RESEARCH ARTICLE

CHÚ THÍCH ẢNH DỰA VÀO VIỆC KHAI THÁC MỐI QUAN HỆ GIỮA CÁC ĐỐI TƯỢNG TRONG ẢNH

Đàng Huỳnh Khánh Đoan, Bùi Nguyễn Quỳnh Như, Trần Quốc Long, Lê Văn Khánh, Trần Tú Quyên, Nguyễn Văn Thịnh

Khoa Công nghệ thông tin, Đại học Sư phạm Thành phố Hồ Chí Minh
{4801104027, 4901104105, 4801104085, 4901104068, 4701104179}@student.hcmue.edu.vn,
thinhnv@hcmue.edu.vn

TÓM TẮT—Bài toán chú thích ảnh tự động đóng vai trò quan trọng trong việc kết nối thông tin thị giác với ngôn ngữ tự nhiên. Tuy nhiên, nhiều mô hình hiện nay vẫn còn hạn chế trong việc biểu diễn đầy đủ ngữ nghĩa do chưa khai thác hiệu quả các mối quan hệ giữa đối tượng trong ảnh. Bài báo này đề xuất một mô hình mới kết hợp đặc trưng vùng đối tượng với thông tin quan hệ được biểu diễn thông qua đồ thị cảnh. Cụ thể, chúng tôi sử dụng RelTR để phát hiện bộ ba quan hệ và Graph Transformer để ánh xạ đồ thị thành véc-tơ ngữ nghĩa. Biểu diễn tích hợp sau đó được đưa vào bộ giải mã LSTM nhằm sinh ra chuỗi mô tả ảnh chính xác và giàu thông tin. Kết quả thực nghiệm trên MS COCO, Flickr30K và Flickr8K cho thấy phương pháp đề xuất đạt hiệu suất vượt trội so với nhiều nghiên cứu gần đây, qua các độ đo BLEU, METEOR và CIDEr. Nghiên cứu này góp phần khẳng định tiềm năng của việc tích hợp thông tin quan hệ vào mô hình chú thích ảnh, hướng đến nâng cao độ chính xác và khả năng diễn đạt trong mô tả sinh ra.

Từ khóa— Chú thích ảnh, phát hiện đối tượng, dự đoán mối quan hệ, đồ thị cảnh, mạng nơ-ron học sâu

I. GIỚI THIỆU

Trong bối cảnh gia tăng mạnh mẽ của dữ liệu hình ảnh, nhu cầu hiểu và khai thác thông tin từ ảnh một cách hiệu quả đã thúc đẩy sự phát triển nhanh chóng của các hệ thống trí tuệ nhân tạo có khả năng diễn giải nội dung thị giác. Một trong những hướng nghiên cứu nổi bật trong lĩnh vực này là bài toán tạo mô tả ảnh tự động (image captioning) – tức quá trình tạo văn bản bằng ngôn ngữ tự nhiên để diễn đạt nội dung của một hình ảnh đầu vào [1]. Bài toán này đòi hỏi sự kết hợp hiệu quả giữa các kỹ thuật nhận diện đặc trưng thị giác và mô hình hóa ngôn ngữ, đóng vai trò cầu nối giữa hai lĩnh vực: thị giác máy tính và xử lý ngôn ngữ tự nhiên. Nhờ tính ứng dụng thực tiễn cao, chú thích ảnh đã được triển khai trong nhiều hệ thống thông minh như truy vấn ảnh ngữ nghĩa, trợ lý ảo, và hỗ trợ người khiếm thị tiếp cận nội dung hình ảnh [2].

Các phương pháp chú thích ảnh truyền thống chủ yếu dựa trên kiến trúc encoder–decoder, trong đó CNN được sử dụng để trích xuất đặc trưng ảnh và LSTM đảm nhiệm vai trò phát sinh chú thích dưới dạng văn bản mô tả cho nội dung hình ảnh [3-5]. Việc tích hợp cơ chế chú ý (attention) đã giúp tăng khả năng tập trung vào các vùng ảnh quan trọng, cải thiện độ chính xác và tính mạch lạc của các câu chú thích [6-8]. Tuy nhiên, hầu hết các mô hình hiện tại chỉ tập trung vào đặc trưng toàn cục hoặc từng vùng riêng lẻ, trong khi thông tin về quan hệ ngữ nghĩa giữa các đối tượng trong ảnh – yếu tố then chốt để hiểu ngữ cảnh – vẫn chưa được khai thác một cách hiệu quả. Một số nghiên cứu gần đây đã bắt đầu tích hợp biểu diễn đồ thị cảnh (scene graph) nhằm mô hình hóa quan hệ giữa các đối tượng [9, 10]. Mặc dù đạt được một số kết quả khả quan, các phương pháp này vẫn còn hạn chế trong việc tích hợp sâu và biểu diễn linh hoạt các quan hệ phức tạp trong ảnh.

Để khắc phục những hạn chế nêu trên, bài báo này đề xuất một mô hình chú thích ảnh mới kết hợp giữa phát hiện đối tượng và mô hình hóa quan hệ ngữ nghĩa bằng cách sử dụng đồ thị cảnh (scene graph) [11] - một dạng biểu diễn có cấu trúc, cho phép thể hiện rõ ràng các đối tượng, thuộc tính và mối quan hệ giữa chúng trong ảnh. Cụ thể, mô hình sử dụng RelTR để trích xuất các bộ ba quan hệ (subject–predicate–object) [12], sau đó áp dụng Graph Transformer để ánh xạ đồ thị thành véc-tơ đặc trưng [13]. Các đặc trưng thu được từ đối tượng và đồ thị được tích hợp và đưa vào bộ giải mã LSTM để sinh chú thích. Mô hình đề xuất nhằm cải thiện khả năng hiểu ngữ cảnh và tạo ra mô tả ảnh có độ chính xác cao và giàu ngữ nghĩa hơn.

Đóng góp chính của bài báo gồm:

- Đề xuất mô hình chú thích ảnh tích hợp đặc trưng vùng đối tượng với biểu diễn quan hệ ngữ nghĩa từ đồ thị cảnh, giúp tăng cường khả năng mô hình hóa ngữ cảnh hình ảnh.
- Thực nghiệm toàn diện trên ba tập dữ liệu MS COCO, Flickr30K và Flickr8K cho thấy hiệu quả vượt trội so với các phương pháp baseline và các mô hình tiên tiến gần đây.

Phần còn lại của bài báo gồm: Phần II trình bày tổng quan các công trình liên quan. Phần III mô tả chi tiết mô hình đề xuất. Phần IV trình bày thực nghiệm và phân tích kết quả. Phần V đưa ra kết luận và định hướng nghiên cứu tương lai.

II. CÁC NGHIÊN CỨU LIÊN QUAN

Trong thời gian gần đây, bài toán tạo chú thích tự động cho hình ảnh đã trở thành một chủ đề nghiên cứu thu hút nhiều sự quan tâm, nhờ tiềm năng ứng dụng trong các lĩnh vực như trợ lý ảo, hỗ trợ người khiếm thị, và hệ thống truy xuất hình ảnh dựa trên ngữ nghĩa [14]. Các phương pháp đầu tiên trong lĩnh vực này thường được xây dựng theo khung mã hóa–giải mã (encoder–decoder), trong đó các đặc trưng ảnh được trích xuất thông qua mạng tích chập (CNN), sau đó được đưa vào các mô hình sinh chuỗi như RNN, LSTM hoặc GRU để tạo ra mô tả bằng ngôn ngữ tự nhiên [3, 4].

Tiêu biểu, Kiros và cộng sự (2014) đã đưa ra một phương pháp đa phương thức kết hợp không gian nhúng ảnh–văn bản [3], trong khi Vinyals và cộng sự (2015) phát triển mô hình NIC huấn luyện end-to-end để tối đa hóa xác suất chuỗi từ có điều kiện theo ảnh đầu vào [4]. Các phương pháp này chủ yếu khai thác đặc trưng toàn cục và chưa xét đến thông tin cấu trúc hoặc quan hệ giữa các đối tượng trong ảnh. Tương tự, Hassain và cộng sự (2019) sử dụng DenseNet để trích xuất đặc trưng ảnh đầu vào và áp dụng cơ chế attention trong mạng LSTM nhằm tạo ra chú thích tập trung vào các vùng quan trọng [15]. Mặc dù mô hình đạt kết quả khả quan trên MS COCO theo các độ đo BLEU-2, 3, 4, nhưng vẫn chưa khai thác sâu mối quan hệ giữa các đối tượng trong ảnh.

Các cải tiến tiếp theo tập trung vào tối ưu hóa hiệu quả mô hình hoặc thay thế các thành phần của kiến trúc truyền thống. Patwari và cộng sự (2021) sử dụng Inception-v3 làm encoder kết hợp với GRU làm decoder có tích hợp attention, cho kết quả tốt trên độ đo BLEU-1 đến 4 [16]. Tuy vậy, việc chỉ dựa vào đặc trưng ảnh trích xuất từ CNN huấn luyện trước khiến các mô hình này khó biểu diễn đầy đủ nội dung ngữ nghĩa ảnh, đặc biệt là các tương tác giữa đối tượng. Một ví dụ khác là nghiên cứu của Reddy và cộng sự (2021) với kiến trúc RCNN-LSTM đơn giản, hướng tới hiệu quả triển khai nhưng chưa cải thiện biểu diễn ngữ nghĩa sâu [17]. Al-Malla và cộng sự (2022) [18] cũng khai thác đặc trưng đối tượng kết hợp attention để mô phỏng cách con người hiểu ảnh [18], nhưng chưa sử dụng cấu trúc đồ thị để mô hình hóa quan hệ trực tiếp giữa các thành phần trong ảnh.

Sự phát triển của cơ chế attention đã mở ra hướng tiếp cận hiệu quả hơn trong mô hình hóa thông tin không gian và ngữ nghĩa. Một số công trình tiêu biểu như Bi-LS-AttM (Xie và cộng sự 2023) [6], Deore và cộng sự (2024) [7], và Rahul và cộng sự (2025) [8] đã tận dụng attention theo hướng top-down hoặc hai chiều để tập trung vào các vùng ảnh liên quan trong quá trình sinh mô tả. Tuy nhiên, các mô hình này vẫn chỉ xử lý từng vùng ảnh riêng biệt mà chưa mô hình hóa rõ ràng mối quan hệ giữa các đối tượng – yếu tố then chốt để hiểu ngữ cảnh sâu và chính xác.

Một nhánh nghiên cứu đáng chú ý trong lĩnh vực chú thích ảnh tập trung vào khai thác thông tin quan hệ giữa các đối tượng, khởi nguồn từ công trình của Lu và cộng sự (2016) [9]. Trong nghiên cứu này, tác giả đề xuất mô hình xác định quan hệ nhị phân giữa các cặp đối tượng bằng cách kết hợp biểu diễn đặc trưng thị giác với embedding từ vựng. Dù chưa tích hợp cơ chế attention và chưa đưa trực tiếp vào pipeline sinh mô tả, hướng tiếp cận này đã đặt nền móng cho các phương pháp dựa trên biểu diễn đồ thị sau này. Tiếp nối ý tưởng đó, Yao và cộng sự (2018) phát triển một hệ thống xây dựng đồ thị thể hiện mối quan hệ ngữ nghĩa và không gian giữa các vùng ảnh [10]. Sau đó, đồ thị được mã hóa bằng mạng neural tích chập trên đồ thị (Graph Convolutional Network – GCN) và kết hợp với mô hình LSTM tích hợp attention để tạo ra mô tả. Thực nghiệm cho thấy phương pháp này mang lại cải thiện đáng kể về điểm số CIDEr, dù phạm vi biểu diễn quan hệ còn giới hạn.

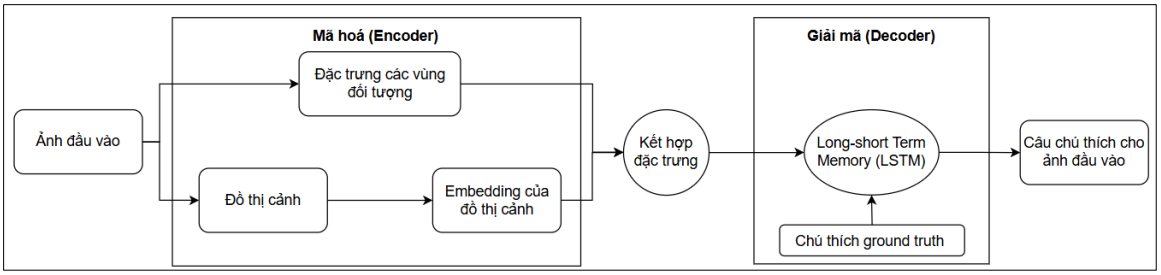
Gần đây, các mô hình dựa trên Transformer đã cho thấy tiềm năng mạnh mẽ trong việc tích hợp đồng thời thông tin thị giác và ngữ nghĩa. GRIT (Nguyen và cộng sự, 2022) là một ví dụ điển hình, trong đó mô hình sử dụng đặc trưng vùng từ Deformable DETR và đặc trưng lưới từ Swin Transformer, kết hợp bằng cross-attention song song [19]. GRIT cho phép huấn luyện end-to-end và đạt kết quả tốt cả về độ chính xác lẫn tốc độ suy luận. Dù chưa biểu diễn quan hệ đối tượng dưới dạng đồ thị, GRIT cho thấy hướng tích hợp ngữ cảnh ảnh vào pipeline mô tả là rất tiềm năng.

Từ việc khảo sát và phân tích các công trình nghiên cứu liên quan, có thể nhận thấy rằng phần lớn các mô hình chú thích ảnh hiện nay tập trung vào việc khai thác đặc trưng vùng và cơ chế attention nhằm cải thiện chất lượng mô tả. Tuy nhiên, chỉ một số ít nghiên cứu thực sự hướng tới việc mô hình hóa và tích hợp thông tin quan hệ giữa các đối tượng trong ảnh – một thành phần thiết yếu để hiểu đầy đủ bối cảnh và biểu đạt ngữ nghĩa toàn diện. Nhằm khắc phục khoảng trống này, nghiên cứu này đề xuất một mô hình chú thích ảnh mới, trong đó đồ thị quan hệ giữa các đối tượng được xây dựng và ánh xạ vào không gian đặc trưng, sau đó tích hợp cùng đặc trưng vùng để làm đầu vào cho bộ giải mã. Hướng tiếp cận này cho phép mô hình khai thác tốt hơn cấu trúc cảnh và quan hệ ngữ nghĩa, từ đó nâng cao độ chính xác và tính diễn đạt của chú thích sinh ra.

III. PHƯƠNG PHÁP ĐỀ XUẤT

Trong nghiên cứu này, một mô hình chú thích ảnh mới được đề xuất như Hình 1. Phương pháp này kết hợp giữa việc phát hiện các vùng đối tượng trong ảnh và phân tích mối quan hệ ngữ nghĩa giữa các đối tượng này nhằm

tạo ra chú thích có độ chính xác và tính mô tả cao. Mô hình đề xuất bao gồm các bước chính sau: đầu tiên, đặc trưng của các vùng đối tượng trong ảnh đầu vào được phát hiện và trích xuất thông qua mô hình Faster R-CNN. Tiếp đó, nhằm khai thác thông tin về mối quan hệ ngữ nghĩa giữa các đối tượng, đồ thị cảnh tương ứng của ảnh được xây dựng và chuyển đổi (embedding) thành véc-tơ đặc trưng. Hai loại đặc trưng này sau đó được tích hợp thành một biểu diễn thống nhất, phản ánh toàn diện nội dung ngữ nghĩa của ảnh. Cuối cùng, biểu diễn này được sử dụng làm đầu vào cho bộ giải mã (Decoder) LSTM để sinh chú thích. Phương pháp đề xuất không chỉ cải thiện khả năng nhận diện đối tượng mà còn tăng cường sự hiểu biết về các mối quan hệ ngữ nghĩa phức tạp giữa chúng. Nhờ đó, các chú thích được tạo ra có chất lượng cao, phản ánh chính xác và đầy đủ nội dung của hình ảnh đầu vào. Nội dung chi tiết của phương pháp bao gồm hai thành phần chính: (A) Bộ mã hóa hình ảnh (Encoder) và (B) Bộ giải mã tạo chú thích (Decoder). Các thành phần này sẽ được mô tả cụ thể trong các tiểu mục tiếp theo.



Hình 1. *Sơ đồ tổng quát của phương pháp chú thích ảnh đề xuất*

A. BỘ MÃ HOÁ HÌNH ẢNH

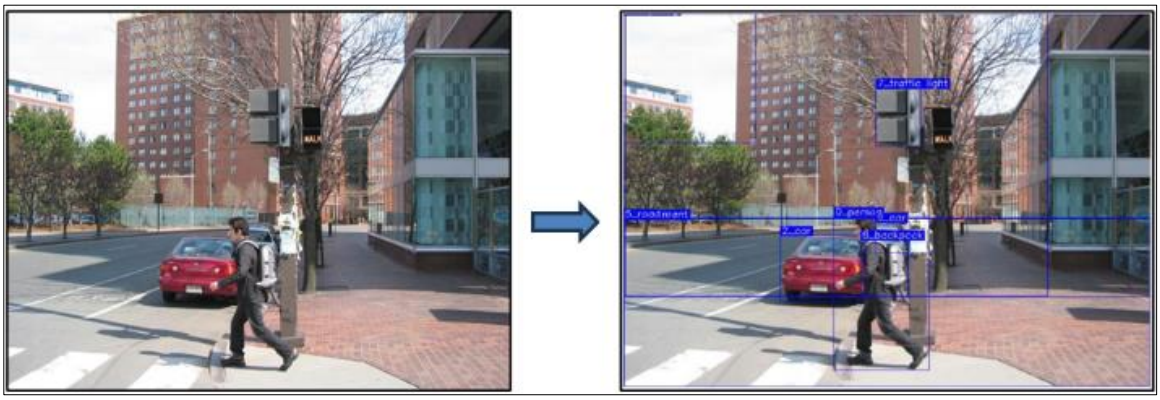
Bộ mã hóa (Encoder) trong hệ thống đề xuất có nhiệm vụ chuyển đổi ảnh đầu vào thành biểu diễn ngữ nghĩa giàu thông tin, phản ánh cả đặc trưng thị giác của từng đối tượng và các mối quan hệ giữa chúng. Kiến trúc của bộ mã hóa bao gồm ba thành phần chính: (1) trích xuất đặc trưng vùng đối tượng bằng mô hình Faster R-CNN, (2) xây dựng đồ thị cảnh biểu diễn các quan hệ giữa đối tượng và (3) ánh xạ đồ thị cảnh sang không gian vector sử dụng Graph Transformer. Các đặc trưng thu được sau đó được tích hợp thành một biểu diễn toàn diện, làm đầu vào cho bộ giải mã LSTM để phát sinh chú thích ảnh.

1. PHÁT HIỆN VÀ TRÍCH XUẤT ĐẶC TRƯNG ĐỐI TƯỢNG BẰNG FASTER R-CNN

Faster R-CNN là mô hình phát hiện đối tượng mạnh mẽ kết hợp giữa Region Proposal Network (RPN) và mạng nơ-ron tích chập (CNN) [20]. Trong phần này, Faster R-CNN được sử dụng để phát hiện các đối tượng trong ảnh và trích xuất đặc trưng vùng tương ứng cho từng đối tượng. Với một ảnh đầu vào I, kết quả thu được là một tập véc-tơ đặc trưng F_I biểu diễn các vùng đối tượng:

$$F_I = \{f_1, f_2, \dots, f_{N_B}\},$$

với f_i là véc-tơ đặc trưng của vùng đối tượng thứ i, N_B là số vùng đối tượng phát hiện trong ảnh.



Hình 2. *Minh họa ảnh đầu vào và kết quả đầu ra của mô hình phát hiện đối tượng trong ảnh*

Hình 2 minh họa kết quả đầu ra của mô hình Faster R-CNN trên một ảnh đầu vào, trong đó các vùng đối tượng được phát hiện được hiển thị cùng với chỉ số và nhãn phân loại tương ứng. Các đặc trưng này đóng vai trò là cơ sở để tích hợp cùng biểu diễn ngữ nghĩa từ đồ thị cảnh ở các bước tiếp theo trong pipeline của mô hình.

2. XÂY DỰNG ĐỒ THỊ CẢNH CHO ẢNH BẰNG RELTR

Để mô hình hóa mối quan hệ giữa các đối tượng, phần này đề xuất xây dựng một đồ thị cảnh $G = (V, E)$ [21]. Trong đó, tập đỉnh $V = \{v_1, v_2, \dots, v_{N_B}\}$ là tập đỉnh đại diện cho các đối tượng được phát hiện trong ảnh. Tập cạnh $E = \{e_{ij}\}$ mô tả mối quan hệ giữa các cặp đối tượng (v_i, v_j). Mỗi quan hệ được biểu diễn dưới dạng bộ ba $(v_i,\ e_{ij},\ v_j)$, với e_{ij} là nhãn quan hệ.

Đồ thị cảnh được xây dựng sử dụng RelTR (Relation Transformer), một mô hình Transformer một giai đoạn được đề xuất bởi Cong và cộng sự [12]. RelTR kế thừa kiến trúc DETR [22] và thực hiện dự đoán trực tiếp các bộ ba quan hệ từ đặc trưng thị giác đầu vào. Mô hình hoạt động theo cơ chế end-to-end, giúp đơn giản hóa pipeline truyền thống gồm nhiều bước trung gian phức tạp. Điểm nổi bật của RelTR là số lượng tham số ít so với phương pháp trước nhưng vẫn đạt hiệu suất cao nhờ thiết kế attention chuyên biệt cho nhận diện quan hệ.

Hình 3 minh họa kiến trúc tổng quát của RelTR. Đầu vào là ảnh đã được mã hóa đặc trưng bằng backbone của DETR. Mô hình sử dụng hai tập truy vấn (queries): Subject queries Q_S và Object queries Q_O. Các truy vấn này là các vector học được đại diện cho các "câu hỏi" tiềm năng về các bộ ba quan hệ.

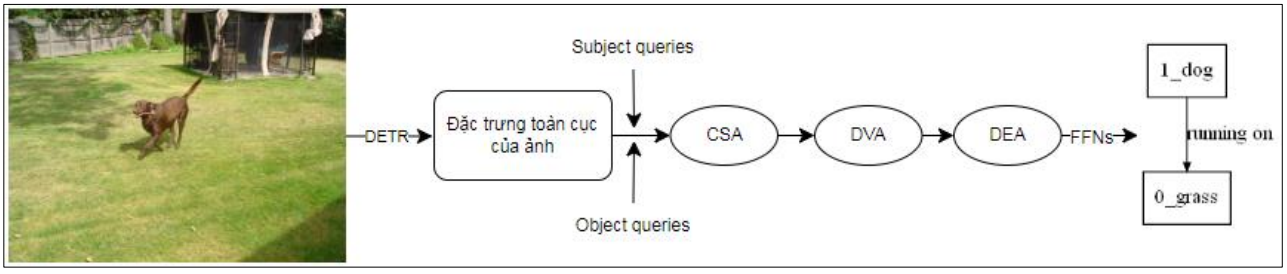
Ba cơ chế attention chính trong RelTR bao gồm:

- **Coupled Self-Attention** (CSA): cho phép các truy vấn chủ thể và đối tượng trong cùng một bộ ba tương tác để cập nhật thông tin ngữ cảnh nội bộ và tương quan giữa các bộ ba khác. Các truy vấn được nối với positional encodings và đưa vào mô-đun multi-head self-attention.
- **Decoupled Visual Attention** (DVA): là cơ chế cross-attention giữa truy vấn và đặc trưng thị giác toàn cục. DVA xử lý Q_S và Q_O một cách độc lập nhằm xác định các vùng ảnh liên quan đến từng đối tượng. Kết quả là các attention heatmaps M_S và M_O chỉ ra các vùng ảnh tương ứng với chủ thể và đối tượng.
- **Decoupled Entity Attention** (DEA): cơ chế này nhằm nâng cao độ chính xác trong việc định vị và phân loại chủ thể và đối tượng. DEA sử dụng cross-attention giữa các truy vấn và kết quả DVA, sau đó kết hợp với hai lớp tuyến tính (Feed-Forward Networks – FFNs) cùng hàm kích hoạt ReLU và lớp chuẩn hóa có kết nối tàn dư (residual connection) để hoàn thiện biểu diễn cuối.

Tiếp theo, các heatmaps M_S, M_O được chuyển đổi thành các vector đặc trưng không gian V_{spa} thông qua một mô-đun convolutional mask head. Sau đó, xác suất nhãn quan hệ (predicate) được tính bằng một mạng MLP:

$$\hat{p}_{prd} = softmax\big(MLP([Q_s, Q_O, V_{spa}])\big)$$

Tập các bộ ba $(v_i,\ e_{ij},\ v_j)$ thu được tạo thành đồ thị cảnh hoàn chỉnh, cung cấp biểu diễn ngữ nghĩa trực quan và đầy đủ cho nội dung ảnh đầu vào. Đồ thị này là thành phần đầu vào thiết yếu cho bước nhúng và tích hợp đặc trưng trong pipeline mô hình chú thích ảnh.



Hình 3. *Kiến trúc tổng quát của phương pháp tạo đồ thị cảnh RelTR*

3. NHÚNG ĐỒ THỊ CẢNH BẰNG GRAPH TRANSFORMER

Sau khi thu thập tập bộ ba quan hệ dạng (v_i, e_{ij}, v_j), đồ thị cảnh $G = (V, E)$ được tiến hành xây dựng, trong đó V là tập các nút đại diện cho đối tượng và E là tập các cạnh có hướng mô tả các mối quan hệ giữa các cặp đối tượng trong ảnh. Tuy nhiên, cấu trúc rời rạc của đồ thị khiến việc sử dụng trực tiếp biểu diễn này làm đầu vào cho các mô hình sinh ngôn ngữ gặp nhiều hạn chế. Do đó, việc chuyển đổi đồ thị sang dạng biểu diễn liên tục trong không gian vector là bước cần thiết nhằm khai thác hiệu quả thông tin cấu trúc.

Sau khi xây dựng đồ thị cảnh $G = (V, E)$, bước tiếp theo là thực hiện chuyển đồ thị thành véc-tơ đặc trưng. Mỗi nút $v_i \in V$ biểu diễn một đối tượng cụ thể trong ảnh đầu vào. Mỗi cạnh $e_{ij} \in E$ thể hiện mối quan hệ giữa cặp đối tượng (v_i, v_j). Mục tiêu là biểu diễn toàn bộ cấu trúc đồ thị dưới dạng véc-tơ đặc trưng để phụ vụ cho bước phát sinh mô tả. Để thực hiện điều này, nghiên cứu sử dụng kiến trúc Graph Transformer được đề xuất bởi Dwivedi và Bresson [13], vốn là một tổng quát hóa của Transformer cổ điển để xử lý dữ liệu dạng đồ thị.

Kiến trúc Graph Transformer kế thừa cơ chế attention từ mô hình Transformer gốc [23], nhưng được điều chỉnh để khai thác cấu trúc liên kết của đồ thị thông qua hai đặc điểm chính: (1) attention được tính theo kết nối cục bộ (local neighborhood) giữa các nút, thay vì toàn bộ đồ thị;(2) mã hóa vị trí của các nút được thực hiện bằng cách sử dụng vector riêng của ma trận Laplacian của đồ thị – gọi là Laplacian Positional Encodings (LapPE) – thay thế cho mã hóa vị trí tuần tự sinusoidal truyền thống.

Cụ thể, mỗi nút $v_i \in V$ được chuyển thành một véc-tơ đặc trưng $y_i \in \mathbb{R}^d$ bằng cách sử dụng hàm nhúng tuyến tính. Các véc-tơ này được cộng với véc-tơ vị trí Laplacian tương ứng (sau khi chiếu tuyến tính) trước khi đưa vào tầng Graph Transformer đầu tiên. Đồng thời, nếu đồ thị có chứa đặc trưng cạnh (edge features), các đặc trưng này cũng được nhúng và duy trì trong pipeline riêng biệt để truyền qua các tầng như mô tả trong kiến trúc mở rộng của Graph Transformer [13].

Trong mỗi tầng của Graph Transformer, biểu diễn của nút được cập nhật thông qua cơ chế multi-head attention giữa các nút lân cận, tiếp theo là mạng feed-forward kết hợp với chuẩn hóa và kết nối tàn dư (residual connections). Các biểu diễn cạnh cũng được cập nhật song song, nếu có. Sự phối hợp giữa attention và positional encoding giúp mô hình học được không chỉ mối quan hệ ngữ nghĩa cục bộ mà còn cả cấu trúc toàn cục của đồ thị.

Quá trình lan truyền thông tin trong đồ thị được thực hiện qua nhiều lớp kế tiếp nhau, giúp mô hình dần học được biểu diễn ngữ nghĩa sâu sắc cho toàn bộ cấu trúc. Kết quả cuối cùng của bước embedding là một véc-tơ biểu diễn tổng thể, ký hiệu là $GE \in \mathbb{R}^d$, phản ánh ngữ nghĩa toàn cảnh của hình ảnh (scene-level semantics). Biểu diễn này sau đó sẽ được tích hợp với các đặc trưng vùng đối tượng, vốn được trích xuất từ mô hình Faster R-CNN, trong bước hợp nhất đặc trưng. Mục tiêu là cung cấp đầu vào phong phú cho bộ giải mã, nhằm nâng cao chất lượng mô tả ảnh sinh ra.

4. KẾT HỢP ĐẶC TRƯNG BẰNG ATTENTIONAL FEATURE FUSION

Sau khi thu được hai nguồn thông tin biểu diễn—(i) đặc trưng vùng đối tượng $\{f_1, f_2, \dots, f_{N_B}\} \in \mathbb{R}^{2048}$, từ mô hình Faster R-CNN và (ii) đặc trưng ngữ nghĩa toàn cục $GE \in \mathbb{R}^{256}$ từ mô hình Graph Transformer—giai đoạn tiếp theo là tích hợp chúng thành một biểu diễn hợp nhất, nhằm cung cấp đầu vào đầy đủ ngữ nghĩa cho bộ giải mã phát sinh chú thích. Thay vì kết hợp đặc trưng bằng phép cộng (+) hay ghép nối đơn gian (concatenation), chúng tôi sử dụng cơ chế Attentional Feature Fusion (AFF) [24], một cơ chế cổng chú ý mềm theo nội dung vừa gọn nhẹ vừa hiệu quả. Quy trình kết hợp đặc trưng bằng AFF được mô tả ngắn gọn qua ba bước sau (đối với từng vùng f_i):

- Bước 1: Chiếu và tính tổng sơ bộ hai nguồn đặt trưng

$$s_i = W_f f_i + GE, W_f \in \mathbb{R}^{256\times2048}, s_i \in \mathbb{R}^{256}$$

- Bước 2: Tính hệ số chú ý

$$\alpha_i = \sigma\big(W_2 ReLU(W_1 \bar{s}_i)\big), \alpha_i \in (0,1)$$
$$\bar{s}_i = \frac{1}{256}\sum_j s_{ij}, W_1 \in \mathbb{R}^{16\times1}, W_2 \in \mathbb{R}^{1\times16}$$

- Bước 3: Hợp nhất mềm

$$z_i = \alpha_i W_f f_i + (1-\alpha_i) GE, z_i \in \mathbb{R}^{256}$$

Tập $Z = \{z_1, z_2, \dots, z_{N_B}\}$ đã cân bằng linh hoạt giữa đặc trưng thị giác (vùng đối tượng) và ngữ nghĩa quan hệ (đồ thị), Z được đưa vào bộ giải mã LSTM để phát sinh chú thích cho hình ảnh. Nhờ hệ số α_i học được, AFF tự điều chỉnh tỷ lệ đóng góp của hai nguồn đặc trưng theo nội dung ảnh, , thiết lập cầu nối chặt chẽ giữa khối xử lý thị giác và ngôn ngữ. Biểu diễn giàu ngữ nghĩa này giúp bộ giải mã tạo ra các chú thích không chỉ chính xác về đối tượng mà còn phản ánh đúng các quan hệ giữa chúng.

B. BỘ GIẢI MÃ PHÁT SINH CHÚ THÍCH

Sau khi đặc trưng của các đối tượng trong ảnh và mối quan hệ giữa chúng được trích xuất và tích hợp thành một biểu diễn ngữ nghĩa toàn diện, bước tiếp theo trong mô hình chú thích ảnh là sử dụng bộ giải mã dựa trên mạng nơ-ron hồi quy LSTM để sinh chú thích mô tả hình ảnh. LSTM là một biến thể của mạng nơ-ron hồi quy (Recurrent Neural Network – RNN), được thiết kế để giải quyết hiệu quả bài toán mô hình hóa chuỗi dài hạn nhờ khả năng duy trì và cập nhật trạng thái bộ nhớ trong suốt quá trình sinh chuỗi. Khác với RNN truyền thống, LSTM có kiến trúc với ba cổng điều khiển (cổng quên, cổng vào, cổng ra) và một trạng thái ô nhớ giúp giảm thiểu hiện tượng mất dần gradient (vanishing gradient), đặc biệt hữu ích trong các tác vụ sinh ngôn ngữ, nơi mỗi từ phụ thuộc vào ngữ cảnh trước đó. Quá trình sinh chú thích từ ảnh được thực hiện bằng một mạng LSTM, bao gồm bốn bước chính: (i) khởi tạo trạng thái từ biểu diễn đặc trưng ngữ nghĩa của ảnh, (ii) tính toán ngữ cảnh chú

ý ở bước t, (iii) cập nhật LSTM và sinh phân phối từ, và (iv) huấn luyện mô hình. Các bước này được mô tả chi tiết như sau:

(1) **Khởi tạo trạng thái LSTM**: Biểu diễn ngữ nghĩa $\bar{z} \in \mathbb{R}^{256}$, kết quả của quá trình kết hợp đặc trưng vùng đối tượng và đặc trưng đồ thị cảnh, được sử dụng để khởi tạo trạng thái ẩn h_0 và trạng thái ô nhớ c_0 của mạng LSTM thông qua các biến đổi phi tuyến:

$$\bar{z} = \frac{1}{N_B}\sum_{1}^{N_B} z_i$$

$h_0 = \tanh(W_h . \bar{z} + b_h), c_0 = \tanh(W_c . \bar{z} + b_c)$

Trong đó $W_h, W_c \in \mathbb{R}^{H\times 256}$, $b_h, b_c \in \mathbb{R}^H$ là các tham số học được trong quá trình huấn luyện, H là số chiều của trạng thái ẩn LSTM.

(2) **Tính ngữ cảnh chú ý ở bước** t: Tại mỗi thời điểm t, bộ giải mã truy vấn tập Z để thu vector ngữ cảnh c_t:

$$e_t^{(i)} = v_a^{\mathrm{T}} tanh(W_z z_i + W_h' h_{t-1}), \alpha_t^{(i)} = softmax\left(e_t^{(i)}\right), c_t = \sum_{i=1}^{N_B} \alpha_t^{(i)} z_i$$

Với $v_a \in \mathbb{R}^H, W_z \in \mathbb{R}^{H\times 256}, W_h' \in \mathbb{R}^{H\times H}$.

(3) **Cập nhật LSTM và sinh phân phối từ**: Input tại bước t là phép ghép giữa embedding của từ trước (e_{t-1}) và vector ngữ cảnh c_t.

$$x_t = [e_{t-1}; c_t], (h_t, c_t) = LSTM(x_t, h_{t-1}, c_{t-1}),$$
$$p(y_t | y_{<t}, I) = softmax(W_o h_t + b_o).$$

Với $W_o \in \mathbb{R}^{|V|\times H}, b_o \in \mathbb{R}^{|V|}$, $|V|$ là số từ trong tập từ vựng của tập dữ liệu.

(4) **Huấn luyện mô hình có giám sát**: Trong giai đoạn huấn luyện, mô hình được tối ưu hóa bằng cách so sánh các từ sinh ra với các từ trong chú thích tham chiếu (ground-truth caption). Hàm mất mát được sử dụng là cross-entropy, được định nghĩa như sau:

$\mathrm{L} = -\sum_{t=1}^{T} \log P(y_t^* | y_{<t}, \mathrm{v})$

Trong đó: y_t^*: là từ đúng tại thời điểm t trong chú thích ground truth, $y_{<t}$: các từ đã sinh trước thời điểm t, T: độ dài chuỗi chú thích.

Quá trình huấn luyện được thực hiện theo cơ chế teacher forcing, trong đó tại mỗi bước thời gian, từ đầu vào là từ đúng từ tập ground-truth chứ không phải từ mô hình sinh ra trước đó.

(5) **Kết quả đầu ra**: Sau quá trình giải mã, LSTM sinh ra một chuỗi từ $\{y_1, y_2, \dots, y_T\}$ tạo thành một mô tả ảnh hoàn chỉnh. Chuỗi từ đầu ra phản ánh không chỉ thông tin về sự hiện diện của các đối tượng trong ảnh, mà còn biểu thị rõ các mối quan hệ và bối cảnh tương ứng giữa chúng, nhờ vào việc khai thác các đặc trưng ngữ nghĩa được tích hợp trong giai đoạn mã hóa.

IV. THỰC NGHIỆM

Dựa trên nền tảng lý thuyết và phương pháp đề xuất, phần này sẽ triển khai chi tiết dữ liệu thực nghiệm, bao gồm các yêu cầu cấu hình cụ thể và tiêu chí đánh giá mô hình. Đồng thời, các thước đo phổ biến trong đánh giá mô hình chú thích hình ảnh sẽ được áp dụng để phân tích hiệu quả của phương pháp. Ngoài ra, phần này cũng trình bày kết quả thực nghiệm, thảo luận về những ưu điểm và hạn chế của phương pháp đề xuất.

A. DỮ LIỆU THỰC NGHIỆM

Mô hình đề xuất được huấn luyện và đánh giá trên ba bộ dữ liệu là MS COCO [25], Flickr30K [26] và Flickr8K [27]. Theo phân chia của Karpathy et al. [28], MS COCO được chia thành 82.783 ảnh huấn luyện, 5.000 ảnh xác thực và 5.000 ảnh kiểm tra. Flickr30K bao gồm 31.783 ảnh với 5 mô tả cho mỗi ảnh, trong đó 1.000 ảnh dùng cho xác thực, 1.000 ảnh cho kiểm tra và phần còn lại cho huấn luyện. Flickr8K gồm 8.092 ảnh, 6.092 ảnh cho huấn luyện, 1.000 ảnh cho xác thực và 1.000 ảnh cho kiểm tra. Trong giai đoạn tiền xử lý, chỉ giữ lại những từ xuất hiện tối thiểu 5 lần để xây dựng từ vựng, thu được 10.010 từ cho MS COCO, 7.414 từ cho Flickr30K và 2.512 từ cho Flickr8K. Độ dài tối đa của mỗi câu chú thích được giới hạn là 20 từ để đảm bảo tính đồng đều trong quá trình huấn luyện.

B. CÁC ĐỘ ĐO ĐÁNH GIÁ

Hiệu quả của mô hình được đánh giá bằng ba độ đo tự động phổ biến là BLEU, METEOR, và CIDEr. Đây là các độ đo được sử dụng rộng rãi trong các nghiên cứu chú thích ảnh, cho phép đánh giá mức độ chính xác, tính tương đồng và sự phù hợp về mặt ngữ nghĩa giữa câu sinh ra bởi mô hình và câu tham chiếu (ground truth).

BLEU (Bilingual Evaluation Understudy) đo lường độ trùng khớp n-gram giữa câu sinh ra bởi mô hình và các câu tham chiếu, với các biến thể BLEU-n (BLEU-1, BLEU-2, v.v.) đánh giá ở các mức độ n-gram khác nhau [29]. Mặc dù đơn giản và hiệu quả trong các bài toán dịch máy, BLEU không xem xét tính ngữ nghĩa hoặc sự mạch lạc trong ngữ cảnh, do đó thường được sử dụng kết hợp với các độ đo bổ sung.

METEOR (Metric for Evaluation of Translation with Explicit ORdering) được thiết kế nhằm khắc phục một số hạn chế của BLEU bằng cách kết hợp nhiều yếu tố: sự trùng khớp về từ gốc (stemming), đồng nghĩa (synonymy), và thứ tự từ [30]. Điểm số METEOR phản ánh tốt hơn mức độ trùng khớp ngữ nghĩa và cấu trúc giữa mô tả sinh và mô tả tham chiếu, đồng thời có tương quan cao hơn với đánh giá của con người.

CIDEr (Consensus-based Image Description Evaluation) đánh giá mức độ đồng thuận giữa câu sinh ra và tập chú thích tham chiếu bằng cách phân tích n-gram đã chuẩn hóa và gán trọng số theo độ đặc trưng (TF-IDF) [31]. Chỉ số này được thiết kế đặc biệt cho bài toán chú thích ảnh, có khả năng phản ánh mức độ phù hợp cả về nội dung lẫn ngữ cảnh ngôn ngữ.

C. CHI TIẾT CÀI ĐẶT

Mô hình được triển khai trên nền tảng PyTorch, với ba thành phần chính: bộ mã hóa vùng đối tượng, bộ dự đoán quan hệ và bộ giải mã sinh chú thích.

Bộ mã hóa hình ảnh sử dụng Faster R-CNN với backbone ResNet-50. Các tầng đầu (conv1, layer1, layer2) được đóng băng để giảm chi phí huấn luyện. Đầu ra từ layer4 (2048 chiều) được đưa vào mô-đun RPN và ROIHead. Các vùng đề xuất được chuẩn hóa bằng RoIAlign thành đặc trưng 256 chiều.

Bộ dự đoán quan hệ sử dụng mô hình RelTR gồm 6 tầng Transformer với chiều ẩn 256, 8 head attention, dropout 0.1 và pre-layer normalization. Mỗi tầng sử dụng FFN ẩn kích thước 2048 và hàm kích hoạt ReLU. Chỉ các quan hệ có độ tin cậy trên 0.3 được giữ lại làm đầu ra của đồ thị cảnh.

Bộ giải mã là một mạng LSTM gồm 4 tầng, chiều ẩn 256. Đặc trưng ảnh (256 chiều) được dùng để khởi tạo trạng thái LSTM. Phần nhúng từ sử dụng BERT (số chiều nhúng = 256). Mô hình được huấn luyện bằng thuật toán Adam (learning rate = 0.0001, betas = [0.9, 0.999]) với hàm mất mát cross-entropy và kỹ thuật teacher forcing.

Các thực nghiệm được thực hiện trên máy tính có cấu hình: Intel Core i5-12450HX, GPU NVIDIA RTX 3050 (6GB VRAM), RAM 12GB.

D. KẾT QUẢ THỰC NGHIỆM

Kết quả định lượng trong **Bảng 1** chứng thấy hiệu quả của hai thành phần chính của mô hình: (i) khai thác ngữ nghĩa quan hệ bằng đồ thị cảnh và (ii) cơ chế hợp nhất *Attentional Feature Fusion*. Ở MS-COCO – tập dữ liệu có bối cảnh phức tạp với nhiều đối tượng tương tác và có quy mô xấp xỉ 118 ngàn ảnh – mô hình đạt 83.2 BLEU-1, 28.8 BLEU-4, 26.3 METEOR và 86.9 CIDEr. Trên Flickr30K (xấp xỉ 31 ngàn ảnh) kết quả vẫn duy trì ở mức cao với 72.8 BLEU-1, 26.2 BLEU-4, 24.7 METEOR và 85.2 CIDEr; điều này chứng tỏ thông tin quan hệ được đồ thị cung cấp giúp mô hình giữ vững độ phong phú ngữ nghĩa dù dữ liệu suy giảm bốn lần. Đáng lưu ý, ở Flickr8K – chỉ có 5 000 ảnh huấn luyện – hệ thống vẫn ghi nhận 80.4 BLEU-1, 26.9 BLEU-4, 23.3 METEOR và 85.1 CIDEr, cho thấy khả năng khái quát hoá mạnh mẽ trong điều kiện dữ liệu rất hạn chế.

Bảng 1. *Hiệu suất của phương pháp chú thích ảnh trên các tập dữ liệu thực nghiệm*

Tập dữ liệu	**BLEU1**	**BLUE4**	**METEOR**	**CIDEr**
MS COCO	83.2	28.8	26.3	86.9
Flickr30K	72.8	26.2	24.7	85.2
Flickr8K	80.4	26.9	23.3	85.1

So sánh với các phương pháp tiêu biểu trên MS-COCO (**Bảng 2**) cho thấy mô hình đề xuất có độ chính xác vượt trội. So với Google NIC, mô hình tăng +16.6 BLEU-1, +4.2 BLEU-4, +2.6 METEOR và +1.4 CIDEr. Với Show, Attend and Tell (SAT) [5], mức cải thiện đạt +11 – 13 BLEU-1 và +3.8 – 4,5 BLEU-4 tuỳ cấu hình Hard-/Soft-ATT. So với kiến trúc hiện đại Bi-LS-AttM, mô hình của chúng tôi vượt +14.4 BLEU-1, +3.6 BLEU-4 và +4.8 METEOR, đồng

thời thiết lập giá trị CIDEr cao nhất trong nhóm so sánh. Những chênh lệch này cho thấy AFF – bằng cách điều tiết mềm tầm quan trọng tương đối giữa đặc trưng vùng và đặc trưng quan hệ – giúp bộ giải mã mô hình hoá tốt hơn các n-gram dài và các từ khóa giàu ngữ nghĩa. Trên tập dữ liệu MS COCO (Bảng 2), mô hình đề xuất thể hiện sự cải thiện đáng kể ở tất cả các chỉ số đánh giá, đặc biệt là các chỉ số phản ánh độ chính xác và tính đa dạng của các chú thích được tạo ra (BLUE4, CIDEr). Điều này chứng tỏ tính hiệu quả của việc kết hợp đặc trưng vùng đối tượng được trích xuất bằng Faster R-CNN cùng với embedding đồ thị cảnh, giúp nâng cao khả năng biểu diễn quan hệ giữa các đối tượng trong hình ảnh.

Bảng 2. *So sánh hiệu suất chú thích ảnh giữa các phương pháp trên tập dữ liệu MS COCO*

Phương pháp	**BLEU1**	**BLUE4**	**METEOR**	**CIDEr**
Google NIC (2015) **[4]**	66.6	24.6	23.7	85.5
Show, attend and tell (Hard-ATT) (2015) [5]	71.8	25.0	23.0	-
Show, attend and tell (Soft-ATT) (2015) [5]	70.7	24.3	23.9	-
Dense_Soft-ATT (2019) [15]	68.3	22.9	22.6	74.3
En-De-Cap (2021) [16]	70.6	24.3	-	
Bi-LS-AttM (2023) [6]	68.8	25.2	21.5	-
Đề xuất của chúng tôi	83.2	28.8	26.3	86.9

Xu hướng tương tự được ghi nhận trên Flickr30 K (Bảng 3). Mô hình đề xuất đạt 72.8 BLEU-1 và 26.2 BLEU-4, cao hơn SAT lần lượt ≈ +6 và ≈ +7 điểm; đồng thời vượt Bi-LS-AttM +8.3 BLEU-1 và +6.0 BLEU-4. Điểm METEOR 24.7 vượt mức 18–19 của các mô hình so sánh tới 5 điểm, trong khi CIDEr 85.2 cách biệt lớn so với Xception. Những kết quả này khẳng định vai trò của đồ thị cảnh – cung cấp ngữ nghĩa quan hệ phong phú – và AFF – đảm bảo hai dòng thông tin vùng-đồ thị được phối hợp tối ưu – đặc biệt hữu dụng khi dữ liệu huấn luyện hạn chế.

Bảng 3. *So sánh hiệu suất chú thích ảnh giữa các phương pháp trên tập dữ liệu Flickr30K*

Phương pháp	**BLEU1**	**BLUE4**	**METEOR**	**CIDEr**
Google NIC (2015) [4]	66.3	18.3	-	-
Show, attend and tell (Hard-ATT) (2015) [5]	66.9	19.9	18.5	-
Show, attend and tell (Soft -ATT) (2015) [5]	66.7	19.1	19.5	-
Xception (2022) [18]	39.9	6.2	12.3	14.8
Bi-LS-AttM (2023) [6]	64.5	20.2	18.6	-
Đề xuất của chúng tôi	72.8	26.2	24.7	85.2

Nhìn chung, kết quả thực nghiệm trên cả ba tập dữ liệu MS COCO, Flickr30K và Flickr8K đều khẳng định rõ ràng hiệu quả của phương pháp đề xuất, qua đó minh chứng cho lợi ích đáng kể của việc tích hợp embedding đồ thị cảnh vào bài toán chú thích ảnh.

V. KẾT LUẬN

Trong nghiên cứu này, chúng tôi đề xuất một mô hình chú thích ảnh tích hợp đặc trưng vùng và quan hệ ngữ nghĩa giữa các đối tượng. Mô hình sử dụng Faster R-CNN để trích xuất đặc trưng vùng ảnh. Tiếp theo, RelTR được dùng để phát hiện các quan hệ giữa các đối tượng. Các quan hệ được tổ chức dưới dạng đồ thị cảnh có cấu trúc rõ ràng.
Graph Transformer được áp dụng để nhúng đồ thị vào không gian biểu diễn liên tục. Đặc trưng tích hợp từ ảnh và đồ thị được đưa vào bộ giải mã LSTM. LSTM sẽ sinh chuỗi mô tả bằng ngôn ngữ tự nhiên từ biểu diễn ngữ nghĩa tổng hợp. Thử nghiệm trên tập MS COCO, Flickr30K và Flickr8K cho kết quả vượt trội so với baseline. Mô hình đạt điểm số cao hơn ở các độ đo BLEU, METEOR và CIDEr. Đặc biệt, điểm METEOR và CIDEr cho thấy khả năng sinh mô tả ngữ nghĩa tốt. Điều này khẳng định hiệu quả của việc tích hợp thông tin quan hệ vào pipeline chú thích. Bên cạnh ý nghĩa học thuật, mô hình đề xuất còn có tiềm năng ứng dụng cao trong các hệ thống thực tế như: hỗ trợ người khiếm thị tiếp cận nội dung hình ảnh thông qua mô tả tự động, cải tiến các hệ thống tìm kiếm hình ảnh theo ngữ nghĩa, hoặc tích hợp vào trợ lý ảo trong các môi trường tương tác đa phương thức. Khả năng sinh chú thích giàu ngữ nghĩa giúp nâng cao mức độ hiểu ngữ cảnh, mở rộng phạm vi ứng dụng trong các bài toán thị giác-ngôn ngữ hiện đại. Trong tương lai, chúng tôi hướng đến việc mở rộng mô hình với các nguồn tri

thức ngữ nghĩa ngoài như ConceptNet hoặc AMR nhằm tăng cường khả năng hiểu ngữ cảnh sâu hơn; đồng thời mở rộng thực nghiệm trên các bộ dữ liệu dễ phát sinh lỗi, làm rõ giới hạn và khả năng khái quát của mô hình trong các tình huống thực tế phức tạp.

VI. LỜI CẢM 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 NCKH của sinh viên năm học 2024 – 2025.

VII. TÀI LIỆU THAM KHẢO

[1] A. S. Al-Shamayleh, O. Adwan, M. A. Alsharaiah, A. H. Hussein, Q. M. Kharma, and C. I. Eke, "A comprehensive literature review on image captioning methods and metrics based on deep learning technique," *Multimedia Tools and Applications,* vol. 83, pp. 34219-34268, 2024.

[2] A. Jamil, K. Mahmood, M. G. Villar, T. Prola, I. D. L. T. Diez, M. A. Samad*, et al.*, "Deep learning approaches for image captioning: Opportunities, challenges and future potential," *IEEE Access,* 2024.

[3] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," *arXiv preprint arXiv:1411.2539,* 2014.

[4] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3156-3164.

[5] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov*, et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*, 2015, pp. 2048-2057.

[6] T. Xie, W. Ding, J. Zhang, X. Wan, and J. Wang, "Bi-LS-AttM: A bidirectional LSTM and attention mechanism model for improving image captioning," *Applied Sciences,* vol. 13, p. 7916, 2023.

[7] S. Deore, T. Bagwan, P. Bhukan, H. Rajpal, and S. Gade, "Enhancing image captioning and auto-tagging through a FCLN with faster R-CNN integration," *Inf. Dyn. Appl,* vol. 3, pp. 12-20, 2023.

[8] M. Rahul, A. Jahnavi, K. Anusha, G. A. Merlyn, and A. V. S. Teja, "Natural Language Processing: Generating Captions from Image."

[9] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei, "Visual relationship detection with language priors," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, 2016, pp. 852-869.

[10] T. Yao, Y. Pan, Y. Li, and T. Mei, "Exploring visual relationship for image captioning," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 684-699.

[11] X. Chang, P. Ren, P. Xu, Z. Li, X. Chen, and A. Hauptmann, "A comprehensive survey of scene graphs: Generation and application," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 45, pp. 1-26, 2021.

[12] Y. Cong, M. Y. Yang, and B. Rosenhahn, "Reltr: Relation transformer for scene graph generation," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 45, pp. 11169-11183, 2023.

[13] V. P. Dwivedi and X. Bresson, "A generalization of transformer networks to graphs. arXiv," *arXiv preprint arXiv:2012.09699,* 2020.

[14] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga, "A comprehensive survey of deep learning for image captioning," *ACM Computing Surveys (CsUR),* vol. 51, pp. 1-36, 2019.

[15] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, H. Laga, and M. Bennamoun, "Attention-based image captioning using DenseNet features," in *Neural Information Processing: 26th International Conference, ICONIP 2019, Sydney, NSW, Australia, December 12–15, 2019, Proceedings, Part V 26*, 2019, pp. 109-117.

[16] N. Patwari and D. Naik, "En-de-cap: An encoder decoder model for image captioning," in *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, 2021, pp. 1192-1196.

[17] S. S. S. Reddy, A. Kumar, and M. Jyaram, "An Image Sentence Generation Based on Deep Neural Network Using RCNN-LSTM Model," in *2021 Sixth International Conference on Image Information Processing (ICIIP)*, 2021, pp. 364-368.

[18] M. A. Al-Malla, A. Jafar, and N. Ghneim, "Image captioning model using attention and object features to mimic human image understanding," *Journal of Big Data,* vol. 9, p. 20, 2022.

[19] V.-Q. Nguyen, M. Suganuma, and T. Okatani, "Grit: Faster and better image captioning transformer using dual visual features," in *European Conference on Computer Vision*, 2022, pp. 167-184.

[20] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE transactions on pattern analysis and machine intelligence,* vol. 39, pp. 1137-1149, 2016.

[21] H. Li, G. Zhu, L. Zhang, Y. Jiang, Y. Dang, H. Hou*, et al.*, "Scene graph generation: A comprehensive survey," *Neurocomputing,* vol. 566, p. 127052, 2024.

[22] Y. Wang, J. Xu, and Y. Sun, "End-to-end transformer based model for image captioning," in *Proceedings of the AAAI conference on artificial intelligence*, 2022, pp. 2585-2594.
[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez*, et al.*, "Attention is all you need," *Advances in neural information processing systems,* vol. 30, 2017.
[24] Y. Dai, F. Gieseke, S. Oehmcke, Y. Wu, and K. Barnard, "Attentional feature fusion," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2021, pp. 3560-3569.
[25] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár*, et al.*, "Microsoft coco captions: Data collection and evaluation server," *arXiv preprint arXiv:1504.00325,* 2015.
[26] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Transactions of the association for computational linguistics,* vol. 2, pp. 67-78, 2014.
[27] M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics," *Journal of Artificial Intelligence Research,* vol. 47, pp. 853-899, 2013.
[28] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3128-3137.
[29] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311-318.
[30] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005, pp. 65-72.
[31] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4566-4575.

IMAGE CAPTIONING BASED ON EXPLOITING RELATIONSHIPS BETWEEN OBJECTS IN IMAGES

Dang Huynh Khanh Doan, Bui Nguyen Nhu Quynh, Tran Quoc Long, Le Van Khanh, Tran Tu Quyen, Nguyen Van Thinh

ABSTRACT—Automatic image captioning is crucial in bridging visual information and natural language. However, existing models often struggle to capture the full semantic content of an image due to the limited exploitation of inter-object relationships. This paper proposes a novel image captioning framework that integrates regional object features with relational information from scene graphs. Specifically, the model employs RelTR to detect subject–predicate–object triplets and utilizes a Graph Transformer to encode the scene graph into rich semantic embeddings. These integrated representations are then fed into an LSTM-based decoder to generate accurate and informative image descriptions. Experimental results on two benchmark datasets, MS COCO, Flickr30K and Flickr8K, demonstrate that the proposed approach outperforms several recent methods across standard evaluation metrics including BLEU, METEOR, and CIDEr. The findings highlight the potential of incorporating relational information into image captioning systems to enhance the generated descriptions' accuracy and expressiveness.

Keywords— Image captioning, object detection, relation prediction, scene graph, deep neural network

Đàng Huỳnh Khánh Đoan là sinh viên năm 3 ngành Công nghệ thông tin, trường Đại học Sư phạm Thành phố Hồ Chí Minh. Hiện đang quan tâm về các nghiên cứu trong lĩnh vực học máy và ứng dụng.

Bùi Nguyễn Quỳnh Như là sinh viên năm 2 ngành Công nghệ thông tin, trường Đại học Sư phạm Thành phố Hồ Chí Minh. Hiện đang quan tâm về các nghiên cứu trong lĩnh vực học máy và ứng dụng.

Trần Quốc Long là sinh viên năm 3 ngành Công nghệ thông tin, trường Đại học Sư phạm Thành phố Hồ Chí Minh. Hiện đang quan tâm về các nghiên cứu trong lĩnh vực học máy và ứng dụng.

Lê Văn Khánh là sinh viên năm 2 ngành Công nghệ thông tin, trường Đại học Sư Phạm Thành phố Hồ Chí Minh. Hiện đang quan tâm về các nghiên cứu trong lĩnh vực học máy và ứng dụng.

Trần Tú Quyên là sinh viên năm 4 ngành Công nghệ thông tin, trường Đại học Sư phạm Thành phố Hồ Chí Minh. Hiện đang quan tâm về các nghiên cứu trong lĩnh vực học máy và ứng dụng.

Nguyễn Văn Thịnh nhận bằng thạc sĩ ngành Hệ thống thông tin tại Trường Đại học Khoa học Tự nhiên – Đại học Quốc gia TP. Hồ Chí Minh năm 2012. Hiện nay, ông là nghiên cứu sinh ngành Khoa học máy tính tại Học viện Khoa học và Công nghệ – Viện Hàn lâm Khoa học và Công nghệ Việt Nam, đồng thời là giảng viên tại Khoa Công nghệ thông tin, Trường Đại học Sư phạm TP. Hồ Chí Minh. Nghiên cứu của ông tập trung vào trí tuệ nhân tạo và các ứng dụng, đặc biệt là học sâu đa phương thức và các mô hình ngôn ngữ–thị giác, hướng đến các bài toán như truy vấn thông tin và tạo chú thích ảnh tự động.