. Các bài toán thành phần của Image Retrieval

a) Trích xuất đặc trưng hình ảnh:

Việc trích xuất đặc trưng là một bước quan trọng trong quá trình tìm kiếm hình ảnh, vì nó giúp giảm chi phí tính toán và làm cho quá trình tìm kiếm trở nên nhanh hơn và hiệu quả hơn. Ngoài ra, việc trích xuất đặc trưng cũng giúp đơn giản hóa quá trình so sánh hình ảnh và tìm kiếm hình ảnh tương tự.

Việc trích xuất đặc trưng của hình ảnh thường được thực hiện bằng một số phương pháp trích xuất đặc trưng phổ biến như SIFT (Scale-Invariant Feature Transform), SURF (Speeded Up Robust Features), HoG (Histogram of Oriented Gradients) hay sử dụng mạng học máy để trích xuất đặc trưng, điển hình là CNN (Convolutional Neural Networks)

b) Đánh chỉ mục cho hình ảnh:

- Trong bài toán Image Retrieval, đánh chỉ mục hình ảnh là một bước quan trọng để xây dựng cơ sở dữ liệu hình ảnh hiệu quả. Bước này giúp chúng ta gán nhãn (label) cho các hình ảnh trong cơ sở dữ liệu dựa trên nội dung của chúng.

- Đánh chỉ mục hình ảnh có thể được thực hiện bằng các hàm hashing sẽ trình bày ở phần sau.

c) Tìm kiếm hình ảnh:

- Bài toán tìm kiếm hình ảnh trong Image Retrieval là bài toán giúp tìm kiếm các hình ảnh trong cơ sở dữ liệu có độ tương đồng lớn so với ảnh query thông qua việc tính toán độ đo tương đồng hoặc độ đo khác biệt giữa chỉ mục từng ảnh trong cơ sở dữ liệu và ảnh query. Sau khi tìm được các ảnh tương đồng, ta sẽ thực hiện xếp hạng các ảnh đó dựa trên giá trị độ đo.

- Các độ đo chúng ta có thể áp dụng trong việc tìm kiếm hình ảnh khá đa dạng, ví dụ như là độ đo Euclidean, khoảng cách Manhattan, độ tương đồng Hamming code, …

2. Các thành phần dữ liệu của bài toán Image Retrieval:

a) Cơ sở dữ liệu:

Là một tập hợp các hình ảnh nhiều đối tượng khác nhau thuộc nhiều lĩnh vực, chủ đề khác nhau như ảnh thời trang, ảnh động thực vật, ảnh phong cảnh, ảnh y khoa … được lưu trữ để sử dụng trong truy xuất, tìm kiếm hình ảnh tuần tự. Xây dựng cơ sở dữ liệu là một bước quan trọng trong việc giải quyết bài toán Image Retrieval, đảm bảo cho hệ thống Image Retrieval có thể hoạt động hiệu quả và chính xác nhất có thể. Để có thể xây dựng cơ sở dữ liệu hoàn chỉnh, chúng ta cần phải trải qua các bước: Thu thập hình ảnh, hoặc từ internet hoặc các nguồn ảnh nội bộ hoặc những nguồn khác hoặc tự thu thập thủ công; Trích xuất đặc trưng: thông qua các phần mềm chương trình để có thể rút trích đặc trưng từ trong máy và lưu trữ dưới dạng vector số học; Gắn nhãn và phân loại thường được thực hiện bởi các mô hình máy học, mục đích là khiến cho việc tìm kiếm, truy xuất hình ảnh trở nên dễ dàng hơn; Lưu trữ và bảo quản dữ liệu – phải đảm bảo việc truy cập, tìm kiếm trong cơ sở dữ liệu phải nhanh chóng, hiệu quả song song với việc bảo mật và toàn vẹn thông tin; và cuối cùng là kiểm tra và đánh giá nhằm xác định mức độ đáp ứng yêu cầu của cơ sở dữ liệu để có thể thực hiện các phương pháp nhằm hoàn thiện cơ sở dữ liệu tốt hơn.

b) Input:

Dữ liệu đầu vào của bài toán Image Retrieval thường có 3 dạng: query image, văn bản mô tả hoặc là cả query image và văn bản mô tả.

Query Image: là hình ảnh được truyền vào để thực hiện truy vấn các hình ảnh tương tự nó trong cơ sở dữ liệu đã có. Query image có thể được truyền vào bằng cách thủ công bởi người dùng hoặc được tự động truyền vào bởi chương trình hệ thống. Query image thường được truyền vào trong hệ thống tìm kiếm bằng phương pháp Content-based Image Retrieval và Semantic-based Image Retrieval.

Văn bản mô tả: là những mô tả cơ bản về các hình ảnh cần tìm kiếm, thường là những câu miêu tả về hình ảnh, các câu bình luận, tiêu đề hay các từ khóa liên quan đến hình ảnh cần tìm kiếm. Văn bản mô tả thường được dùng để tìm kiếm hình ảnh bằng phương pháp Word-based Image Retrieval và Semantic-based Image Retrieval.

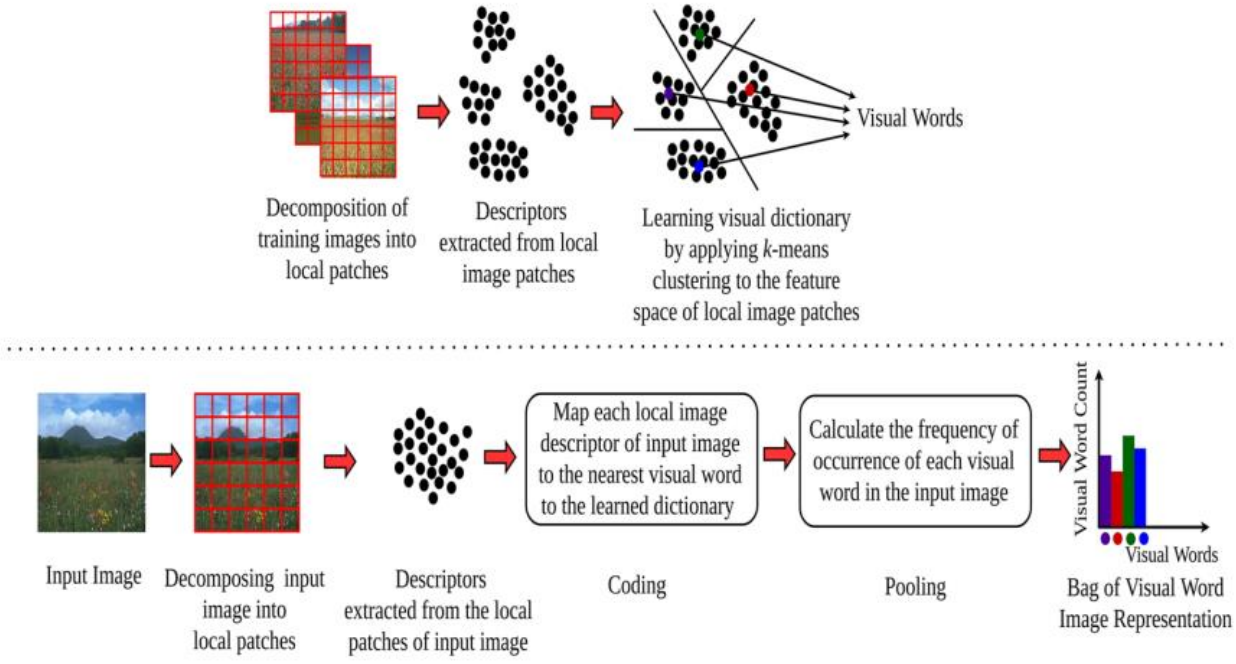
c) Output:

Tập hợp các hình ảnh từ cơ sở dữ liệu tương đồng với query image nhất hoặc gần với mô tả văn bản nhất. Sự tương đồng giữa hình ảnh query thường được tính toán bởi độ đo tương đồng hoặc độ đo khác biệt. Các hình ảnh này thường được sắp xếp dựa trên mức độ tương đồng giữa chúng và input, độ tương đồng càng lớn thì sẽ được sắp xếp ở vị trí cao nhất.

III. Những phương pháp truyền thống trong bài toán IR

1. Content-based image retrieval

Phương pháp Bag of Visual Words:



Hình ảnh trên là sơ đồ thực hiện của phương pháp BoVW.

Bag of Visual Words (BoVW) là một trong những phương pháp truyền thống trong bài toán image retrieval (cụ thể hơn là CBIR). Đây là một phương pháp dựa trên đặc trưng và áp dụng kỹ thuật phân tích nhóm từ (clustering) để biểu diễn hình ảnh.

Các bước thực hiện của phương pháp BoVW gồm:

Decomposing input image into local patches:

Giai đoạn này đề cập đến việc chia các hình ảnh ban đầu thành nhiều phần (patches) nhỏ hơn để tiện cho việc trích xuất đặc trưng và xây dựng histogram BoVW. Việc chia nhỏ các hình ảnh này có thể được thực hiện bằng cách sử dụng các phương pháp như:

- Sliding window Phương pháp này sử dụng một cửa sổ trượt trên toàn bộ ảnh để chia nó thành các phần bằng nhau. Việc chọn kích thước và bước trượt phù hợp là quan trọng để đảm bảo rằng tất cả các vùng hình ảnh quan trọng đều được bao phủ.
- Random sampling: Random Sampling: Phương pháp này chọn ngẫu nhiên một số lượng các vùng hình ảnh trên toàn bộ ảnh để trích xuất đặc trưng. Việc chọn số lượng và vị trí của các vùng này có thể được điều chỉnh để đảm bảo rằng không bỏ sót bất kỳ đặc trưng nào quan trọng.

Việc chia nhỏ hình ảnh thành các phần nhỏ hơn giúp giảm chi phí tính toán cho việc trích xuất đặc trưng, đồng thời tăng tính cục bộ (locality) của các đặc trưng được trích xuất, giúp mô tả nội dung của hình ảnh tốt hơn. Tuy nhiên, việc chia nhỏ hình ảnh cũng có thể dẫn đến mất mát thông tin về mối quan hệ giữa các đặc trưng toàn cục (global features) của hình ảnh. Do đó, việc chọn kích thước và phương pháp chia nhỏ hình ảnh phù hợp là rất quan trọng trong phương pháp BOVW để đảm bảo độ chính xác và hiệu quả của kết quả tìm kiếm ảnh.

Feature extraction

Trích xuất các đặc trưng của hình ảnh: Đây là bước quan trọng nhất trong phương pháp BoVW. Các đặc trưng này là các local feature như các cạnh, màu sắc, hình dạng, v.v. Các đặc trưng này được trích xuất từ hình ảnh bằng các phương pháp như SIFT, SURF, HoG, v.v.

Clustering

Bước này của mô hình BoVWxác định K yếu tố đại diện được gọi là visual word bằng cách áp dụng K-Means clustering cho một tập gồm N (N > K) vector đặc trưng cục bộ.

Building visual words cho hình ảnh (INDEXING):

Ở giai đoạn trước đó, tổng số mô tả đặc trưng được trích xuất là rất lớn. Để giải quyết vấn đề này, các mô tả đặc trưng được phân cụm bằng cách áp dụng thuật toán KMeans, để tạo ra một visual word. Mỗi cụm được coi là một từ hình ảnh khác nhau trong từ vựng, được biểu diễn bởi điểm trung tâm của cụm đó.

Đầu tiên, BoVW thường định nghĩa tập dữ liệu huấn luyện bao gồm các hình ảnh được biểu diễn bằng V = v_1, v_2, ..., v_n, trong đó v là các đặc trưng hình ảnh được trích xuất. Sau đó, sử dụng thuật toán phân cụm như KMeans, chia không gian bộ mô tả thành K vùng không trùng lắp với trọng tâm là W được biểu diễn bằng W = w_1, w_2, ..., w_K, trong đó K là số cụm. Với một v_i, đặt $h_i(j)$ biểu thị giá trị thứ j của vector mã hóa nhị phân h_i và nó chỉ định visual word (từ trực quan) w_j mà v_i được đánh chỉ mục. Sau đó, thuật toán K-means cực tiểu hóa một hàm chi phí có dạng:

$$C = \sum_{i=1}^{n} \sum_{j=1}^{k} h_i(j) \left\| v_i - w_j \right\|^2$$

Note: Hàm chi phí ở biểu thức trên là tổng bình phương khoảng cách của mỗi vector mô tả v_i với từ trực quan gần nhất của nó w_j.

Để tối thiểu hóa hàm chi phí C sử dụng cách tiếp cận two-step iterative.

Bước đầu tiên: tối thiểu hóa hàm chi phí C theo $h_i(j)$ bằng cách giữ cố định giá trị của các visual words. Bước thứ hai: tối thiểu hóa C theo visual words bằng cách giữ giá trị của $h_i(j)$ cố định.

Đầu tiên, xem xét việc ước tính giá trị của $h_i(j)$. Vì hàm C là tuyến tính theo $h_i(j)$, nên một closed form solution có thể thực hiện được. Giá trị của $h_i(j)$ cho mỗi đặc trưng cục bộ v_i có thể được tính toán bằng cách gán giá trị của $h_i(j)$ bằng 1 cho visual work w_j để cho ra được chi phí $\left\| v_i - w_j \right\|^2$ nhỏ nhất. Có thể định nghĩa như sau:

$$h_i(j) = \begin{cases} 1, & if\ j = \arg\min_j \left\| v_i - w_j \right\|^2 \\ 0, & otherwise \end{cases}$$

Tiếp theo, xem xét việc tối ưu hóa visual words với giá trị $h_i(j)$ cố đinh. Hàm chi phí C có bản chất là bậc 2 và các visual words giúp giảm thiểu hàm chi phí C bằng cách cho đạo hàm của C theo w_j bằng 0. Có thể định nghĩa như sau:

$$C' = 2\sum_{i=1}^{N} h_i(j)(v_i - w_j) = 0$$

Từ đó, bộ visual words có thể được tính như sau:

$$w_j = \frac{\Sigma_{i=1}^{N} h_i(j) v_i}{\Sigma_{i=1}^{N} h_i(j)}$$

Toàn bộ quy trình được lặp lại cho đến khi giá trị hàm chi phí hội tụ (nghĩa là chênh lệch trong giá trị hàm C từ bước này sang bước khác nhỏ hơn một số ε được xác định trước) hoặc nó đạt đến số lần lặp mong muốn.

Coding:

Khi một visual dictionary đã được xây dựng, bước tiếp theo là mã hóa các visual features đã được trích xuất từ các vùng cục bộ của một hình ảnh theo visual words đã học. Đối với một bộ mô tả hình ảnh cục bộ v_i, vector mã hóa h_i tương ứng có K chiều theo visual dictionary W được xác định bằng cách giải bài toán tối ưu sau đây:

$$\arg\min_{h_i} \|v_i - Wh_i\|^2$$

$$s.t\ \|h_i\|_0 = 1, \|h_i\|_1 = 1$$

Trong đó $\|h_i\|_0$ là ràng buộc 0-sparsity giới hạn các phần tử khác 0 trong vector mã hóa h_i. Tương tự $\|h_i\|_1$ là ràng buộc 1-sparsity và nó tính tổng của tất cả các hệ số trong mỗi vector mã hóa h_i. Kết quả là, chỉ có một phần tử trong mỗi vector mã hóa được active tại một thời điểm và phần tử khác 0 này được tìm thấy bằng cách xác định visual words gần nhất (w_i) trong directory tương ứng với vector đặc trưng cục bộ (v_i) đã cho trước. Thuật toán ánh xạ các bộ mô tả hình ảnh tới các visual words tương ứng trong visual dictionary được gọi là một encoder (bộ mã hóa).

Pooling:

Bước pooling tổng hợp các vector mã hóa liên quan đến bộ mô tả cục bộ trong một hình ảnh để tạo ra một biểu diễn toàn cục. Trong sum pooling, tần suất xuất hiện của các visual word riêng lẻ trong một hình ảnh được gộp lại với nhau để tạo thành một mô tả duy nhất. Điều này dẫn đến một histogram visual word cho từng ảnh trong cơ sở dữ liệu và được biến đến phổ biến như là biểu diễn hình ảnh dựa trên túi các visual word (bag of visual word). Đối vơi một hình ảnh cho trước, thành phần thứ j của biểu diễn BoVW B = [b_1, b_2, ..., b_k] có thể được tính toán như sau:

$$b_j = \sum_{i=1}^{n} h_{ij}$$

Trong đó N là số mô tả cục bộ trong hình ảnh cho trước, h_i là vector mã hóa tương ứng với mô tả thứ i của bộ v_i và b là kết quả sau bước pooling.

Kết quả tổng hợp B được chuẩn hóa để làm cho biểu diễn này bất biến đối với số lượng mô mả cục bộ được trích xuất từ các hình ảnh đã cho. L2-norm được sử dụng làm kỹ thuật chuẩn hóa cho mục đích truy xuất hình ảnh và được tính toán như sau:

$$B = \frac{B}{\sqrt{\sum_{i=1}^{N} {b_i}^2}}$$

Image comparision:

Sau khi biểu diễn dựa trên BoVW, hình ảnh query và từng hình ảnh trong cơ sở dữ liệu, chúng phải được so sánh để xác định mức độ tương đồng giữa các hình ảnh. Các hàm khoảng cách là phương pháp đơn giản và được sử dụng rộng rãi để đánh giá sự tương đồng giữa các biểu diễn dựa trên BoVW. Bằng cách sử dụng một hàm khoảng cách phù hợp, một danh sách sắp xếp được tạo ra theo thứ tự tăng dần của khoảng cách tính toán giữa biểu diễn dựa trên BoVW được tạo ra từ các hình ảnh trong cơ sở dữ liệu và hình ảnh query cho trước.

2. Sơ lược về semantic based image retrieval

Input: Ảnh đầu vào hoặc câu mô tả ảnh.

Output: Các ảnh tương tự với ảnh đầu vào hoặc câu mô tả ảnh theo độ tương đồng ngữ nghĩa, bao gồm các đối tượng, vật phẩm, hoạt động, v.v. có trong ảnh hoặc câu mô tả.

Phương pháp này thường sử dụng các mô hình học sâu để biểu diễn các đặc trưng của ảnh và sử dụng các thông tin ngữ nghĩa để phân loại và tìm kiếm các ảnh có nội dung tương tự.

Các bước để triển khai hệ thống tìm kiếm ảnh dựa trên nội dung ngữ nghĩa có thể được mô tả như sau:

- Xây dựng cơ sở dữ liệu ảnh: Trước hết, chúng ta cần xây dựng một cơ sở dữ liệu ảnh với các ảnh đã được gán nhãn và phân loại. Các ảnh này có thể được lấy từ các nguồn khác nhau như bộ dữ liệu ImageNet, Flickr, hoặc ảnh từ website.
- Biểu diễn ảnh dưới dạng vector đặc trưng: Để biểu diễn các đặc trưng của ảnh, chúng ta cần sử dụng các mô hình học sâu như Convolutional Neural Networks (CNN) để trích xuất các đặc trưng từ ảnh. Các đặc trưng này có thể được biểu diễn dưới dạng các vector đặc trưng, ví dụ như vector đặc trưng của ảnh được huấn luyện bởi mô hình ResNet50.
- Phân loại các đối tượng trong ảnh: Chúng ta có thể sử dụng các mô hình như Object Detection để phát hiện và phân loại các đối tượng trong ảnh, từ đó cung

cấp thêm thông tin ngữ nghĩa về các đối tượng trong ảnh. Ví dụ, chúng ta có thể sử dụng mô hình YOLO để phân loại các đối tượng trong ảnh.

- Tính toán độ tương đồng giữa các ảnh: Để tìm kiếm các ảnh có nội dung tương tự, chúng ta có thể sử dụng các phương pháp tính toán độ tương đồng giữa các vector đặc trưng của các ảnh. Các phương pháp này có thể bao gồm cosine similarity, Euclidean distance hoặc độ tương tự cos.

IV. Content-based image retrieval using deep learning

Content-based image retrieval (CBIR) là một trong những nghiên cứu rộng rãi trong lĩnh vực thị giác máy tính. CBIR nhằm tìm kiếm hình ảnh thông qua phân tích nội dung hình ảnh của chúng, do đó việc image representation là điểm trọng yếu của CBIR. Trong những thập kỷ qua, đã có nhiều low-level feature descriptor được đề xuất cho image representation , từ các đặc trưng toàn cục, chẳng hạn như đặc trưng màu sắc , đặc trưng cạnh, đặc trưng cấu trúc, đến các biểu diễn đặc trưng cục bộ gần đây, chẳng hạn như BoVW sử dụng các mô tả đặc trưng cục bộ (ví dụ như SIFT và SURF, vv.) đã nêu ở phần trên. Các phương pháp CBIR truyền thống thường chọn các hàm khoảng cách cứng trên một số đặc trưng cấp thấp đã được trích xuất và tìm kiếm thông qua độ đo, chẳng hạn như khoảng cách Euclide hoặc độ tương đồng Cosin.

1. Deep learning

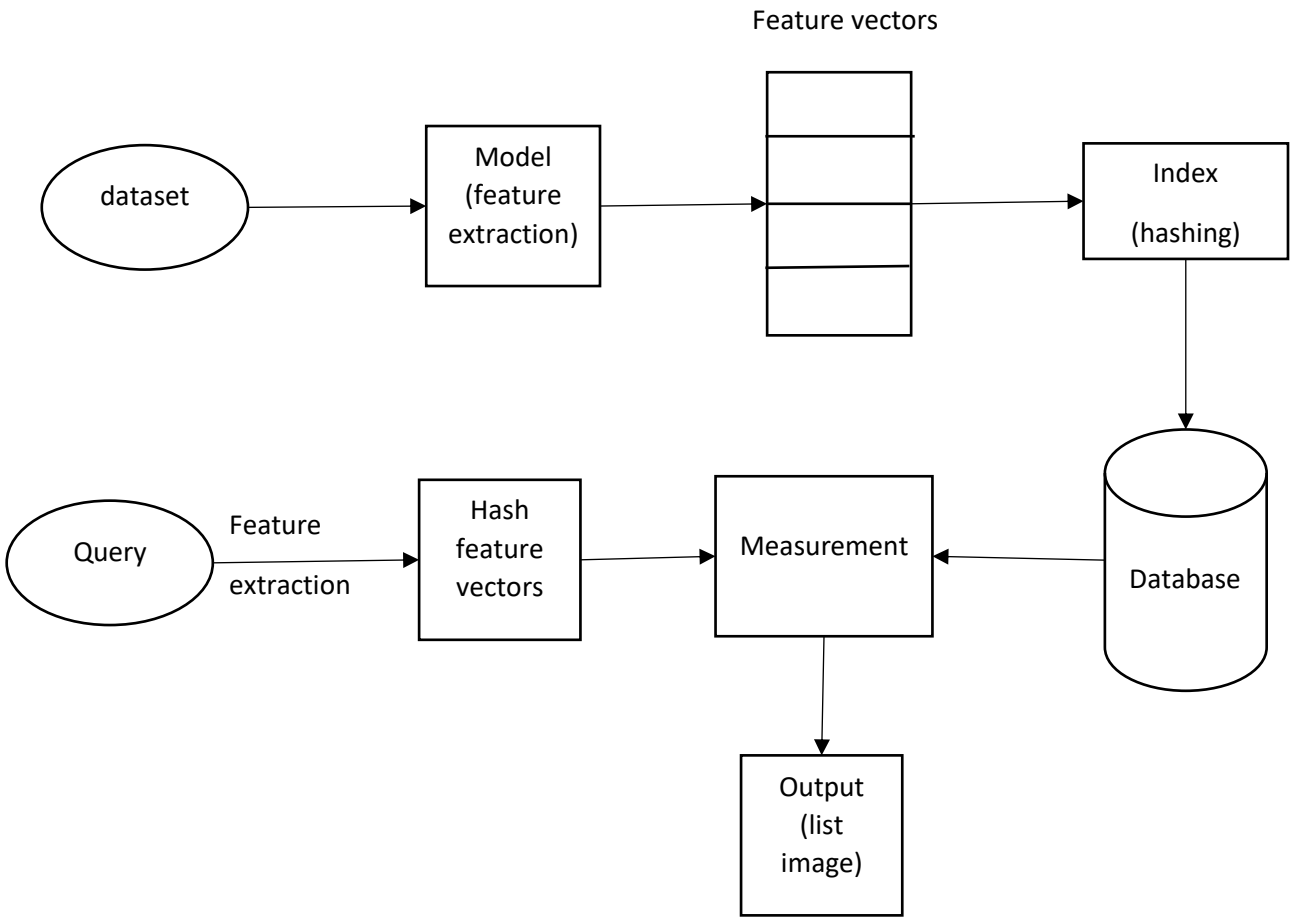
Deep learning đề cập đến một class các kỹ thuật học máy (machine learning techniques), trong đó nhiều layer của các giai đoạn xử lý thông tin trong hierarchical architectures được khai thác cho pattern classification và feature hoặc representation learning. Nó nằm ở nhiều lĩnh vực nghiên cứu, bao gồm mạng nơ-ron, graphical modeling, tối ưu hóa, nhận dạng mẫu và xử lý tín hiệu, v.v.

Ý tưởng cơ bản của Deep learning bắt nguồn từ nghiên cứu về mạng neural nhân tạo. Các mạng neural feed-forward với hidden layer là một ví dụ về các mô hình có kiến trúc sâu (deep architecture). Thuật toán lan truyền ngược (Back-propagation), phổ biến trong thập niên 1980, đã trở thành một thuật toán nổi tiếng để học trọng số của các mạng này. Ví dụ, LeCun et al. đã thành công trong việc áp dụng mạng nơ-ron tích chập (CNN) sâu và được giám sát bằng lan truyền ngược để nhận dạng chữ số. Gần đây, nó trở thành một chủ đề nóng trong cả thị giác máy tính và học máy, trong đó các kỹ thuật học sâu đạt được hiệu suất tốt nhất cho nhiều nhiệm vụ khác nhau. Mạng nơ-ron tích chập (CNN) sâu được đề xuất đã đạt được vị trí đầu tiên trong nhiệm vụ phân loại ảnh của ILSVRC-2012 1. Mô hình được huấn luyện trên hơn một triệu ảnh và đạt được tỷ lệ lỗi top-5 thấp nhất là 15,3% trên 1.000 classes. Sau đó, một số nghiên cứu gần đây đã đạt được kết quả tốt hơn bằng cách cải tiến các mô hình CNN. Tỷ lệ lỗi top-5 đã giảm xuống 13,24% bằng cách huấn luyện mô hình một cách **đồng thời** vừa phân loại, định vị và phát hiện đối tượng.

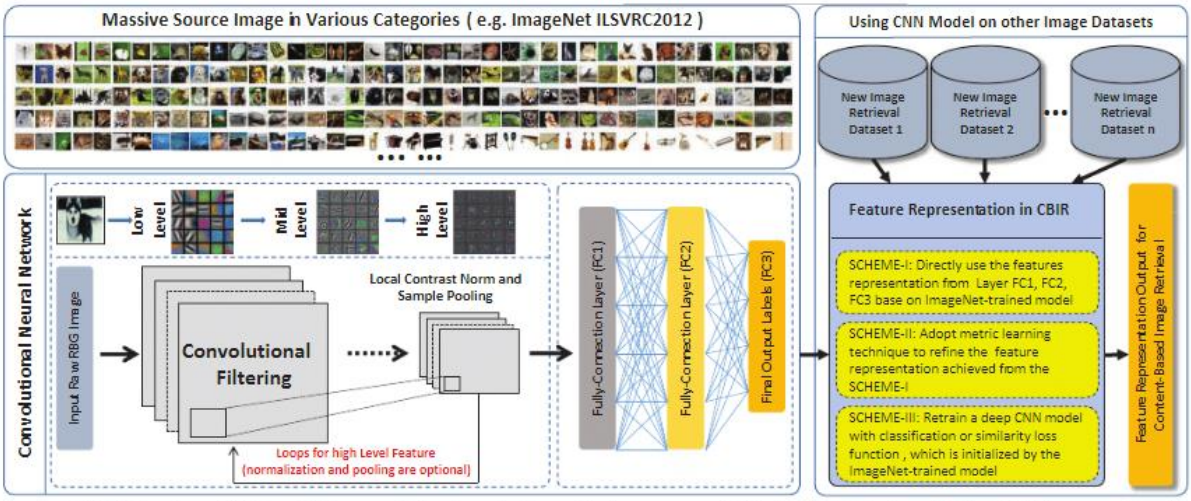
Trong những năm gần đây, kỹ thuật học sâu đã được đề xuất và nghiên cứu rộng rãi, ví dụ như Deep Belief Network (DBN), Boltzmann Machines (BM), Restricted Boltzmann Machines (RBM), Deep Boltzmann Machine (DBM), Deep Neural Networks (DNN), vv. Trong số các kỹ thuật khác nhau, mô hình mạng neural tích chập sâu, một kiến trúc sâu phân biệt và thuộc loại DNN, đã đạt được hiệu suất tốt nhất trên các tác vụ và cuộc thi trong thị giác máy tính và nhận dạng hình ảnh. Cụ thể, mô hình CNN bao gồm nhiều lớp tích chập và pooling, được xếp chồng lên nhau. Lớp tích chập chia sẻ nhiều trọng số và lớp pooling lấy mẫu con của đầu ra của layer tích chập và giảm rate dữ liệu từ layer bên dưới. Việc chia sẻ trọng số trong lớp tích chập, cùng với pooling được lựa chọn một cách thích hợp, trang bị cho CNN một số tính chất "bất biến" (ví dụ: translation invariance).

2. Deep learning for CBIR

Work flow:



2.1 Framework



(a) Training deep CNN models in an existing domain (ImageNet) (b) Adopting trained model for CBIR in a new domain

Figure 1: A Framework of Deep Learning with Application to Content-based Image Retrieval.

Framework deep learning thực hiện feature extraction được đề xuất cho CBIR, bao gồm hai giai đoạn:

(i) huấn luyện một mô hình học sâu từ một bộ sưu tập lớn dữ liệu huấn luyện;

(ii) áp dụng mô hình học sâu đã được huấn luyện để học các biểu diễn đặc trưng cho các tác vụ CBIR trong một miền dữ liệu mới.

Hình 1 cung cấp một cái nhìn tổng quan về framework deep learning cho CBIR. Đối với việc triển khai học sâu CNNs. Mô hình này đã được huấn luyện thành công trên tập dữ liệu "ILSVRC-2012" từ ImageNet và được tìm thấy có hiệu suất tốt nhất với 1.000 categories và hơn 1 triệu hình ảnh huấn luyện.

Nói chung, mạng tích chập sâu, như được thể hiện trong Hình 1(a), bao gồm hai phần chính:

1) các lớp tích chập và lớp max-pooling

2) fully connected layer và output layer.

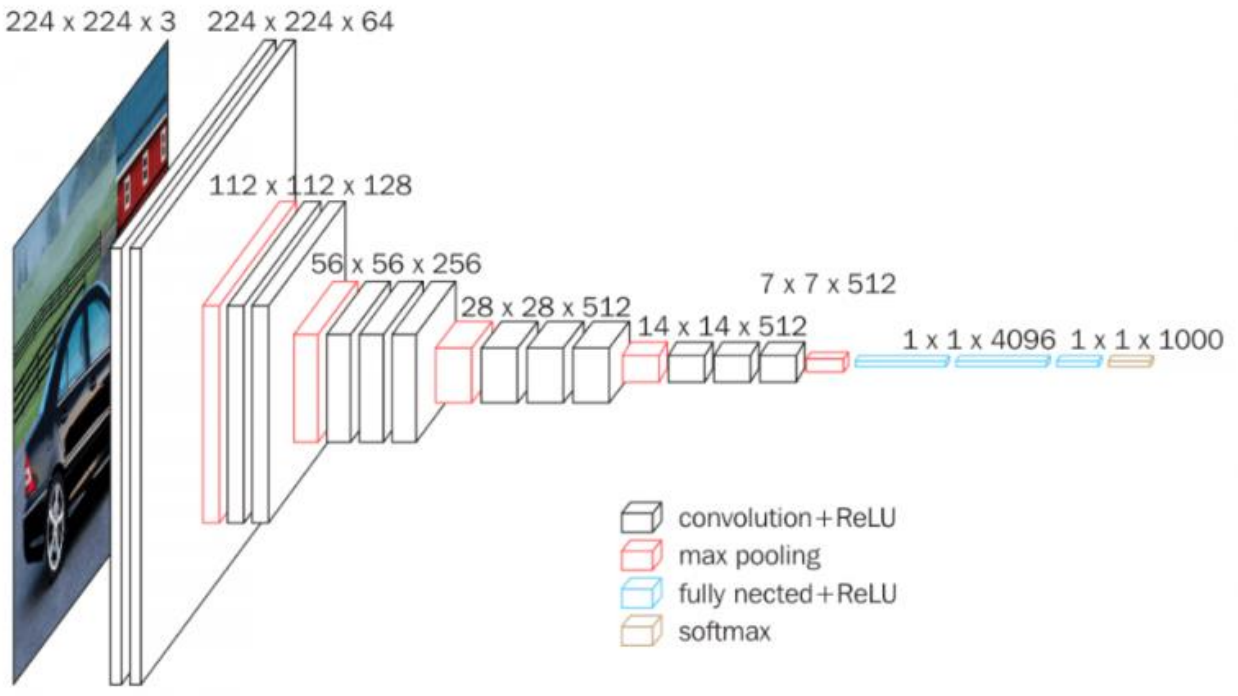
Cụ thể, lớp đầu tiên là lớp đầu vào, sử dụng các mean-centered raw RGB pixel trong intensity value. Để giảm thiểu overfiting, hai kỹ thuật augmentation dữ liệu được thực hiện: đầu tiên, các hình ảnh đầu vào được tạo ra bằng cách translation và horizontal reflections bằng cách trích xuất random 224 × 224 patches từ 256 × 256 image và mạng sẽ được huấn luyện trên các patches được trích xuất này; thứ hai, để capture được tính không đổi (invariance) trong chiếu sáng (illumination) và màu sắc, they add random multiples of the principle components of the RGB pixel values throughout the dataset.

Sau các lớp tích chập là hai lớp fully connected với 4.096 neuron, được gọi là "FC1" và "FC2". Lớp đầu ra cuối cùng, được cung cấp bởi lớp "FC2", là một lớp softmax 1000 chiều, tạo ra một phân bố trên 1000 nhãn lớp trong ImageNet. Trong toàn bộ mạng nơ-ron tích chập sâu, có khoảng 60 triệu tham số tổng cộng. Train mạng nơ-ron tích chập sâu dựa trên bộ dữ liệu huấn luyện ILSVRC-2012 của ImageNet, chứa khoảng 1,2 triệu hình ảnh. Việc huấn luyện một mô hình với tỷ lệ lỗi 0,424 trên testing set (50.000 hình ảnh) mất khoảng 200 giờ, gần với tỷ lệ lỗi 0,407.

Cấu trúc phân cấp và tham số hóa đầy đủ của DCNNs đã đưa đến thành công của chúng trong nhiều tác vụ thị giác máy tính đáng kể. Đối với việc truy xuất hình ảnh, có một số mô hình phổ biến được sử dụng để trích xuất đặc trưng, bao gồm VGG, GoogLeNet, ResNet và NetVLAD.

2.2. Các model thường dùng cho feature extraction:

a. VGGNet



Được truyền cảm hứng bởi AlexNet, VGGNet có hai phiên bản được sử dụng rộng rãi: VGG-16 và VGG-19. Là một mô hình CNN, VGG Net có kiến trúc bao gồm nhiều lớp convolutional layer và pooling layer. Các lớp này được xếp chồng lên nhau để tạo thành một kiến trúc sâu. Mô hình này được sử dụng để trích xuất đặc trưng từ hình ảnh, giúp phân loại và tìm kiếm ảnh dễ dàng hơn.

Khi đưa hình ảnh vào mô hình, VGG Net sử dụng các bộ lọc để tìm kiếm các đặc trưng của hình ảnh. Bộ lọc trong mô hình VGG Net được biểu diễn dưới dạng một ma

trận 2 chiều (3x3 hoặc 1x1) có các giá trị thực (đây là kích thước nhỏmà vẫn đảm bảo tính chính xác trong việc phát hiện các đặc trưng trên hình ảnh).Mỗi bộ lọc này được áp dụng trên từng phần trong ảnh bằng phép tích chập (convolution) để tìm kiếm các đặc trưng của hình ảnh. Mỗi lớp convolutional trong mô hình VGG Net sử dụng nhiều bộ lọc khác nhau để tìm kiếm các đặc trưng khác nhau của hình ảnh. Các bộ lọc này được đào tạo để phát hiện các đặc trưng cụ thể của hình ảnh như đường nét, góc cạnh, hoặc các khu vực tối sáng khác nhau.

Quá trình tích chập diễn ra như sau:

Đầu tiên, bộ lọc được di chuyển qua toàn bộ bức ảnh theo từng bước (stride) để tính tích chập trên từng vùng của ảnh.

Tại mỗi vùng trên ảnh, tích chập được tính bằng cách nhân từng giá trị trong ma trận bộ lọc với các giá trị tương ứng trong vùng đó trên ảnh, sau đó cộng tất cả các kết quả lại với nhau.

Kết quả của phép tích chập được lưu trữ trong một ma trận mới gọi là bản đồ đặc trưng (feature map).

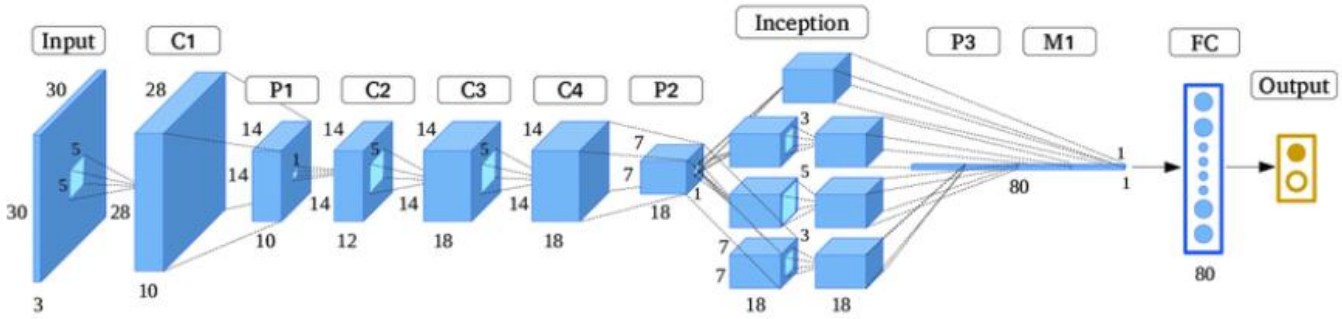
Khi áp dụng nhiều bộ lọc khác nhau lên ảnh, các bản đồ đặc trưng sẽ bao gồm nhiều thông tin về các đặc trưng của ảnh, ví dụ như các cạnh, nét, hay các đối tượng khác trong ảnh.

Sau đó, kết quả feature map từ lớp convolutional được truyền qua lớp pooling để giảm kích thước của hình ảnh và giữ lại những đặc trưng quan trọng.

Các tham số của mô hình VGG Net được đào tạo bằng phương pháp lan truyền ngược (backpropagation), trong đó mô hình sử dụng gradient descent để điều chỉnh các trọng số của lớp convolutional và fully connected layer.

Khi áp dụng mô hình VGG Net vào tác vụ content-based image retrieval, ta đưa hình ảnh vào mô hình để trích xuất đặc trưng. Sau đó, các đặc trưng này sẽ được sử dụng để so sánh với các đặc trưng của các hình ảnh khác và tìm ra những hình ảnh tương đồng nhất.

b. GoogleNet



GoogleNet architecture

GoogleNet, hay còn được gọi là Inception v1, là một trong những mô hình CNN nổi tiếng được phát triển bởi Google. Mô hình này có cấu trúc rất đặc biệt với nhiều lớp tích chập song song chồng lên nhau để tăng hiệu suất tính toán và độ chính xác.

GoogleNet sử dụng một kiến trúc được gọi là Inception module, trong đó có nhiều lớp tích chập có kích thước khác nhau được kết hợp với nhau để tìm ra đặc trưng của ảnh. Bên cạnh đó, mô hình còn sử dụng kỹ thuật global average pooling để giảm kích thước của đầu ra trước khi đưa vào lớp fully connected.

Trong content based image retrieval, GoogleNet cũng được sử dụng để trích xuất đặc trưng của ảnh và đưa vào một mô hình máy học để tìm kiếm các hình ảnh tương đồng. Các thử nghiệm đã chứng minh rằng GoogleNet có khả năng tìm kiếm ảnh tốt và độ chính xác cao.

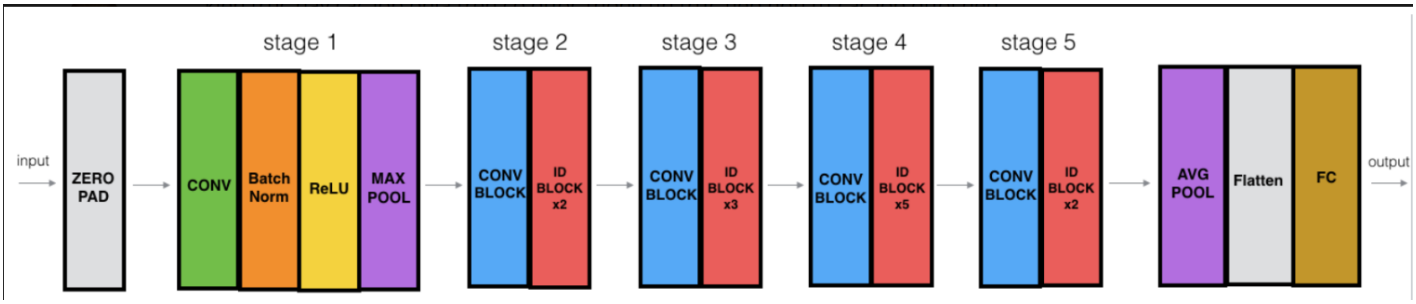
Kiến trúc của GoogleNet, hay còn được gọi là Inception v1, bao gồm một chuỗi các module Inception. Mỗi module Inception gồm 4 nhánh, trong đó có một tập hợp các bộ lọc tích chập khác nhau (1x1, 3x3, 5x5) được kết hợp song song và cuối cùng được ghép lại thành một đầu ra duy nhất (đặc trưng) cho mỗi module. So với AlexNet và VGGNet, GoogLeNet sâu hơn và rộng hơn nhưng có ít tham số hơn trong 22 lớp của nó, dẫn đến hiệu quả học tập cao hơn. Kiến trúc sâu hơn của GoogleNet có lợi cho việc học các đặc trưng trừu tượng cấp cao để giảm khoảng cách ngữ nghĩa.

Để giảm số lượng trọng số, GoogleNet sử dụng các bộ lọc 1x1 để thực hiện phép chiếu xuống chiều sâu, giúp giảm số lượng kênh đặc trưng và tăng tốc tính toán.

Ngoài ra, GoogleNet cũng sử dụng một số kỹ thuật khác như kết hợp giữa dropout và L2 regularization để tránh overfitting, sử dụng global average pooling thay vì fully connected layer ở cuối cùng để giảm số lượng trọng số và tính toán, và sử dụng auxiliary classifiers ở các layer trung gian để giúp giảm gradient vanishing và tăng tốc độ hội tụ.

GoogleNet là một trong những mô hình CNN hiệu quả nhất và đã được ứng dụng trong nhiều bài toán, trong đó có content-based image retrieval.

c. Resnet



Cuối cùng, ResNet được phát triển bằng cách thêm nhiều lớp tích chập để trích xuất đặc trưng trừu tượng hơn. Kết nối trượt được thêm giữa các lớp tích chập để giải quyết vấn đề độ dốc bị mất tích khi huấn luyện mạng này.

Mạng ResNet (Residual Network) là một trong những kiến trúc CNN phổ biến nhất hiện nay. Nó được sử dụng để trích xuất đặc trưng ảnh trong tác vụ content-based image retrieval.

Tương tự như VGGNet, ResNet cũng sử dụng các lớp tích chập để trích xuất các đặc trưng của ảnh. Tuy nhiên, ResNet sử dụng một kiến trúc đặc biệt có tên là "residual block" để giúp việc huấn luyện mô hình dễ dàng hơn và tránh hiện tượng "vanishing gradient".

Sau khi trích xuất đặc trưng của ảnh bằng ResNet, ta có thể sử dụng các kỹ thuật vector hóa như PCA, t-SNE, hoặc LDA để giảm chiều dữ liệu và biểu diễn ảnh dưới dạng một vector đặc trưng. Vector này sau đó có thể được sử dụng để so sánh và tìm kiếm các ảnh tương đồng trong tác vụ content-based image retrieval.

Residual block trong ResNet là một kiến trúc đặc biệt của các lớp tích chập (convolutional layers) được sử dụng để xây dựng mạng ResNet. Mỗi residual block bao gồm hai nhánh song song, một là một chuỗi các lớp tích chập (convolutional layers) và một là một hàm đường cong đơn giản để ánh xạ đầu vào trực tiếp vào đầu ra. Sự xuất hiện của hàm đường cong này cho phép việc học được một hàm đồng nhất cho mỗi residual block, mà đây là một trong những yếu tố quan trọng giúp giảm sự phân tán thông tin trong quá trình lan truyền ngược và làm cho quá trình học mạng nhanh hơn và hiệu quả hơn.

Một residual block được định nghĩa như sau:

Input

↓

Convolution

↓

Batch Normalization

↓

ReLU Activation

↓

Convolution

↓

Batch Normalization

↓

Skip Connection

↓

ReLU Activation

↓

Output

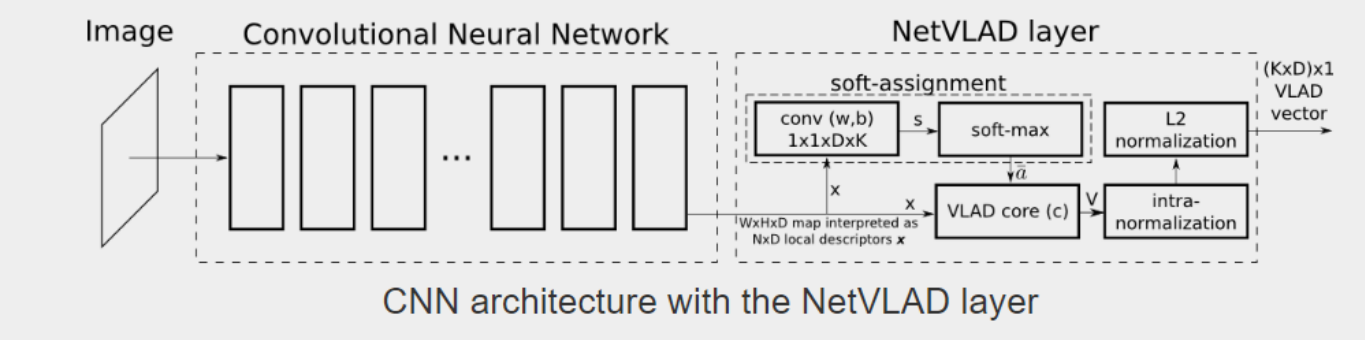
Trong đó, Skip Connection là kết nối trực tiếp từ đầu vào của residual block tới đầu ra của nó, với mục đích tránh mất mát thông tin khi ánh xạ đầu vào tới đầu ra.

Ta có thể hiểu nó như một "shortcut" hoặc "bypass connection" trong quá trình học. Thay vì truyền dữ liệu thẳng từ lớp này sang lớp khác, ResNet sử dụng một đường dẫn ngắn (shortcut) để truyền dữ liệu trực tiếp từ lớp này sang lớp khác, nằm ở khoảng cách nhiều lớp so với lớp hiện tại.

Việc này giúp cho việc huấn luyện mạng sâu trở nên dễ dàng hơn bởi vì nó giảm hiện tượng vanishing gradient và cải thiện khả năng học của mô hình. Trong mỗi residual block, đầu vào sẽ được truyền qua một số lớp convolutional và activation functions, sau đó đường dẫn ngắn sẽ được kết nối trực tiếp với đầu ra của residual block. Cuối cùng, đầu ra của residual block sẽ được truyền tiếp cho các lớp convolutional và activation functions tiếp theo.

Vanishing gradient là một vấn đề phổ biến xảy ra trong quá trình huấn luyện các mạng neural sâu, khi độ dốc của hàm mất mát giảm dần khi lan truyền ngược qua các lớp đầu tiên của mạng. Trong quá trình lan truyền ngược, độ dốc được tính bằng cách sử dụng đạo hàm của hàm mất mát theo các tham số của mạng, và độ dốc này được truyền ngược qua từng lớp để cập nhật các tham số. Tuy nhiên, khi đạo hàm giảm dần theo độ sâu của mạng, các lớp đầu tiên nhận được một lượng độ dốc rất nhỏ, dẫn đến khó khăn trong việc cập nhật các tham số và làm chậm tốc độ hội tụ của mạng. Vấn đề này được gọi là "vanishing gradient" vì độ lớn của độ dốc giảm dần và có thể trở nên rất nhỏ đến mức không thể cập nhật được tham số của mạng.

d. NetVLAD



CNN architecture with the NetVLAD layer

NetVLAD (Network of VLAD) là một kiến trúc mạng nơ-ron tích chập (CNN) được sử dụng để trích xuất đặc trưng của hình ảnh và tìm kiếm các hình ảnh tương đồng trong lĩnh vực tìm kiếm ảnh dựa trên nội dung (content-based image retrieval - CBIR).

NetVLAD được đặc trưng bởi phương pháp biểu diễn đặc trưng của hình ảnh được gọi là vector VLAD (Vector of Locally Aggregated Descriptors). Vector VLAD là một phương pháp biểu diễn đặc trưng của hình ảnh bằng cách tổng hợp các mẫu đặc trưng cục bộ của hình ảnh. Nó là một dạng phương pháp pooling trong CNN, tuy nhiên khác với các phương pháp pooling thông thường, vector VLAD sử dụng một số lượng lớn các mẫu đặc trưng cục bộ để tính toán.

NetVLAD kết hợp giữa hai thành phần chính: mô hình CNN để trích xuất đặc trưng của hình ảnh và vector VLAD để biểu diễn đặc trưng của các hình ảnh. Mô hình CNN được sử dụng để trích xuất các đặc trưng cục bộ của hình ảnh, trong đó các đặc trưng này được sử dụng để tính toán vector VLAD.

Sau khi tính toán được vector VLAD của một hình ảnh, nó có thể được sử dụng để tìm kiếm các hình ảnh tương đồng trong cơ sở dữ liệu. NetVLAD sử dụng khoảng cách Euclid để tính toán độ tương đồng giữa các vector VLAD của các hình ảnh và trả về các hình ảnh có độ tương đồng cao nhất.

NetVLAD đã đạt được hiệu suất tốt trên nhiều bộ dữ liệu ảnh khác nhau và đã được sử dụng trong các ứng dụng thực tế, chẳng hạn như phân loại hình ảnh, tìm kiếm hình ảnh và nhận dạng đối tượng.

Cách tính toán vector VLAD:

Để tính toán vector VLAD của một hình ảnh, trước hết ta cần trích xuất các đặc trưng cục bộ từ hình ảnh bằng một mạng CNN được huấn luyện trước đó. Sau đó, các đặc trưng cục bộ này sẽ được sử dụng để tính toán vector VLAD của hình ảnh.

Cụ thể, quá trình tính toán vector VLAD gồm các bước sau đây:

Chuẩn bị bộ từ điển (vocabulary): Bộ từ điển bao gồm các vector đặc trưng được trích xuất từ một số lượng lớn các hình ảnh khác nhau trong tập dữ liệu. Các vector đặc trưng này sẽ được sử dụng để biểu diễn các đặc trưng cục bộ của hình ảnh cần tính toán vector VLAD.

Gán nhãn từng đặc trưng cục bộ: Với mỗi đặc trưng cục bộ được trích xuất từ hình ảnh cần tính toán vector VLAD, ta sẽ gán nhãn cho nó bằng cách tìm kiếm vector từ điển gần nhất với đặc trưng đó.

Tính toán trọng số cho từng nhóm đặc trưng cục bộ: Sau khi đã gán nhãn cho từng đặc trưng cục bộ, ta sẽ tính toán trọng số cho từng nhóm đặc trưng cục bộ tương ứng với mỗi vector từ điển. Trọng số này được tính bằng cách lấy tổng của sự khác biệt giữa đặc trưng cục bộ và vector từ điển, chia cho tổng của sự khác biệt giữa đặc trưng cục bộ và tất cả các vector từ điển.

Tính toán vector VLAD: Cuối cùng, ta sẽ tính toán vector VLAD cho hình ảnh bằng cách lấy tổng của các vectơ hiệu giữa đặc trưng cục bộ và vector từ điển, với trọng số tương ứng được áp dụng cho mỗi nhóm đặc trưng cục bộ.

Vector VLAD là một vector có chiều dài bằng với số lần gán nhãn các đặc trưng cục bộ và số chiều của mỗi vector từ điển. Nó biểu diễn một cách tổng thể đặc trưng của hình ảnh dưới dạng một phương trình tuyến tính.

2.3 Indexing

Vấn đề đặt ra là sau khi rút trích đặc trưng từ dataset, số lượng vector đặc trưng là quá nhiều. Từ đó, nếu sử dụng phương pháp so sánh thông thường ta phải quét hết tập vector đặc trưng của dataset để xem cái nào tương đồng với ảnh query, điều đó tốn quá nhiều thời gian và chi phí.

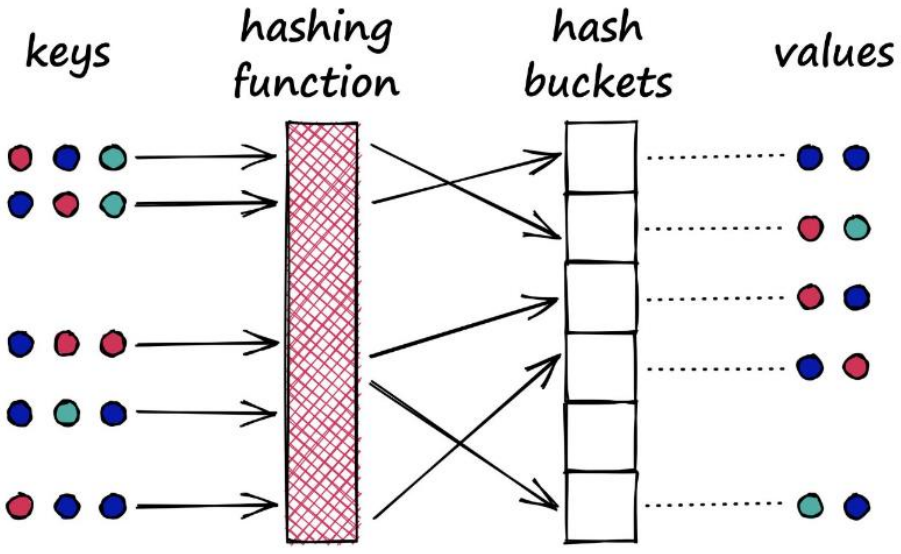
Từ đó, ngoài việc rút trích đặc trưng, CBIR còn tập trung vào việc đánh chỉ mục (feature indexing) để tạo điều kiện thuận lợi cho tốc độ tìm kiếm. Vì một trong những tính năng chính trong hệ thống CBIR là thời gian phản hồi, nên tầm quan trọng của việc đánh chỉ mục trở nên tích cực hơn và đặc biệt là trong những tập dataset có quy mô lớn.

Những phương pháp đánh chỉ mục được sử dụng rộng rãi trong CBIR hashing based indexing.

Hash based indexing:

Phương pháp đánh chỉ mục dựa vào hàm hashing (hàm băm): chuyển đổi hình ảnh thành không gian Hamming, trong đó dữ liệu tương tự sẽ được ánh xạ thành các mã nhị phân tương ứng.

Các kỹ thuật băm như Locality sensitive hashing (LSH) được sử dụng rộng rãi. Đây là kỹ thuật băm sử dụng một hàm băm h để phân chia hình ảnh trong cơ sở dư liệu vào từng ô. Tất cả các vector đặc trưng có cùng giá trị đầu ra sau khi trải qua hàm băm h sẽ được đặt vào cùng một ô. Số lượng ô phụ thuộc vào hàm băm và phạm vi đầu vào và các giá trị đầu ra. Khi đưa ra một hình ảnh truy vấn, áp dụng hàm băm cho nó để trả ra kết quả và ánh xạ tới 1 ô đã có sẵn trong database.



Locality sensitive hashing

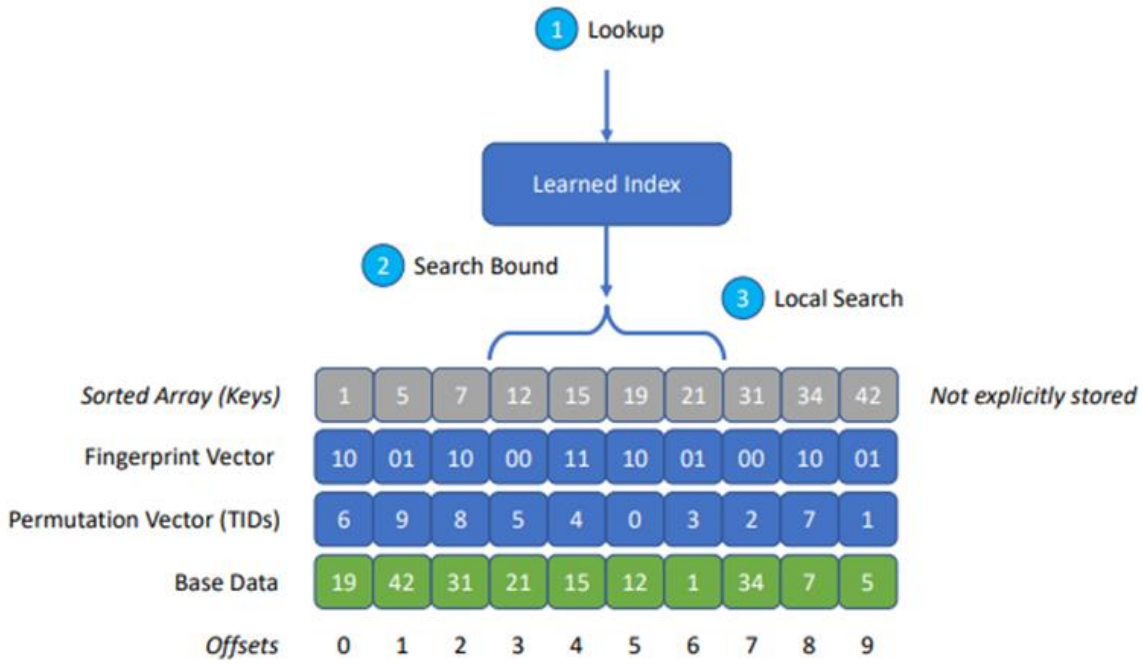
Ngoài ra, chúng ta cũng có thể sử dụng một kỹ thuật băm khác để thực hiện việc đánh chỉ mục, đó là Spectral Hashing:

Spectral Hashing là một phương pháp clustering và giảm chiều cho các vector đặc trưng trong các ảnh. Việc thực hiện giảm chiều dữ liệu được tiến hành bằng cách sử dụng bài toán phân tích giá trị riêng của ma trận đại diện cho sự tương đồng giữa các vector đặc trưng có trong ảnh.

Các bước thực hiện đánh chỉ mục dựa trên phương pháp Special Hashing:

1. Tính ma trận Gram K từ ma trận đặc trưng X – ma trận với mỗi dòng là một vector đặc trưng cho một ảnh sau khi chuẩn hóa, bằng cách tính tích vô hướng của các cặp vector đặc trưng như sau:
   $K_{ij}=x_i \ .x_j$,với x_i,x_j là vector đặc trưng thứ i,thứ j
2. Áp dụng các phương pháp tính kernel như Gaussian Kernel, Laplacian Kernel, Polynominal Kernel vào ma trận K nhằm mục đích chuyển các vector đặc trưng vào không gian mới nhiều chiều hơn nhằm giúp việc phân biệt các nhóm vector đặc trưng gần nhau dễ dàng hơn. Việc này không bắt buộc, thêm bước này sẽ giúp tăng độ chính xác của phương pháp Special Hashing
3. Tìm giá trị riêng và vector riêng của ma trận Gram K bằng thuật toán Eigenvalue Decomposition.
4. Lấy một số vector riêng ứng với các giá trị riêng lớn nhất bằng các thuật toán như PCA. Sau khi chọn xong các vector riêng chúng ta sẽ tiến hành sắp xếp chúng vào tạo thành ma trận μ có kích thước MxK, với K là số vector đặc trưng được chọn và M là số chiều vector
5. Tính toán hàm băm bằng cách cho ma trận X nhân với ma trận μ. Kết quả là một ma trận có kích thước NxK, với N là số lượng hình ảnh, cũng là số hang, số lượng vector đặc trưng tạo thành ma trận X. Nếu cần thiết, có thể chuẩn hóa hàm băm để tăng tính đống đều của việc phân tán ảnh trong không gian băm
6. Lưu trữ hàm băm thành một hash table, một cấu trúc dữ liệu có thể truy cập nhanh và lưu trữ các ảnh. Khi tiến hành truy vấn, các ảnh đầu vào sẽ được chuyển hóa thành giá trị băm và so sánh với các giá trị có trong hàm băm để xác định kết quả của bài toán truy vấn.

Learned Secondary Index (LSI)



Learned Secondary Index và quy trình tra cứu

Gồm 3 phần: permutation vector, learned index, fingerprint vector

- Permutation vector: biểu diễn nén dữ liệu chưa được sắp xếp d bằng cách pernutation vector p sẽ được thiết lập phần tử thứ i của tập dữ liệu d là d[p[i]]. Permutation vector xây dựng learned index trên dữ liệu đã được sắp xếp và sau đó ánh xạ dự đoán từ learned index đến dữ liệu chưa được sắp xếp(d).

- Learned index: ánh xạ lookup key đến bounded search range. Để triển khai bước này thì sử dụng PLEX. PLEX xây dựng một mô hình spline trên hàm phân phối tích lũy (CDF) của dữ liệu và giới hạn tìm kiếm trong phạm vi do người dùng xác định.

- Fingerprint vector: để tra cứu thì cần tìm kiếm trong toàn bộ giá trị trả về ở Learned index điều này dẫn đến nhiều giá trị không liên quan cũng được quét qua, Để giảm thiểu điều này thì Fingerprint vector sẽ lưu mỗi giá trị bằng một thước cố định. Khi tra cứu, nếu fingerprint trong lookup có trong fingerprint vector thì mới truy cập vao permutation vector để chỉ đên base data, ngược lại thì sẽ bỏ qua việc truy cập để tiết kiệm bộ nhớ cache.

2.4. Measurement

Sau khi index xong, tiến hành dùng các độ đo cứng (Euclid, Cosin hay Mahalanobis) để so sánh sự tương đồng hoặc khác biệt giữa ảnh query và các bộ mô tả trong cơ sở dữ liệu để đưa ra những bức ảnh giảm dần về mức độ tương đồng.

Trong CBIR, đặc trưng được trích xuất từ mỗi hình ảnh sẽ được lưu trong tập dữ liệu đặc trưng và truy xuất kết quả từ database. Để đánh giá hiệu quả thì ta sử dụng độ đo để đo sự giống nhau giữa ảnh truy vấn và tất cả ảnh được lưu trong dataset.

- Euclidean: là khoảng cách giữa hai điểm trong không gian có n chiều.

Công thức: $d_e = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2}$

- Cosin: là độ đo giữa hai vector khác không, được tính bằng cách lấy tích vô hướng của hai vector chia cho tích có hướng.

Công thức: $d_c = 1 - \cos(\theta) = 1 - \frac{a.b}{\|a\|\|b\|} = 1 - \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2}\sqrt{\sum_{i=1}^{n} b_i^2}}$

- Mahalanobis: tính khoảng cách giữa hai vector a,b có ma trận covariance c.

Công thức: $d_m = \sqrt{(a_i - b_j)^T c^{-1}(a_i - b_j)}$

Ví dụ: cho 3 hình ảnh và 3 vector tương ứng với mỗi ảnh. Vị trí trong vector lần lượt đại diện cho Ngân, Minh, Nhung.

⇨ Sau khi trích xuất đặc trưng:

Từ vector của hình ảnh và vector đặc trưng, tính khoảng cách Euclidean, khoảng cách Cosin, khoảng cách Mahalanobis.

⇨ Ngân

Vector a = $\begin{bmatrix} 0.9 \\ 0.07 \\ 0.03 \end{bmatrix}$

Vector b = $\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$

- Khoảng cách Euclidean

$d_e = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2}$

$= \sqrt{(0.9-1)^2 + (0.07-0)^2 + (0.03-0)^2}$

= 0.126

- Khoảng cách Cosin

$d_c = 1 - \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2}\sqrt{\sum_{i=1}^{n} b_i^2}}$

$= 1 - \frac{0.9\times1+0.07\times0+0.03\times0}{\sqrt{0.9^2+0.07^2+0.03^2}\sqrt{1^2+0^2+0^2}}$

= 0.004

- Khoảng cách Mahalanobis

Covariance matrix $c^{-1} = \begin{bmatrix} 0.2412 & 0.2833 & 9.6667 \\ 0 & 0.3333 & 12.6667 \\ 0 & 0 & 6.7111 \end{bmatrix}$

$$d_m = \sqrt{[0.9 \quad 0.07 \quad 0.03] \times \begin{bmatrix} 0.2412 & 0.2833 & 9.6667 \\ 0 & 0.3333 & 12.6667 \\ 0 & 0 & 6.7111 \end{bmatrix} \times \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}}$$

$= \sqrt{0.0058} = 0.08$

- Nhận xét: khoảng cách euclid và mahalanobis có giá trị nhỏ => hình ảnh có độ tương đồng lớn, tuy cosin ra kết quả ngược lại (giá trị lớn gần bằng một) nhưng theo ý nghĩa của độ đo cosin thì đó là một kết quả tốt (khá tương đồng với query).

Tuy nhiên, hàm tương đồng/khoảng cách cứng cố định có thể không tối ưu đối với tasks phức tạp của CBIR do thách thức lớn của khoảng cách ngữ nghĩa giữa các đặc trưng cấp thấp được trích xuất bởi máy tính và nhận thức cao của con người.

Do đó, trong những năm gần đây, xuất hiện thêm nhiều nghiên cứu trong việc design các độ đo distance/similarity trên một số đặc trưng cấp thấp bằng cách khai thác các kỹ thuật học máy. Trong số các kỹ thuật này, một số nghiên cứu đã tập trung vào việc learning to hashing or compact codes.

Một cách khác để tăng cường biểu diễn đặc trưng là học đo khoảng cách (DML).

Distance metric learning

Distance metric learning cho việc tìm kiếm hình ảnh đã được nghiên cứu rộng rãi trong machine learning.

Về định dạng dữ liệu huấn luyện, hầu hết các nghiên cứu DML hiện có thường làm việc với hai loại dữ liệu:

- các ràng buộc cặp (pairwise constraints) trong đó các ràng buộc must-link và cannot-link được cung cấp
- các ràng buộc triplet chứa một cặp tương tự và một cặp không tương tự. Cũng có các nghiên cứu trực tiếp sử dụng các class labels cho DML theo một typical machine learning scheme, như thuật toán Large Margin Nearest Neighbor (LMNN).

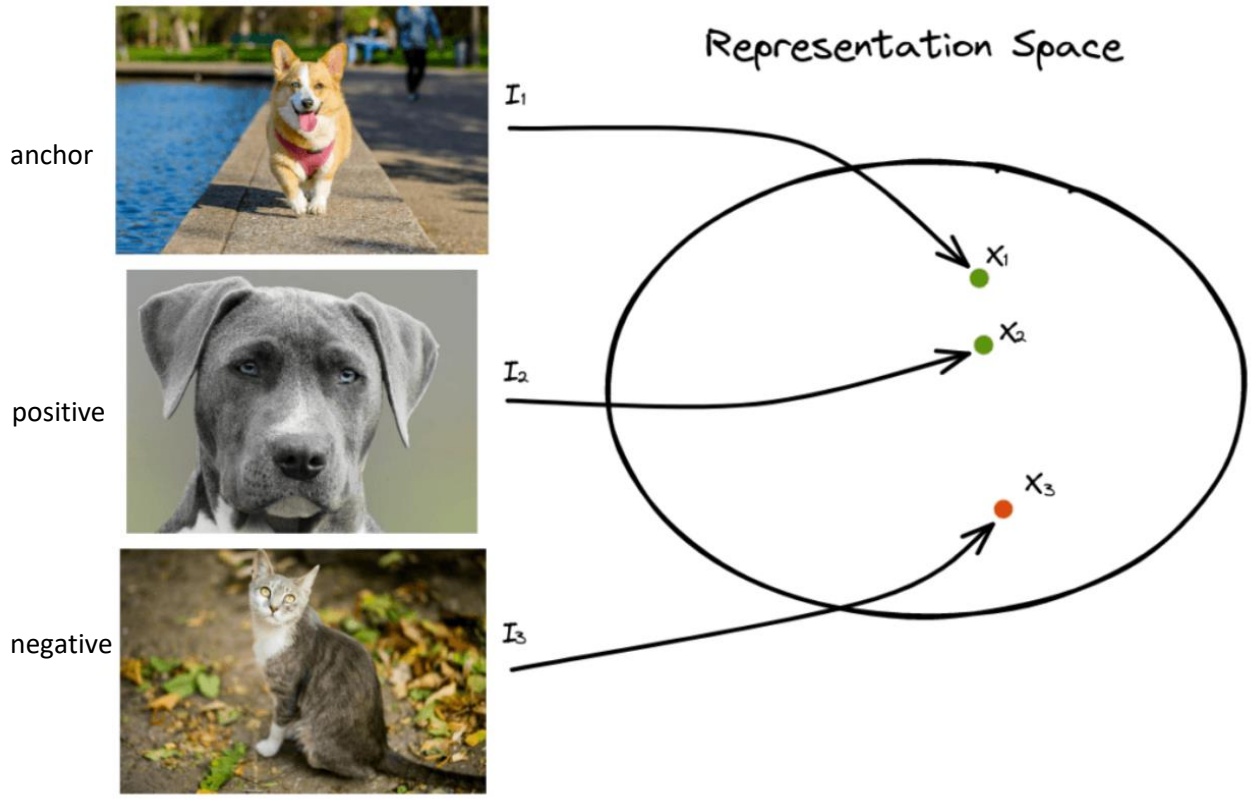
Về các pháp học khác nhau, các kỹ thuật metric learning thường được phân loại thành hai nhóm:

- Global supervised approaches: học một metric trong một bối cảnh toàn cục bằng cách đồng thời thỏa mãn tất cả các ràng buộc.
- Các phương tiện giám sát cục bộ (local supervised approaches) học một metric theo cách cục bộ bằng cách chỉ thỏa mãn các ràng buộc cục bộ đã cho từ neighboring information.

Về learning methodology, hầu hết các nghiên cứu DML hiện có thường sử dụng các phương pháp học dạng batch, trong đó thường giả định toàn bộ tập dữ liệu huấn luyện đã được cung cấp trước khi thực hiện nhiệm vụ học và huấn luyện mô hình từ scratch. Khác với các phương pháp học dạng batch, để xử lý dữ liệu quy mô lớn, các thuật toán DML đã được nghiên cứu gần đây.

a. **Triplet network:**

Mô hình này học cách biểu diễn một ảnh dưới dạng một vector đặc trưng sao cho những ảnh cùng loại có khoảng cách nhỏ hơn so với các ảnh không cùng loại.



Triplet network thường được huấn luyện trên một tập các bộ ba ảnh (triplets) gồm ảnh anchor, ảnh positive và ảnh negative. Trong đó, ảnh anchor là ảnh đang được xét, ảnh positive là một ảnh cùng loại với anchor và ảnh negative là một ảnh không cùng loại với anchor. Mục đích của mô hình là học cách biểu diễn một ảnh anchor sao cho khoảng cách giữa vector đại diện của anchor và vector đại diện của ảnh positive là nhỏ hơn so với khoảng cách giữa vector đại diện của anchor và vector đại diện của ảnh negative.

Cụ thể, quá trình huấn luyện triplet network diễn ra như sau:

- Đầu vào của mô hình là một bộ ba ảnh anchor, positive và negative.

- Mỗi ảnh được đưa qua một mạng convolutional để trích xuất ra vector đặc trưng tương ứng.
- Từ ba vector đặc trưng tương ứng với ba ảnh, ta tính khoảng cách giữa vector đại diện của anchor và vector đại diện của positive (có thể sử dụng khoảng cách Euclid hoặc khoảng cách cosine).
- Tính khoảng cách giữa vector đại diện của anchor và vector đại diện của negative.
- Huấn luyện mô hình để minimize khoảng cách giữa anchor và positive và maximize khoảng cách giữa anchor và negative (thông qua một hàm mất mát như hàm margin loss).

  Bộ ba ta có ảnh anchor (a), positive (p) và negative (n):

  (a,p,n) chia thành các cặp: (a,p); (a,n)

  Hàm triplet loss function:

  $$L = \max(d(a,n) - d(a,p) + \text{margin}, 0)$$

  - margin là một hằng số (thường được đặt là 1), đảm bảo khoảng cách giữa anchor và negative lớn hơn khoảng cách giữa anchor và positive một lượng tối thiểu.

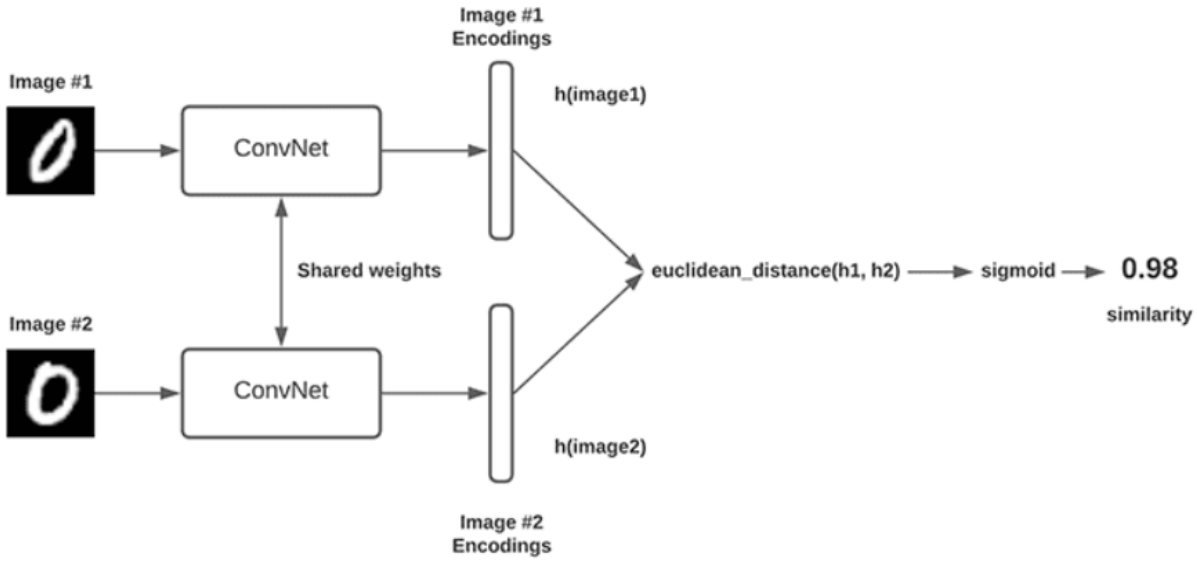
  Tối ưu hóa bằng gradient descent: stochastic gradient descent (SGD)

=> Sử dụng phương pháp backpropagation để tính toán gradient của hàm loss đối với các tham số trong mạng và cập nhật các tham số này bằng cách trừ gradient nhân với một hệ số learning rate.

b. Siamese network

Siamese network là một kiến trúc mạng nơ-ron được sử dụng trong các tác vụ so sánh và đo lường sự tương đồng giữa các cặp dữ liệu. Thay vì chỉ đưa vào một dữ liệu và đưa ra kết quả, siamese network nhận vào hai dữ liệu và đưa ra một giá trị đo lường tương đồng giữa chúng.

Kiến trúc siamese network thường được xây dựng bằng cách tạo ra hai mạng nơ-ron trùng nhau và chia sẻ trọng số giữa chúng. Hai dữ liệu đầu vào sẽ được đưa qua hai mạng nơ-ron này để rút trích đặc trưng, sau đó kết quả sẽ được đưa vào một hoặc nhiều lớp để tính toán độ tương đồng giữa chúng.



- Input: hai ảnh xi, xj

- Mỗi ảnh sẽ đi qua ConvNet của Siamese Network để xử lý, sau đó mỗi ConvNet sẽ trả về một vector (h1 và h2)

- Tính khoảng cách Euclidean giữa h1 và h2.

- Dùng hàm Sigmoid để ép giá trị khoảng cách Euclidean về khoảng [0, 1]. Kết quả của hàm Sigmoid càng gần 1 thì hai hình ảnh càng giống nhau.

Siamese loss

$$L_{Siam}(x_i, x_j) = \frac{1}{2} s(x_i, x_j) D(x_i, x_j) + \frac{1}{2}\left(1 - S(x_i, x_j)\right) \max(0, m - D(x_i, x_j))$$

Trong đó, S(xi, xj) là Sigmoid function, D(xi, xj) là khoảng cách Euclidean giữa xi và xj.

3. Evaluation methodology:

Bộ dữ liệu thử nghiệm là yếu tố cần thiết để đánh giá hiệu suất của các hệ thống tìm kiếm hình ảnh dựa trên nội dung (CBIR). Nó bao gồm một tập hình ảnh truy vấn và những đánh giá liên quan tới mỗi truy vấn. Ngoài ra, các độ đo đánh giá hiệu suất cũng quan trọng để xác định tính hiệu quả của các hệ thống CBIR khác nhau. Do đó, các bộ dữ liệu thử nghiệm phổ biến, các độ đo chuẩn được sử dụng để đánh giá hiệu quả của một hệ thống CBIR điển hình được trình bày sơ lược trong phần này.

3.1. Dataset used

Trong quá khứ, đã có một số bộ dữ liệu thử nghiệm có hình ảnh groundtruth nằm trong một tập hợp các truy vấn đã được xác định trước để đánh giá các framework khác nhau cho hệ thống CBIR. Trong số đó, có các bộ datasets sau:

1. INRIA holiday (Jegou et al. 2008): Bộ dữ liệu này bao gồm 1.491 hình ảnh độ phân giải cao của các địa điểm khác nhau trên toàn vũ trụ. Các hình ảnh trong bộ sưu tập này có độ phân giải là 570x760 hoặc 1020x760 và chủ yếu bao gồm các loại cảnh tự nhiên.
2. Scene-15 (Lazebnik et al. 2006): Bộ dữ liệu này gồm chủ yếu là 4.485 hình ảnh được nhóm lại thành 15 danh mục. Tổng cộng, có từ 210 đến 410 hình ảnh trong mỗi danh mục và tất cả đều có độ phân giải cố định là 250x300 pixel. Hầu hết các hình ảnh trong bộ sưu tập Scene-15 có ngữ cảnh phân biệt rõ ràng giữa nền và mặt trước.
3. Oxford (Philbin et al. 2007): Bao gồm 5.062 hình ảnh các tòa nhà nằm tại 11 địa điểm khác nhau trong thành phố Oxford, Anh. Các hình ảnh trong tập dữ liệu này khó để phân biệt các mặt tiền của các tòa nhà tương tự nhau. Tất cả các hình ảnh trong bộ sưu tập này có độ phân giải cố định là 1020 x 760.
4. GHIM-10K (Liu et al. 2015): Bao gồm 10.000 hình ảnh trong 20 danh khác nhau. Mỗi danh mục chứa 500 hình ảnh màu trong định dạng JPEG với độ phân giải là 300 x 400 hoặc 400 x 300.
5. IAPR TC-12 (Grubinger et al. 2006): Là một tập dữ liệu hình ảnh phổ biến được chọn để đánh giá hệ thống truy vấn. Bao gồm 20.000 hình ảnh được thu thập từ các địa điểm khác nhau trên toàn cầu, bao gồm các loại hình ảnh cảnh tự nhiên khác nhau. Tất cả các hình ảnh trong bộ sưu tập này đều có định dạng JPEG với kích thước cố định là 360 x 480 điểm ảnh. Một tính năng thú vị của bộ sưu tập này là có nhiều hình ảnh có nội dung hình ảnh giống nhau, tuy nhiên chúng khác nhau về phông nền, điều kiện ánh sáng và góc nhìn.
6. SUN-397 (Huiske et al. 2010): được hình thành dưới dạng các nhóm hình ảnh trong 397 danh mục ngữ nghĩa khác nhau. Số lượng hình ảnh trong mỗi danh mục có thể khác nhau, nhưng có ít nhất 100 hình ảnh trong mỗi danh mục và tổng số hình ảnh trong toàn bộ dữ liệu là 108.754 hình ảnh. Các hình ảnh trong cùng một danh mục ngữ nghĩa có thể xuất hiện trong các ngữ cảnh khác nhau và diện mạo của chúng có thể thay đổi theo các ngữ cảnh này. Do đó, việc tìm kiếm và truy xuất các hình ảnh tương tự đối với một truy vấn cho trước là một nhiệm vụ đầy thách thức.

3.2. Performance measures used

CBIR được đánh giá dựa trên danh sách kết quả truy vấn thứ tự tương ứng với mỗi truy vấn được cung cấp.

Giả sử I = { I_1, I_2, ..., I_N} là tập hợp N hình ảnh được biểu diễn bởi low-level feature hoặc high-level feature { $\mathbb{f}_1$, $\mathbb{f}_2$, ..., $\mathbb{f}_N$}. Để truy vấn hình ảnh, vector đặc trưng $\mathbb{f}_d$ được trích xuất từ mỗi hình ảnh trong cơ sơ dữ liệu $\mathcal{J}_d$ được so sánh với đặc trưng $\mathbb{f}_q$ của truy vấn q bằng cách sử dụng phương pháp đo khoảng cách phù hợp $d(\mathbb{f}_q, \mathbb{f}_d)$. Sau đó một danh sách trả về R_q được tạo ra bằng cách sắp xếp các hình ảnh trong cơ sở dữ liệu dựa trên khoảng cách của chúng đến hình ảnh truy vấn sao cho $d(\mathbb{f}_q, \mathbb{f}_d) < d(\mathbb{f}_q, \mathbb{f}_{d+1})$ đúng cho mỗi cặp hình ảnh $\mathcal{J}_d$ và $\mathcal{J}_{d+1}$ trong $\mathcal{J}$. Để đánh giá chất lượng của danh sách trả về R_q đối với truy vấn q cho trước, một tập hợp các độ đo hiệu suất được định nghĩa dựa trên thống kê sau đây:

1. True positives (tp_k): số lượng hình ảnh được xác định đúng nằm ở vị trí từ 1 đến k trong R_q. $tp_k = \sum_{p=1}^{k} rel_p$ trong đó rel_p thuộc {0,1} biểu thị tính đúng đắn của hình ảnh xuất hiện ở vị trí p trong danh sách trả về R_q. Nếu hình ảnh ở vị trí p của danh sách trả về R_q là đúng cho truy vấn q thì $rel_p = 1$ và ngược lại.
2. False postivies (fp_k): là số lượng kết quả sai xuất hiện ở vị trí từ 1 đến k trong danh sách trả về R_q, được tính bằng công thức $fp_k = \sum_{p=1}^{k}(1 - rel_p)$
3. False negatives (fn_k): số lượng hình ảnh có liên quan nhưng lại bị bỏ qua, không được xuất hiện trong danh sách trả về. Công thức: $fn_k = \sum_{p=1}^{k} rel_p - tp_k$
4. True negatives (tn_k): số lượng hình ảnh không liên quan và hệ thống bỏ qua, không xuất hiện trong danh sách trả về. Công thức: $tn_k = \sum_{p=1}^{k}(1 - rel_p) - fp_k$

Dựa trên các số liệu thống kê trên, các cách đánh giá CBIR được định nghĩa cụ thể như sau:

- Precision (pr_k): độ chính xác của một hệ thống truy xuất cho truy vấn q được tính bằng tỷ lệ của số lượng hình ảnh liên quan trong k vị trí đầu tiên của danh sách trả về R_q với tổng số hình ảnh được truy xuất. Công thức:
$$pr_k = \frac{tp_k}{tp_k + fp_k}$$
- Recall (re_k): recall là tỉ lệ số lượng hình ảnh liên quan ở k vị trí đầu tiên của kết quả tìm kiếm với tổng số hình ảnh có liên quan cho truy vấn q. Công thức:
$$re_k = \frac{tp_k}{tp_k + n}$$
- Mean average precision (mAP) : độ chính xác trung bình. Giả sử $|Q|$ biểu thị số lượng hình ảnh trong truy vấn của tập truy vấn Q, ta có công thức:
$$mAP = \frac{1}{|Q|}\sum_{q \in Q} AP(q)$$

Trong đó $AP(q)$ là độ chính xác trung bình cho truy vấn $q \in Q$, công thức:

$$AP(q) = \frac{1}{NG_q} \sum_{k=1}^{NG_q} (pr_k \delta_k)$$

Trong đó NG_q biểu thị số lượng hình ảnh thực của truy vấn q và δ_k là hàm indicator (0 hoặc 1) đại diện cho mức độ liên quan của một hình ảnh tại vị trí k.

Cách truyền thống là sử dụng 11-point interpolated average precision. Với mỗi truy vấn, interpolated precision (độ chính xác nội suy) được đo ở 11 recall levels 0.0, 0.1, 0.2, ..., 1.0. Với 11-point interpolated average precision ký hiệu P_{int} tại recall level r_i được tính là độ chính xác cao nhất được ghi nhận với mọi recall value k giữa r_i và 1. Được biểu diễn như sau:

$$P_{int}(r_i) = \max_{r_i \leq r \leq 1} precision(r)$$

V. Thực nghiệm

Dataset: Content Based Image Retrieval Dataset
https://www.kaggle.com/datasets/theaayushbajaj/cbir-dataset

Tập dataset bao gồm 3 class:

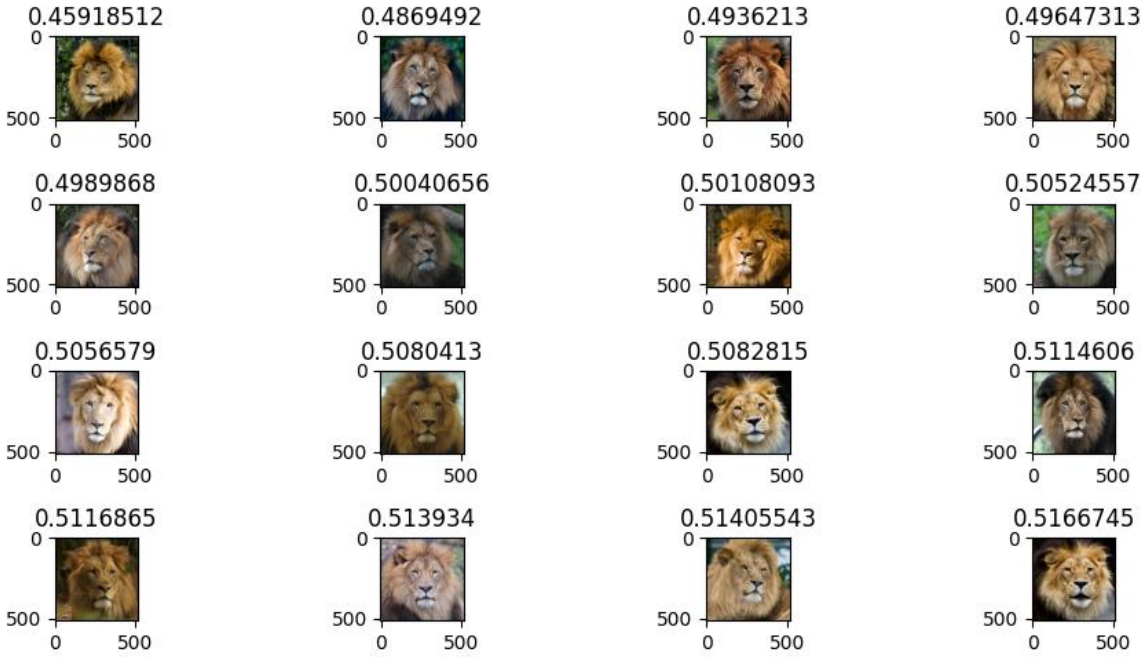
- Hổ
- Cáo
- Sư tử
- Báo

Sử dụng model VGG16 pre-trained.

Một số kết quả:

- Query:

- Output: 16 ảnh giống nhất

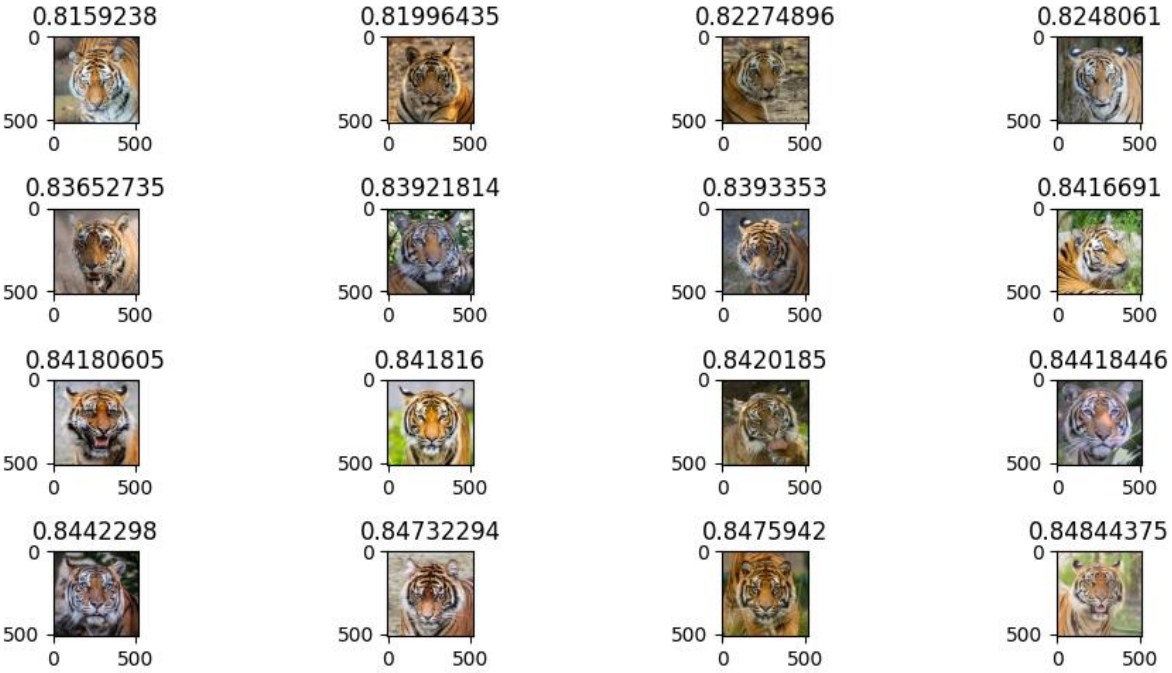


- Query

- Output

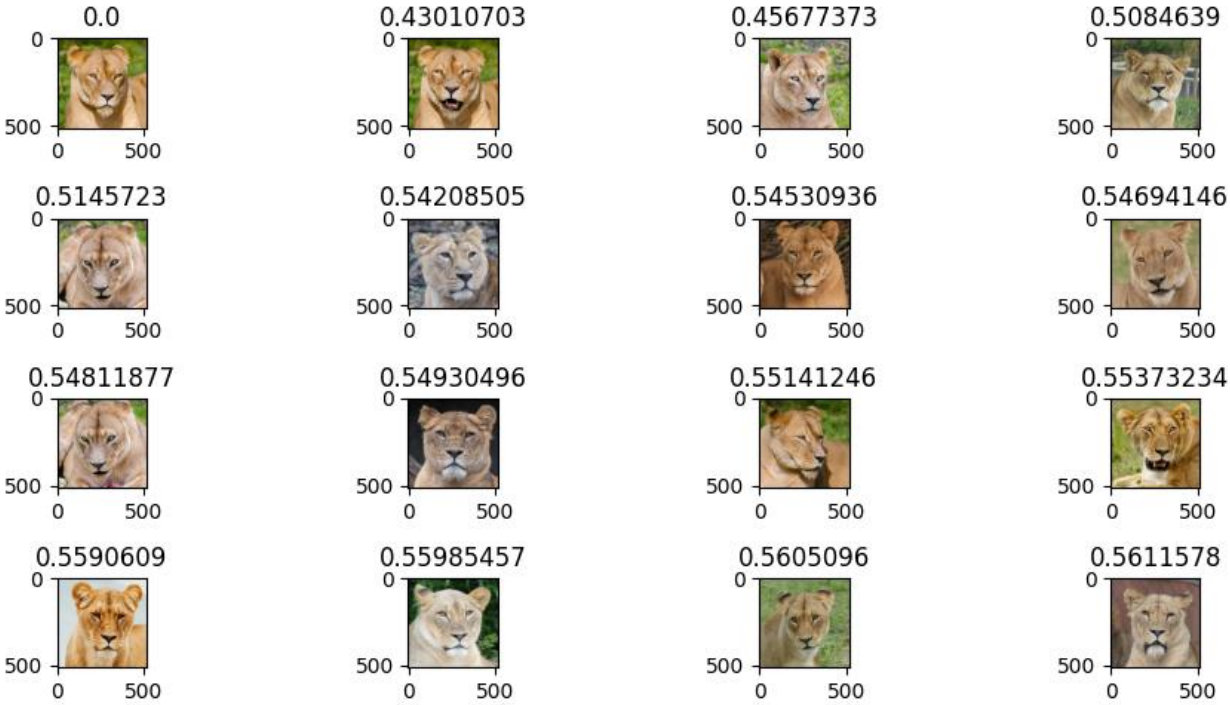
- Query

- Output



- Query: ta lấy 1 ảnh thẳng từ trong tập dataset:

- Output: như ta thấy, có một distance là 0.0 vì nó đã nằm trong tập data base sẵn rồi. Nên khoảng cách từ nó đến chính nó bằng 0.



VI. Tài liệu tham khảo

1. "A Survey on Image Retrieval Methods", A. Malcom Marshall and Dr.S.Gunasekaran. https://www.researchgate.net/publication/265298224_A_Survey_on_Image_Retrieval_Methods
2. "Deep image retrieval: a survey", Chen, W.; Liu, Y.; Wang, W.; Bakker, E.M.; Georgiou, T.K.; Fieguth, P.; Liu, L.; Lew, M.S.K. https://scholarlypublications.universiteitleiden.nl/handle/1887/3166004
3. "A Survey on Semantic-Based Image Retrieval", Jia Li, et al. (2021): https://www.mdpi.com/2079-9292/10/2/255
4. "Content-Based Image Retrieval: A Comprehensive Review", S. Smeulders, et al. (2000): https://link.springer.com/article/10.1023/A:1008189014710
5. "Content-Based Image Retrieval using Bag-of-Visual Words", Mohammed Alkhawlani, Mohammed Elmogy, Hazem Elbakry. https://bom.so/JAnLpj
6. "A Learned Secondary Index Structure", Andreas Kipf, Dominik Horn, Pascal Pfeil. https://arxiv.org/pdf/2205.05769.pdf

7. "Evaluation Of Distance Measures In Content Based Image Retrieval", Nehal M. Varma; Anamika Choudhary. https://ieeexplore.ieee.org/document/8821957

8. "Content Based Image Indexing and Retrieval", Avinash N Bhute, B. B. Meshram. https://www.researchgate.net/publication/259604534_Content_Based_Image_Indexing_and_Retrieval/link/5c6c3f8c92851c1c9dee88b8/download

9. "Deep Learning for Content-Based Image Retrieval: A Comprehensive Study", Ji Wan, Dayong Wang, Steven C.H. Hoi, Pengcheng Wu, Jianke Zhu, Yongdong Zhang, Jintao Li. https://g2.by/XNKP

10. "Enhanced bag of visual words representations for content based image retrieval: a comparative study", K.S. Arun, V.K. Govindan, S.D. Madhu Kumar https://sci-hub.ru/10.1007/s10462-019-09715-6

11. "Content-Based Image Retrieval using Deep Learning", Anshuman Vikram Singh https://scholarworks.rit.edu/cgi/viewcontent.cgi?article=9984&context=theses

12. "Content-Based Image RetrievalUsing Convolutional Neural Networks", Ouhda Mohamed, El Asnaoui Khalid, Ouanan Mohammed, and Aksasse Brahim https://www.researchgate.net/publication/325155660_Content-Based_Image_Retrieval_Using_Convolutional_Neural_Networks/link/5c07aa74299bf169ae3368e0/download

13. "Scalable Database Indexing and Fast Image Retrieval Based on Deep Learning and Hierarchically Nested Structure Applied to Remote Sensing and Plant Biology", Pouria Sadeghi-Tehran,* Plamen Angelov, Nicolas Virlet, and Malcolm J. Hawkesford https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8320911/