


THÔNG TIN CHUNG CỦA BÁO CÁO

- Link YouTube video của báo cáo (tối đa 5 phút):

<https://youtu.be/Q9KdboRsygs>

- Link slides (dạng .pdf đặt trên Github):

<https://github.com/kvt0012/CS2205.APR2023/blob/main/Khi%C3%AAm%20Tr%E1%BA%A7n%20V%C4%A9nh%20-%20xCS2205.DeCuong.FinalReport.Template.Slide.pdf>

<ul style="list-style-type: none">• Họ và Tên: Trần Vĩnh Khiêm• MSHV: 220101041 	<ul style="list-style-type: none">• Lớp: CS2205.APR2023• Tự đánh giá (điểm tổng kết môn): 9/10• Số buổi vắng: 0• Số câu hỏi QT cá nhân: 16/16• Số câu hỏi QT của cả nhóm: 5/5• Link Github: https://github.com/kvt0012/CS2205.APR2023/
--	--

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI: 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.

TÊN ĐỀ TÀI TIẾNG ANH: ENHANCING VISION-LANGUAGE MODELS FOR VISUAL QUESTION ANSWERING WITH LOW-RESOURCES VIA PROMPT-BASED LEARNING.

TÓM TẮT (Tối đa 400 từ)

Các mô hình Hình ảnh-Ngôn ngữ (VL) được huấn luyện quy mô lớn đã chứng tỏ khả năng học các bài toán mới chỉ với một số ví dụ rất nhỏ và có khả năng tổng quát hóa cho các bài toán mới mà không cần điều chỉnh lại. Tuy nhiên, kích thước rộng lớn và tốc độ suy luận chậm của chúng đặt ra thách thức cho triển khai trong thực tế. Để giải quyết hạn chế này, chúng tôi đề xuất một phương pháp mang tên PromptVQA để học tập tài nguyên thấp cho các bài toán hình ảnh – ngôn ngữ, cụ thể là bài toán hỏi đáp trực quan (VQA) dựa trên câu gợi ý, với kích thước tương đối nhỏ hơn so với một số các mô hình học từ gần đây. Phương pháp của chúng tôi bao gồm việc tiền huấn luyện một mô hình biến đổi dựa trên chuỗi sử dụng mô hình ngôn ngữ với tiền tố (PrefixLM) và mô hình ngôn ngữ được ẩn (MaskedLM). Ngoài ra, chúng tôi nghiên cứu tác động của các câu gợi ý đa dạng đối với các tác vụ học từ một số ví dụ.

GIỚI THIỆU

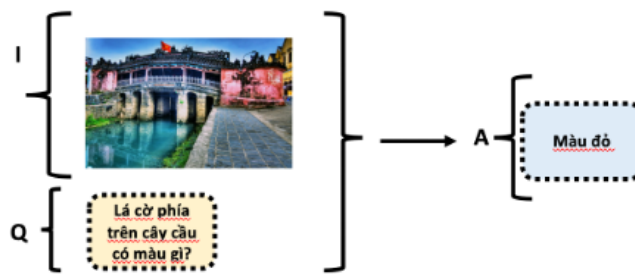
Mô hình hình ảnh – ngôn ngữ là một loại mô hình trong lĩnh vực xử lý ngôn ngữ tự nhiên và thị giác máy tính. Mô hình này nhằm kết hợp thông tin từ cả hai phương thức – hình ảnh và ngôn ngữ- để hiểu và tạo ra các mô tả, câu chú thích hoặc trả lời câu hỏi về hình ảnh.

Sự điều chỉnh lại (fine-tune) các mô hình ngôn ngữ lớn (LLMs) đã đạt được kết quả mạnh mẽ trong nhiều lĩnh vực bao gồm các tác vụ hình ảnh và ngôn ngữ [1]. Các LLMs như vậy có thể học một bài toán mới chỉ với một vài ví dụ hoặc tổng quát hóa cho một bài toán mới mà không cần điều chỉnh lại trên bất kỳ ví dụ huấn luyện nào, tức là học bằng cách Few-shot (sử dụng ít dữ liệu) hoặc Zero-shot (không có dữ liệu). Few-shot learning (Học ít dữ liệu)

vượt qua những thách thức của việc học có giám sát dựa trên dữ liệu nhiều, trong đó việc thu thập dữ liệu được gán nhãn bởi con người tốn kém và chậm. Tuy nhiên, các mô hình few-shot learning gần đây như GPT3 [2], Frozen [3] và PICA [4] quá lớn để triển khai trên các máy tính nhỏ và vừa do kích thước mô hình khổng lồ của chúng.

Trong nghiên cứu này, chúng tôi tìm hiểu việc học ít dữ liệu trong bài toán hình ảnh-ngôn ngữ bằng phương pháp mà chúng tôi đề xuất gọi là PromptVQA, một mô hình hình ảnh-ngôn ngữ kích thước vừa, trong đó chúng tôi điều chỉnh lại mô hình mà không có hoặc chỉ với một số ít mẫu dữ liệu huấn luyện. Đối với PromptVQA, chúng tôi tiền huấn luyện một mô hình transformer encoder/decoder tuần tự với mô hình ngôn ngữ PrefixLM [5] và mô hình ngôn ngữ MaskedLM [6]. Cài đặt này mang tính thực tế hơn vì có thể tiết kiệm phần cứng tính toán và đồng thời, việc thu thập một số lượng lớn các ví dụ huấn luyện chất lượng trong thực tế thì rất tốn kém. Trong cài đặt học ít dữ liệu như vậy, các câu prompt là quan trọng và đã được chứng minh hiệu quả trong các nhiệm vụ học ít dữ liệu văn bản [7, 8]. Trong đề tài này, chúng tôi tập trung chủ yếu vào bài toán hỏi đáp trực quan. Bài toán hỏi đáp trực quan được đề xuất vào năm 2015 bởi Antol và các cộng sự [9], bài toán được định nghĩa như sau: Cho một hình ảnh I, một câu hỏi Q liên quan đến hình ảnh I. Mục tiêu là xây dựng một hệ thống có thể xác định, trích xuất thông tin thông qua các yếu tố ngôn ngữ câu hỏi Q và trích xuất thông tin từ hình ảnh I dựa trên việc hiểu mối liên hệ giữa câu hỏi và hình ảnh, cụ thể:

- **Đầu vào (Input):** Một hình ảnh I và câu hỏi Q liên quan đến I.
- **Đầu ra (Output):** câu trả lời A theo thông tin có trong đầu vào.



Hình 1 Ví dụ về bài toán hỏi đáp trực quan

Nhằm mở rộng thành công cho bài toán hỏi đáp trực quan, chúng tôi trả lời những câu hỏi sau đối với việc học hình ảnh-ngôn ngữ có nguồn lực hạn chế dựa trên câu gợi ý. Chúng tôi nhằm mục đích trả lời những câu hỏi sau trong nghiên cứu này thông qua các thí nghiệm trên nhiều bộ dữ liệu VQA.

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

Để trả lời những câu hỏi này, chúng tôi khám phá các định dạng câu gợi ý khác nhau bao trên các bộ dữ liệu học thuật về học sâu hợp nhất giữa hình ảnh và ngôn ngữ. Mục tiêu của chúng tôi bao gồm:

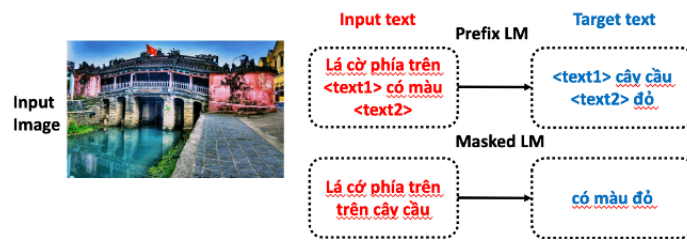
- 1/ Đề xuất một mô hình transformer encoder/decoder tuần tự với mô hình PrefixLM và mô hình MaskedLM nhằm tiết kiệm với phần cứng tính toán và đảm bảo hiệu suất mô hình khi so sánh với các mô hình ngôn ngữ lớn.
- 2/ Phân tích hiệu suất bị ảnh hưởng ra sao bởi các prompts khác nhau, bao gồm cả hand-craft prompt, noisy prompt và irrelevant prompt (**RQ1**).
- 3/ So sánh hiệu suất giữa các mô hình với các kích thước dữ liệu huấn luyện khác nhau và các prompts khác nhau. (**RQ2**).
- 4/ 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

Phương pháp

- Khảo sát, tìm hiểu các mô hình tiền huấn luyện lớn và các mô hình hình ảnh – ngôn ngữ đạt kết quả State-of-the-art trong tác vụ VQA sử dụng ít hoặc không cần dữ liệu huấn luyện trên bài toán VQA trên các hội nghị A/A* về ngôn ngữ và hình ảnh như CVPR, ACL, ICCV, EMNLP.
- Đề xuất mô hình PromptVQA bao gồm mô hình sequential transformer encoder/decoder với mô hình PrefixLM và mô hình MaskedLM nhằm hướng dẫn mô hình trả lời đúng mục tiêu với mục đích tiết kiệm với phần cứng tính toán và đảm bảo hiệu suất mô hình.



Hình 2 Ví dụ về prefix và masked

- Tiến hành đánh giá hiệu suất mô hình đề xuất với các mô hình tiền huấn luyện lớn và các mô hình hình ảnh – ngôn ngữ đạt kết quả State-of-the-arts (SOTA) trên các bộ dữ liệu VQA sử dụng ít hoặc không cần dữ liệu huấn luyện bằng độ đo Accuracy và Training Time trên bộ dữ liệu VQA [9].

Nội dung 2: Thử nghiệm một số prompts trên bài toán hỏi đáp trực quan và phân tích xem hiệu suất bị ảnh hưởng ra sao bởi các câu gợi ý khác nhau.

Phương pháp

- Tìm hiểu các loại prompts khác nhau như hand-crafted prompts, noisy prompts và irrelevant prompts.
- Nghiên cứu tác động của các prompt khác nhau đối với hiệu suất của mô hình đề xuất và các mô hình state-of-the-arts trong các tác vụ zero-shot và few-shot nhằm nghiên cứu xem các prompts có hữu ích trong hỏi đáp trực quan hay không.
- Tiến hành tạo ra các mẫu prompt dựa trên kết quả đã tìm hiểu. Một ví dụ về mẫu prompt như sau: "câu hỏi: [Q], câu trả lời: <text1>" như là một gợi ý đầu vào và "<text1> [A]" là một gợi ý đầu ra cho việc trả lời câu hỏi bằng hình ảnh như 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 [10] với độ đo accuracy và training time.

Nội dung 3: Huấn luyện các mô hình với các kích thước dữ liệu huấn luyện khác nhau và các câu hỏi khác nhau, sau đó so sánh hiệu suất giữa chúng.

Phương pháp

- 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 [10].

- 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 4: Tiền huấn luyện mô hình đề xuất và thử nghiệm mô hình đề xuất trên các tác vụ khác nhau mà không cần dữ liệu huấn luyện và chỉ có một ít dữ liệu huấn luyện

Phương pháp

- 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 [11] và OKVQA [12] tương ứng với mỗi tác vụ.
- Đánh giá và so sánh mô hình đề xuất và các mô hình LLMs với các tác vụ khác nhau trên các bộ dữ liệu như GQA [11] và OKVQA [12] mà không cần huấn luyện lại mô hình và chỉ huấn luyện với một ít dữ liệu.
- Tiến hành tiền huấn luyện mô hình đề xuất và các mô hình LLMs với các mục tiêu tiền huấn luyện khác nhau như suy luận trực quan và hỏi đáp trực quan dựa trên cơ sở tri thức.
- Đồng thời, tiến hành đánh giá với các mẫu prompt khác nhau đã được tạo ra ở **nội dung 2** với cách đánh giá như tương tự như trên bộ VQA ở **nội dung 3**.

KẾT QUẢ MONG ĐỢI

- Phương pháp đề xuất đạt kết quả cao hơn hoặc bằng với các mô hình SOTA ở thời điểm hiện tại trên độ đo Accuracy, Training Time và cần dữ liệu đầu vào ít hơn, nhờ đó là giảm kích thước của 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 với các prompt đã thiết kế. (**RQ2**).
- Chứng minh được tiền huấn luyện với các mục tiêu huấn luyện khác nhau sẽ giúp cải thiện hiệu suất tổng quát và khả năng tổng quát hoá của các tác vụ khác của VQA trên độ đo accuracy. (**RQ3**).

TÀI LIỆU THAM KHẢO (*Định dạng DBLP*)

- [1]. 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.
- [2]. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D: **Language models are few-shot learners**. NeurIPS 2020: 1877-1901.
- [3]. Maria Tsimpoukelli, Jacob Menick, Serkan Cabi, S. M. Ali Eslami, Oriol Vinyals, Felix Hill: **Multimodal Few-Shot Learning with Frozen Language Models**. NeurIPS 2021: 200-212.
- [4]. Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, Lijuan Wang: **An Empirical Study of GPT-3 for Few-Shot Knowledge-Based VQA**. AAAI 2022: 3081-3089.
- [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]. Tianyu Gao, Adam Fisch, Danqi Chen: **Making Pre-trained Language Models Better Few-shot Learners**. ACL 2021: 3816-3830.
- [8]. Timo Schick, Hinrich Schütze: **It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners**. NAACL 2021: 2339-235.
- [9]. Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh: **VQA: Visual Question Answering**. ICCV 2015: 2425-2433.
- [10]. 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.
- [11]. Drew A. Hudson and Christopher D. Manning: **GQA: A new dataset for real-world visual reasoning and compositional question answering**. CVPR 2019: 6700–6709.
- [12]. Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi: **OK-VQA: A visual question answering benchmark requiring external knowledge**. CVPR 2019: 3195– 3204.