Hướng tiếp cận mới trên bộ dữ liệu UIT-ViIC: Hỏi đáp trực quan

1st Trương Phước Bảo Khanh
University of Information Technology, Ho Chi Minh City
Vietnam National University Ho Chi Minh City
Ho Chi Minh City, Vietnam
20520579@gm.uit.edu.vn

2nd Trần Thị Thu Hà
University of Information Technology, Ho Chi Minh City
Vietnam National University Ho Chi Minh City
Ho Chi Minh City, Vietnam
20521273@gm.uit.edu.vn

***Tóm tắt nội dung*—Hỏi đáp trực quan (Visual Question Answering-VQA) là bài toán kết hợp thị giác máy tính và xử lý ngôn ngữ tự nhiên. Trong bài báo này, chúng tôi trình bày phương pháp xây dựng bộ dữ liệu VQA tiếng Việt bằng cách sinh tự động các cặp câu hỏi – câu trả lời từ những chúthích của trên bộ dữ liệu chú thích ảnh có sẵn UIT-ViIC. Chúng tôi mô tả các công cụ đã sử dụng, phân tích những trở ngại gặp phải trong quá trình xây dựng và cách khắc phục.**

***Index Terms*—NLP, Computer Vision, VQA, Question Generation**

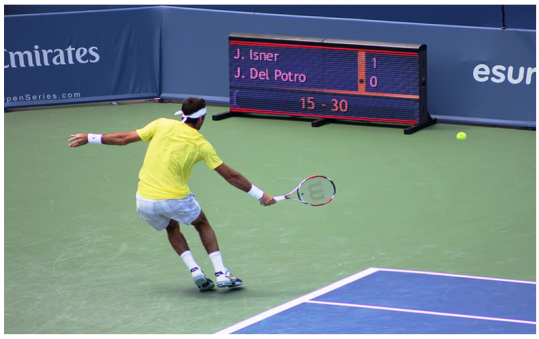
I. Giới thiệu

Được đề xuất từ năm 2015 [1], bài toán trả lời câu hỏi trực quan (Visual Question Answering) kết hợp hai lĩnh vực quan trọng của học máy (Machine Learning) là thị giác máy tính (Computer Vision) và xử lý ngôn ngữ tự nhiên (Natural Language Processing). Dựa vào một hình ảnh và một câu hỏi ngôn ngữ tự nhiên về hình ảnh đó, mô hình phải đưa ra một câu trả lời tương ứng bằng ngôn ngữ tự nhiên. Do câu hỏi có thể tập trung vào các vùng khác nhau của hình ảnh (tiền cảnh – foreground, hậu cảnh – background, ngữ cảnh – context hoặc các chi tiết khác) nên đòi hỏi mô hình vừa phải nhận biết được các bộ phận của ảnh, vừa phải kết hợp các bộ phận đó với câu hỏi và suy luận ra câu trả lời. Các nghiên cứu bằng tiếng Việt về bài toán này hiện nay gặp một trở ngại chung, đó là thiếu các bộ dữ liệu huấn luyện và kiểm thử bằng tiếng Việt. Do khác biệt về ngữ pháp và văn phạm nên việc huấn luyên trên bộ dữ liệu tiếng Anh hay ngôn ngữ khác rồi áp dụng vào tiếng Việt là không khả thi. Từ ý tưởng của Changpinyo et al [2], trong bài báo này, chúng tôi đã thử nghiệm phương pháp sinh các cặp câu hỏi - câu trả lời từ những bộ dữ liệu chú thích ảnh tiếng Việt, hay ở đây là bộ dữ liệu UIT-ViIC [1].

II. Các công trình nghiên cứu liên quan

A. Sinh câu hỏi trong lĩnh vực thị giác máy tính

Theo các công trình nghiên cứu trước đây, với đầu vào là một bức ảnh hoặc video, câu hỏi được sinh ra (mà không biết trước câu trả lời) cho người dùng trả lời, nhằm mục đích tăng cường dữ liệu huấn luyện. Các công trình nghiên cứu gần đây có các hướng tiếp cận khác bằng cách thay đổi đầu vào là các chú thích ảnh/ video, sinh ra các cặp câu hỏi-câu trả lời bằng phương pháp "template-based" [3], sau đó ghép với ảnh tương ứng để thu được bộ dữ liệu VQA.



Chú thích: Vận động viên tennis nam đang cầm vợt đứng ngoài biên đánh bóng

VQA

Câu hỏi	Câu trả lời
Ai đang chơi bóng ở ngoài biên đánh bóng?	Vận động viên tennis nam
Vận động viên tennis nam đang làm gì?	Cầm vợt đánh bóng
Ai đang làm gì với vợt đỡ bóng?	Vận động viên

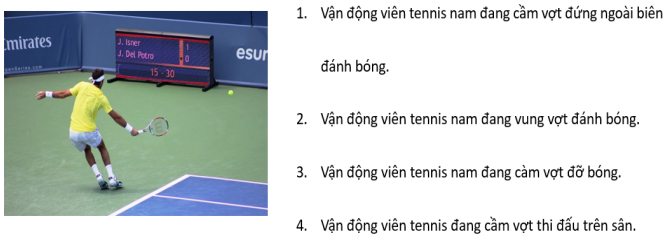
Hình 1: Ví dụ về một điểm dữ liệu bộ dữ liệu UIT-ViIC sau khi áp dụng phương pháp sinh cặp câu hỏi - câu trả lời trực quan

B. Phương pháp học chuyển giao trong bài toán VQA

Các nghiên cứu chỉ ra có sự liên quan giữa hai bài toán sinh chú thích ảnh (Image Captioning) và hỏi đáp trực quan (VQA). Fisch et al [4] đã tìm ra cách sinh chú thích từ các cặp câu VQA. Yang et al [5] sử dụng mô hình GPT-3, bằng một số ví dụ về VQA, đã trả lời được các câu hỏi được sinh ra từ chú thích ảnh. Tuy nhiên, những cách này đều cần fine-tunning bằng dữ liệu VQA. Điều mà chúng tôi mong muốn là tạo ra bộ dữ liệu VQA mà không cần đến bất kì dữ liệu VQA nào khác.

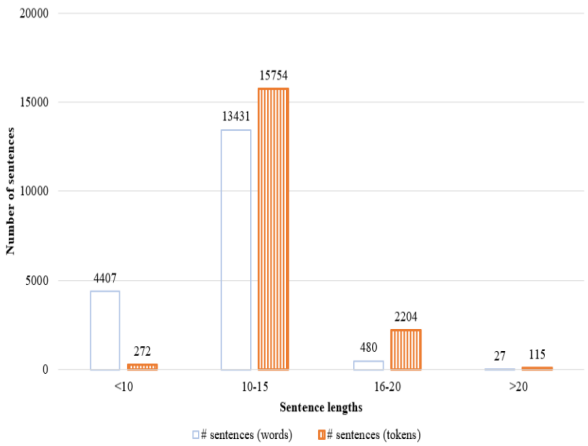
III. Bộ dữ liệu

Trong bài báo này, chúng tôi sử dụng bộ dữ liệu bộ dữ liệu UIT-ViIC [1], một bộ dữ liệu cho việc đánh giá tự động sinh chú thích hình ảnh bằng tiếng Việt, bộ dữ liệu này bao gồm 19,250 chú thích tiếng Việt cho 3,850 hình ảnh từ bộ dữ liệu Microsoft COCO [6] liên quan đến các môn thể thao chơi với bóng (5 câu chú thích cho mỗi hình ảnh - ví dụ ở hình [2]).



Hình 2: Một điểm dữ liệu của bộ dữ liệu UIT-ViIC

Bộ dữ liệu được phát triển bởi nhóm nghiên cứu NLP của trường đại học Công nghệ thông tin - ĐHQG thành phố Hồ Chí Minh. Chúng tôi đã dùng hình ảnh kết hợp với chú thích hình ảnh để sinh ra các cặp câu hỏi và câu trả lời.



Hình 3: Trực quan độ dài của câu từ bộ dữ liệu UIT-ViIC

IV. Phương pháp tiếp cận

Với đầu vào là bộ dữ liệu Image Captioning gồm ảnh và chú thích. Với mỗi chú thích, chúng tôi sẽ tiến hành trích xuất (IV-A) để thu được các câu trả lời ban đầu. Sau đó, các câu trả lời này và chú thích ban đầu sẽ được dùng để sinh câu hỏi (IV-B). Các câu hỏi này cùng với chú thích sẽ là đầu vào cho việc sinh câu trả lời (IV-C). Cuối cùng, các câu trả lời lúc sau sẽ được đánh giá độ tương đồng với các câu trả lời ban đầu và tiến hành lọc bỏ. Phương pháp này được mô tả tổng quan ở hình [4].

A. Trích xuất câu trả lời VQA từ câu chú thích của ảnh

1) Gán nhãn từ loại (POS tagging): Chúng tôi đã sử dụng thư viện VNCore NLP [7] gán nhãn từ loại cho câu chú thích, nghĩa là đánh dấu một từ trong văn bản dựa theo bối cảnh và định nghĩa của từ đó.

Sau khi đã có các từ loại chúng tôi đã lọc ra lấy các danh từ (N), danh từ riêng (Np), danh từ chỉ loại (Nc), số từ (M), động từ (V) để làm một câu trả lời để trả lời câu hỏi về người nào, sự vật nào, hành động nào và số lượng nào được nhắc đến trong chú thích của hình ảnh.

2) Cấu trúc cây (Phrase structure tree): Cấu trúc cây [8] dùng để biểu diễn cấu trúc ngữ pháp của một câu bằng cách sử dụng một cây nhị phân. Mỗi nút trong cây đại diện cho một cụm từ (phrase) hoặc một từ loại (part of speech) và các nút con đại diện cho các thành phần cấu thành nên cụm từ đó. Cấu trúc cây giúp biểu diễn mối quan hệ giữa các từ và cụm từ trong câu kết hợp với nhau để tạo thành ý nghĩa của câu.

Chúng tôi đã sử dụng cây cấu trúc để lấy các cụm danh từ, cụm động từ làm câu trả lời và đảm bảo các cụm từ đó đã bao gồm các cây con của các cụm từ (Hình [5])

B. Sinh câu hỏi VQA

Sinh câu hỏi VQA [9] là đặt câu hỏi dựa trên những gì có trong hình ảnh và chú thích hình ảnh đã cung cấp.

Chúng tôi đã tiến hành fine-tuning mô hình pre-trained ViT5 base [10] dựa trên kiến trúc Transformer cho các tác vụ xử lý ngôn ngữ tự nhiên. Từ những gì đã có trong hình ảnh được biểu diễn bởi chú thích hình ảnh và câu trả lời đã được trích xuất phần trên chứa những gì có trong ảnh để tạo ra câu hỏi. Tuy nhiên, bộ dữ liệu UIT-ViIC mà chúng tôi sử dụng không có chứa những câu hỏi nên chúng tôi đã có một giải pháp từ một bộ dữ liệu hỏi đáp UIT-ViQuAD 2.0 [11] được phát triển bởi nhóm UIT NLP để mô hình tạo câu hỏi dựa trên bộ dữ liệu này. Đầu vào của mô hình ViT5 là cặp câu trả lời - ngữ cảnh và đầu ra là câu hỏi (Hình [6]).

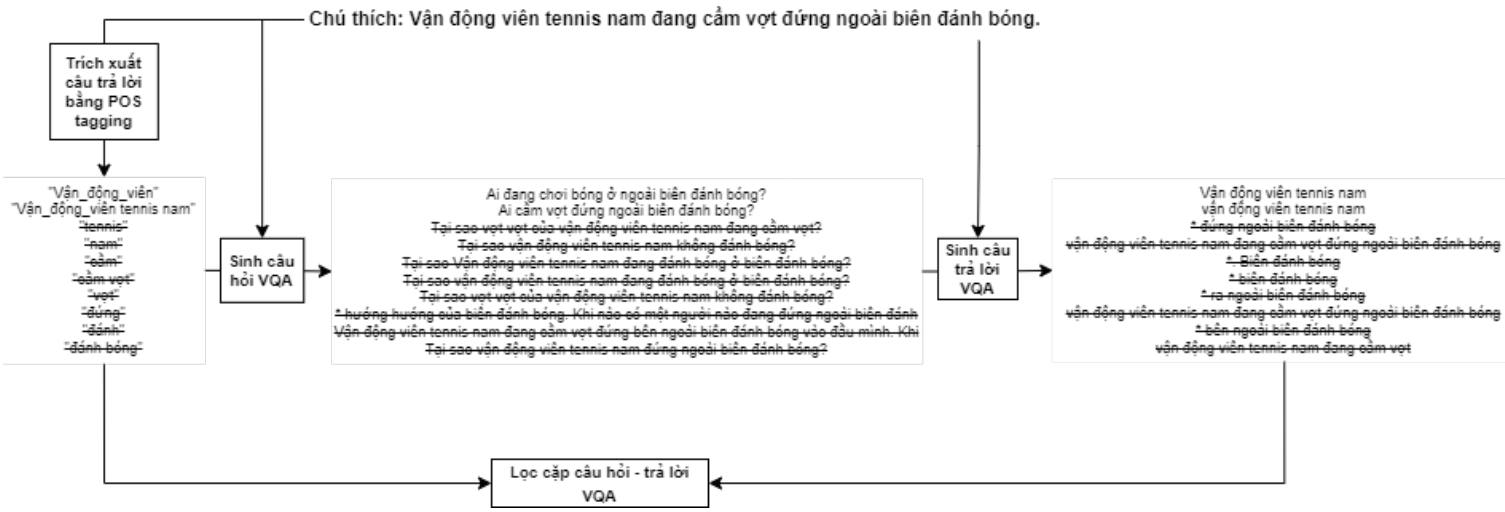
C. Sinh câu trả lời VQA

Ở đây, chúng tôi sử dụng cùng một phương pháp với cách tạo sinh câu hỏi VQA, bằng cách thay đổi đầu vào là cặp câu hỏi - ngữ cảnh.

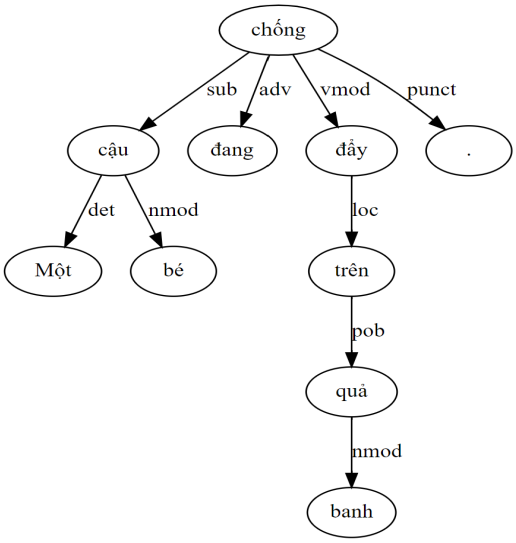
Việc sinh tự động các cặp câu hỏi – câu trả lời dựa vào các chú thích ảnh có sẵn theo phương án đem lại nhiều lợi ích: (i) Có thể tận dụng số lượng có sẵn các chú thích ảnh từ các bộ dữ liệu; (ii) Việc sinh tự động giúp hạn chế tối đa việc cần đến sự can thiệp của con người, cho phép sinh ra số lượng lớn mẫu dữ liệu VQA trong thời gian ngắn với chi phí tối thiểu.

V. Kết quả

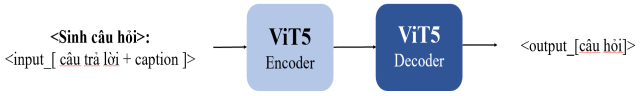
Chúng tôi đã tạo ra được bộ câu hỏi và câu trả lời tương ứng cho bộ dữ liệu UIT-ViIC. Tuy nhiên vì vấn đề bộ dữ liệu chúng tôi train để sinh ra câu hỏi và câu trả lời là bộ dữ liệu khác - bộ dữ liệu UIT-ViQuAD 2.0, bộ dữ liệu này vì có các



Hình 4: Phương pháp tiếp cận bài toán



Hình 5: Biểu diễn cây cấu trúc của một chú thích hình ảnh



Hình 6: Sử dụng mô hình pre-trained ViT5 sinh ra câu hỏi

cặp câu hỏi và câu trả lời dựa trên một bài viết ngữ cảnh từ Wikipedia nên sẽ có cách đặt câu hỏi sẽ liên quan nhiều đến nguyên nhân và kết quả - câu hỏi tại sao, điều gì khiến,... Còn đối với câu hỏi mà chúng tôi mong muốn sẽ liên quan đến ai, cái gì, làm gì nhiều hơn nên với kết quả đạt được chúng tôi tiến hành lọc và đánh giá bằng phương pháp F1 score token-level trong khoảng [0; 1] và trên 0.65 là đạt (Bảng [I]).

VI. Hướng phát triển

Vì thời gian nghiên cứu có hạn, chúng tôi nhận thấy còn một số hạn định trong phương pháp như số lượng câu hỏi trong bộ dữ liệu là chưa nhiều, chưa đa dạng, cũng như chưa có tính suy luận cao như mục tiêu mà chúng tôi hướng tới là bộ dữ liệu Open-ended VQA, mà chỉ dừng lại ở VQA tiêu chuẩn. Dự định trong tương lai sẽ áp dụng các mô hình VQA baselines và SOTA multi modal fusion để có thể đánh giá phương pháp, bộ dữ liệu chúng tôi được chính xác, khách quan hơn.

VII. Kết luận

Trong ngữ cảnh nghiên cứu bằng tiếng Việt, một trở ngại chung là thiếu các bộ dữ liệu huấn luyện và kiểm thử bằng tiếng Việt. Vì khác biệt về ngữ pháp và văn phạm, không khả thi để huấn luyện trên các bộ dữ liệu tiếng Anh hoặc các ngôn ngữ khác rồi áp dụng vào tiếng Việt.

Tổng kết lại, bài báo đã giới thiệu về hướng tiếp cận mới trong việc xây dựng bộ dữ VQA dựa trên bộ dữ liệu UIT-ViIC, hay mở rộng ra hơn là các bộ dữ liệu sinh chú thích ảnh trong tiếng Việt.

Như vậy, nghiên cứu này đã đưa ra một hướng tiếp cận mới trong việc xây dựng bộ dữ liệu và mô hình cho bài toán Trả lời câu hỏi trực quan bằng tiếng Việt, góp phần khắc phục trở ngại và mở ra tiềm năng phát triển trong lĩnh vực này.

Tài liệu

[1] Antol, Stanislaw, et al. "Vqa: Visual question answering." Proceedings of the IEEE international conference on computer vision. 2015.
[2] Changpinyo, Soravit, et al. "All you may need for vqa are image captions." arXiv preprint arXiv:2205.01883 (2022).
[3] Lyu, Chenyang, et al. "Improving unsupervised question answering via summarization-informed question generation." arXiv preprint arXiv:2109.07954 (2021).
[4] Fisch, Adam, et al. "Capwap: Captioning with a purpose." arXiv preprint arXiv:2011.04264 (2020).
[5] Yang, Zhengyuan, et al. "An empirical study of gpt-3 for few-shot knowledge-based vqa." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 36. No. 3. 2022.

Bảng I: Ví dụ về đánh giá các cặp câu hỏi được sinh ra

Câu trả lời	**Câu hỏi đã tạo**	**Câu trả lời đã được**	**Kết quả đánh giá**
Vận đông viên	Ai đang chơi bóng ở ngoài biên đánh bóng?	Vận động viên tennis nam	0.7 (đạt)
Vận động viên tennis nam	Ai đang làm gì với vợt đỡ bóng?	Vận động viên tennis nam	1 (đạt)
Đánh bóng	Vận động viên tennis nam đang làm gì?	Vận động viên tennis nam đang cầm vợt	0 (không đạt)
càm vợt	Tại sao vận động viên tennis nam đang càm vợt đỡ bóng?	vận động viên tennis nam đang càm vợt đỡ bóng	0 (không đạt)
cầm vợt đánh bóng	Vận động viên tennis đang làm gì?	Biên đánh bóng	0.5 (không đạt)

[6] Lin, Tsung-Yi, et al. "Microsoft coco: Common objects in context." Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. Springer International Publishing, 2014.

[7] Thanh Vu, Dat Quoc Nguyen, Dai Quoc Nguyen, Mark Dras, and Mark Johnson. 2018. VnCoreNLP: A Vietnamese Natural Language Processing Toolkit. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, pages 56–60, New Orleans, Louisiana. Association for Computational Linguistics.

[8] Junru Zhou and Hai Zhao. 2019. Head-Driven Phrase Structure Grammar Parsing on Penn Treebank. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 2396–2408, Florence, Italy. Association for Computational Linguistics.

[9] Lu, Y. et al. (2022). RTN: Reinforced Transformer Network for Coronary CT Angiography Vessel-level Image Quality Assessment. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2022. MICCAI 2022. Lecture Notes in Computer Science, vol 13431. Springer, Cham. https://doi.org/10.1007/978-3-031-16431-6_61

[10] Long Phan, Hieu Tran, Hieu Nguyen, và Trieu H. Trinh. 2022. ViT5: Pretrained Text-to-Text Transformer for Vietnamese Language Generation. arXiv preprint arXiv:2205.06457.

[11] Kiet Van Nguyen, Ngan Luu-Thuy Nguyen, Son T. Luu, Tin Van Huynh, và Luan Thanh Nguyen. 2020. A Vietnamese Dataset for Evaluating Machine Reading Comprehension. arXiv preprint arXiv:2009.14725.