TĂNG CƯỜNG MÔ HÌNH HÌNH ẢNH-NGÔN NGỮ CHO HỎI ĐÁP TRỰC QUAN VỚI TÀI NGUYÊN HẠN CHẾ THÔNG QUA HỌC DỰA TRÊN CÂU GỢI Ý.

Trần Vĩnh Khiêm - 220101041

UIT-HCM

Tóm tắt

- Lớp: CS2205.APR2023
- Link Github: https://github.com/kvt0012/CS2205.APR2023
- Link YouTube video: https://youtu.be/u470_0VtTi4
- Ảnh + Họ và Tên:

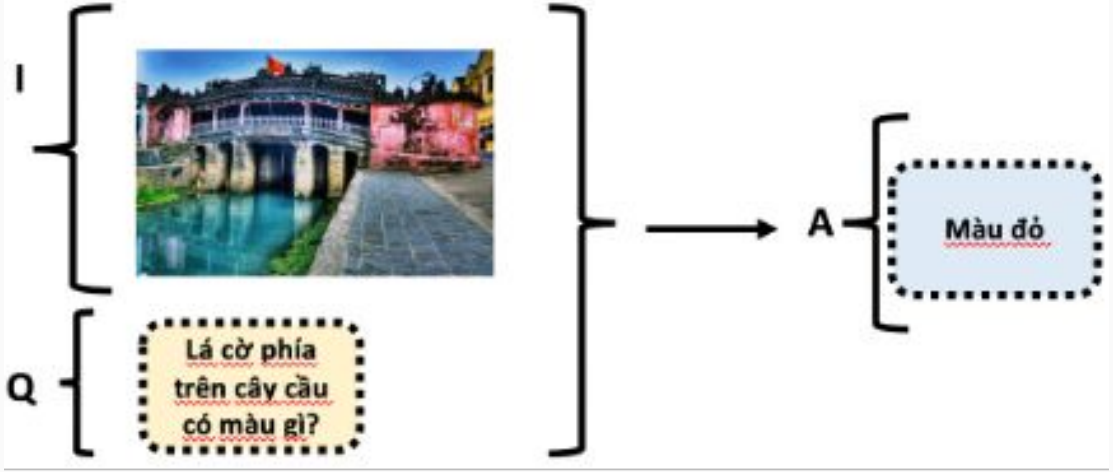
Trần Vĩnh Khiêm

Giới thiệu

Bài toán VQA được đề xuất bởi Antol và các cộng sự [1].

Input: Một hình ảnh I và câu hỏi Q liên quan đến I.

Output: Câu trả lời A theo thông tin có trong đầu vào.



Lý do chọn đề tài và câu hỏi nghiên cứu

- Thu thập dữ liệu và gán nhãn bởi con người tốn kém và chậm
- Các mô hình ngôn ngữ lớn gần đây đạt được nhiều thành tựu trên nhiều bài toán, bao gồm bài toán VQA [2]. Tuy nhiên, chúng quá lớn để triển khai trên máy tính vừa và nhỏ.
- Việc thiết kế prompt không hiệu quả làm giảm độ chính xác của mô hình
- Các câu prompt đã được chứng minh hiệu quả trong các tác vụ học ít dữ liệu văn bản [3,4]

Câu hỏi nghiên cứu (Research question)

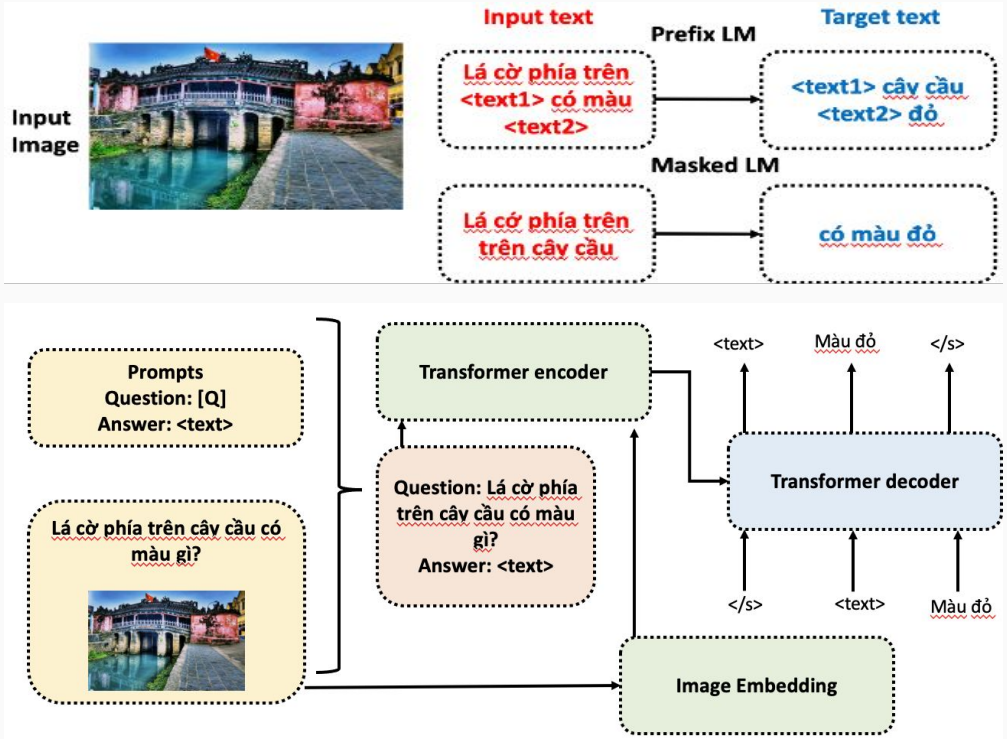
- **RQ1**: Việc thiết kế prompts có ảnh hưởng như thế nào đến zero-shot learning và few-shot learning trên bài toán mới?
- **RQ2**: Việc thiết kế prompts có quan trọng dữ liệu đầu vào có kích thước lớn?
- **RQ3**: Các mục tiêu tiền huấn luyện khác nhau có ảnh hưởng đến zero-shot learning và few-shot learning như thế nào?

Mục tiêu

- Đề xuất một mô hình sequential transformer encoder/decoder với mô hình PrefixLM [5] và mô hình MaskedLM [6] nhằm tiết kiệm với phần cứng tính toán và đảm bảo hiệu suất mô hình khi so sánh với các mô hình ngôn ngữ lớn.
- Phân tích hiệu suất bị ảnh hưởng ra sao bởi các prompts khác nhau **(RQ1)**.
- So sánh hiệu suất giữa các mô hình với các kích thước dữ liệu huấn luyện khác nhau và các prompts khác nhau. **(RQ2).**
- Thử nghiệm mô hình trên các tác vụ khác nhau của VQA sử dụng zero-shot và few-shot **(RQ3).**

Nội dung và Phương pháp

Nội dung 1: Đề xuất mô hình PromptVQA cho bài toán hỏi đáp trực quan



Nội dung và Phương pháp

Nội dung 2: Thử nghiệm nhiều loại prompt và đề xuất một số prompt template

- Tìm hiểu các loại prompts khác nhau như hand-crafted prompts, noisy prompts và irrelevant prompts và nghiên cứu tác động của các prompt trên các hội nghị A/A*.
- Tiến hành tạo ra các mẫu prompt (prompt template) để phù hợp với prefix và masked đã nêu ở **nội dung 1.**
- Tiến hành đánh giá các mẫu prompt có ảnh hưởng đến tác vụ như VQA như thế nào trên bộ dữ liệu VQA [7] với độ đo accuracy và training time.

Nội dung và Phương pháp

Nội dung 3: Huấn luyện mô hình với nhiều kích thước khác nhau

- Tiến hành huấn luyện các mô hình với các kích thước khác nhau của dữ liệu huấn luyện bao gồm **n-shot**, với n từ 0 tới 10 và với các kiểu prompt khác nhau bao gồm **no prompt, prompt P1, P2, P3…PN** theo các mẫu đã thiết kế ở **nội dung 2**, sau đó so sánh hiệu suất giữa chúng trên bộ dữ liệu VQA [7].
- Đánh giá các mô hình đề xuất và mô hình LLMs theo từng shot và prompt khác nhau trên độ đo accuracy và training time.

Nội dung và Phương pháp

Nội dung 4: Tiền huấn luyện và đánh giá mô hình trên các tác vụ khác nhau

- Nghiên cứu các mục tiêu tiền huấn luyện ảnh hưởng đến các tác vụ khác nhau của VQA như **suy luận trực quan và hỏi đáp trực quan dựa trên cơ sở tri thức** trên các bộ dữ liệu chuẩn tại các **hội nghị CVPR như GQA [8] và OKVQA [9]** tương ứng với mỗi tác vụ.
- Đánh giá mô hình đề xuất và các mô hình LLMs đạt SOTA với các tác vụ.
- Tiến huấn luyện mô hình đề xuất và các mô hình LLMs với các mục tiêu tương ứng với từng tác vụ. Sau đó, tiến hành đánh giá các mẫu prompt ở **nội dung 2** với cách đánh giá tương tự như ở **nội dung 3**.

Kết quả dự kiến

- Phương pháp đề xuất đạt kết quả cao hơn hoặc bằng với các mô hình LLMs SOTA ở thời điểm hiện tại trên độ đo Accuracy và cần dữ liệu đầu vào ít hơn → giảm kích thước mô hình.
- Một tập dữ liệu bao gồm các mẫu prompt được thiết kế sẵn. Chứng minh được việc thiết kế các mẫu prompt có thể giúp mô hình thích ứng nhanh với các tác vụ mới và gia tăng hiệu suất trên độ đo accuracy và giảm training time **(RQ1).**
- Chứng minh được mô hình đề xuất có hiệu suất cao trên độ đo accuracy kể cả khi không cần dữ liệu huấn luyện và chỉ có một ít dữ liệu huấn luyện. **(RQ2).**
- Chứng minh được tiền huấn luyện với các mục tiêu khác nhau sẽ giúp cải thiện hiệu suất và khả năng tổng quát hoá của các tác vụ khác của VQA trên độ đo accuracy. **(RQ3).**

Tài liệu tham khảo

[1]. Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh: **VQA: Visual Question Answering**. ICCV 2015: 2425-2433.
[2]. Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, Ilya Sutskever: **Learning Transferable Visual Models From Natural Language Supervision**. ICML 2021: 8748-8763.
[3]. Tianyu Gao, Adam Fisch, Danqi Chen: **Making Pre-trained Language Models Better Few-shot Learners**. ACL 2021: 3816-3830.
[4]. Timo Schick, Hinrich Schütze: **It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners**. NAACL 2021: 2339-235.
[5]. Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J. Liu: **Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer**. J. Mach. Learn. Res. 21: 140:1-140:67 (2020)
[6]. Jaemin Cho, Jie Lei, Hao Tan, Mohit Bansal: **Unifying Vision-and-Language Tasks via Text Generation**. ICML 2021:1931-1942.
[7]. Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh: **Making the V in VQA matter: Elevating the role of image understanding in visual question answering**. CVPR 2017: 6325–6334.
[8]. Drew A. Hudson and Christopher D. Manning: **GQA: A new dataset for real-world visual reasoning and compositional question answering**. CVPR 2019: 6700–6709.
[9]. Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi: **OK-VQA: A visual question answering benchmark requiring external knowledge**. CVPR 2019: 3195– 3204.