THÔNG TIN CHUNG CỦA NHÓM

- Link YouTube video của báo cáo:
  https://youtu.be/9eom5zfc0Ck
- Link slides:
  https://github.com/mainbt/CS519.O11/blob/a7378bf09aae737eb49ffbd6c8bc0fd28b3a2bc1/MAI%20NGUY%E1%BB%84N%20B%C3%99I%20THANH%20-%20xCS519.DeCuong.FinalReport.Template.Slide.pdf

• Họ và Tên: Nguyễn Bùi Thanh Mai
• MSSV: 21522320
 
	• Lớp: CS519.O11
	• Tự đánh giá (điểm tổng kết môn): 9/10
	• Số buổi vắng: 1
	• Số câu hỏi QT cá nhân: 7
	• Link Github: https://github.com/mainbt/CS519.O11/
	• Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm:
	  ○ Lên ý tưởng cho đề tài
	  ○ Soạn đề cương, slide và poster
	  ○ Làm video YouTube

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI

CLIP SHOE: MỘT PHƯƠNG PHÁP TÌM KIẾM GIÀY BẰNG VĂN BẢN VÀ HÌNH ẢNH DỰA TRÊN CLIP

TÊN ĐỀ TÀI TIẾNG ANH

CLIP SHOE: A CLIP-BASED APPROACH FOR TEXT AND IMAGE-BASED SHOE RETRIEVAL

TÓM TẮT

Trong bối cảnh thị trường giày ngày càng phát triển và đa dạng, nhu cầu về một hệ thống tìm kiếm giày hiệu quả, chính xác nhằm nâng cao trải nghiệm người dùng là cực kỳ cần thiết. 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. Tuy nhiên, do hình ảnh hoặc câu truy vấn mà khách hàng thực hiện trên web mang tính chất đa dạng và không định hướng, có thể chưa gặp qua khi huấn luyện mô hình, nên để đảm bảo độ chính xác, mô hình cần phải nhạy và thích ứng với dữ liệu mới. Học không giám sát (zero-shot learning) trở thành một khía cạnh quan trọng cần xem xét trong tình huống này. Trong đó, CLIP nổi bật như một lựa chọn phù hợp nhờ việc đã được huấn luyện trên một tập dữ liệu lớn trên Internet và đã chứng minh được khả năng zero-shot đáng chú ý trên nhiều nhiệm vụ, vượt xa một số mô hình khác. Đã có một số nghiên cứu trước đây tinh chỉnh mô hình CLIP để áp dụng trong lĩnh vực thời trang, nhưng chúng thường chỉ tập trung chủ yếu vào mảng quần áo và bỏ qua các chi tiết nhỏ như giày. Vì vậy, trong nghiên cứu này, chúng tôi tập trung vào việc xây dựng CLIP Shoe - một mô hình được tinh chỉnh cho tác vụ tìm kiếm giày dựa trên CLIP, với mục tiêu đạt được độ chính xác tìm kiếm cao nhất có thể khi áp dụng cho một hệ thống tìm kiếm giày trên các trang web thực tế. Mô hình cũng sẽ được đánh giá trên tập dữ liệu lớn về giày như UT Zappos50K,

nhằm so sánh và chứng minh độ chính xác của nó so với các mô hình khác, để đảm bảo tính chính xác của kết quả.

GIỚI THIỆU

Trong những năm gần đây, sự phát triển nhanh chóng của thương mại điện tử và các nền tảng thời trang trực tuyến đã tạo ra nhu cầu về hệ thống tìm kiếm giày hiệu quả và chính xác. Nhưng hầu hết các trang web thương mại hiện nay lại chỉ cho phép tìm kiếm giày dựa trên từ khóa hoặc lựa chọn từ các danh mục có sẵn. Phương pháp đó có thể hạn chế và không đáp ứng đầy đủ nhu cầu của người dùng trong lĩnh vực tìm kiếm ảnh nói chung và lĩnh vực tìm kiếm ảnh về giày nói riêng.

Một cách tiếp cận khác để cải thiện trải nghiệm tìm kiếm ảnh nói chung là tìm kiếm ảnh dựa trên nội dung bằng cả văn bản và hình ảnh. Do trên thực tế, các hình ảnh và câu truy vấn mà người dùng cung cấp cho hệ thống tìm kiếm trên web thường đa dạng và chưa được gặp qua khi huấn luyện mô hình, nên không chỉ cần mô hình phải hiểu và liên kết được văn bản và hình ảnh mà còn phải có khả năng học zero shot tốt để đảm bảo độ chính xác của kết quả tìm kiếm. Đã có nhiều nghiên cứu trước đây về liên kết ngữ nghĩa giữa hình ảnh và văn bản [2, 3], nhưng chúng lại cần lượng dữ liệu lớn để tinh chỉnh hoặc huấn luyện cho tác vụ cụ thể và học zero shot không tốt.

Một giải pháp khác là sử dụng mô hình có khả năng học zero shot tốt như CLIP [1] để tinh chỉnh. CLIP là mô hình được huấn luyện trên lượng dữ liệu rất lớn, 400 triệu cặp hình ảnh và văn bản trên nhiều lĩnh vực khác nhau, nên có khả năng học zero shot vượt trội trên nhiều tác vụ, thậm chí vượt qua nhiều mô hình được huấn luyện sẵn cho tác vụ đó. Mặc dù có một số nghiên cứu đã tinh chỉnh CLIP để áp dụng trong thời trang, nhưng chúng thường chỉ tập trung chủ yếu vào mảng quần áo và bỏ qua các chi tiết nhỏ như giày. Vậy nên trong đề tài này, chúng tôi sẽ nghiên cứu và áp dụng phương pháp tinh chỉnh CLIP cho tác vụ tìm kiếm giày, dựa trên các nghiên cứu trước đó về tinh chỉnh CLIP cho ngành thời trang [4, 5]. Mô hình được tinh chỉnh gọi là CLIP Shoe, có khả năng nhận ảnh giày làm truy vấn và trả về kết quả tìm kiếm phù hợp. Cụ thể:

Input: Một đoạn văn bản mô tả giày hoặc một ảnh giày bất kỳ.

Output: Một danh sách các ảnh giày, được sắp xếp theo mức độ tương đồng với ảnh truy vấn, từ cao đến thấp. Các ảnh giày được lấy từ một cơ sở dữ liệu giày có sẵn.

MỤC TIÊU

- 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.
- Nghiên cứu tinh chỉnh mô hình CLIP theo các phương pháp được nghiên cứu trong các bài báo áp dụng CLIP trong lĩnh vực thời trang [4, 5] cho tác vụ tìm kiếm giày bằng hình ảnh và văn bản sao cho đảm bảo độ chính xác cao nhất.

NỘI DUNG VÀ PHƯƠNG PHÁP

Quá trình xây dựng CLIP Shoe sẽ gồm hai giai đoạn chính như sau:

Giai đoạn 1: Thu thập dataset để huấn huyện mô hình

Theo như các tác giả của mô hình FashionCLIP [4] và OpenFashionCLIP [5], họ đã sử dụng khoảng từ 700k đến 800k cặp hình ảnh - văn bản để tiến hành tinh chỉnh mô hình dựa trên CLIP. Vậy nên, chúng tôi cũng sẽ 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. Chúng tôi sẽ 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. Chúng tôi sẽ chia dataset thành hai tập: tập huấn luyện (720k cặp) và tập kiểm tra (80k cặp) để đánh giá tính chính xác của mô hình sau này.

Giai đoạn 2: Xây dựng mô hình CLIP Shoe dựa trên dữ liệu đã thu thập được

Đầu tiên, chúng tôi sẽ nghiên cứu chi tiết về cấu trúc và phương pháp huấn luyện của mô hình CLIP [1]. Dựa trên cơ sở đó, chúng tôi tìm cách 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 [4] và OpenFashionCLIP [5]. Chúng tôi cũng sẽ thử tìm

hiểu thêm các nghiên cứu về truy vấn dựa trên văn bản và hình ảnh tương tự để tìm thêm các giải pháp tối ưu độ chính xác của mô hình và áp dụng vào đề tài này.
Sau đó, chúng tôi sẽ tiến hành quá trình tinh chỉnh mô hình CLIP Shoe trên dữ liệu đã thu thập được. Quá trình này sẽ 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. Kỹ thuật huấn luyện này chính là kỹ thuật contrastive learning.
Cuối cùng, để đánh giá về khả năng học zero shot và khả năng tìm kiếm ảnh giày bằng hình ảnh của mô hình, chúng tôi sẽ sử dụng bộ dữ liệu UT Zappos50K để đánh giá hiệu suất của mô hình đã được tinh chỉnh, sử dụng hai độ đo là Acc@k và R@k. Do UT Zappos50K chỉ có hình ảnh giày chứ không có văn bản mô tả, nên để đánh giá về khả năng tìm kiếm ảnh giày bằng văn bản, chúng tôi sẽ sử dụng tập kiểm tra được chia trước đó và sử dụng hai độ đo Acc@k và R@k để đánh giá. Kết quả sẽ được so sánh với các mô hình khác như CLIP [1], FashionCLIP [4], và OpenFashionCLIP [5] để đưa ra được kết luận về độ chính xác của mô hình và thực hiện điều chỉnh thêm để tăng độ chính xác. Cụ thể, tôi sẽ nghiên cứu thêm về kỹ thuật prompt engineering để tìm được để tìm được cách viết những câu truy vấn tốt nhất. Theo như trong nghiên cứu về mô hình OpenFashionCLIP [5], hiệu suất của mô hình được cải thiện hơn khi áp dụng kỹ thuật prompt engineering.

KẾT QUẢ MONG ĐỢI

- Bộ dữ liệu gồm 800k cặp hình ảnh và văn bản tương ứng được thu thập và xử lý từ các trang web thương mại điện tử lớn.
- Báo cáo về các kỹ thuật và phương pháp được sử dụng trong nghiên cứu này, cùng với các kết quả đánh giá mô hình dựa trên cơ sở so sánh với một số mô hình khác để đề xuất ra các hướng nghiên cứu tiếp theo.

TÀI LIỆU THAM KHẢO

[1]. Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, Ilya Sutskever:
Learning Transferable Visual Models From Natural Language Supervision. ICML 2021: 8748-8763

[2]. Jiasen Lu, Dhruv Batra, Devi Parikh, Stefan Lee:
ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. NeurIPS 2019: 13-23

[3]. Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, Kai-Wei Chang:
VisualBERT: A Simple and Performant Baseline for Vision and Language. CoRR abs/1908.03557 (2019)

[4]. Patrick John Chia, Giuseppe Attanasio, Federico Bianchi, Silvia Terragni, Ana Rita Magalhães, Diogo Gonçalves, Ciro Greco, Jacopo Tagliabue:
FashionCLIP: Connecting Language and Images for Product Representations. CoRR abs/2204.03972 (2022)

[5]. Giuseppe Cartella, Alberto Baldrati, Davide Morelli, Marcella Cornia, Marco Bertini, Rita Cucchiara:
OpenFashionCLIP: Vision-and-Language Contrastive Learning with Open-Source Fashion Data. ICIAP (1) 2023: 245-256