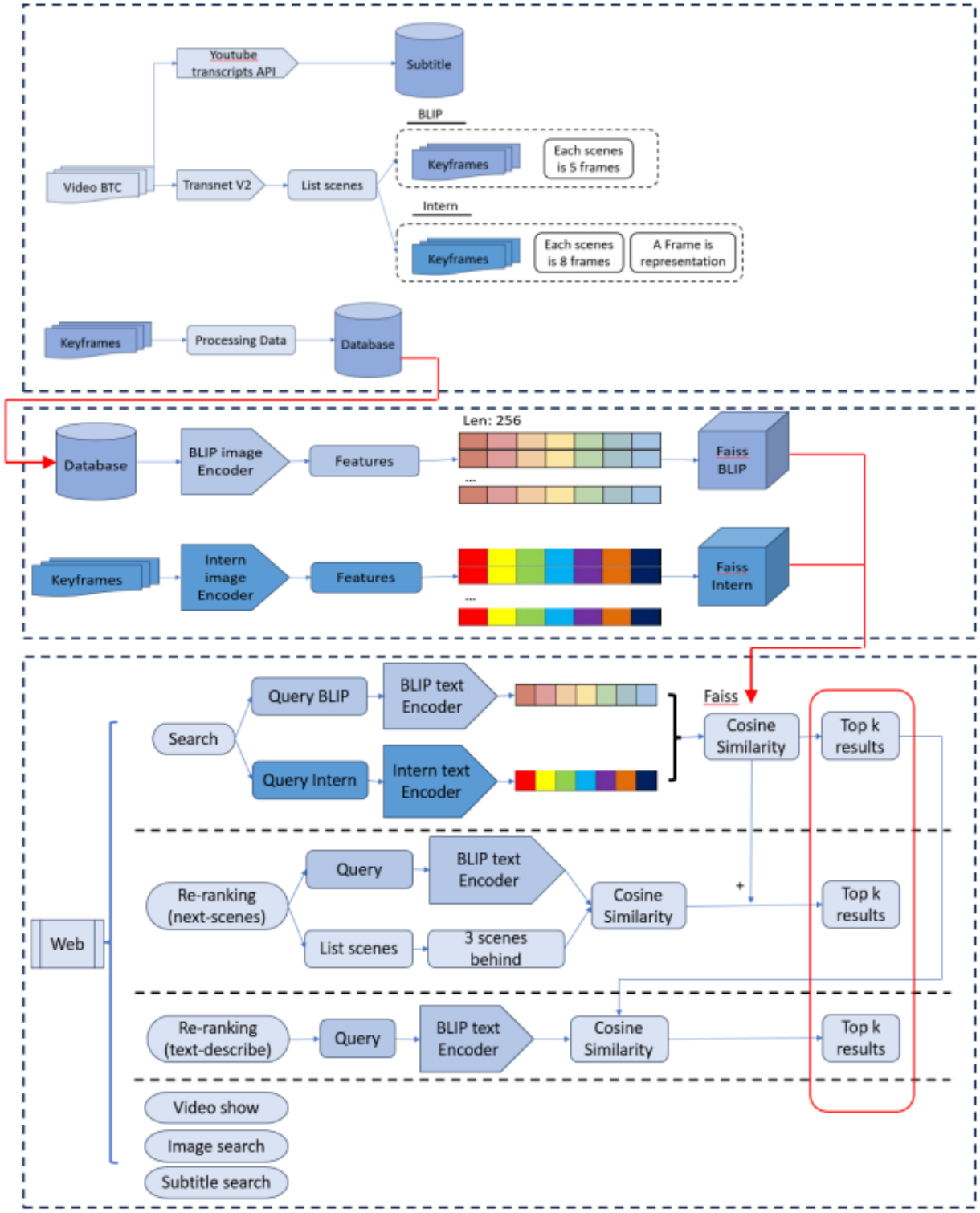
AI Challeng 2023

Team SPK_Sandbox

1. Baseline



2. Mô tả giải pháp

Giải pháp của nhóm gồm 3 giai đoạn như sau:

+ Processing Data

+ Encoding keyframes

+ Building a retrieval system

2.1 Processing Data

Ngoài keyframes do BTC cung cấp, nhóm cũng lấy thêm keyframes khác từ video bằng cách chia các video thành nhiều khung cảnh(scenes), sau đó mỗi scenes nhóm sẽ lấy 5 keyframes trong đó để tạo thành dataset của nhóm. Đối với model Intern, nhóm sẽ lấy 8 keyframes để encode, sau đó lấy 1 keyframes làm đại diện cho 8 keyframes.

Để có thể chia video thành các scenes, mà mỗi scenes là thông tin về một khung cảnh trong video. Nhóm sử dụng model Transnet V2. Input sẽ là video và kết quả trả về được các keyframes, cũng như là thông tin về các scenes có trong video.

Vì lấy cố định 5 keyframes mỗi scenes nên đối với các scenes ngắn sẽ cho ra nhiều tấm ảnh trùng nhau. Vì thế nhóm sẽ lọc bớt dataset bằng model CleanVision nhằm giảm bớt các ảnh trùng, ảnh không có thông tin, ảnh bị tối …. Ngoài ra nhóm cũng sẽ nhìn lại data và xóa bớt các ảnh không mang nhiều ý nghĩa như các ảnh giới thiệu 60s, ảnh MC đang nói, …

Cuối cùng là nhóm sẽ tiến hành lấy phụ đề của các video, đối với các video không có phụ đề, nhóm dùng model ASR của *vietai* để lấy được âm thanh từ video.

2.2 Encoding keyframes

Sau khi xử lý xong data, nhóm tiến hành encode để lấy được các vector đặc trưng của các keyframes và lưu chúng lại. Nhóm dùng model BLIP (Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation) là mô hình được huấn luyện bởi cặp dữ liệu hình ảnh - văn bản. BLIP khác và đặc biệt ở chỗ có khả năng tận dụng bộ dữ liệu khổng lồ và kết hợp các kiến trúc chuyên môn trong việc “hiểu” hình ảnh-văn bản và tạo sinh. Vì nó có khả năng xử lý những tác vụ Vision-Language.

Sau khi có và lưu được các vector đặc trưng của keyframes. Ta cho văn bản cần truy vấn vào model, sẽ được vector có độ dài là 256. Dùng cosine similarity ta sẽ cho ra được các ảnh có độ tương thích với văn bản. Việc sử dụng cosine similarity cũng có một ưu điểm là nó sẽ giúp mình truy vấn được các keyframes chuyển cảnh (mức contrast thấp).

Để tăng thời gian truy vấn, nhóm sử dụng Faiss (Facebook AI Similarity Search), là một thư viện mã nguồn mở được phát triển bởi Facebook AI Research (FAIR) để hỗ trợ việc tìm kiếm và đánh chỉ mục các dữ liệu lớn dựa trên sự tương đồng. Thư viện này được thiết kế đặc biệt để xử lý các nhiệm vụ tìm kiếm tương đồng nhanh chóng và hiệu quả trên các bộ dữ liệu lớn, chẳng hạn như hình ảnh, văn bản, âm thanh, và các dạng dữ liệu khác. Faiss có nhiều hình thức tìm kiếm, trong đó có cosine similarity, vì thế nhóm quyết định sử dụng Faiss.

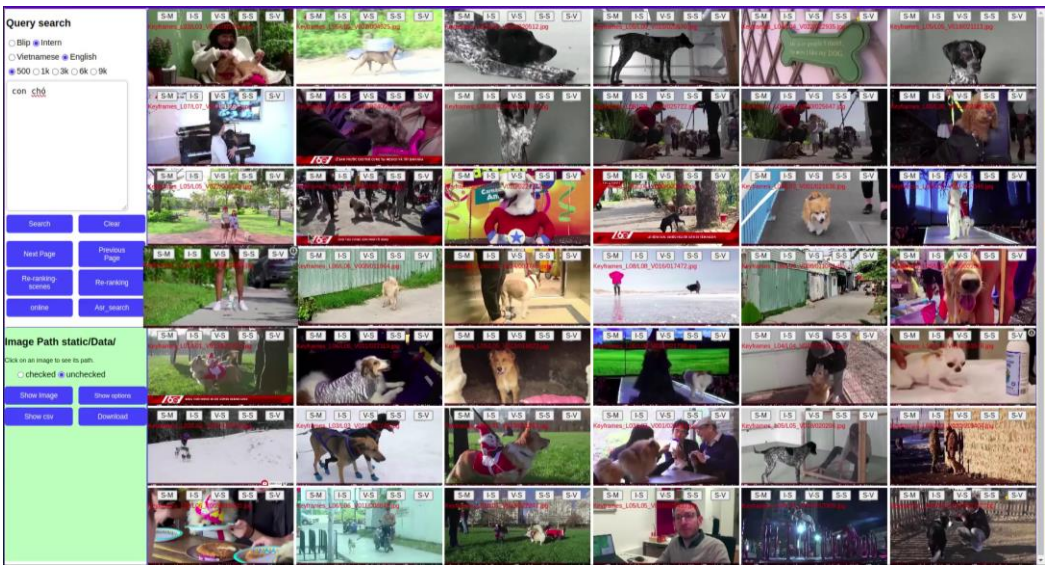
Các vector đặc trưng của ảnh sẽ được thành file.bin. Việc này sẽ giúp cho hệ thống truy vấn tiết kiệm rất nhiều thời gian. Nhóm sẽ lưu thành hai file.bin cho BLIP và cho Intern

Ngoài ra, nhóm cũng tận dụng thêm thông tin từ data đó là Speech (giọng nói) để hỗ trợ cho hệ thống truy vấn. Mục đích của hệ thống này là giúp có thể tổng hợp hoặc tìm kiếm các ảnh có các nguồn thông tin mà hai model trên không thể tìm ra như là các tên riêng của khu vực, quốc gia, và một số thông tin không phổ biến khác. Kết quả là trả về được các keyframes có thông tin như đã được đề cập. Lúc này nhóm sẽ sử dụng BLIP để tiếp tục tìm kiếm trên kết quả vừa trả về.

Để lấy được thông tin Speech. Nhóm lấy phụ đề từ các video trên youtube, trong phụ đề này sẽ có cả thông tin về thời gian, từ đó nhóm có thể truy ngược lại vị trí của frames. Đối với các video không có phụ đề, nhóm sẽ dùng model ASR để lấy âm thanh.

2.3 Building a retrieval system

Về hệ thống truy vấn, web của nhóm sẽ gồm các chức năng như sau:



- *Tìm kiếm bằng text query:* Người dùng nhập thông tin và text box, và sau đó bấm search để tìm kiếm. Có các options có thể chọn như chọn model search (BLIP hoặc Intern), có dịch sang tiếng anh (Vietnamese, English) hay không hay số lượng kết quả trả về (500, 1000, 3000, 6000, 9000)

- *Tìm kiếm bằng đường dẫn ảnh từ nguồn trên mạng:* Người dùng nhập đường dẫn vào và bấm online. Kết quả sẽ trả về ảnh có độ tương đồng với ảnh có đường dẫn vừa nhập

- *Re-ranking:* Sau khi search xong, có thể nhập câu text và tìm kiếm lại trên những kết quả vừa được trả về. Phù hợp với các ảnh có thể mô tả bằng nhiều hình thức. Được dùng bởi model BLIP

- *Re-ranking-scenes:* Sau khi search xong, có thể nhập câu text và sắp xếp lại kết quả với tiêu chí là tìm những ảnh nằm phía sau các ảnh trong kết quả có độ tương đồng gần với câu text vừa nhập. Kết hợp với giá trị cosine của lần tìm kiếm ban đầu để trả về các ảnh có độ tương đồng nhất với câu text có dạng <“mô tả”, chuyển cảnh là “mô tả”>. Được dùng bởi model BLIP

- *Tìm kiếm qua ASR search:* nhập câu text vào và ta có thể trả về các ảnh có xuất hiện từ đó trong file âm thanh. Kết quả trả về không cố định, thời gian trả về nhanh chậm tùy vào độ xuất hiện của từ đó. Vì thế chỉ nên dùng cho các trường hợp tên riêng, hoặc các câu đặc biệt. Sau khi có kết quả ta có thể search trên chúng bằng re-ranking.

- *Các nút bấm chuyển trang; show image, show options; clear:* Giúp chuyển trang để xem ảnh; hiển thị chỉ ảnh hoặc hiển thị các options có trên ảnh; xóa text box.

- *Các options trên mỗi tấm ảnh:* Hỗ trợ trong việc tìm kiếm và xác nhận lại độ chính xác của kết quả. Bao gồm thông tin đường dẫn của ảnh đó cùng với các chức năng như sau:

+S-M: submit frames được chọn, khi bấm "checked" ta sẽ submit toàn bộ ảnh có trong scenes của frames được chọn

+I-S: Tìm kiếm ảnh thông qua ảnh được chọn, trả về các ảnh có độ tương đồng

+V-S: Show tất cả frames thuộc 12 scenes phía trước và 12 scenes phía sau của frames được chọn

+S-S: show duy nhất frame được chọn, giúp nhìn rõ được bức hình

+S-V: show video của frame được chọn, bao gồm 9 video, gồm 4 video thuộc 4 scenes nằm trước và 4 video thuộc 4 scenes nằm sau frame được chọn.

- *Show file csv:* Show kết quả sau khi submit, có thể thêm hoặc xóa các cột, đặt tên file và tải file xuống.