AN IMAGE CAPTIONING MODEL INTEGRATING KNOWLEDGE GRAPHS AND DEEP LEARNING

Nguyen Do Thai Nguyen*, Nguyen Van Tuan, Nguyen Ngoc Phu Ty, Nguyen Huu Minh Quan
Ho Chi Minh City University of Education

ARTICLE INFO

Received: 17/4/2025
Revised: 16/6/2025
Published: 27/6/2025

KEYWORDS

Scene graph
Knowledge graph
Image captioning
Deep learning
Scene graph generation

ABSTRACT

This study proposes a novel image captioning model that integrates knowledge graphs and deep learning to enhance semantic understanding and generate more accurate image descriptions. The research aims to address the limitations of conventional captioning approaches that often overlook the relationships between entities within an image. Our method involves generating scene graphs from input images, which are then enriched with external knowledge from structured knowledge graphs to generate semantically rich captions. The model is trained and evaluated on standard datasets, including MSCOCO and Visual Genome. Experimental results demonstrate that the proposed model outperforms existing baselines in terms of BLEU 41.3 and METEOR 31.6, especially in complex scenes with multiple entities. Furthermore, the use of knowledge graph augmentation significantly improves the contextual relevance and informativeness of the generated captions. This research contributes to advancing multi-objects image captioning and highlights the potential of combining symbolic knowledge with deep learning models for comprehensive scene understanding.

MỘT MÔ HÌNH MÔ TẢ HÌNH ẢNH KẾT HỢP ĐỒ THỊ TRI THỨC VÀ MẠNG HỌC SÂU

Nguyễn Đỗ Thái Nguyên*, Nguyễn Văn Tuấn, Nguyễn Ngọc Phú Tỷ, Nguyễn Hữu Minh Quân
Trường Đại học Sư phạm Thành phố Hồ Chí Minh

THÔNG TIN BÀI BÁO

Ngày nhận bài: 17/4/2025
Ngày hoàn thiện: 16/6/2025
Ngày đăng: 27/6/2025

TỪ KHÓA

Đồ thị ngữ cảnh
Đồ thị tri thức
Mô tả ảnh
Học sâu
Tạo đồ thị cảnh

TÓM TẮT

Nghiên cứu này đề xuất một mô hình mô tả ảnh tích hợp đồ thị tri thức và học sâu nhằm nâng cao khả năng hiểu ngữ nghĩa và tạo ra các mô tả hình ảnh chính xác hơn. Mục tiêu của nghiên cứu là khắc phục những hạn chế của các phương pháp mô tả ảnh truyền thống, vốn thường bỏ qua mối quan hệ giữa các thực thể trong ảnh. Phương pháp của chúng tôi bao gồm việc tạo đồ thị ngữ cảnh từ ảnh đầu vào bằng mạng học sâu, sau đó được bổ sung tri thức bên ngoài từ các đồ thị tri thức có cấu trúc để tạo ra các mô tả giàu ngữ nghĩa. Mô hình được huấn luyện và đánh giá trên các tập dữ liệu chuẩn, bao gồm MSCOCO và Visual Genome. Kết quả thực nghiệm cho thấy mô hình đề xuất vượt trội hơn so với các phương pháp cơ sở hiện có với BLEU4 là 41,3 và METEOR là 31,6, đặc biệt trong các ảnh phức tạp có nhiều thực thể. Hơn nữa, việc bổ sung tri thức từ đồ thị giúp cải thiện đáng kể mức độ liên kết ngữ cảnh và tính thông tin của các mô tả được tạo ra. Nghiên cứu này góp phần thúc đẩy việc nghiên cứu mô hình mô tả ảnh đa đối tượng và làm nổi bật tiềm năng của việc kết hợp tri thức biểu tượng với các mô hình học sâu để hiểu ảnh một cách toàn diện.

DOI: https://doi.org/10.34238/tnu-jst.12614

* Corresponding author. *Email: nguyenndt@hcmue.edu.vn*

1. Giới thiệu

Mô tả ảnh tự động (Image Captioning) là bài toán tạo văn bản mô tả nội dung ngữ nghĩa của hình ảnh một cách chính xác, đóng vai trò quan trọng trong nhiều ứng dụng như tìm kiếm ảnh, hỗ trợ người khiếm thị, hệ thống giám sát thông minh, và phân tích nội dung đa phương tiện. Các mô hình học sâu, đặc biệt là sự kết hợp giữa mạng nơ-ron tích chập (CNN) để trích xuất đặc trưng ảnh và mạng nơ-ron hồi quy (RNN hoặc LSTM) để sinh văn bản, đã đạt được nhiều kết quả đáng kể trong lĩnh vực này [1], [2]. Tuy nhiên, các mô hình truyền thống thường chỉ khai thác đặc trưng trực quan bề mặt, thiếu khả năng biểu diễn mối quan hệ giữa các thực thể trong ảnh, từ đó tạo ra các mô tả còn hạn chế về chiều sâu ngữ nghĩa và ngữ cảnh.

Để giải quyết vấn đề này, các nghiên cứu gần đây tập trung vào việc tích hợp tri thức bổ sung từ đồ thị tri thức (Knowledge Graphs – KG) hoặc đồ thị ngữ cảnh (Scene Graphs – SG). Đồ thị ngữ cảnh là cấu trúc biểu diễn các thực thể (như người, vật thể) và mối quan hệ giữa chúng (như "người-đang-cưỡi-ngựa"), thường được trích xuất tự động từ ảnh thông qua các mô hình học máy. Những đồ thị này sau đó được liên kết với nguồn tri thức bên ngoài như DBpedia hoặc ConceptNet nhằm mở rộng phạm vi hiểu biết và tăng tính ngữ nghĩa cho mô hình sinh mô tả [3] - [5].

Nghiên cứu của Santiesteban và cộng sự [6] cho thấy rằng việc kết hợp đồ thị tri thức giúp cải thiện đáng kể độ chính xác trong mô tả hình ảnh, đặc biệt là trong các ngữ cảnh phức tạp có nhiều thực thể. Zhao và Wu [5] đề xuất mô hình sử dụng đồ thị tri thức đa phương thức, tăng cường khả năng nhận diện ngữ cảnh và liên kết giữa các thực thể, tuy nhiên mô hình này yêu cầu tài nguyên huấn luyện rất lớn và khó triển khai ở quy mô rộng. Osman và cộng sự [7] cũng tổng quan các mô hình chú ý (attention-based models) trong mô tả ảnh và khẳng định rằng tích hợp tri thức là một trong những xu hướng đầy tiềm năng cho tương lai.

Đối với ngôn ngữ ít tài nguyên như tiếng Việt, việc phát triển dữ liệu và mô hình cho mô tả ảnh còn hạn chế. Phạm Anh Cường và cộng sự [8] đã giới thiệu bộ dữ liệu KTVIC – một tập dữ liệu mô tả ảnh tiếng Việt đầu tiên ở lĩnh vực đời sống – mở ra cơ hội nghiên cứu và ứng dụng cho ngôn ngữ bản địa. Tuy nhiên, vẫn còn thiếu các nghiên cứu kết hợp đồ thị tri thức và học sâu trong bối cảnh tiếng Việt.

Từ thực trạng đó, nghiên cứu này đề xuất một mô hình mô tả ảnh kết hợp đồ thị tri thức với mạng học sâu nhằm nâng cao chất lượng chú thích, đặc biệt đối với các ảnh có nhiều đối tượng và quan hệ phức tạp. Mặc dù mô hình chưa được thực nghiệm trực tiếp trên dữ liệu tiếng Việt như KTVIC, chúng tôi đã tiến hành huấn luyện và đánh giá trên các tập dữ liệu chuẩn như MSCOCO [9] và Visual Genome [10], nhằm kiểm chứng hiệu quả ngữ nghĩa và khả năng tổng quát hóa trong nhiều bối cảnh khác nhau. Phương pháp đề xuất mang tính nền tảng và có thể mở rộng cho các nghiên cứu ứng dụng cho ngữ cảnh tiếng Việt trong tương lai.

2. Phương pháp nghiên cứu

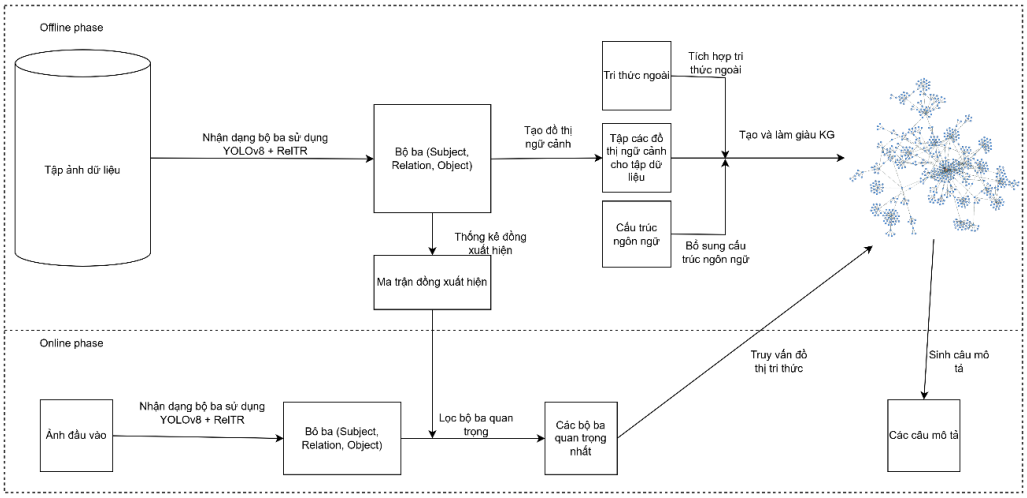
Phương pháp mô tả ảnh đề xuất trong nghiên cứu này được tổ chức thành hai giai đoạn chính: Giai đoạn Offline và Giai đoạn Online. Mục tiêu của hai giai đoạn là tối ưu hiệu quả xử lý, vừa tận dụng tri thức thu được từ dữ liệu huấn luyện, vừa đảm bảo khả năng phản hồi nhanh trong giai đoạn sử dụng thực tế. Trong đó, giai đoạn Offline thực hiện xử lý tri thức và học sâu từ trước, còn giai đoạn Online sẽ sinh mô tả ảnh theo thời gian thực dựa trên các kết quả đã được chuẩn bị. Hình 1 minh họa cho mô hình đề xuất của bài báo.

2.1. Giai đoạn Offline: Trích xuất tri thức và xây dựng đồ thị tri thức

Giai đoạn Offline được tiến hành một lần trước khi triển khai hệ thống chính thức, gồm các bước:

(1) Chuẩn bị dữ liệu

Hệ thống sử dụng tập dữ liệu ảnh quy mô lớn như Visual Genome và MSCOCO. Các ảnh trong Visual Genome và MSCOCO cung cấp annotation chi tiết về đối tượng và quan hệ giữa chúng.



Hình 1. *Mô hình mô tả ảnh sử dụng đồ thị tri thức và mạng học sâu*

(2) Nhận dạng bộ ba bằng mạng học sâu

Sử dụng mô hình YOLOv8 để phát hiện đối tượng, sau đó dùng mạng học sâu RelTR [11] để trích xuất các bộ ba (chủ thể, quan hệ, đối tượng) từ mỗi ảnh. Mỗi bộ ba thể hiện một mối quan hệ ngữ nghĩa trong ảnh.

(3) Tạo đồ thị ngữ cảnh và thống kê đồng xuất hiện

Từ các bộ ba, hệ thống tạo đồ thị ngữ cảnh cho từng ảnh. Đồng thời, tiến hành thống kê tần suất đồng xuất hiện của các cặp đối tượng trong toàn bộ tập dữ liệu, hình thành một bảng đồng xuất hiện (co-occurrence matrix), phản ánh tần suất xuất hiện đồng thời của các cặp đối tượng trong cùng một ảnh trên toàn bộ tập dữ liệu huấn luyện. Cấu trúc bảng là một ma trận vuông n×n với mỗi hàng và cột đại diện cho một lớp đối tượng; ô (i, j) chứa số lượng ảnh mà hai đối tượng thuộc lớp i và j cùng xuất hiện. Ma trận này được dùng để đánh giá mức độ liên kết ngữ nghĩa giữa các đối tượng trong quá trình lọc thông tin.

(4) Tích hợp tri thức ngoài

Các thực thể trong bộ ba được ánh xạ (entity linking) vào đồ thị tri thức như ConceptNet hoặc DBpedia. Quá trình này thu được thông tin liên quan như phân loại, thuộc tính, hành vi, và liên kết ngữ nghĩa giữa các thực thể. Những thông tin này được tích hợp vào đồ thị ngữ cảnh để tạo thành một đồ thị tri thức tăng cường ngữ nghĩa (Semantic-enhanced KG).

(5) Xây dựng cơ sở cấu trúc ngôn ngữ để làm giàu đồ thị tri thức

Dựa trên loại quan hệ và kiểu thực thể, hệ thống xác định các mẫu câu phù hợp (các cấu trúc mô tả ngôn ngữ tự nhiên) và gắn liên kết này vào đồ thị tri thức để phục vụ pha Online. Sáu loại cấu trúc ngôn ngữ được định nghĩa và áp dụng như sau:

- Cấu trúc Động từ đơn: Dùng khi quan hệ chỉ là một động từ cơ bản. Ví dụ: quan hệ là "hit" → "The [subject] hits [the object]".
- Cấu trúc Động từ tiếp diễn: Dùng khi quan hệ chỉ hành động đang diễn ra. Ví dụ: quan hệ là "riding" → "The [subject] is riding the [object]".
- Cấu trúc Động từ với giới từ: Dùng với các quan hệ có kèm giới từ như "on", "sit on", "behind", … Ví dụ: "The [subject] sits on the [object]".
- Cấu trúc Động từ với trạng từ: Mô tả chi tiết hành động như "quickly running" → "The [subject] is quickly running".
- Cấu trúc Mô tả thuộc tính: Sử dụng "has/have" để chỉ thuộc tính đối tượng. Ví dụ: "The man has a red hat".
- Cấu trúc Mô tả trạng thái: Áp dụng cho mẫu "to be + adjective", như "The girl is happy".

Mỗi bộ ba được liên kết với một hoặc nhiều cấu trúc ngôn ngữ tương ứng trong KG. Các cấu trúc ngôn ngữ này đóng vai trò định hướng quá trình tạo câu, giúp mô hình thể hiện chính xác ý nghĩa ngữ cảnh và tăng cường khả năng diễn đạt linh hoạt khi mô tả nội dung hình ảnh.

Toàn bộ quá trình ở giai đoạn Offline nhằm mục tiêu tổ chức lại tri thức một cách hệ thống, giúp mô hình không chỉ dựa vào đặc trưng trực quan mà còn hiểu được bối cảnh ngữ nghĩa của ảnh. Đây là nền tảng then chốt để mô hình có thể sinh ra các mô tả chính xác, tự nhiên và phù hợp hơn với ngữ cảnh thực tế trong giai đoạn Online.

2.2. Giai đoạn Online: Mô tả ảnh thời gian thực dựa trên KG

Giai đoạn Online là quá trình sinh mô tả từ ảnh đầu vào theo thời gian thực, gồm các bước sau:

(1) Phát hiện đối tượng và bộ ba từ ảnh mới

Ảnh đầu vào được xử lý bởi YOLOv8 và mạng học sâu RelTR để sinh ra các bộ ba mô tả quan hệ giữa các đối tượng.

(2) Lọc bộ ba quan trọng

Tập các bộ ba sinh ra từ ảnh đầu vào thường bao gồm cả những quan hệ không rõ ràng hoặc ít giá trị mô tả. Để lựa chọn các quan hệ có ý nghĩa, mô hình sử dụng ma trận đồng xuất hiện, được xây dựng trong giai đoạn Offline từ tập dữ liệu huấn luyện, phản ánh tần suất hai đối tượng xuất hiện đồng thời trong ảnh. Mỗi bộ ba được đánh giá dựa trên giá trị liên kết của cặp đối tượng tương ứng trong ma trận. Những bộ ba có điểm cao hơn được xem là nổi bật và được ưu tiên giữ lại, với tối đa ba bộ ba mỗi ảnh. Cách tiếp cận này giúp tăng tính tập trung cho mô tả và loại bỏ thông tin không cần thiết. Tuy nhiên, trong một số trường hợp, việc giới hạn số lượng có thể dẫn đến bỏ sót các chi tiết bổ trợ quan trọng và đây là điểm hạn chế cần được cân nhắc trong các phiên bản mở rộng của mô hình.

(3) Truy vấn đồ thị tri thức

Các bộ ba được dùng làm truy vấn vào KG đã xây dựng từ pha Offline để thu lại thông tin tri thức và cấu trúc câu phù hợp. Các thông tin này bao gồm: tên đối tượng mở rộng, thuộc tính, cách diễn đạt phù hợp, hành vi ngữ cảnh phổ biến...

(4) Sinh câu mô tả

Dựa trên thông tin truy vấn được từ KG và tập cấu trúc ngôn ngữ đã chuẩn hóa, hệ thống tiến hành sinh câu mô tả cho ảnh. Mỗi bộ ba được ánh xạ với các cấu trúc câu phù hợp, được lựa chọn theo quy trình như sau:

- Khai thác tri thức mở rộng từ kết quả truy vấn, bao gồm từ đồng nghĩa, thuộc tính ngữ nghĩa, hành vi phổ biến và các mẫu diễn đạt tương đương. Nhờ đó, mô hình tăng tính linh hoạt và tránh lặp lại cách diễn đạt.
- Đối chiếu bộ ba với tập cấu trúc câu đã được định nghĩa và làm giàu trong KG, bao gồm sáu loại cấu trúc ngôn ngữ (động từ đơn, tiếp diễn, giới từ, trạng từ, thuộc tính, trạng thái). Hệ thống ưu tiên lựa chọn cấu trúc phù hợp nhất dựa trên:
  - Loại quan hệ (hành động, trạng thái, thuộc tính…),
  - Bối cảnh ngữ nghĩa mở rộng (ví dụ: quan hệ "riding" thường gắn với hành vi nên sẽ chọn cấu trúc tiếp diễn),
  - Sự xuất hiện phổ biến của cấu trúc trong tập huấn luyện.
- Nếu có nhiều cấu trúc phù hợp, hệ thống chọn cấu trúc có mức ưu tiên cao hơn theo thứ tự đã định sẵn trong KG. Nếu không tìm thấy cấu trúc phù hợp, hệ thống sẽ sử dụng mẫu dự phòng "The [subject] is [relation] the [object]" nhằm đảm bảo mọi ảnh đều sinh được mô tả hợp lệ.

Với cách kết hợp giữa tri thức ngữ nghĩa và cấu trúc ngôn ngữ phong phú từ KG, mô hình tạo ra các mô tả có độ đa dạng cao, phản ánh chính xác nội dung ảnh trong nhiều ngữ cảnh khác nhau.

Quy trình hai giai đoạn giúp tách biệt giai đoạn học và giai đoạn triển khai, từ đó tận dụng tri thức trong giai đoạn Offline để tăng tốc và tối ưu chất lượng mô tả trong giai đoạn Online. Mô hình không chỉ đảm bảo khả năng phản hồi thời gian thực, mà còn giữ được chiều sâu ngữ nghĩa trong các mô tả được sinh ra.

3. Kết quả và bàn luận

3.1. Kết quả thực nghiệm

Mô hình mô tả ảnh tích hợp đồ thị tri thức và mạng học sâu được huấn luyện và kiểm thử trên hai tập dữ liệu chuẩn là MSCOCO [9] và Visual Genome[10]. Các chỉ số đánh giá bao gồm BLEU-1 đến BLEU-4 và METEOR – những chỉ số phổ biến trong lĩnh vực xử lý ngôn ngữ tự nhiên để đo lường độ chính xác, tính ngữ nghĩa và mức độ trùng khớp giữa mô tả sinh ra và mô tả tham chiếu. Kết quả thực nghiệm được thể hiện trong Bảng 1.

Bảng 1. *Kết quả đánh giá mô hình đề xuất trên hai tập dữ liệu*

Tập dữ liệu	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
MSCOCO [9]	82,2	70,1	58,5	41,3	31,6
Visual Genome [10]	81,4	69,3	56,7	39,8	30,7

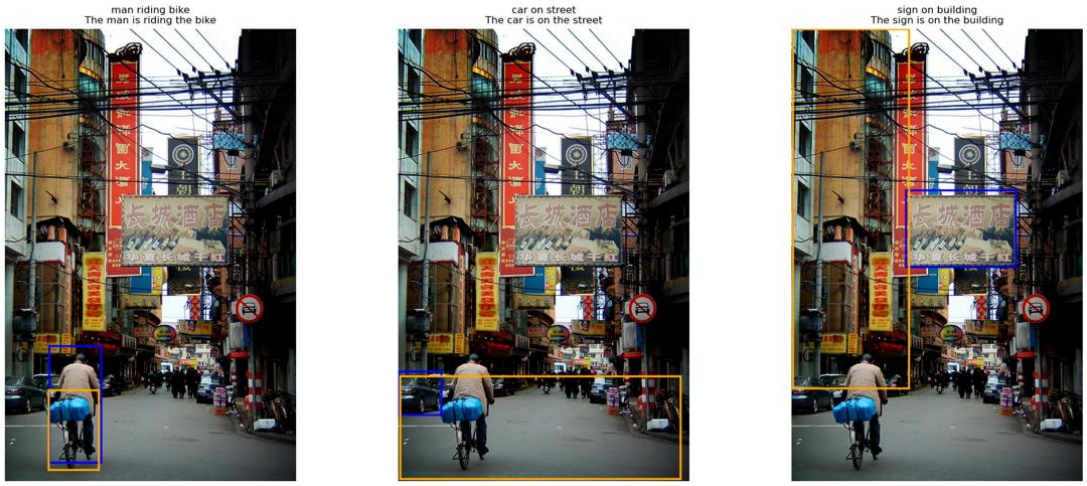
Kết quả cho thấy mô hình hoạt động ổn định và hiệu quả cao trên cả hai tập dữ liệu, với BLEU-4 lần lượt là 41,3 và 39,8 có mức cải thiện đáng kể so với các mô hình truyền thống không sử dụng tri thức nền.

Để đánh giá sâu hơn về hiệu quả, mô hình được so sánh với hai phương pháp nổi bật gần đây: mô hình Concept-based của Osman [12] và mô hình Graph-based SSE của Zhao [13]. Dữ liệu so sánh được thể hiện trong Bảng 2.

Bảng 2. *So sánh với các phương pháp khác*

Mô hình	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
Concept-based [12]	75,8	59,6	45,3	34,1	27,4
Graph-based SSE [13]	81,6	-	-	40,8	30,0
Mô hình đề xuất	82,2	70,1	58,5	41,3	31,6

Kết quả so sánh cho thấy mô hình đề xuất vượt trội về mặt ngữ nghĩa và cú pháp. Đặc biệt, chỉ số BLEU-4 – phản ánh chất lượng tổng thể của câu mô tả – đạt giá trị cao hơn rõ rệt so với các phương pháp còn lại. Bên cạnh đó, chỉ số METEOR đạt 31,6, cho thấy mô hình không chỉ tạo ra các câu có cấu trúc phù hợp, mà còn giữ được ý nghĩa sâu sắc và độ tự nhiên cao. Hình 2 và 3 trình bày một ví dụ minh họa kết quả mô tả hình ảnh.

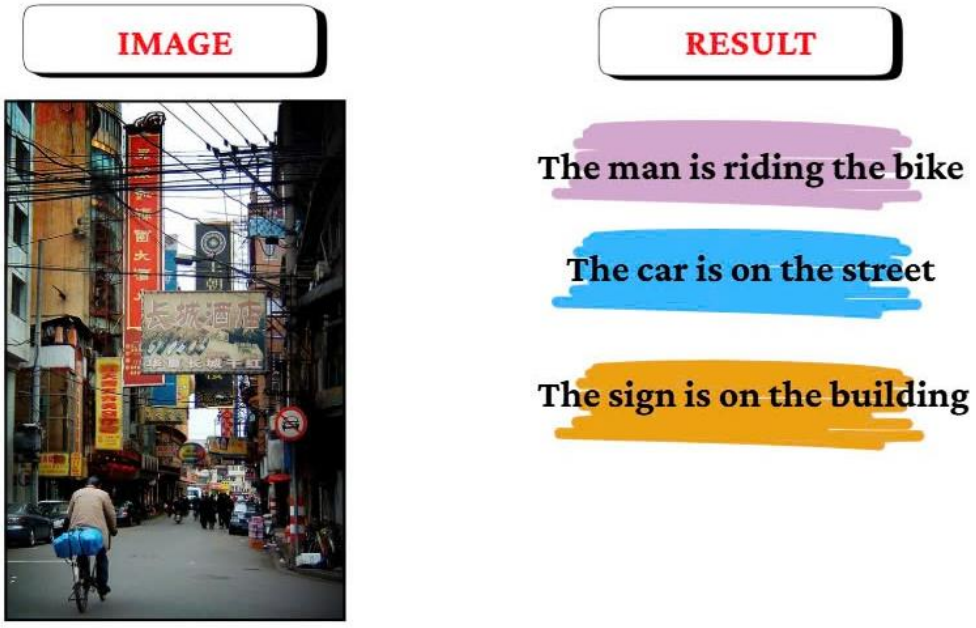


Hình 2. *Lọc lại các bộ ba quan trọng và sinh câu mô tả cho bộ ba*

3.2. Phân tích và thảo luận

Hiệu quả của mô hình phần lớn đến từ khả năng kết hợp tri thức nền thông qua đồ thị tri thức vào quá trình sinh mô tả. Việc ánh xạ các thực thể trong ảnh sang các khái niệm có trong đồ thị như ConceptNet hoặc DBpedia giúp hệ thống hiểu được bản chất, hành vi và mối liên hệ ngữ

nghĩa giữa các thực thể, từ đó tạo ra mô tả giàu tính logic và bám sát ngữ cảnh. Điều này đặc biệt hữu ích trong các ảnh có nhiều đối tượng và quan hệ phức tạp, nơi các mô hình truyền thống dễ mắc lỗi mô tả chung chung hoặc sai lệch.



Hình 3. *Các câu mô tả cho hình ảnh*

Một yếu tố quan trọng khác là cách tổ chức hệ thống theo hai pha: pha Offline tập trung vào việc xây dựng đồ thị tri thức và cấu trúc câu từ trước, giúp giảm tải tính toán tại pha Online, nơi việc truy vấn và sinh mô tả được thực hiện trong thời gian thực. Nhờ vậy, hệ thống có thể phản hồi nhanh chóng mà vẫn duy trì được độ chính xác cao và chiều sâu ngữ nghĩa trong các mô tả được sinh ra.

Tuy nhiên, vẫn còn tồn tại một số hạn chế đáng lưu ý. Việc ánh xạ thực thể từ ảnh vào đồ thị tri thức có thể gặp sai lệch trong các trường hợp tên đối tượng ngắn, đa nghĩa hoặc không rõ ràng (chẳng hạn như từ "bat" có thể là "dơi" hoặc "gậy bóng chày"). Ngoài ra, mô hình hiện tại chưa xử lý tốt các mô tả mang tính trừu tượng, ẩn dụ hoặc cảm xúc, những trường hợp đòi hỏi tri thức nền phức tạp và ngữ cảnh sâu hơn.

Tổng thể, mô hình đề xuất cho thấy tiềm năng rõ rệt trong việc kết hợp giữa đồ thị tri thức và mạng học sâu nhằm tạo ra các mô tả ảnh chính xác, ngữ nghĩa và phù hợp với ngữ cảnh thực tế. Đây là một hướng nghiên cứu đầy triển vọng trong các ứng dụng xử lý ảnh, tìm kiếm đa phương tiện và phân tích ngữ nghĩa ảnh.

4. Kết luận

Nghiên cứu đã đề xuất một phương pháp mô tả ảnh tích hợp giữa đồ thị tri thức và mạng học sâu, triển khai theo hai pha: xử lý tri thức ngoại tuyến và mô tả ảnh trực tuyến theo thời gian thực. Hệ thống kết hợp đồ thị ngữ cảnh với tri thức từ các đồ thị như ConceptNet để tạo ra mô tả ảnh sát ngữ cảnh và giàu ngữ nghĩa hơn. Kết quả trên các tập dữ liệu chuẩn cho thấy mô hình đề xuất vượt trội so với các phương pháp truyền thống về BLEU và METEOR. Trong tương lai, nghiên cứu có thể tiếp tục mở rộng theo hướng cải thiện ánh xạ tri thức, tăng cường mô hình sinh ngôn ngữ, và phát triển các ứng dụng đa ngữ hoặc bản địa hóa mô tả.

Lời cám ơn

Nghiên cứu này được tài trợ bởi nguồn ngân sách khoa học và công nghệ Trường Đại học Sư phạm Thành phố Hồ Chí Minh trong đề tài nghiên cứu của sinh viên năm học 2024 – 2025.

TÀI LIỆU THAM KHẢO/ REFERENCES

[1] M. Chohan, A. Khan, M. S. Mahar, S. Hassan, A. Ghafoor, and M. Khan, "Image Captioning using Deep Learning: A Systematic Literature Review," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 11, no. 5, 2020, doi: 10.14569/IJACSA.2020.0110537.

[2] S. He, W. Liao, H. R. Tavakoli, M. Yang, B. Rosenhahn, and N. Pugeault, "Image captioning through image transformer," *Proceedings of the Asian conference on computer vision*, 2020, doi: 10.48550/arXiv.2004.14231.

[3] X. Yang, H. Zhang, and J. Cai, "Autoencoding and distilling scene graphs for image captioning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 5, pp. 2313-2327, 2020, doi: 10.1109/TPAMI.2020.3042192

[4] R. Li, S. Zhang, D. Lin, K. Chen, and X. He, "From Pixels to Graphs: Open-Vocabulary Scene Graph Generation with Vision-Language Models," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 28076-28086, doi: 10.1109/CVPR52733.2024.02652.

[5] W. Zhao and X. Wu, "Boosting entity-aware image captioning with multi-modal knowledge graph," *IEEE Transactions on Multimedia*, vol. 26, pp. 2659 – 2670, 2023, doi: 10.1109/TMM.2023.3301279.

[6] S. S. Santiesteban, S. Atito, M. Awais, Y. S. Song, and J. Kittler, "Improved Image Captioning Via Knowledge Graph-Augmented Models," *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, doi: 10.1109/ICASSP48485.2024.10447637.

[7] A. Osman, M. A. W. Shalaby, M. M. Soliman, and K. M. Elsayed, "A survey on attention-based models for image captioning," *International Journal of Advanced Computer Science and Application*, vol. 14, no. 2, 2023, doi: 10.14569/IJACSA.2023.0140249.

[8] A. C. Pham, V. Q. Nguyen, T. H. Vuong, and Q. T. Ha, "KTVIC: A Vietnamese Image Captioning Dataset on the Life Domain," *arXiv preprint arXiv:2401.08100*, 2024, doi: 10.48550/arXiv.2401.08100.

[9] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*, 2014, pp. 740–755.

[10] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L. J. Li, D. A. Shamma, M. S. Bernstein, and F. F. Li, "Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations," *Int. J. Comput. Vision*, vol. 123, pp. 32-73, 2017.

[11] Y. Cong, M. Y. Yang, and B. Rosenhahn, "RelTR: Relation Transformer for Scene Graph Generation," *in IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 11169-11183, 2023, doi: 10.1109/TPAMI.2023.3268066.

[12] A. Osman, M. A. W. Shalaby, and M. M. Soliman, "Novel concept-based image captioning models using LSTM and multi-encoder transformer architecture," *Sci. Rep.*, vol. 14, no. 1, 2024, doi: 10.1038/s41598-024-69664-1.

[13] F. Zhao, Z. Yu, T. Wang, and L. Yi, "Image Captioning Based on Semantic Scenes," *Entropy*, vol. 26, no. 10, 2024, Art. no. 876, doi: 10.3390/e26100876.