Nhóm 8

BÁO CÁO TIẾN ĐỘ

1. Những việc đã làm được

- Hoàn thành các yêu cầu chính:
  + Tìm hiểu về mô hình CLIP.
  + Xây dựng chương trình cho phép đưa vào 1 câu truy vấn, tìm ra các ảnh trong 1 thư mục chứa đối tượng có trong câu truy vấn
  + Tạo CSDL ảnh và tập các câu truy vấn, đánh giá chương trình đã được xây dựng với độ đo độ chính xác và độ triệu hồi.
- Cài đặt thành công và sử dụng được mô hình CLIP
  + Sử dụng mã nguồn chính thức của OpenAI tại Colab Notebook gốc.
  + Dùng mô hình ViT-B/32 từ clip để encode văn bản và ảnh.
- Xây dựng được chương trình tìm kiếm ảnh theo truy vấn văn bản
  + Cho phép người dùng nhập một câu truy vấn.
  + Duyệt toàn bộ ảnh trong một thư mục.
  + So sánh vector embedding giữa truy vấn và từng ảnh bằng cosine similarity.
  + Trả về danh sách ảnh có độ tương đồng cao nhất.
- Tạo cơ sở dữ liệu ảnh và tập truy vấn
  + Thư mục ảnh bao gồm nhiều đối tượng khác nhau: mèo, chó, xe, trái cây...
  + Các truy vấn ví dụ:
    "a cat sitting on the floor"
    "a dog running"
    "a red car"
  + Mỗi ảnh được gắn nhãn để đánh giá kết quả sau truy vấn.
- Đánh giá độ chính xác (Precision) và độ triệu hồi (Recall)
  + Với mỗi truy vấn, so sánh ảnh trả về với nhãn thật.
  + Tính Precision = Số ảnh đúng / Tổng ảnh trả về.
  + Tính Recall = Số ảnh đúng / Tổng ảnh đúng có trong tập.

*** Kết quả đạt được**

- CLIP hoạt động rất tốt với các truy vấn mô tả trực tiếp như: "a cat", "a car", "a dog".
- Precision cao với những ảnh rõ ràng, ít bị nhiễu.
- Recall giảm nếu ảnh có nhiều chi tiết không liên quan hoặc mô tả quá mơ hồ.

2. Những khó khăn gặp phải

- Việc xử lý ảnh có kích thước lớn làm chậm quá trình tính embedding.
- Code khá dài vẫn chưa được tối ưu
- Kết quả khi truy vấn đôi khi chưa thực sự chính xác
- Cần thêm thời gian để tối ưu hóa tốc độ xử lý hàng loạt.

3. Câu hỏi

- Yêu cầu cụ thể của cô là phải thực hiện được khoảng bao nhiêu câu truy vấn ạ?
- Hiện tại thì khi truy vấn mô hình đưa ra top kết quả đúng với câu truy vấn, nhưng vì là top kết quả nên những kết quả như top 2 hoặc top 3 vẫn sai và không đúng chính xác với truy vấn, có cách nào để khắc phụ triệt để đều này không ạ?
- Làm thế nào để cải thiện kết quả truy vấn cho mô tả dài hoặc nhiều đặc điểm vậy ạ?
- Với truy vấn trong ảnh có nhiều đối tượng thì làm thế nào để khoanh vùng đối tượng ạ?
- Có nên thêm bộ lọc ngữ nghĩa để rút gọn câu truy vấn trước khi embedding không ạ?

4. Nội dung chính

4.1. Tìm hiểu về mô hình Clip (2đ)

- Khái niệm:

“CLIP là viết tắt của “Contrastive Language-Image Pre-Training”. Đây là một mô hình pre-training cho các bài toán liên quan đến xử lý ngôn ngữ và hình ảnh. Mục tiêu của CLIP là tạo ra một khả năng hiểu biết bề ngoài của thế giới từ việc đào tạo một mô hình học sâu trên tập dữ liệu lớn của văn bản và hình ảnh.”

- Khả năng:

+ **Phân loại hình ảnh:** CLIP có thể nhận diện và phân loại hình ảnh chỉ dựa trên mô tả văn bản. Điều đặc biệt ở đây là CLIP không cần được huấn luyện chuyên biệt cho từng loại hình ảnh mà vẫn có thể phân loại chúng chính xác. Điều này giúp mô hình trở nên linh hoạt và dễ dàng áp dụng trong nhiều tình huống khác nhau.

+ **Tìm kiếm đa phương tiện:** Một trong những điểm mạnh của CLIP là khả năng kết nối thông tin giữa văn bản và hình ảnh. Người dùng có thể mô tả bằng ngôn ngữ tự nhiên và CLIP sẽ tìm kiếm nội dung hình ảnh tương ứng. Điều này

rất hữu ích trong các ứng dụng tìm kiếm nội dung hoặc sáng tạo nội dung trực quan.
- Ứng dụng
+ Phân loại nội dung: Mô hình này được sử dụng trên các nền tảng trực tuyến để phân loại hình ảnh và video dựa trên nội dung.
Ví dụ, các trang mạng xã hội có thể áp dụng CLIP để nhận diện và gắn thẻ nội dung, giúp quản lý và sắp xếp dữ liệu hiệu quả hơn.
+ Gợi ý hình ảnh: Trong lĩnh vực thiết kế đồ họa và quảng cáo, CLIP hỗ trợ tìm kiếm và đề xuất các hình ảnh phù hợp với ý tưởng sáng tạo của người dùng. Điều này giúp rút ngắn thời gian tìm kiếm nội dung và cải thiện chất lượng sản phẩm.
+ Tìm kiếm thông minh: CLIP cho phép người dùng nhập mô tả văn bản để tìm kiếm hình ảnh hoặc video liên quan. Tính năng này đặc biệt hữu ích trong các công cụ tìm kiếm nội dung đa phương tiện, giúp người dùng dễ dàng tiếp cận thông tin trực quan một cách nhanh chóng và chính xác.

Dựa trên chương trình
https://colab.research.google.com/github/openai/clip/blob/master/notebooks/Interacting_with_CLIP.ipynb

4.2. Xây dựng chương trình cho phép đưa vào 1 câu truy vấn, tìm ra các ảnh trong 1 thư mục chứa đối tượng có trong câu truy vấn (6đ).

4.3. Tạo CSDL ảnh và tập các câu truy vấn, đánh giá chương trình đã được xây dựng với độ đo độ chính xác và độ triệu hồi. (2đ)