


THÔNG TIN CHUNG CỦA NHÓM

- Link YouTube video của báo cáo (tối đa 5 phút):
<https://youtu.be/vykQfiSjowg?si=V6AQQhGq2YoEugKd>
- Link slides (dạng .pdf đặt trên Github của nhóm):
(ví dụ: <https://github.com/anh-ngn/CS519.O11/doc.pdf>)
- Mỗi thành viên của nhóm điền thông tin vào một dòng theo mẫu bên dưới
- Sau đó điền vào Đề cương nghiên cứu (tối đa 5 trang), rồi chọn Turn in

<ul style="list-style-type: none">• Họ và Tên: Nguyễn Trần Việt Anh• MSSV: 21520006 	<ul style="list-style-type: none">• Lớp: CS519.O11• Tự đánh giá (điểm tổng kết môn): 8.5/10• Số buổi vắng: 1• Số câu hỏi QT cá nhân: ??• Số câu hỏi QT của cả nhóm: ??• Link Github: https://github.com/anh-ngn/CS519.O11/• Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm:<ul style="list-style-type: none">○ Lên ý tưởng đề tài○ Viết báo cáo, slide, poster.○ Làm video YouTube
-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

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

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

IMPROVING THE SEMANTIC REPRESENTATION OF IMAGES IN PIC2WORD MODEL WITH IMAGE CAPTIONING

TÓM TẮT *(Tối đa 400 từ)*

Trong những năm gần đây, dưới thời đại của việc bùng nổ dữ liệu (information overload) và dữ liệu lớn (big data), nhu cầu truy vấn thông tin nói chung hay truy vấn hình ảnh nói riêng là một nhu cầu rất quan trọng trong cuộc sống của mỗi người chúng ta trong cuộc sống hiện đại.

Sự ra đời của mô hình CLIP với khả năng nổi bật trong việc giải quyết tới các bài toán liên quan tới mối quan hệ giữa hình ảnh và văn bản, đã mở ra một bước ngoặt lớn trong rất nhiều lĩnh vực, đặc biệt là lĩnh vực Information Retrieval.

Cùng với sự ra đời của bài toán Zero-shot Composed Image Retrieval (ZS-CIR), sự ra đời của mô hình Pic2Word đã giải quyết được thách thức về dữ liệu của bài toán Composed Image Retrieval (CIR). Tuy nhiên, mô hình Pic2Word vẫn còn một số hạn chế về mặt biểu diễn ngữ nghĩa của hình ảnh do mô hình chỉ biểu diễn mỗi hình ảnh bằng 1 token duy nhất.

Trong đề tài này, chúng tôi sẽ tập trung vào việc cải tiến mô hình Pic2Word thông qua việc sử dụng một mô hình Image Captioning để biểu diễn ảnh một cách đầy đủ về mặt ngữ nghĩa.

GIỚI THIỆU *(Tối đa 1 trang A4)*

Ngày nay, với sự phát triển mạnh mẽ của các thiết bị di động và máy ảnh số, cùng với sự bùng nổ của các nền tảng mạng xã hội, lượng ảnh chia sẻ trên internet đang ngày càng tăng lên một cách đáng kể. Điều này tạo nên một nguồn thông tin lớn và đa

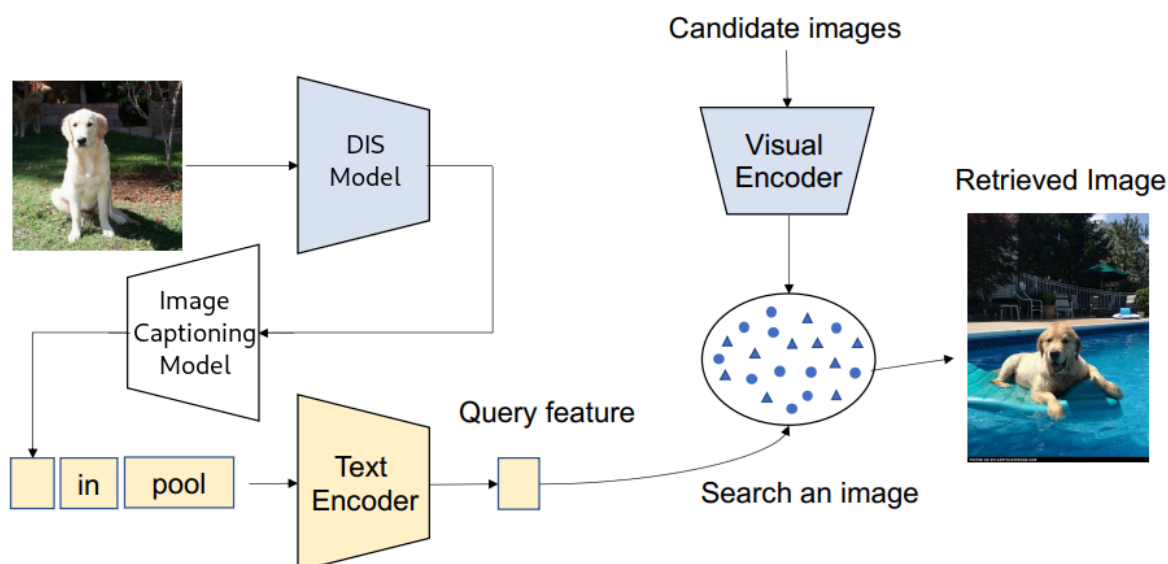
dạng, biến nhu cầu truy vấn hình ảnh trở thành một nhu cầu không thể thiếu trong cuộc sống hằng ngày của mỗi người chúng ta.

Nhằm đáp ứng nhu cầu này, rất nhiều bài toán đã được đặt ra, trong số đó có bài toán Composed Image Retrieval (CIR)[1] liên quan tới việc truy xuất hình ảnh dựa trên thông tin hỗn hợp bao gồm hình ảnh và văn bản. CIR yêu cầu một lượng lớn dữ liệu đã gán nhãn, gồm các bộ 3 (hình ảnh đầu vào, mô tả đầu vào, và hình ảnh mục tiêu). Tuy nhiên, tùy vào từng yêu cầu cụ thể của bài toán mà mô tả đầu vào sẽ khác nhau, do đó, việc thu thập dữ liệu sẽ rất tốn chi phí và giảm khả năng khái quát hoá. Để giải quyết những thách thức trên, bài toán Zero-shot Composed Image Retrieval (ZS-CIR)[2] được nhóm tác giả của mô hình Pic2Word[3] đề xuất nhằm loại bỏ việc thu thập các bộ 3 nhãn dữ liệu.

Mô hình Pic2Word đã mở ra một cách tiếp cận mới trong việc giải quyết bài toán ZS-CIR bằng việc ánh xạ một hình ảnh thành một token, sao cho token đó có thể được kết hợp linh hoạt và liên mạch vào mô tả đầu vào thông qua một mapping network được huấn luyện thông qua Contrastive Loss, sau đó so sánh mô tả đầu vào sau khi được kết hợp với hình ảnh với các hình ảnh trong database bằng Text Encoder và Image Encoder của mô hình CLIP.

Cùng với đó, các mô hình như mPLUG[4], OFA[5], và GIT[6] đã đạt được thành công trong Image Captioning[7] với độ chính xác cao và khả năng biểu diễn hình ảnh trong văn bản. Ngoài ra, các mô hình như InSPyReNet[8] và IS-Net[9] cũng đạt được kết quả rất tốt trong bài toán Dichotomous Image Segmentation (DIS)[10] với độ chính xác ấn tượng.

Trong đề tài này, chúng tôi đề xuất một mô hình được mở rộng từ mô hình Pic2Word, thay vì chỉ biểu diễn hình ảnh bằng 1 token duy nhất, chúng tôi sẽ sử dụng một mô hình Dichotomous Image Segmentation để xoá background của ảnh, loại bỏ các thông tin dư thừa của ảnh sau đó đưa vào một mô hình Image Captioning để trích xuất phần biểu diễn của hình ảnh đó dưới dạng một hoặc nhiều token. Cuối cùng, thông tin này sẽ được kết hợp với mô tả đầu vào và đưa vào Text Encoder của mô hình CLIP để so sánh với các hình ảnh sau khi được đưa vào Image Encoder.



(Hình 1: Kiến trúc tổng quát của mô hình được đề xuất.)

MỤC TIÊU

(Viết trong vòng 3 mục tiêu, lưu ý về tính khả thi và có thể đánh giá được)

1. 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.
2. 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.
3. Xây dựng một hệ thống truy vấn hình ảnh hỗn hợp.

NỘI DUNG VÀ PHƯƠNG PHÁP

(Viết nội dung và phương pháp thực hiện để đạt được các mục tiêu đã nêu)

Nội dung 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 để có thể hiểu rõ các mô hình, nắm rõ được cách cài đặt.
- Phân tích, đánh giá, so sánh 3 mô hình Image Captioning mPLUG, OFA và

GIT.

- Phân tích, đánh giá, so sánh 2 mô hình InSPyReNet và IS-Net.

Nội dung 2: Cài đặt và đánh giá mô hình.

Phương pháp thực hiện:

- Cài đặt mô hình được đề xuất với từng mô hình Image Captioning và Dichotomous Image Segmentation.
- 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.

Nội dung 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 MONG ĐỢI

(Viết kết quả phù hợp với mục tiêu đặt ra, trên cơ sở nội dung nghiên cứu ở trên)

- Xây dựng thành công mô hình được đề xuất, đạt được hiệu suất tốt.
- Cài đặt hoàn chỉnh mô hình được đề xuất.
- Viết báo cáo về phương pháp, quá trình thực hiện, kết quả đã đạt được.
- Xây dựng thành công hệ thống truy xuất hình ảnh hỗn hợp (CIR) với mô hình được đề xuất.

KẾ HOẠCH THỰC HIỆN:

- **Tuần 1-2:** Nghiên cứu các mô hình Pic2Word, CLIP, các mô hình truy xuất hình ảnh hỗn hợp (CIR) hiện có:

Kết quả dự kiến: Hiểu rõ các mô hình, nắm được cách cài đặt, đánh giá hiệu quả của các mô hình.

- **Tuần 3-5:** Nghiên cứu các mô hình mPLUG, OFA, GIT, InSPyReNet và IS-Net:

Kết quả dự kiến: Hiểu rõ các mô hình, nắm rõ được cách cài đặt, đánh

giá hiệu quả các mô hình và so sánh điểm mạnh, yếu của các mô hình.

- **Tuần 6-10:** Cài đặt mô hình được đề xuất, đánh giá kết quả và tìm hướng cải thiện mô hình.

Kết quả dự kiến: Cài đặt thành công mô hình, cải thiện mô hình để mô hình có thể đạt được kết quả tốt.

- **Tuần 11-13:** Đánh giá mô hình được đề xuất trên các dataset và độ đo khác nhau, phân tích những điểm yếu, cải thiện mô hình.

Kết quả dự kiến: Đánh giá tính hiệu quả của mô hình được đề xuất, so sánh kết quả với các mô hình khác và cải thiện mô hình nếu có thể.

- **Tuần 14-15:** Xây dựng hệ thống truy xuất hình ảnh hỗn hợp dựa trên mô hình đã cài đặt.
- **Tuần 16-17:** Viết báo cáo kết quả đã đạt được, phương pháp đã thực hiện.

TÀI LIỆU THAM KHẢO (*Định dạng DBLP*)

- [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
- [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.