BÀI THU HOẠCH

CHỦ ĐỀ: IMAGE RETRIEVAL

A. Thông tin cá nhân

MSSV: 19127273

Huỳnh Thị Mỹ Thanh

B. Bài thu hoạch

I. Problem statement

- Input: tùy vào từng loại bài toán mà có nhiều loại input khác nhau như văn bản mô tả hoặc 1 ảnh query.
- 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. 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.
- Có 3 loại bài toán retrieval thường gặp: Content-based image retrieval, Semantic-based image retrieval và Text-based image retrieval.

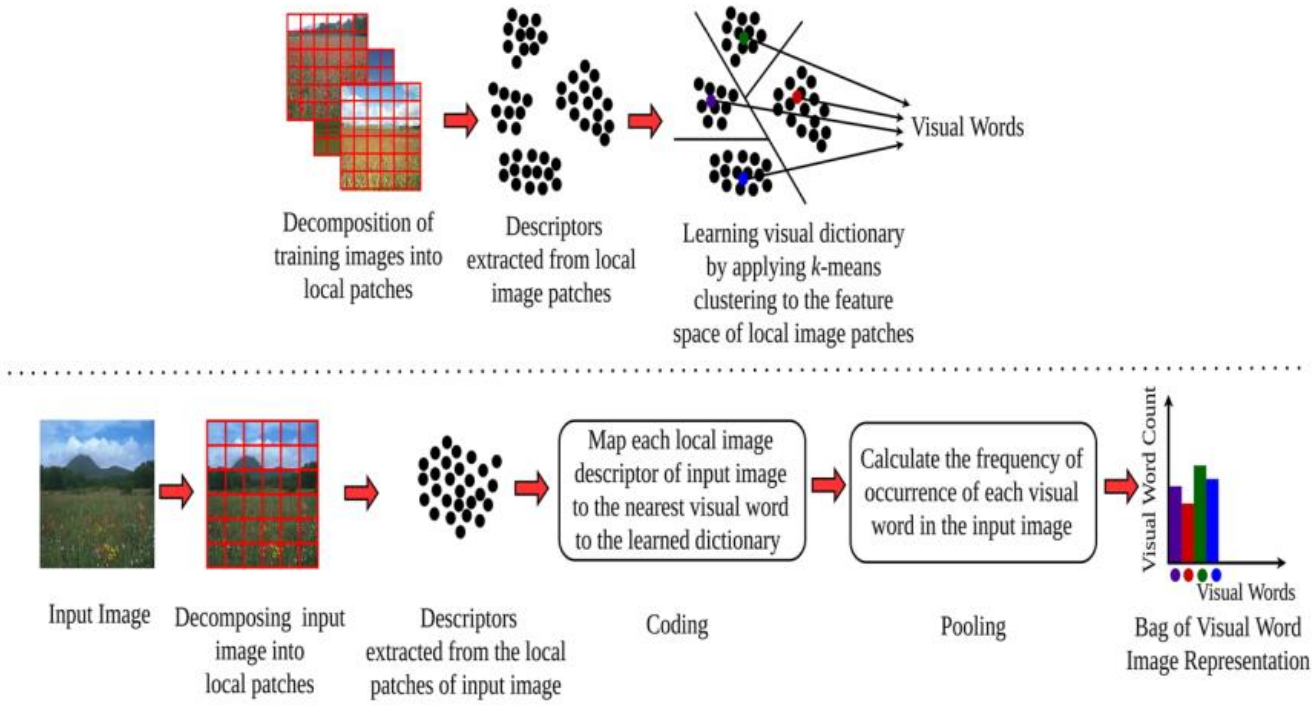
II. Challenge

- Thách thức về dataset: 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.
- 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.
- 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.

III. Methods

*Sử dụng local feature:

Bag of visual words:

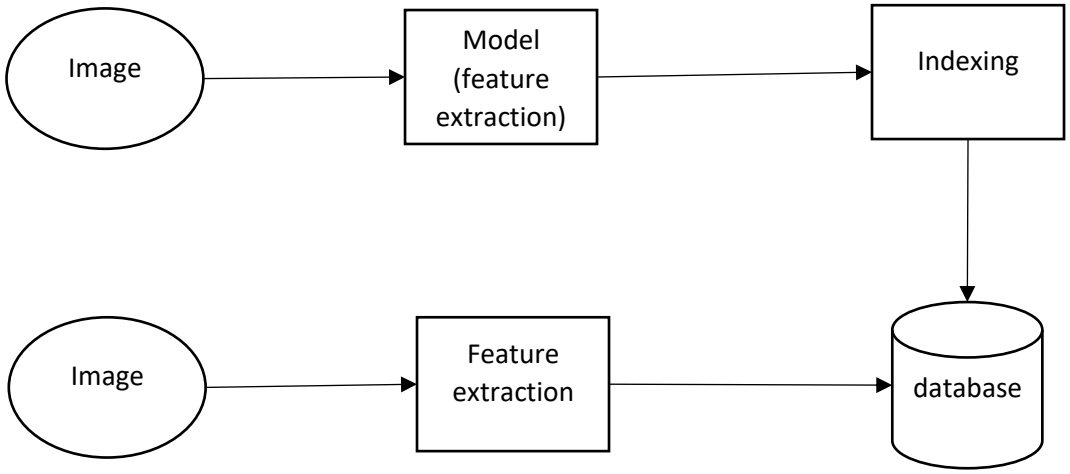


Các bước thực hiện gồm:

- 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, random sampling. 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.
- 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. Cho một tập hợp các đặc trưng ảnh cục bộ đã được trích xuất ở bước trước đó, thuật toán K-means sẽ chia bộ mô tả đó thành K vùng khác nhau với mỗi vùng sẽ có một trọng tâm (điểm trung bình) để đại diện cho mỗi vùng khác nhau.

- Coding: mã hóa các đặc trưng hình ảnh thành biểu diễn dựa trên bag of visual words.
- Pooling: gộp các biểu diễn dựa trên bag of visual words lại để tạo ra biểu diễn tổng thể của hình ảnh, từ đó có thể sử dụng để tìm kiếm và so sánh các hình ảnh dựa trên nội dung.
- Sau khi ảnh query đưa vào cũng trải qua quá trình tương tự (phải extract feature) và xem nó thuộc words nào, sau đó lấy các tập hình ảnh có cùng chỉ mục trong database đã được huấn luyện ra.

***Sử dụng global feature (áp dụng kỹ thuật deep learning):**



Chủ yếu gồm 2 phần chính:

a. **Model feature extraction:**
   - 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.
   - GooleNet: 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.

- 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.
- 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).

b. Indexing:

Sau khi trích xuất đặc trưng từ các hình ảnh trong training dataset, ta thu được một số lượng lớn các bộ mô tả (vector đặc trưng) của các hình ảnh đó. Một phương pháp thủ công để tìm kiếm hình ảnh đó là sử dụng các độ đo cứng thông thường như Euclid, Cosin, Mahalanobis để đánh giá sự tương đồng. Nhưng nếu ở trong tập dataset quá lớn (giả sử 1 triệu ảnh) thì ta phải quét qua hết tất cả 1 triệu vecotr đặc trưng đó để đo, như vậy là quá tốn tài nguyên và thời gian.

Nên trong Image Retrieval có một phương pháp được đề ra gọi là indexing (đánh chỉ mục). Trong bước này, tập vector đặc trưng đã rút trích được sẽ được hashing để gom lại những đặc trưng giống nhau lại cùng 1 nhóm, ta sẽ dùng 1 số hoặc 1 chữ để đại diện cho nhóm đó.

Và tương tự khi đưa vào ảnh query, sau khi rút trích đặc trưng ta sẽ xem nó thuộc nhóm nào bằng cách dùng hàm hash vào vector đặc trưng đã rút được. Sau đó, tiến hành trả về tập hình ảnh có cùng chỉ mục. Từ đó giúp cho việc tìm kiếm diễn ra nhanh hơn.

Bài toán này khá giống như Classification, mà label của nó chính là từng nhóm đại diện.

IV. Conclusion

CBIR là một lĩnh vực nghiên cứu hứa hẹn trong lĩnh vực xử lý hình ảnh, cho phép tìm kiếm hình ảnh dựa trên nội dung của chúng, thay vì dựa trên các thông tin nhãn hoặc từ khóa.

Tuy nhiên, vẫn còn nhiều thách thức trong CBIR, bao gồm độ chính xác của kết quả tìm kiếm, độ nhạy cảm đối với biến đổi hình ảnh (ví dụ: dịch chuyển, xoay, đổi kích thước), và khả năng tự động học và cải thiện kết quả tìm kiếm theo phản hồi của người dùng.

Hướng phát triển trong CBIR có thể bao gồm nghiên cứu và phát triển các phương pháp hashing tiên tiến, kết hợp các kỹ thuật học máy và học sâu để cải thiện tính chính xác và độ nhạy cảm của hệ thống tìm kiếm. Ngoài ra, nghiên cứu về tính bảo mật của các phương pháp hashing trong CBIR cũng là một hướng phát triển tiềm năng để đảm bảo tính riêng tư và an toàn của dữ liệu hình ảnh.