CẢI THIỆN KHẢ NĂNG BIỂU DIỄN NGỮ NGHĨA CỦA HÌNH ẢNH TRONG MÔ HÌNH PIC2WORD BẰNG PHƯƠNG PHÁP SINH CHÚ THÍCH HÌNH ẢNH

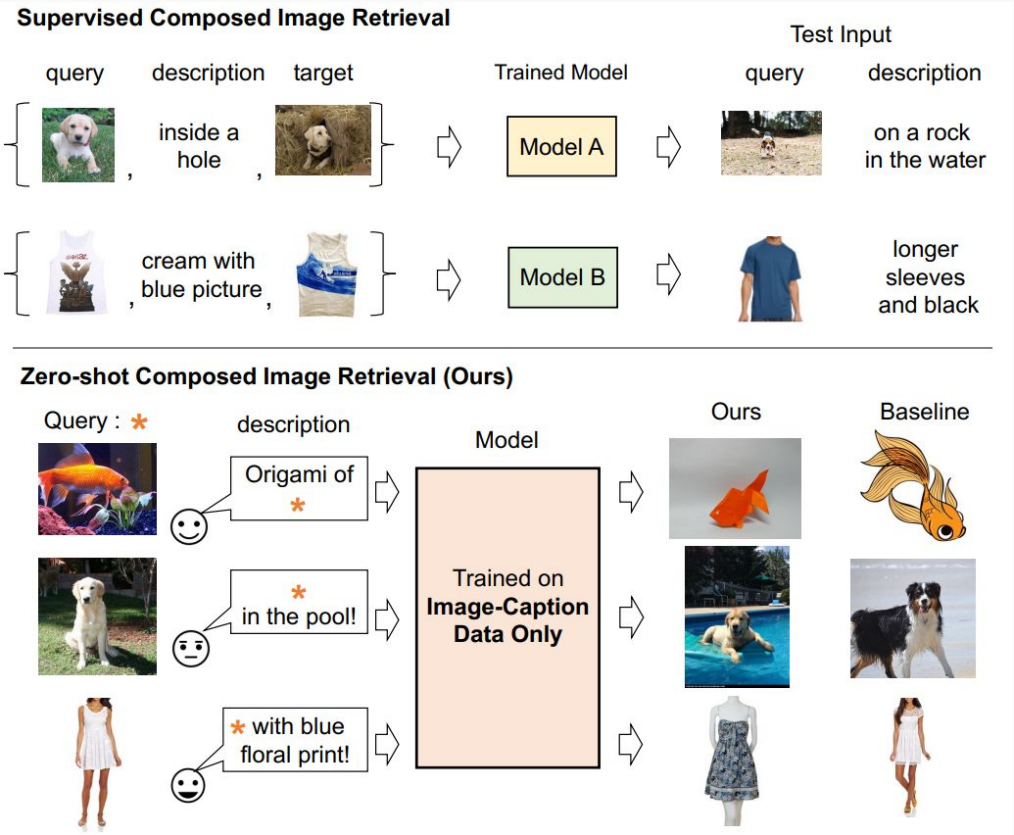
Nguyễn Trần Việt Anh - 21520006

Tóm tắt

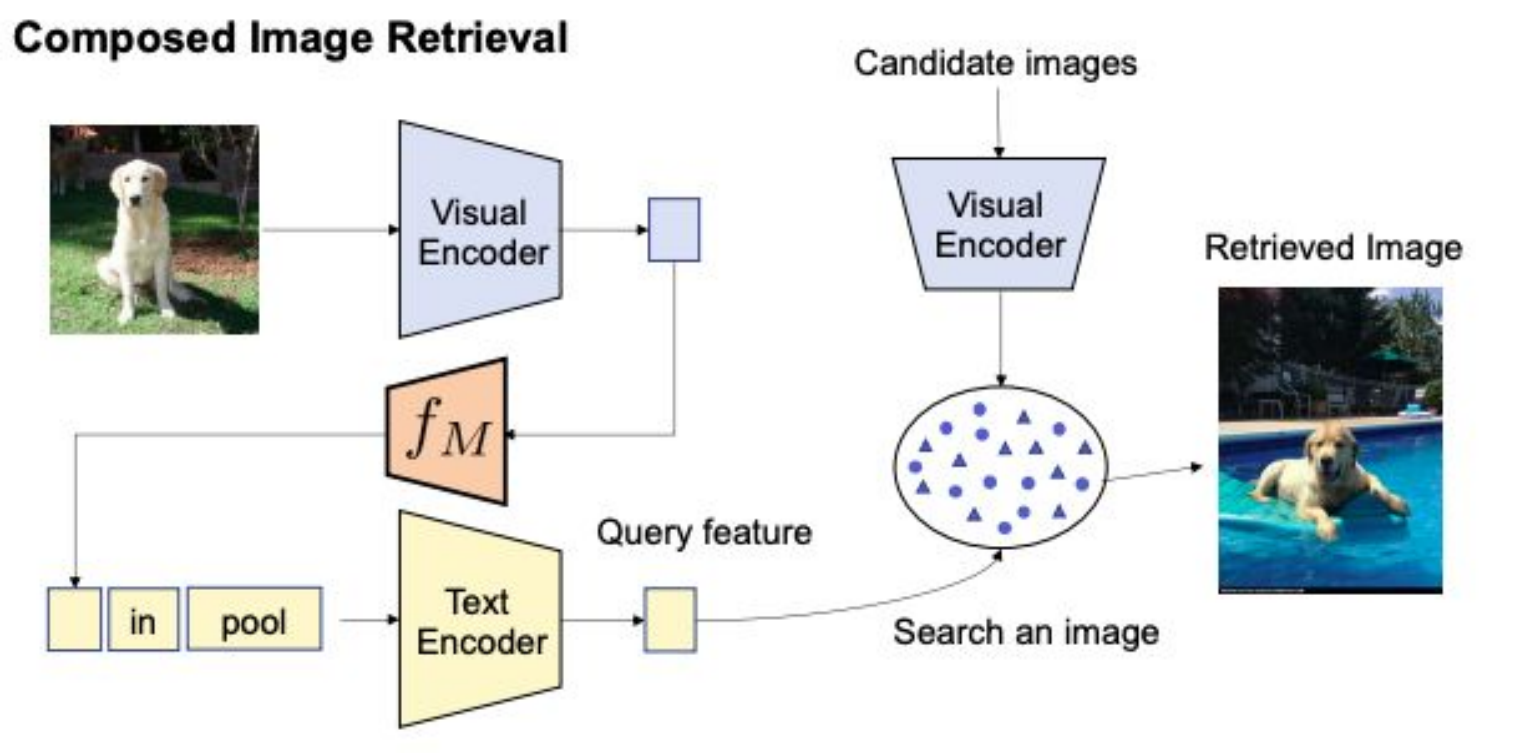
- Lớp: CS519.O11
- Link Github của nhóm: https://github.com/anh-ngn/CS519.O11
- Link YouTube video:
- Thành viên:

Nguyễn Trần Việt Anh - 21520006

Giới thiệu


Supervised Composed Image Retrieval
query
description
target
Trained Model
Test Input
query
description
inside a hole
Model A
on a rock in the water
cream with blue picture
Model B
longer sleeves and black
Zero-shot Composed Image Retrieval (Ours)
Query : *
description
Origami of *
* in the pool!
* with blue floral print!
Model
Trained on Image-Caption Data Only
Ours
Baseline

Giới thiệu



=> Hạn chế: chỉ biểu diễn mỗi hình ảnh bằng 1 token duy nhất.

Mục tiêu

- Nghiên cứu các mô hình Pic2Word, CLIP, mPLUG, OFA, GIT, InSPyReNet và IS-Net tìm hiểu cách cài đặt, thử nghiệm các mô hình để đánh giá hiệu quả của các mô hình.
- Xây dựng mô hình đã đề xuất, đánh giá mô hình trên các tập dataset khác nhau và so sánh mô hình với các mô hình hiện có để đánh giá mô hình và tìm hướng cải thiện.
- Xây dựng một hệ thống truy vấn hình ảnh hỗn hợp dựa trên mô hình được đề xuất.

Nội dung và Phương pháp

Giai đoạn 1: Nghiên cứu các mô hình Pic2Word, CLIP, mPLUG, OFA, InSPyReNet và IS-Net.

Phương pháp thực hiện:

- Nghiên cứu các mô hình Pic2Word, CLIP, mPLUG, OFA, GIT, InSPyReNet và IS-Net.
- Phân tích, đánh giá, so sánh các mô hình Image Captioning mPLUG, OFA và GIT..
- Phân tích, đánh giá, so sánh 2 mô hình InSPyReNet và IS-Net.

Kết quả dự kiến:

- Hiểu rõ các mô hình, nắm rõ được cách cài đặt của các mô hình.
- Phân tích ưu, nhược điểm giữa các mô hình.

Nội dung và Phương pháp

Giai đoạn 2: Cài đặt và đánh giá mô hình.

Phương pháp thực hiện:

- Cài đặt mô hình được đề xuất với các phương pháp khác nhau.
- Thử nghiệm các phương pháp huấn luyện, hàm mục tiêu khác nhau.
- Huấn luyện mô hình và đánh giá mô hình trên nhiều độ đo, dataset khác nhau. Từ đó rút ra được các điểm mạnh, điểm yếu và hướng cải thiện của mô hình.

Kết quả dự kiến:

- Cài đặt thành công mô hình.
- Đánh giá tính hiệu quả của mô hình được đề xuất, so sánh kết quả với các mô hình hiện có.

Nội dung và Phương pháp

Giai đoạn 3: Xây dựng hệ thống truy vấn thông tin từ mô hình đã cài đặt.

Phương pháp thực hiện:

- Xây dựng một hệ thống truy vấn hình ảnh hỗn hợp (CIR) trên nền tảng web, để có thể đánh giá mô hình một cách trực quan.

Kết quả dự kiến:

- Xây dựng thành công một hệ thống truy vấn ảnh hỗn hợp (CIR) trên nền tảng web.

Kết quả dự kiến

- Mã nguồn mô hình được đề xuất, bao gồm chú thích và tài liệu hướng dẫn sử dụng.
- Báo cáo về phương pháp, kĩ thuật đã được sử dụng trong nghiên cứu, kết quả mà mô hình đã đạt được trên các dataset và các độ đo khác nhau.
- Mã nguồn hệ thống truy xuất hình ảnh hỗn hợp (CIR) trên nền tang web với mô hình được đề xuất.

Tài liệu tham khảo

- [1]. Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Effective conditioned and composed image retrieval combining clip-based features. In CVPR, pages 21466–21474, 2022
- [2]. Kuniaki Saito, Kihyuk Sohn , Xiang Zhang , Chun-Liang Li , Chen-Yu Lee, Kate Saenko , Tomas Pfister. Pic2Word: Mapping Pictures to Words for Zero-shot Composed Image Retrieval.
- [3]. Kuniaki Saito, Kihyuk Sohn , Xiang Zhang , Chun-Liang Li , Chen-Yu Lee, Kate Saenko , Tomas Pfister. Pic2Word: Mapping Pictures to Words for Zero-shot Composed Image Retrieval.
- [4]. Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, Hehong Chen, Guohai Xu, Zheng Cao, Ji Zhang, Songfang Huang, Fei Huang, Jingren Zhou, Luo Si. mPLUG: Effective and Efficient Vision-Language Learning by Cross-modal Skip-connections

Tài liệu tham khảo

- [5]. Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, Hongxia Yang. OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework
- [6] JianFeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu. GIT: A Generative Image-to-text Transformer for Vision and Language.
- [7] Towards Data Science. Image Captioning in Deep Learning.
- [8] Taehun Kim, Kunhee Kim, Joonyeong Lee, Dongmin Cha, Jiho Lee, Daijin Kim. Revisiting Image Pyramid Structure for High Resolution Salient Object Detection.
- [9] Xuebin Qin, Hang Dai, Xiaobin Hu, Deng-Ping Fan, Ling Shao, and Luc Van Gool. Highly Accurate Dichotomous Image Segmentation
- [10] Xuebin Qin, Hang Dai, Xiaobin Hu, Deng-Ping Fan, Ling Shao, Luc Van Gool. Highly Accurate Dichotomous Image Segmentation.