XÂY DỰNG HỆ THỐNG TRUY VẤN VIDEO HIỆU SUẤT CAO KẾT HỢP ĐA PHƯƠNG THỨC

Nguyễn Minh Đức - 21520730

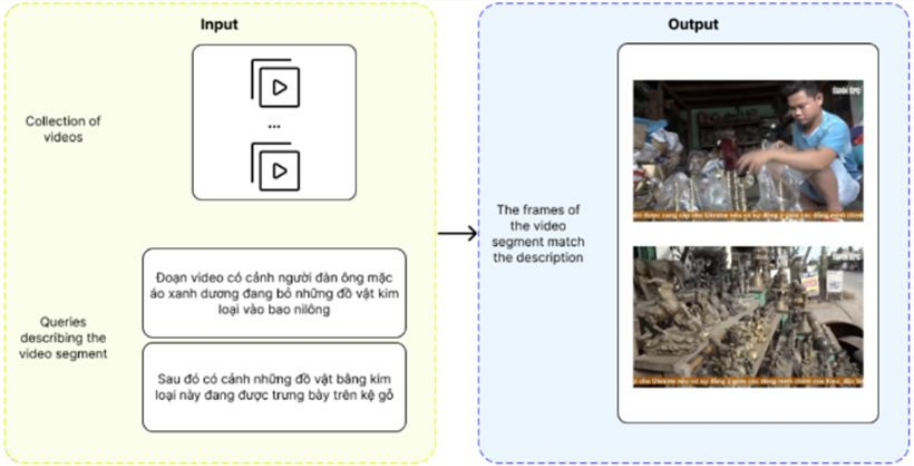
Tóm tắt

- Lớp: CS519.O11
- Link Github của nhóm: https://github.com/minhducnguyen2602/CS519.O11
- Link YouTube video: https://youtu.be/_2d8icVe9E0

Nguyễn Minh Đức - 21520730

Giới thiệu

- Dữ liệu video là nguồn dữ liệu lớn, cần phải khai thác triệt để
- Truy vấn sự kiện trong video là một vấn đề vô cùng thách thức => Cần phải có phương pháp hợp lý và tận dụng tối đa nguồn dữ liệu



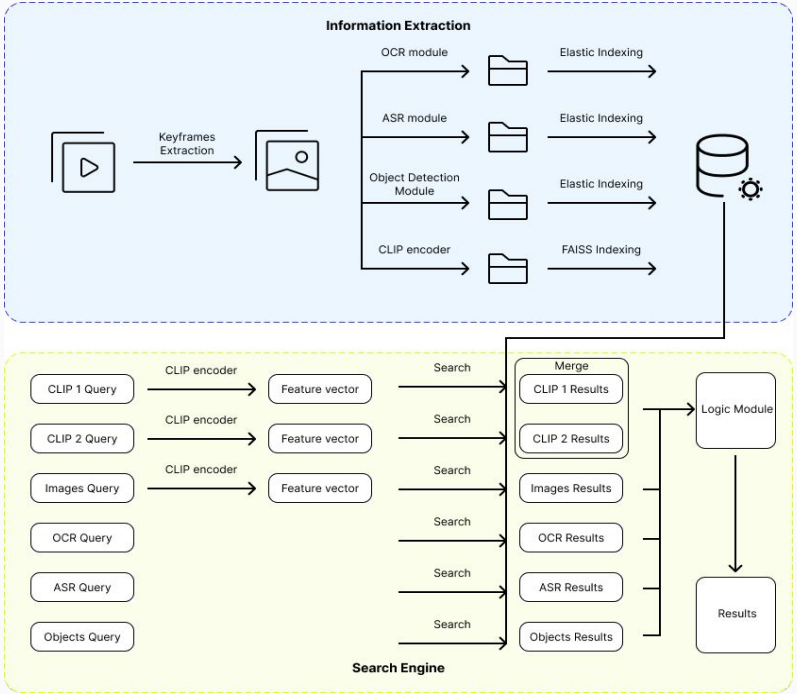
Hình: Mô tả input, output

Mục tiêu

- Tìm hiểu tổng quan về bài toán truy vấn video và xác định hướng tiếp cận
- Xây dựng hệ thống truy vấn video bằng cách sử dụng và kết hợp nhiều phương thức truy vấn khác nhau như truy vấn dựa vào nội dung hình ảnh, nội dung âm thanh, các đối tượng, bằng hình ảnh cụ thể. Kết quả trả về của hệ thống phải có độ chính xác cao và thời gian xử lý nhanh (các kết quả trả về được xếp hạng theo độ liên quan từ cao đến thấp so với câu truy vấn, thời gian thực thi nhỏ hơn 500ms)
- Xây dựng ứng dụng với giao diện thân thiện với người dùng, dễ sử dụng.

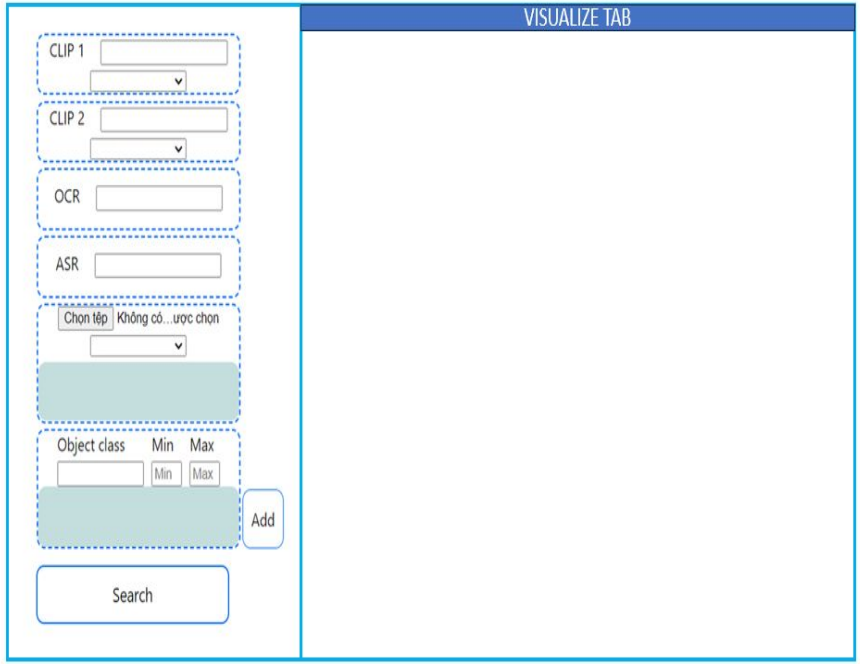
Nội dung và Phương pháp

- Thiết kế hệ thống của ứng dụng:
  - Nghiên cứu sử dụng các pretrained Vision-Language Models, ASR, OCR, Object detection để trích xuất các đặc trưng từ video.
  - Nghiên cứu, đề xuất các thuật toán để kết hợp các phương thức tìm kiếm, tối ưu thời gian tìm kiếm, tối ưu kết quả tìm kiếm.



Nội dung và Phương pháp

- Thiết kế giao diện của ứng dụng:
  - Nghiên cứu, tìm hiểu các framework để xây dựng ứng dụng web đạt hiệu quả cao.



Kết quả dự kiến

- Xây dựng được một hệ thống truy vấn từ video, dễ dàng thêm dữ liệu, xử lý kết quả nhanh và chính xác.
- Xây dựng được một ứng dụng web hoàn chỉnh có thể truy vấn trên video dựa vào mô tả dưới dạng text hoặc ảnh.
- Kết quả dự kiến đạt được sẽ là R@1 = 0.6 và R@10 = 0.9 với R@K là là tỷ lệ số lượng truy vấn mà hệ thống trả về kết quả chính xác trong top K đầu tiên được lựa chọn, thời gian thực hiện truy vấn nhanh, rơi vào khoảng 300-500ms.
- Hệ thống có thể được mang đi ứng dụng trong thực tế, ví dụ ở trường học, sân bay, … để hỗ trợ các công việc liên quan đến tìm kiếm người, vật, sự kiện trong lượng lớn video.

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]** Junnan Li, Dongxu Li, Caiming Xiong, Steven C. H. Hoi: BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. ICML 2022: 12888-12900
- **[3]** Darwin Bautista, Rowel Atienza: Scene Text Recognition with Permuted Autoregressive Sequence Models. ECCV (28) 2022: 178-196
- **[4]** Minghui Liao, Zhaoyi Wan, Cong Yao, Kai Chen, Xiang Bai: Real-time Scene Text Detection with Differentiable Binarization. CoRR abs/1911.08947 (2019)
- **[5]** Nguyen Nguyen, Thu Nguyen, Vinh Tran, Minh-Triet Tran, Thanh Duc Ngo, Thien Huu Nguyen, Minh Hoai: Dictionary-Guided Scene Text Recognition. CVPR 2021: 7383-7392
- **[6]** Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, Ilya Sutskever: Robust Speech Recognition via Large-Scale Weak Supervision. ICML 2023: 28492-28518
- **[7]** Zhuofan Zong, Guanglu Song, Yu Liu: DETRs with Collaborative Hybrid Assignments Training. ICCV 2023: 6725-6735
- **[8]** Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, Lei Zhang: Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. CoRR abs/2303.05499 (2023)