BÁO CÁO ĐỒ ÁN CUỐI KỲ

Môn học: CS2205.RM - PHƯƠNG PHÁP NCKH

Lớp: CS2205.CH1702 - APR2023

GV: PGS.TS. Lê Đình Duy

Trường ĐH Công Nghệ Thông Tin, ĐHQG-HCM

UIT-HCM

M^2: MESHED-MEMORY TRANSFORMER cho Bài toán mô tả ảnh

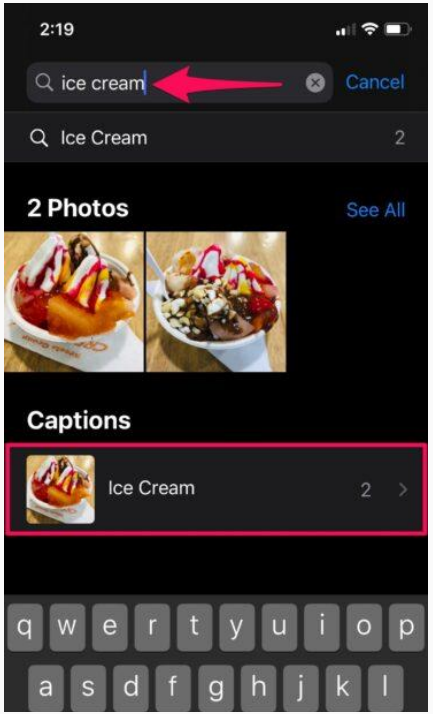
Nguyễn Minh Châu - 220101028

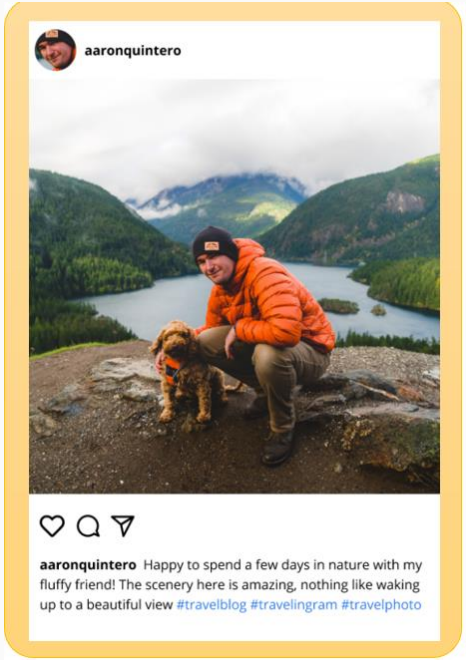
Tóm tắt

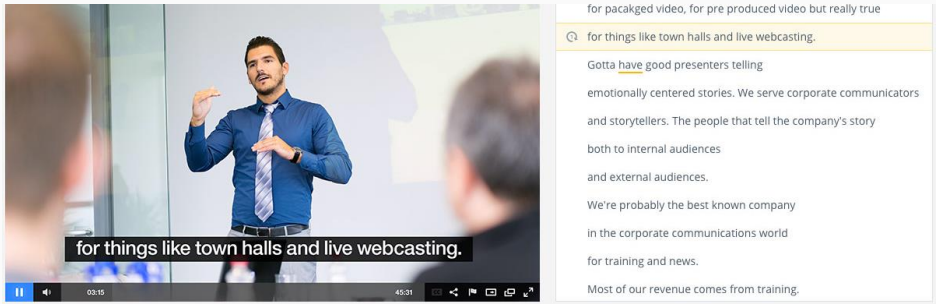
- Họ tên: Nguyễn Minh Châu
- MSHV: 220101028
- Link Github của nhóm: https://github.com/chauminhnguyen/CS2205.APR2023.git

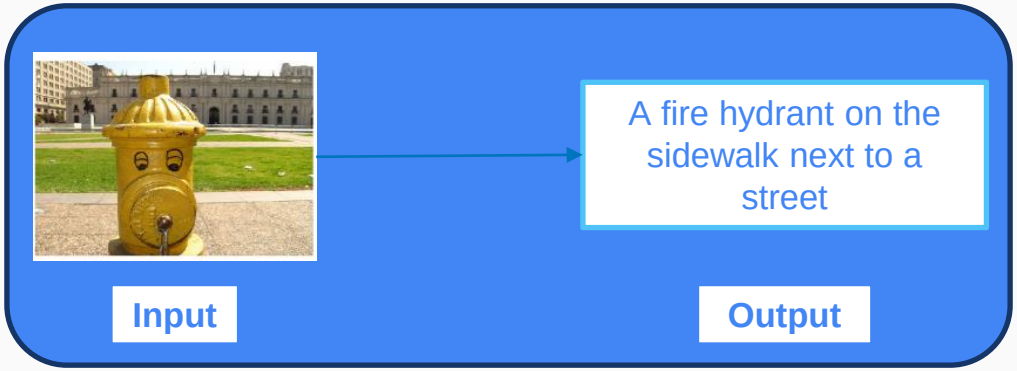
Giới thiệu

Giới thiệu bài toán









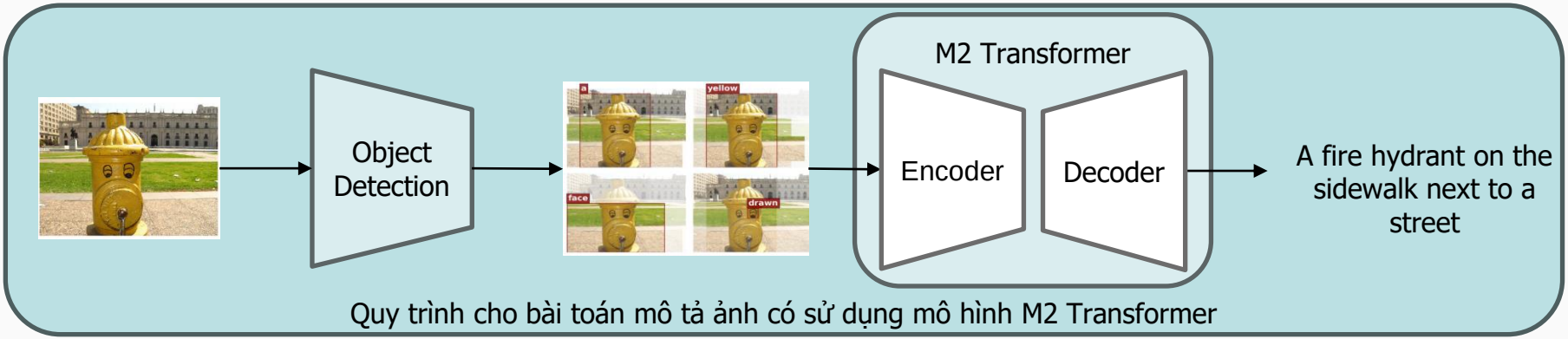
Mục tiêu

- Nghiên cứu kiến trúc Transformer và các biến thể của kiến trúc này nhằm **cải thiện** để phục vụ cho bài toán **mô tả ảnh**.
- **Thiết kế** mô hình M2 Transformer **dựa trên mô hình Transformer** để phục vụ cho bài toán mô tả ảnh
- **Huấn luyện** mô hình M2 Transformer và đánh giá với các mô hình mô tả ảnh khác.

Nội dung và Phương pháp

Nội dung

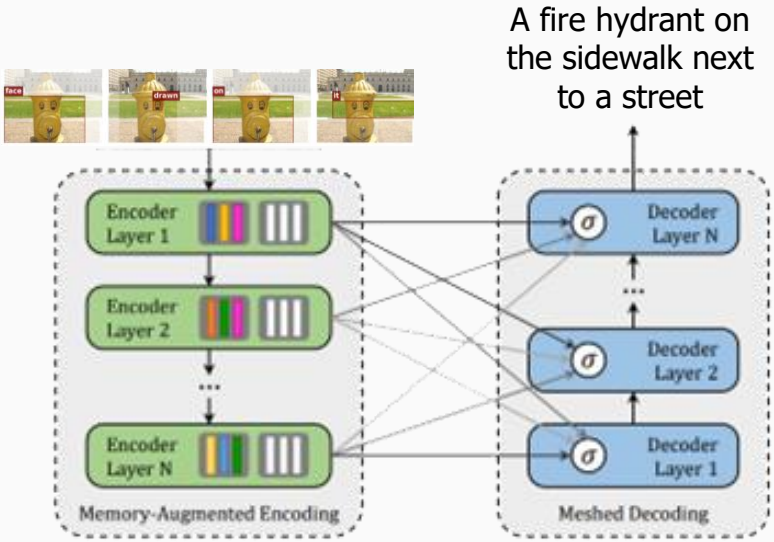
- Sử dụng bộ dữ liệu mô tả ảnh nổi tiếng COCO Captions và nocaps để huấn luyện mô hình và đánh giá với các phương pháp khác.
- Sử dụng mô hình trích xuất đặc trưng để trích xuất các đối tượng trong ảnh.
- Sử dụng mô hình dự trên Transformer là M2 Transformer để xử lý hai tác vụ: Nhận và chuyển vector 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).



Quy trình cho bài toán mô tả ảnh có sử dụng mô hình M2 Transformer

Nội dung và Phương pháp

- Ở mô hình M2 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 M2 Transformer xử lý bài toán mô tả ảnh có độ hiệu quả tốt hơn.



Mô hình M2 Transformer

Nội dung và Phương pháp

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 M2 Transformer.
- Tìm hiểu về mô hình Transformer, mô-đun mã hóa và giải mã.
- Huấn luyện mô hình M2 Transformer trên hai bộ dữ liệu COCO Captions, nocaps và đánh giá dựa trên 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 dự kiến

- Bản báo cáo về phương pháp M2 Transformer, kết quả thực nghiệm, và kết quả so sánh giữa các mô hình mô tả ảnh khác.
- Chương trình giao diện người dùng phục vụ việc dễ dàng tương tác.

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