THÔNG TIN CHUNG CỦA NHÓM

- Link Youtube video báo cáo:

  https://www.youtube.com/watch?v=xCVeW2S7xGY

- Link slides:

  https://github.com/chauminhnguyen/CS2205.APR2023/blob/master/Nguyen_Minh_Chau_xCS519.DeCuong.FinalReport.AIO.pdf

• Họ và Tên: Nguyễn Minh Châu
• MSSV: 220101028
 
	• Lớp: CS2205.CH1702
	• Tự đánh giá: 9
	• Số buổi vắng: 0
	• Số câu hỏi QT cá nhân:
	• Số câu hỏi QT của cả nhóm:
	• Link Github:
	• Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm:
	  ○ Lên ý tưởng
	  ○ Viết báo cáo
	  ○ Làm video Youtube

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI

M²: MÔ HÌNH TRANSFORMER TÍCH HỢP BỘ NHỚ CẶP

TÊN ĐỀ TÀI TIẾNG ANH

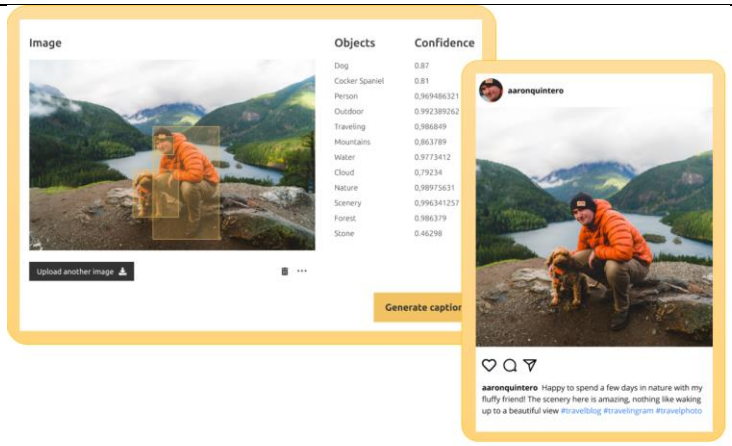
M²: MESHED-MEMORY TRANSFORMER

TÓM TẮT

Bài toán mô tả ảnh liên quan đến việc sử dụng ngôn ngữ tự nhiên để mô tả nội dung của một bức ảnh. Để giải quyết bài toán này, Thị giác máy tính và Xử lý ngôn ngữ tự nhiên phải được kết hợp. Chúng tôi đề xuất sử dụng mô hình Transformer, một mô hình đã chứng tỏ hiệu quả của nó trong cả hai lĩnh vực, để xây dựng mô hình M^2 - Meshed-Memory Transformer cho bài toán mô tả ảnh. Quá trình giải quyết bắt đầu bằng việc trích xuất vector ảnh từ các đối tượng từ ảnh ban đầu bằng một mô hình phát hiện đối tượng. Mô hình M^2 học được cách liên kết giữa hai vector ảnh rút trích từ mô hình phát hiện đối tượng và các vector văn bản trong quá trình huấn luyện. Điều này cho phép đưa cả hai loại biểu diễn này vào cùng một không gian vector, tạo điều kiện thuận lợi cho việc mô hình học cách mô tả ảnh. Chúng tôi dự tính sẽ so sánh mô hình M^2 Transformer trên bộ dữ liệu mô tả ảnh có tên COCO Captions và nocaps.

GIỚI THIỆU

Bài toán mô tả ảnh với đầu vào là bức ảnh và đầu ra là một chuỗi văn bản miêu tả nội dung có trong bức ảnh đó một cách tự động. Việc phát triển các giải pháp cho bài toán mô tả ảnh nhằm giúp các mạng xã hội có thể xác định nội dung ảnh mà người dùng đăng tải, từ đó giải quyết được các vấn đề về người dùng đăng tải ảnh với thông tin sai sự thật, hoặc nội dung không phù hợp với chuẩn mực xã hội. Hay bài toán mô tả ảnh góp phần giúp người dùng mạng dễ tìm kiếm được hình ảnh phù hợp cho nhu cầu của họ thông qua chỉ vài dòng gõ phím, đồng thời giúp người dùng đa ngôn ngữ thông qua khả năng mô tả đa ngôn ngữ.



Mô hình mô tả ảnh (trái) tự động gắn tag (phải) cho ảnh trên instagram (Source: businesswaretech)

Mô tả ảnh là bài toán sử dụng nội dunghình ảnh và ngôn ngữ tự nhiên để biểu diễn nội dung của bức ảnh đầu vào, được xem là bài toán khó có sư kết hợp của cả hai lĩnh vực. Những năm trước đây, các nghiên cứu về bài toán mô tả ảnh sử dụng những mô hình CNN để lấy đặc trưng của ảnh và mô hình RNN như LSTM hay GRU để rút trích đặc trưng ảnh. Tuy nhiên, việc xử lý một các rời rạc như vậy sẽ tạo nên tính không thống nhất giữa hai đặc trưng ảnh và văn bản, từ đó dẫn đến hiệu quả kém trong bài toán cần sự liên kết chặt chẽ nhưng bài toán mô tả ảnh.

Gần đây, Transformer [1] đã có những bước tiến vượt bậc trong cả hai lĩnh vực Thị giác máy tính như mô hình Vision Transformer [2] hay các mô hình Bert và GPT-2 bên Xử lý ngôn ngữ tự nhiên [3, 4]. Trong nghiên cứu này, chúng tôi nghiên cứu mô hình Transformer trong bài toán mô tả ảnh là sự kết hợp của hai lĩnh vực trên.

Input: Ảnh đầu vào.

Output: Chuỗi văn bản miêu tả nội dung của bức ảnh đầu vào.

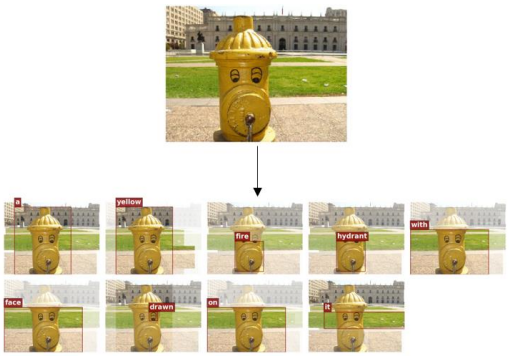
MỤC TIÊU

- Nghiên cứu về mô hình Transformer và các biến thể khác của mô hình này.
- Xây dựng một mô hình biến thể của Transformer phù hợp nhất để phục vụ cho bài toán mô tả ảnh.
- Thí nghiệm mô hình biến thể trên bộ dữ liệu mô tả ảnh nổi tiếng COCO Captions và nocaps đồng thời so sánh các mô hình khác.

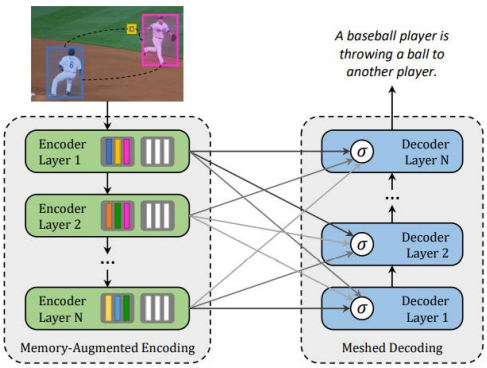
NỘI DUNG VÀ PHƯƠNG PHÁP

1. Nội dung:

- Sử dụng bộ dữ liệu mô tả ảnh nổi tiếng COCO Captions và nocaps để huấn luyện mô hình và đánh giá với các phương pháp khác [1, 5, 6].
- Sử dụng mô hình trích xuất đặc trưng để trích xuất các đối tượng trong ảnh.
- Sử dụng mô hình dựa trên Transformer [1] để xử lý hai tác vụ: Nhận và chuyển Vector đối tượng trong ảnh vào không gian của Transformer bằng mô-đun mã hóa (Encoder) và từ các Vector của Mô-đun mã hóa để tạo mô tả bằng Mô-đun giải mã (Decoder).
- Mô hình Transformer hoạt động theo cách tính toán song song, không tính toán từ tuần tự từng chữ một như các mô hình RNN nên mô hình có thể tiềm năng trong lĩnh vực cho ảnh. Ở mô hình M^2 Transformer sẽ bao gồm hai phần Mô-đun mã hóa (Encoder) và Mô-đun giải mã (Decoder) sẽ được kết nối chéo với nhau, sự liên kết giữa ảnh và văn bản được thể hiện rõ hơn qua trọng số của các kết nối chéo giữa các mô đun xử lý ảnh (Encoder) và mô đun xử lý văn bản (Decoder) với nhau, nhằm giúp mô hình M^2 Transformer xử lý bài toán mô tả ảnh có độ hiệu quả tốt hơn.



Mô hình Faster RCNN rút trích
các vector vật thể trong ảnh đầu vào



Tổng quan về mô hình M^2 Transformer
trong bài toán mô tả ảnh nhận đầu vào
là các vector vật thể trong ảnh đầu vào

2. Phương pháp:

- Thu thập và tìm hiểu về bộ dữ liệu lớn COCO Captions và nocaps để huấn luyện và đánh giá mô hình.
- Tìm hiểu về mô hình trích xuất đặc trưng đối tượng của ảnh, đầu ra của mô hình trích xuất đặc trưng ảnh sẽ là đầu vào của mô hình M^2 Transformer.
- Tìm hiểu về mô hình Transformer [1], mô-đun mã hóa và giải mã.
- Huấn luyện mô hình M^2 Transformer trên hai bộ dữ liệu COCO Captions, nocaps và đánh giá sử dụng độ đo BLEU, METEOR, ROUGE, CIDEr, SPICE.
- Tiến hành tinh chỉnh mô hình và đánh giá chất lượng đầu vào và đầu ra của hệ thống thông qua độ hiệu quả tập dữ liệu.
- Xây dựng giao diện người dùng và triển khai hệ thống tự động mô tả ảnh trên một nền tảng phù hợp.

KẾT QUẢ MONG ĐỢI

1. Bản báo cáo về phương pháp M^2 Transformer, kết quả thực nghiệm, và kết quả so sánh giữa các mô hình mô tả ảnh khác.
2. Chương trình minh họa cho người dùng phục vụ việc dễ dàng tương tác.

KẾ HOẠCH THỰC HIỆN

- Tuần 1 – 6:
  - Tìm hiểu mô hình Transformer, cụ thể về hai mô-đun mã hóa và giải mã để giải quyết hai vấn đề về rút trích đặc trưng hình ảnh đầu vài và sinh mô tả cho ảnh đầu vào.
  - Kết quả dự kiến:
    - Tài liệu khảo sát về mô hình Transformer và các biến thể của các mô hình Transformer.
    - Tài liệu về các dữ liệu lớn về mô tả ảnh, như COCO Captions, nocaps.
    - Tài liệu về độ đo liên quan bài toán mô tả ảnh, cụ thể CIDEr, SPICE.
- Tuần 7 – 12:

- Huấn luyện mô hình Transformer truyền thống, các biến thể Transformer và mô hình M2 Transformer dựa trên tập dữ liệu COCO Captions và nocaps.
- Kết quả dự kiến:
  - Bảng kết quả so sánh giữa các mô hình truyền thống Transformer, các biến thể Transformer và mô hình M2 Transformer.

- Tuần 13 – 16:
  - Xây dựng chương trình demo để dễ dàng thực thi.
  - Kết quả dự kiến:
    - Chương trình minh họa

TÀI LIỆU THAM KHẢO

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin:Attention Is All You Need. CoRR abs/1706.03762 (2017)

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby:An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR 2021

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova:BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL-HLT (1) 2019: 4171-4186

[4] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, “Language Models are Unsupervised Multitask Learners”, 2020

[5] Lun Huang, Wenmin Wang, Jie Chen, Xiaoyong Wei:Attention on Attention for Image Captioning. ICCV 2019: 4633-4642

[6] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, Lei Zhang:Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. CVPR 2018: 6077-6086