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

Nguyễn Bùi Thanh Mai - 21522320

# Tóm tắt

- Lớp: CS519.O11
- Link Github của nhóm: https://github.com/mainbt/CS519.O11
- Link YouTube video: https://youtu.be/9eom5zfc0Ck

Nguyễn Bùi Thanh Mai

# Giới thiệu

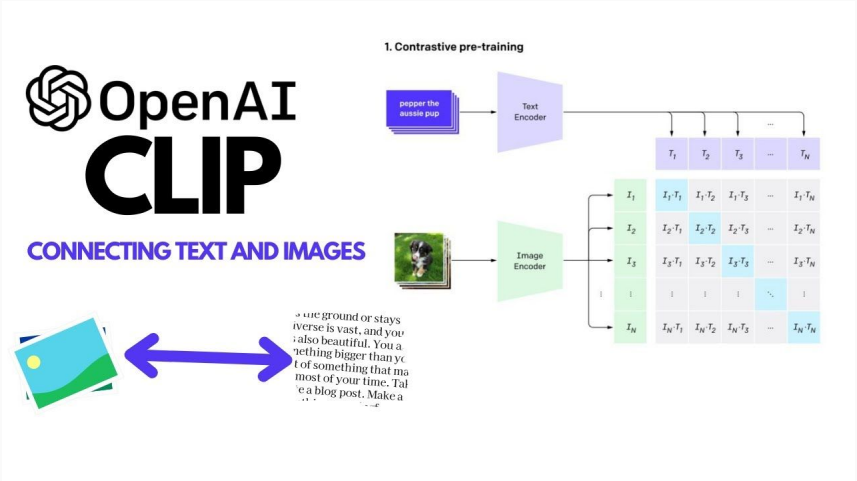
- Nhu cầu về hệ thống tìm kiếm giày hiệu quả và chính xác
  ➔ Shoe CLIP - Phương pháp tìm kiếm ảnh giày dựa trên nội dung bằng cả văn bản và hình ảnh dựa trên CLIP


*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.

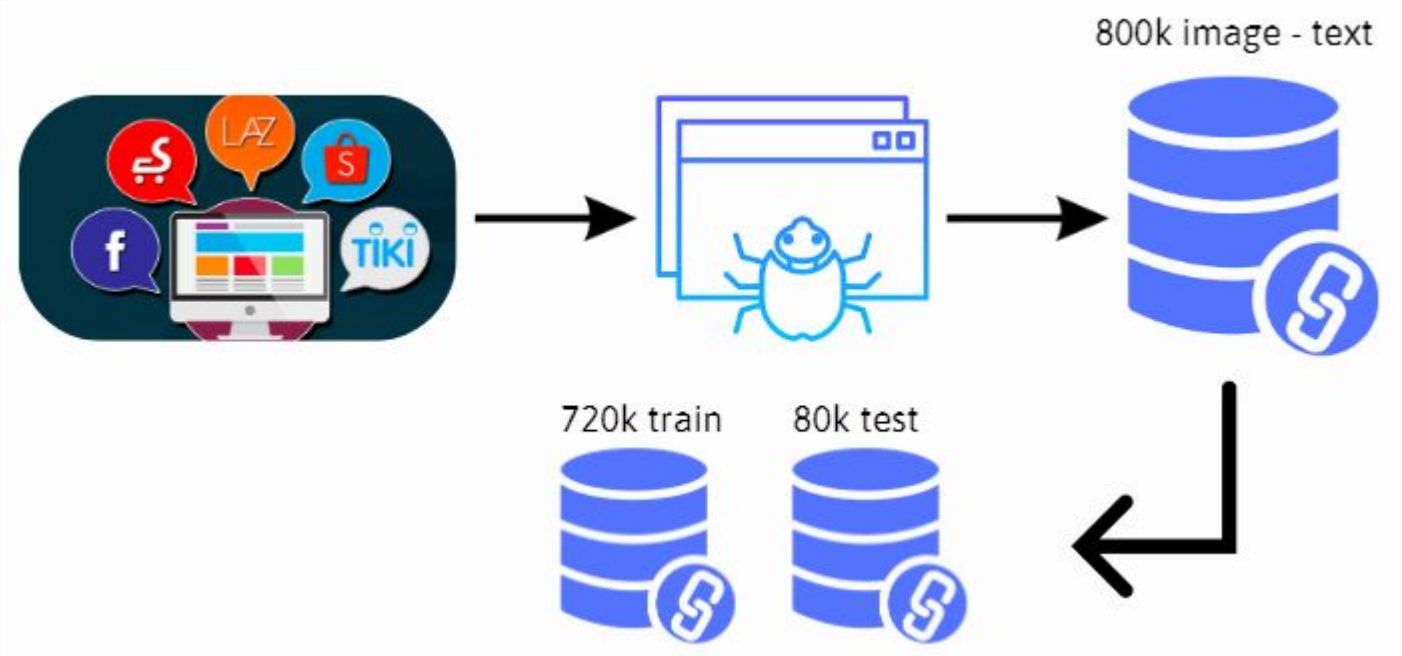
# Mục tiêu

- Thu thập dataset gồm 800k cặp hình ảnh về giày và văn bản mô tả tương ứng trên Zalora, Shopee, Lazada, Tiki, Sendo.
- Tinh chỉnh mô hình CLIP cho tác vụ tìm kiếm giày

# Nội dung và Phương pháp

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

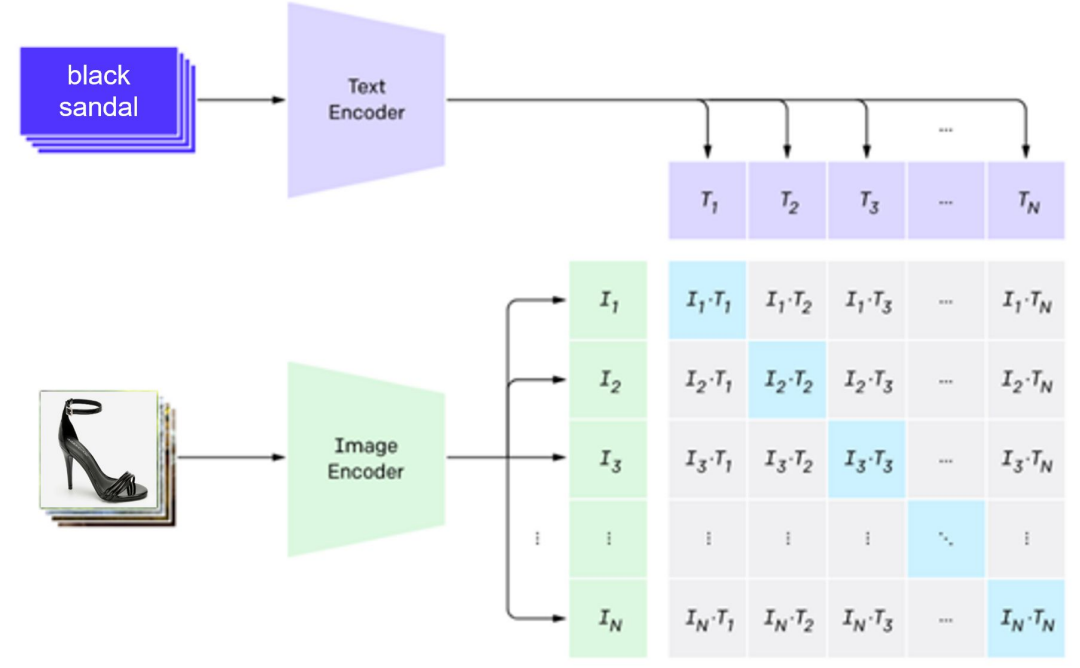


# Nội dung và Phương pháp

- **Giai đoạn 2: Xây dựng mô hình CLIP Shoe dựa trên dữ liệu đã thu thập được**
  1. Tinh chỉnh mô hình CLIP trên tập huấn luyện.
  2. Sử dụng kỹ thuật contrastive learning để tối ưu hóa hàm mất mát tương phản.
  3. Đánh giá hiệu suất của mô hình trên tập kiểm tra và bộ dữ liệu UT Zappos50K bằng các độ đo Acc@k và R@k.
  4. So sánh kết quả với các mô hình khác và áp dụng kỹ thuật prompt engineering để cải thiện hiệu suất.

# Nội dung và Phương pháp

Kiến trúc mô hình CLIP Shoe (tương tự mô hình CLIP):



# Kết quả dự kiến

- 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