CLIP Shoe: Một phương pháp tìm kiếm giày bằng văn bản và hình ảnh dựa trên CLIP

Nguyễn Bùi Thanh Mai[1,2]

[1] Trường Đại học Công nghệ Thông tin, ĐHQG-HCM [2] 21522320@gm.uit.edu.vn

What ?

Chúng tôi đề xuất CLIP Shoe - một mô hình được tinh chỉnh cho tác vụ tìm kiếm giày dựa trên CLIP:

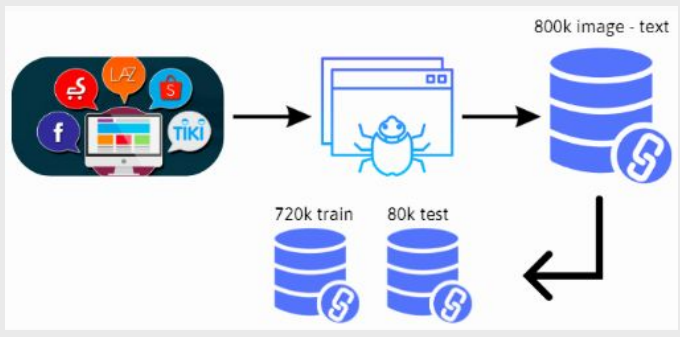
- Mô hình có kiến trúc tương tự như CLIP và được finetune trên CLIP
- Mô hình được huấn luyện trên một dataset do chúng tôi tự thu thập gồm 800k cặp hình ảnh - văn bản
- Được đánh giá trên UT Zappos 50K

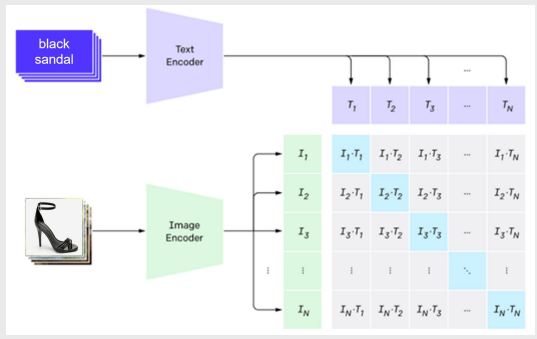
Why ?

- Phương pháp tìm kiếm giày dựa trên nội dung ảnh bằng câu truy vấn là một văn bản hoặc một ảnh cung cấp trải nghiệm mua sắm đa chiều và linh hoạt cho người dùng hơn là phương pháp truy vấn truyền thống dựa trên từ khóa hoặc lựa chọn từ các danh mục có sẵn.
- Đa số các nghiên cứu chỉ tập trung tinh chỉnh mô hình CLIP cho quần áo và thường bỏ qua các chi tiết nhỏ như giày.

Overview

Thu thập dataset → Huấn luyện mô hình → Đánh giá kết quả

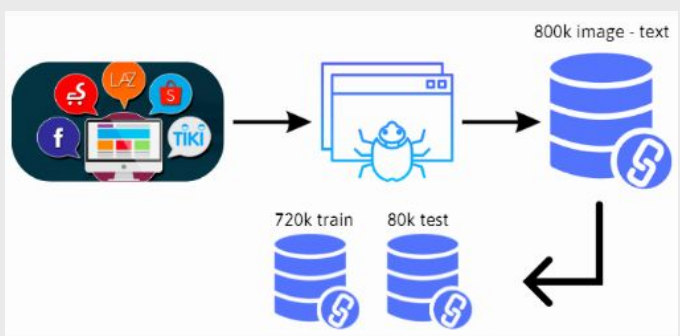




Description

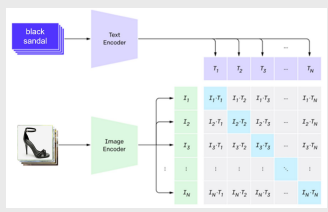
1. Thu thập dataset

- Thu thập dataset gồm 800k cặp hình ảnh về giày và văn bản mô tả tương ứng trên các trang web thương mại điện tử lớn như Zalora, Shopee, Lazada, Tiki, Sendo để huấn luyện mô hình.
- Sử dụng các công cụ web scraping để tự động lấy dữ liệu từ các trang web này, và lọc ra những cặp hình ảnh và văn bản có chất lượng cao, có độ phong phú và đa dạng về kiểu dáng, màu sắc, thương hiệu… của giày.
- Chia dataset thành hai tập: tập huấn luyện (720k cặp) và tập kiểm tra (80k cặp).



2. Huấn luyện mô hình

- Tinh chỉnh mô hình CLIP trên dữ liệu đã thu thập được từ giai đoạn trước theo các phương pháp dùng để xây dựng mô hình FashionCLIP và OpenFashionCLIP. Quy trình huấn luyện như sau:
- Bắt đầu bằng việc khởi tạo tham số mô hình bằng bộ tham số của mô hình CLIP đã được pretrained trước đó.
- Sau đó, mô hình sẽ được huấn luyện để tối đa hóa sự tương tự của các cặp (hình ảnh, văn bản) tương đồng và tối thiểu hóa sự tương tự của các hình ảnh và văn bản không tương đồng thông qua việc huấn luyện cùng lúc một bộ mã hóa hình ảnh và một bộ mã hóa văn bản để tối ưu hóa một hàm mất mát tương phản.



3. Đánh giá mô hình

Sử dụng bộ dữ liệu UT Zappos50K và tập test gồm 80k cặp văn bản - hình ảnh để đánh giá hiệu suất của mô hình, sử dụng hai độ đo là Acc@k và R@k.

NII

Nguyễn Bùi Thanh Mai – Trường Đại học Công nghệ Thông tin, Đại học Quốc gia TP.HCM
Email : 21522320@gm.uit.edu.vn