TRUY VẤN THÔNG TIN THỊ GIÁC

TRUY VẤN TRANG PHỤC THỜI TRANG

Võ Nguyễn Hoàng Kim, Trần Thanh Ngân, Lâm Thanh Ngọc

Giảng viên hướng dẫn: Võ Hoài Việt, Nguyễn Trọng Việt, Phạm Minh Hoàng

GIỚI THIỆU

Truy vấn trang phục thời trang (Fashion Image Retrival) 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.

PHÁT BIỂU BÀI TOÁN

Đầu vào: Hệ thống nhận vào là một ảnh về thời trang mà người dùng muốn truy vấn.
Đầu ra: 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.

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
- 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.

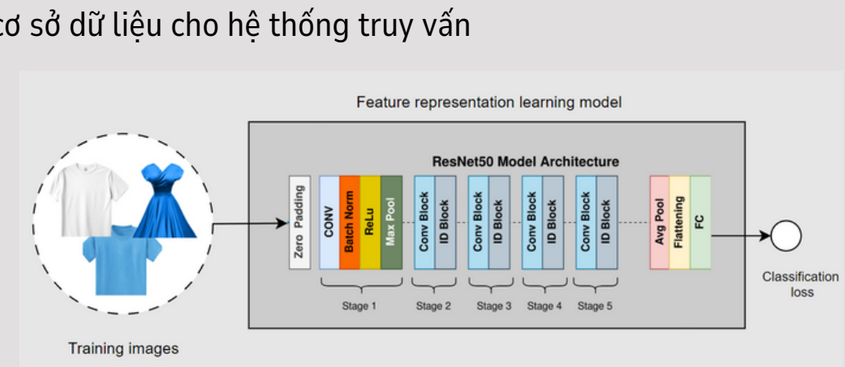
TẬP DỮ LIỆU

FashionIQ: Tập dữ liệu 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 chú thích được thu thập từ cộng đồng, mô tả các thay đổi từ hình ảnh tham chiếu đến mục tiêu.

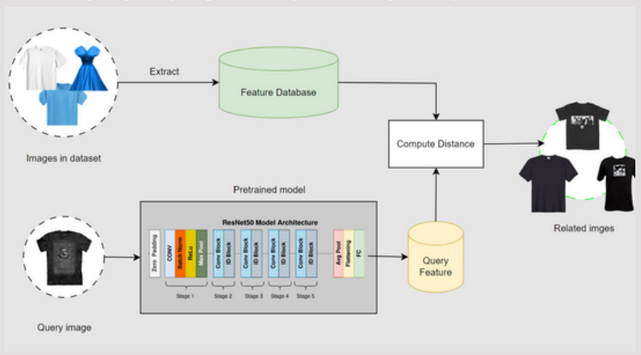
PHƯƠNG PHÁP

Việc xây dựng một hệ thống truy vấn được chia thành 2 bước chính: Chuẩn bị cơ sở dữ liệu và Truy vấn

Chuẩn bị cơ sở dữ liệu: Hệ thống sử dụng bộ ảnh huấn luyện của tập dữ liệu FashionIQ để huấn luyện mô hình ResNet50 trong việc rút trích đặc trưng ảnh. Mô hình sử dụng Cross-Entropy để làm hàm mất mát. Mô hìnhsau khi huấn luyện được sử dụng để rút trích đặc trưng của toàn bộ ảnh trong tập dữ liệu FashionIQ và lưu lại để làm cơ sở dữ liệu cho hệ thống truy vấn

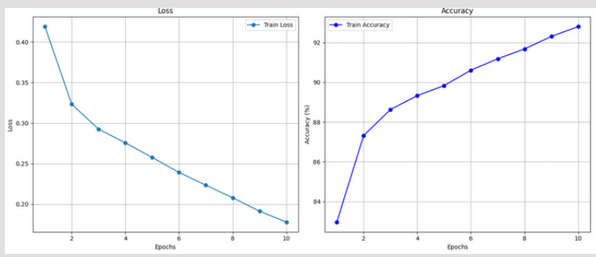


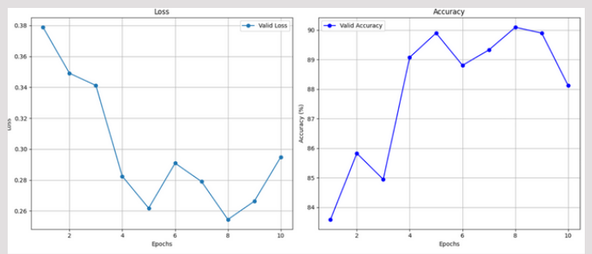
Truy vấn: Hệ thống nhận vào ảnh truy vấn và nhãn tương ứng của nó, thực hiện rút trích đặc trưng trên ảnh truy vấn bằng mô hình đã huấn luyện. Sử dụng các độ đo như Euclidean hay Cosine để tính toán độ dị biệt giữa đặc trưng truy vấn và cơ sở dữ liệu đặc trưng. Cuối cùng, hệ thống sẽ trả ra top-k ảnh kết quả được truy vấn có độ dị biệt thấp nhất so với ảnh truy vấn.



KẾT QUẢ VÀ ĐÁNH GIÁ

Việc huấn luyện mô hình ResNet50 được thực hiện với số lượng epochs là 10, Cross-Entropy làm hàm mất mát, bộ tối ưu hóa Adam với hệ số học là 0.001.





Kiểm thử mô hình bằng bộ dữ liệu Test, độ chính xác mà mô hình đạt được là 0.91.

Thực hiện truy vấn một ảnh bất kỳ với nhãn tương ứng được cung cấp, kết quả trả về được đánh dấu bởi các khung màu xanh và đỏ, tương ứng cho các kết quả truy vấn đối với ảnh đúng và sai.

Với số lượng ảnh truy vấn là 15, hệ thống trả ra kết quả đến với 13 ảnh đúng và 2 ảnh sai. Khi này, AP mà mô hình đạt được là 0.854 với độ chính xác là 0.86.

CẢI TIẾN

Với phương pháp ban đầu, việc so khớp đặc trưng truy vấn với toàn bộ đặc trưng có trong cơ sở dữ liệu tiềm ẩn nhiều vấn đề (tiêu tốn thời gian, tài nguyên và chi phí tính toán).

Vì thế, nhóm đề xuất bổ sung một lớp SVM ở bước truy vấn, giúp phân loại ảnh truy vấn thuộc về loại thời trang nào trước khi đem so khớp.

- Huấn luyện mô hình trên tập dữ liệu FashionIQ và 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.

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 đem 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. 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.

Kết quả

```
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
```

Phương pháp ban đầu

```
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
```

Phương pháp cải tiến

Đánh giá

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.
Các giá trị được ổn định và không thay đổi khi giá trị k ngày càng tăng. Tuy nhiên, 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.

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.

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.