GIẢI PHÁP ZERO-SHOT VÀ TRAINING FREE CHO BÀI TOÁN COMPOSED IMAGE RETRIEVAL

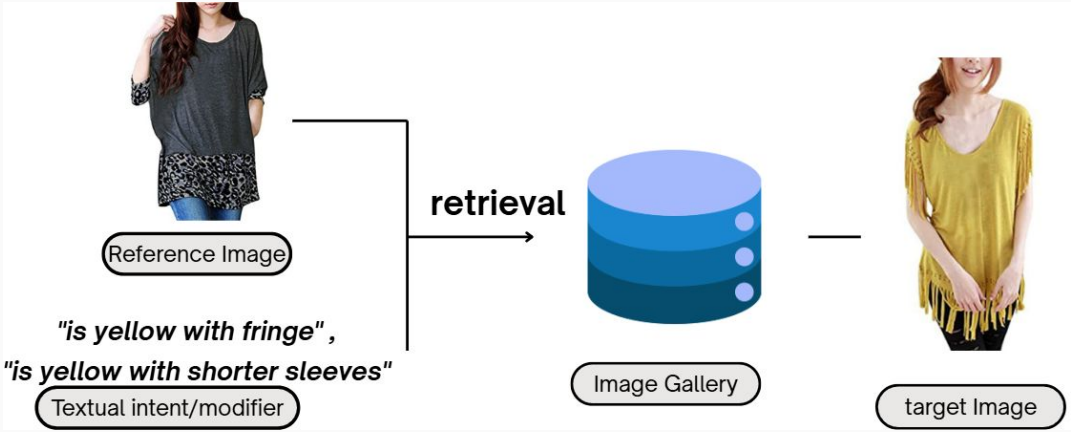
Lê Minh Quân - 240101065

Tóm tắt

- Lớp: CS2205.CH201
- Link Github của nhóm:
- Link YouTube video:
- Ảnh + Họ và Tên của các thành viên
- Tổng số slides không vượt quá 10

Giới thiệu

- CIR là bài toán truy hồi với đầu vào gồm (ảnh + văn bản) và đầu ra là ảnh vừa tương đồng với input image vừa có các chi tiết được miêu tả trong input text.
- Kết hợp ảnh và văn bản để đầu vào thể hiện *User need* chi tiết hơn.



Giới thiệu

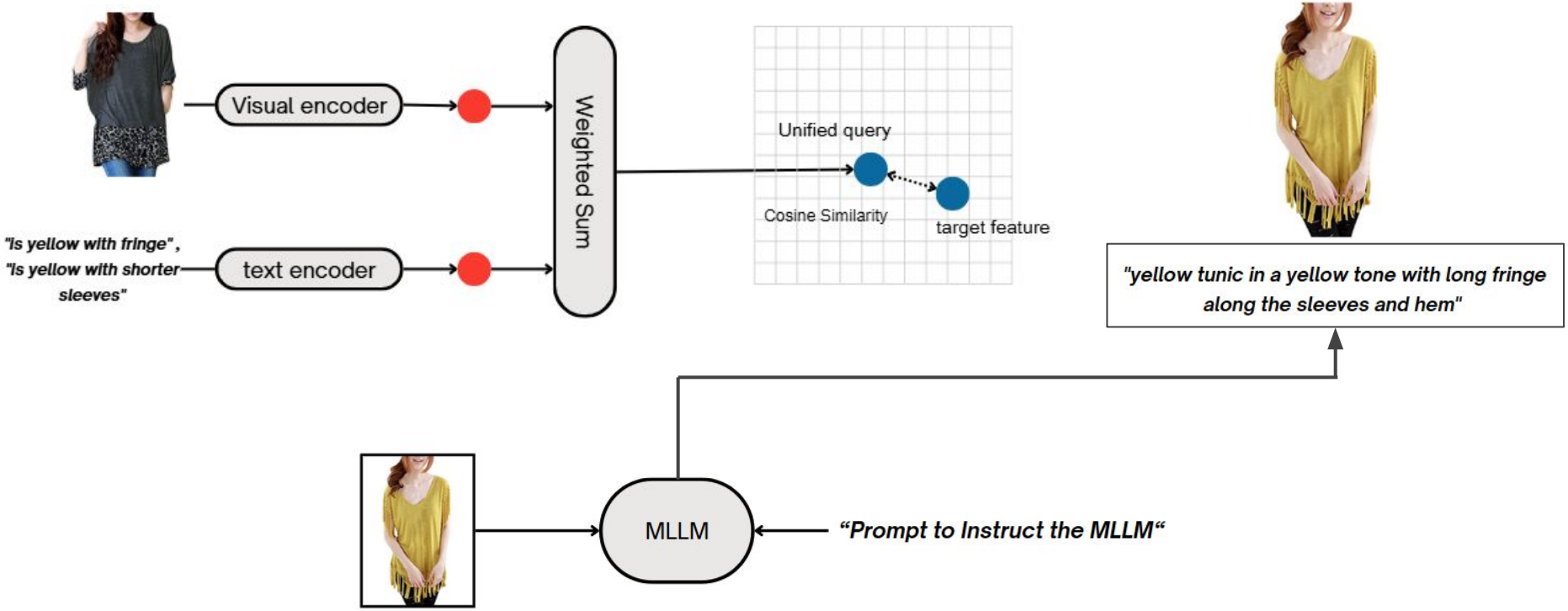
- Các phương pháp về CIR trước đây [2, 3] dùng supervised learning, huấn luyện với bộ dữ liệu dạng triplet.
- Điểm yếu của supervised CIR: (1) khó xây dựng triplet data; (2) Khó chuyển domain

Mục tiêu

- Xây dựng một hệ thống CIR có khả năng zero-shot, đạt kết quả cao hơn hoặc bằng với các hệ thống CIR trước w/o training/fine-tuning.
- Tăng cường hiệu suất truy hồi của hệ thống bằng cách áp dụng Multimodal Large Language model (MLLM).
- Thử nghiệm và đánh giá hệ thống trên các dữ liệu benchmark phổ biến được dùng cho CIR.

Nội dung và Phương pháp

Hệ thống đề xuất [1]:



Nội dung và Phương pháp

Dữ liệu:

- FashionIQ: 30.134 điểm dữ liệu từ 77.684 hình ảnh thời trang
- CIRR: 21.552 ảnh đa dạng thể loại

Khảo sát pre-trained models:

- VLMs cho feature extraction: CLIP [4], BLIP [5],...
- MLLMs để tạo caption cho ảnh: Gemini-2.0 [6], GPT-4o [7],...

Đánh giá:

- Recall@k: Chấm điểm cho những trường hợp hình ảnh có liên quan xuất hiện trong top K kết quả tìm kiếm.

Kết quả dự kiến

- Một pipeline training-free CIR, dễ triển khai và không yêu cầu bất cứ bước huấn luyện mô hình nào.
- Chứng minh được độ hiệu quả của phương pháp kết hợp đặc trưng ảnh và văn bản qua phép weighted average đơn giản.
- Một bảng đánh giá khoa học và chi tiết với các phương pháp đã có trên các bộ dữ liệu thông dụng như FashionIQ và CIRR

Tài liệu tham khảo

[1]. Wu, R.D., Lin, Y.Y. and Yang, H.F.: Training-free Zero-shot Composed Image Retrieval via Weighted Modality Fusion and Similarity. International Conference on Technologies and Applications of Artificial Intelligence. pp. 77-90 (2024)

[2]. Liu, Z., Rodriguez-Opazo, C., Teney, D., Gould, S.: Image retrieval on real-life images with pre-trained vision-and-language models. ICCV. pp. 2125–2134 (2021)

[3]. Baldrati, A., Bertini, M., Uricchio, T., Del Bimbo, A.: Effective conditioned and composed image retrieval combining clip-based features. CVPR. pp. 21466–21474 (2022)

[4] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. ICML. pp. 8748–8763 (2021)

[5]. Li, J., Li, D., Xiong, C., Hoi, S.: BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. ICML. pp. 12888 12900 (2022)

Tài liệu tham khảo

[6]. Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)

[7]. Liu, H., Li, C., Wu, Q. and Lee, Y.J: Visual instruction tuning. Advances in neural information processing systems, 36, pp.34892-34916 (2023)