MÔ HÌNH TÌM KIẾM ẢNH DỰA TRÊN ĐỒ THỊ NGỮ CẢNH

Nguyễn Phương Nam[1,2*], Văn Thế Thành[2], Lê Mạnh Thạnh[1]

[1] Khoa Công nghệ thông tin, Trường Đại học Khoa học, Đại học Huế

[2] Khoa Công nghệ thông tin, Trường Đại học Sư phạm Thành phố Hồ Chí Minh

*Email: npnam@hueuni.edu.vn

Ngày nhận bài: 19/01/2024; ngày hoàn thành phản biện: 20/02/2024; ngày duyệt đăng: 5/3/2024

TÓM TẮT

Đồ thị ngữ cảnh mô tả đối tượng, thuộc tính và mối quan hệ giữa các đối tượng trong một ngữ cảnh nhất định của ảnh. Trong các nhiệm vụ tìm kiếm cần sự hiểu biết về mối quan hệ ngữ nghĩa và cách sắp xếp giữa các đối tượng trong không gian ảnh, sử dụng đồ thị ngữ cảnh có thể cải thiện hiệu suất và khả năng của các hệ thống tìm kiếm ảnh. Bài nghiên cứu này trình bày tóm tắt một số phương pháp tìm kiếm ảnh dựa trên đồ thị ngữ cảnh, những đóng góp và hạn chế của các kỹ thuật xác định mối quan hệ đối tượng trong bài toán xây dựng đồ thị ngữ cảnh cho ảnh. Trên cơ sở những đánh giá này, mô hình DPTree được đề xuất nhằm dự đoán mối quan hệ giữa các đối tượng dựa trên tập luật cây quyết định và làm cơ sở cho việc trích xuất tập quan hệ bộ ba từ ảnh, từ đó phát triển mô hình cho bài toán tìm kiếm ảnh dựa trên đồ thị ngữ cảnh.

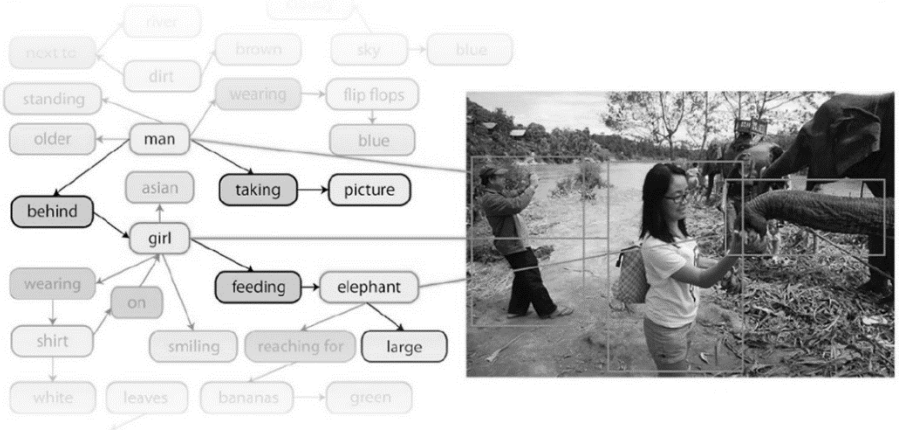
Từ khóa: Đồ thị ngữ cảnh, tìm kiếm ảnh, tìm kiếm ảnh theo ngữ nghĩa.

1. MỞ ĐẦU

Với sự bùng nổ dữ liệu, lượng thông tin tăng trưởng khổng lồ đang gây khó khăn cho các chương trình tìm kiếm, truy xuất, biểu diễn và liên kết thông tin [1], trong đó có dữ liệu ảnh. Để giải quyết vấn đề tìm kiếm ảnh tương tự nhanh chóng, hiệu quả từ tập dữ liệu ảnh lớn là một thách thức trong lĩnh vực thị giác máy tính.

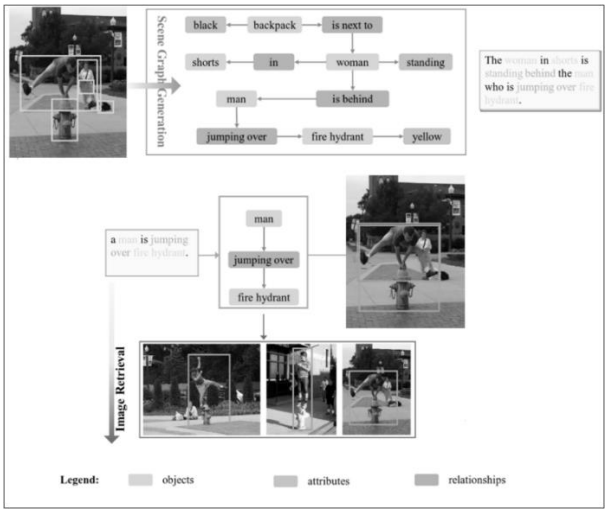
Đồ thị ngữ cảnh (scene graph) là cấu trúc dữ liệu được sử dụng cho việc mô tả đối tượng, thuộc tính và các mối quan hệ của đối tượng trong một ngữ cảnh nhất định. Một đồ thị ngữ cảnh có thể biểu diễn ngữ nghĩa chi tiết của các tập dữ liệu ngữ cảnh, là công cụ hiệu quả để mã hóa hình ảnh thành các yếu tố ngữ nghĩa trừu tượng mà không phụ thuộc vào lớp và thuộc tính của đối tượng hay mối quan hệ giữa các đối tượng. Ngoài tìm kiếm ảnh, đồ thị ngữ cảnh được ứng dụng cho nhiều tác vụ ứng dụng trong

xử lý ảnh như chú thích ảnh, tạo ảnh, hỏi-đáp bằng hình ảnh và phát hiện các mối quan hệ giữa các đối tượng trong ảnh [2], [3], [4]. Hình 1 minh họa về đồ thị ngữ cảnh của ảnh.



Hình 1. Đồ thị ngữ cảnh của ảnh [5]

Tuy còn nhiều thách thức liên quan đến phương pháp xây dựng nhưng so với những cách tiếp cận khác, khi được sử dụng một cách hiệu quả đồ thị ngữ cảnh có thể cải thiện đáng kể hiệu suất và khả năng của các hệ thống tìm kiếm ảnh. Đặc biệt trong các nhiệm vụ tìm kiếm cần sự hiểu biết về mối quan hệ ngữ nghĩa và cách sắp xếp giữa các đối tượng trong không gian ảnh, cho phép hiểu một cách toàn diện hơn về nội dung ảnh ngoài việc chỉ nhận dạng các đối tượng riêng lẻ; hoặc khi cần sự hiểu biết về thông tin ngữ cảnh tổng thể của ảnh nhằm cải thiện tính liên quan và độ chính xác của các truy vấn phức tạp được hiệu quả hơn. Hình 2 minh họa mô hình tìm kiếm ảnh ngữ nghĩa bằng đồ thị ngữ cảnh.



Hình 2. Minh họa tìm kiếm ảnh ngữ nghĩa bằng đồ thị ngữ cảnh [11]

2. CÁC CÔNG TRÌNH TÌM KIẾM ẢNH TIẾP CẬN ĐỒ THỊ NGỮ CẢNH

J. Johnson và cộng sự (2015) [6] phát triển mô hình miền điều kiện ngẫu nhiên (CRF - Conditional Random Field) cho việc tìm kiếm ảnh ngữ nghĩa dựa trên đồ thị ngữ cảnh. Đây là một trong những công trình nghiên cứu cơ bản đầu tiên sử dụng đồ thị ngữ cảnh cho bài toán tìm kiếm ảnh. Nhóm tác giả thiết kế mô hình nhằm nội suy đồ thị ngữ cảnh cơ sở của ảnh đầu vào và sử dụng đồ thị ngữ cảnh này để tìm kiếm ảnh tương tự về ngữ nghĩa. Tuy nhiên, do sử dụng đồ thị ngữ cảnh của mỗi ảnh cụ thể làm nền tảng nên độ phủ trong kết quả tìm kiếm vẫn chưa cao.

Zellers và cộng sự (2018) [7] phân tích rằng mỗi yếu tố của ngữ cảnh mô tả đều có các cấu trúc đồng nhất và phát triển mô hình cấu trúc mạng neural MOTIFNET đề xuất kiến trúc phân rã đồ thị ngữ cảnh thành các pha dự đoán khung, nhãn và mối liên hệ giữa các phân đoạn trong ảnh. Công trình thực hiện xây dựng đồ thị ngữ cảnh dựa trên các mối quan hệ thường gặp nhất giữa các cặp đối tượng trong tập dữ liệu Visual Genome và gọi cấu trúc biểu diễn này là mô-típ (motif). Công trình cho kết quả thực nghiệm phân tích đồ thị ngữ cảnh tốt hơn so với các phương pháp trước đây nhất nhờ khả năng phân tích các cấu trúc biểu diễn quan hệ ở mức toàn cục.

Tang và cộng sự (2019) [8] đề xuất mô hình cây ngữ cảnh trực quan VCTree nhằm thực hiện suy luận hình ảnh ở mức cao như tạo đồ thị ngữ cảnh và hỏi-đáp trực quan (VQA). Công trình sử dụng VCTree và thuật giải LSTM để mã hóa ngữ cảnh cho phép việc truyền thông điệp cụ thể hơn và mã hóa hiệu quả các mối quan hệ song song, phân cấp giữa các đối tượng bằng cấu trúc cây nhị phân. Tuy nhiên do sử dụng ma trận trọng số trong quá trình tạo cấu trúc cây nên sự hiệu quả khi dự đoán mối quan hệ giữa đối tượng phụ thuộc vào độ chính xác của ánh xạ từ quan hệ sang trọng số.

P. Tian và cộng sự (2021) [9] thực hiện quá trình tạo chú thích hình ảnh dựa trên đồ thị ngữ ảnh. Quá trình tạo ra đồ thị ngữ cảnh được xây dựng dựa trên các mối quan hệ đặc trưng của các đối tượng và phân loại mối quan hệ dựa trên mạng neural học sâu. Từ các mối quan hệ đặc trưng, nhóm tác giả tổng hợp tạo thành đồ thị cho ngữ cảnh của hình ảnh. Đây là một bài toán chú thích hình ảnh, nhóm tác giả chưa thực hiện quá trình tìm kiếm hình ảnh với một ảnh đầu vào do đó chưa thực hiện quá trình truy hồi và tìm kiếm tập ảnh tương tự, ngữ nghĩa cho ảnh đầu vào.

Jiang và cộng sự (2023) [10] đề xuất phương pháp HierMotif kết hợp xác suất Bayes để suy luận mối quan hệ giữa các đối tượng trong ngữ cảnh của ảnh bằng cách sử dụng cấu trúc phân cấp, phân loại đối tượng và mối quan hệ thành nhóm loại (super-categories) riêng biệt. Mục tiêu là giải quyết vấn đề mất cân bằng giữa các lớp đối tượng của tập dữ liệu trong quá trình xây dựng đồ thị ngữ cảnh. Với sự kế thừa sử dụng mô-típ của Zellers [7], công trình đã trích xuất được những ngữ cảnh rộng có trong ảnh,

tuy nhiên vẫn chưa giải quyết được việc dự đoán các mối quan hệ phức tạp hoặc các ngữ cảnh có sự nhập nhằng về đối tượng.

Mô hình DR-Net [12] kết hợp mạng nơ-ron sâu và CRF đưa ra mô hình mối quan hệ dựa trên thống kê nhằm nâng cao hiệu quả việc dự đoán các mối liên kết ngữ nghĩa trực quan. Mô hình SG-CRF [13] giải quyết sự nhập nhằng giữa chủ ngữ và vị ngữ trong mối quan hệ bộ ba bằng đề xuất một mạng ngữ nghĩa dựa trên CRF với chuỗi các lớp quan hệ xác định chủ ngữ và vị ngữ các đối tượng trong ảnh. Mô hình ViP-CNN [14] đề xuất xây dựng đồ thị ngữ cảnh dựa trên xử lý mối quan hệ trực quan giữa các đối tượng bằng cụm từ mô tả với ba đặc trưng: tên mô tả đối tượng, không gian và ngữ nghĩa. Mô hình này huấn luyện kết hợp đồng thời các đặc trưng mô tả nhằm phân tích sự tương tác và phát hiện sự phụ thuộc lẫn nhau. Cấu trúc truyền thông điệp theo cụm từ được đề xuất trong mô hình ViP-CNN để mô hình hóa sự phụ thuộc thông tin lẫn nhau giữa các đặc trưng mô tả cục bộ. Factorizable Net [15] là mô hình xây dựng đồ thị ngữ cảnh dựa trên đồ thị con. Đồ thị ban đầu được phân rã thành các đồ thị con và sau đó được gom nhóm, mỗi đồ thị con gồm có các đối tượng và tập mối quan hệ giữa chúng. Hướng tiếp cận của mô hình này là trích xuất các vùng ảnh có đối tượng bằng kiến trúc mạng RPN và gom nhóm lại thành từng cặp để xây dựng một đồ thị liên thông mạnh trong đó mỗi cặp đối tượng kết nối với nhau bằng cặp cạnh có hướng. Sau đó các vùng ảnh được kết nối với nhau thành đồ thị con bằng cách kết nối các cạnh.

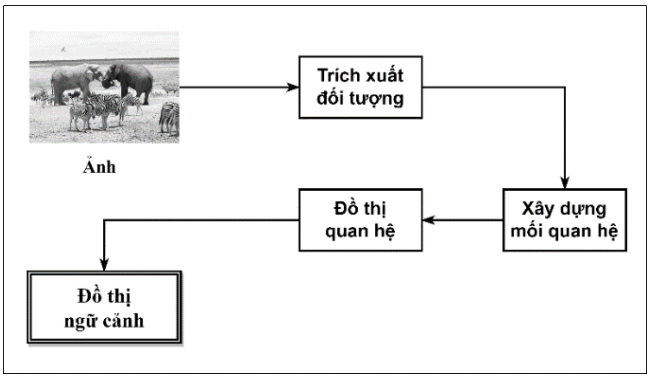
Phần lớn các công trình tập trung vào kỹ thuật xây dựng đồ thị ngữ cảnh cho ảnh ở mức toàn cảnh; chưa tập trung nghiên cứu mối quan hệ của từng cặp đối tượng cũng như cấu trúc dữ liệu mô tả mối quan hệ bộ ba, cấu trúc có thể sinh ra tập luật cho việc dự đoán các mối quan hệ mới, chưa được gán nhãn; đồng thời vẫn chưa áp dụng các kỹ thuật khai thác dữ liệu trên các dữ liệu mô tả như: kỹ thuật phân lớp, gom cụm, mạng neural đặc trưng, …

3. PHƯƠNG PHÁP XÂY DỰNG ĐỒ THỊ NGỮ CẢNH

3.1. Tổng quan xây dựng đồ thị ngữ cảnh

Xây dựng đồ thị ngữ cảnh có hai xu hướng tiếp cận chính: (1) hướng tiếp cận bằng phương pháp dưới-lên: hướng này được chia thành hai pha là phát hiện đối tượng và nhận diện mối quan hệ theo cặp đối tượng [16]; (2) hướng tiếp cận bằng phương pháp trên-xuống: hướng tiếp cận thực hiện đồng thời hai tác vụ phát hiện đối tượng và nhận diện mối quan hệ [14], [17]. Việc thực hiện nội suy và các tác vụ trực quan ở cấp cao thường bao gồm việc nhận diện đối tượng, dự đoán vị trí tương đối dựa trên tọa độ và phát hiện hay nhận diện mối quan hệ vị từ từng cặp giữa các đối tượng. Các bước chính cho quy trình xây dựng đồ thị ngữ cảnh [15], [17], [18] có thể trình bày tổng quát như sau: phát hiện đối tượng nhằm xác định các đỉnh của đồ thị (bao gồm các thông tin

thuộc tính, đặc trưng của đối tượng); dự đoán, định nghĩa mối quan hệ bộ ba giữa các đỉnh đối tượng, từ đó xác định cạnh của đồ thị; các bước được thực hiện lặp lại cho mỗi ảnh và tập ảnh để tạo đồ thị ngữ cảnh. Hình 3 là sơ đồ khối mô tả quá trình xây dựng đồ thị ngữ cảnh từ ảnh.



Hình 3. Sơ đồ khối xây dựng đồ thị ngữ cảnh

3.2. Phương pháp xây dựng đồ thị ngữ cảnh

Các thuật toán học sâu hiện nay giúp phát hiện, phân loại các đối tượng trong ảnh với độ chính xác cao nhưng việc hiểu, suy luận, xác định được mối quan hệ giữa các đối tượng trong một ngữ cảnh nhất định đối với máy tính vẫn đang là bài toán phức tạp, vẫn còn một khoảng cách so với con người. Nhiệm vụ dự đoán, biểu diễn mối quan hệ được xem là nền tảng cơ bản cho việc biểu diễn ảnh một cách trực quan ở mức độ ngữ nghĩa, đó có thể mối quan hệ bộ ba, quan hệ về vị trí không gian, thuộc tính, so sánh hoặc các tác động tương hỗ giữa hai đối tượng… Đồ thị ngữ cảnh là một trong kỹ thuật hiệu quả cho nhiệm vụ đó. Vì được xây dựng bằng việc phân tích mối quan hệ giữa nhiều đối tượng trong tập dữ liệu ảnh, nên đồ thị ngữ cảnh cần xem xét sự liên quan giữa nhiều đối tượng với nhau không chỉ tập trung vào một đối tượng cụ thể.

Dự đoán và nội suy mối quan hệ giữa các đối tượng trong ảnh là yêu cầu cốt lõi trong mô hình xây dựng đồ thị ngữ cảnh cho ảnh. Thuật toán này đươc phát biểu tổng quát như sau:

- Đầu vào là hai đối tượng trong ảnh.
- Nội suy mối quan hệ giữa đối tượng.
- Tạo ra một bộ ba quan hệ <đối tượng, vị từ, đối tượng>.
- Đầu ra là tập bộ ba cho tất cả đối tượng được xác định trong ảnh.

Qua khảo sát các công trình nghiên cứu liên quan, kỹ thuật nội suy mối quan hệ đang được thực hiện theo hai hướng là sử dụng kiến trúc mạng neural học sâu và kiến trúc cây phân cấp. Đối với kiến trúc mạng học sâu, vấn đề cần huấn luyện với lượng

dữ liệu đủ lớn để xác định mối quan hệ nhưng hiện chưa có tập dữ liệu như thế hoặc phân phối có mức độ giảm mật độ thấp (long-tail distribution). Bên cạnh đó, khi có sự tăng trưởng thêm phân lớp thì cần phải huấn luyện lại mô hình học sâu. Với kiến trúc phân cấp như C-Tree [19], KD-Tree [20], RS-Tree [21]… dù đã giải quyết được vấn đề đáp ứng tập dữ liệu huấn luyện nhỏ, có thể mở rộng khi có sự tăng trưởng phân lớp nhưng thời gian, chi phí huấn luyện vẫn còn cao do cần tối ưu hóa giá trị trọng số hoặc phải huấn luyện lại mô hình khi có phân lớp mới.

Dựa trên những đánh giá trên, mô hình DPTree được đề xuất kết hợp sử dụng mô hình cây quyết định và phương pháp thống kê nhằm sinh ra tập luật nội suy mối quan hệ bộ ba giữa hai đối tượng. Trong mô hình này, các đối tượng trên ảnh đầu được trích xuất và tạo ra các vector đặc trưng tương ứng, từ đó làm đầu vào cho mô hình cây quyết định kết hợp hàm Gaussian và phân tích Bayes để nội suy mối quan hệ. Từ tập dữ liệu hình ảnh Visual Genome, các đặc trưng đối tượng được thống kê, huấn luyện và phân loại như một nút (node) trong mô hình học sâu. Mối quan hệ giữa cặp đối tượng được quyết định ở nút lá sau khi huấn luyện mô hình học sâu liên kết các nút con mang đặc trưng trên. Hình 4 minh họa mô hình DPTree dự đoán mối quan hệ giữa hai đối tượng trong ảnh.

Mô hình thực hiện qua hai pha gồm: (1) pha nội suy tập luật quyết định mối quan hệ, (2) pha dự đoán mối quan hệ giữa các đối tượng từ một ảnh đầu vào. Quá trình này được mô tả cụ thể như sau:

(1) Pha nội suy tập luật quyết định:

- Bước 1: Xác định khung giới hạn (bounding box), trích xuất, phân lớp đối tượng trên mỗi ảnh từ tập dữ liệu ảnh.

+ Bước 1a: Trích xuất và xây dựng tập đặc trưng đối tượng qua mạng neural.

+ Bước 1b: Ghép từng cặp đối tượng có thứ tự trên ảnh.

- Bước 2: Tổng hợp từ tập vector đặc trưng cho từng cặp đối tượng để tạo tập vector đặc trưng ghép và thực hiện tính toán hàm Gaussian trả về xác suất trên phân phối Gaussian.

- Bước 3: Thống kê từ tập dữ liệu ảnh mối quan hệ giữa các đối tượng và thực hiện tính toán hàm mật độ xác suất của phân phối Gaussian.

- Bước 4: Tổng hợp từ tập vector đặc trưng ghép và xác suất mối quan hệ qua phân phối Gaussian để xác định phân phối xác suất vector đặc trưng cho mối quan hệ.

- Bước 5: Thực hiện huấn luyện và phân loại mối quan hệ với mỗi đặc trưng bằng mô hình cây quyết định; quá trình huấn luyện này được xem như một quy trình khép kín trong một nút cho từng đặc trưng.

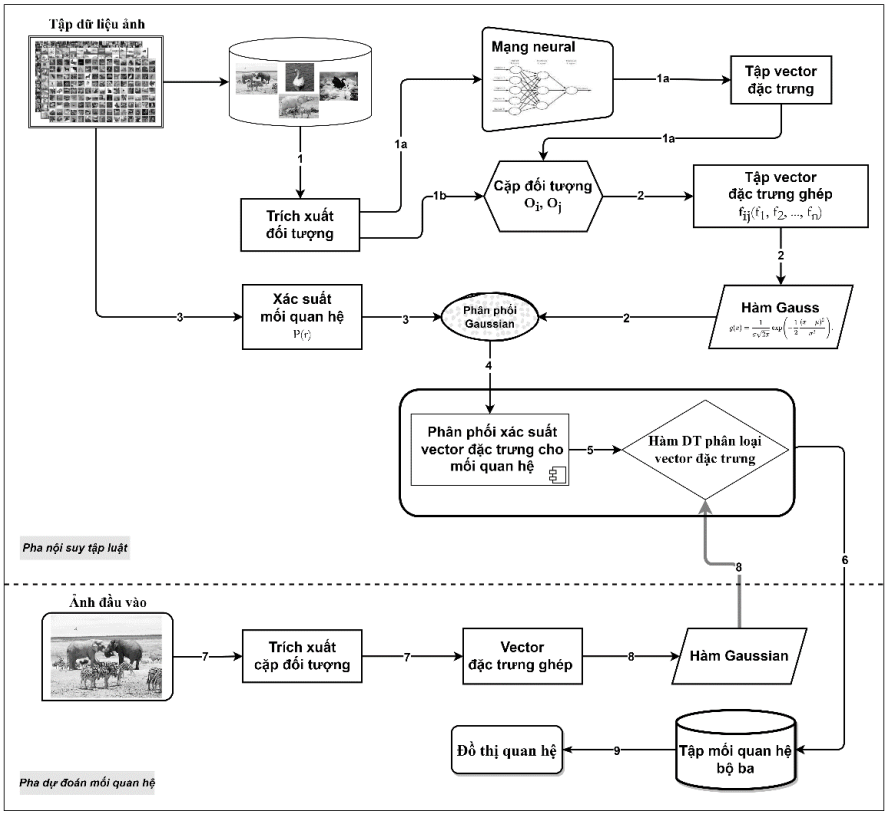
- Bước 6: Từ mô hình huấn luyện nội suy mối quan hệ, xây dựng tập các quan hệ bộ ba giữa các đối tượng trong ảnh.

(2) Pha dự đoán mối quan hệ:

- Bước 7: Với mỗi ảnh đầu vào thực hiện trích xuất đối tượng và xác định vector đặc trưng ghép từng cặp đối tượng.

- Bước 8: Thực hiện tính toán hàm Gauss trên vector đặc trưng ghép và đưa vào mô hình hàm DT (decision tree) phân loại vector đặc trưng đã huấn luyện.

- Bước 9 : Kết xuất kết quả dự đoán mối quan hệ đối tượng; xây dựng tập quan hệ bộ ba làm cơ sở cho việc xây dựng đồ thị quan hệ và đồ thị ngữ cảnh.



Hình 4. Mô hình dự đoán mối quan hệ đối tượng DPTree

4. KẾT LUẬN

Bài nghiên cứu đã trình bày khảo sát một số công trình liên quan về tìm kiếm ảnh ngữ nghĩa dựa trên đồ thị ngữ cảnh; hướng tiếp cận và phân tích các kỹ thuật xây dựng đồ thị ngữ cảnh cho ảnh và đề xuất mô hình DPTree dự đoán mối quan hệ giữa các đối tượng sử dụng mô hình cây quyết định và phương pháp thống kê nhằm sinh ra tập luật nội suy. Việc xây dựng đồ thị ngữ cảnh để đạt được mức ngữ nghĩa bậc cao

trong bối cảnh thực tế vẫn còn nhiều thách thức về độ phức tạp, sự nhập nhằng trong định nghĩa mối quan hệ tương hỗ giữa hai đối tượng, và việc phân loại các lớp đối tượng chưa được gán nhãn. Với xu hướng nghiên cứu và ứng dụng của đồ thị ngữ cảnh ngày càng mở rộng thì những khảo sát phương pháp xây dựng đồ thị ngữ cảnh trong khuôn khổ bài nghiên cứu còn chưa bao quát hết các khía cạnh liên quan. Trong tương lai, nhóm tác giả sẽ mở rộng khảo sát các phương pháp xây dựng đồ thị ngữ cảnh với các tri thức dẫn mở (prior knowledge) và các hướng tiếp cận trên các bộ dữ liệu khác nhau.

TÀI LIỆU THAM KHẢO

[1] Hoàng Hữu Hạnh and Lê Mạnh Thạnh, Giáo trình Web ngữ nghĩa. NXB Giáo dục, 2012.

[2] S. Lee, J.-W. Kim, Y. Oh, and J. H. Jeon, "Visual Question Answering over Scene Graph," in 2019 First International Conference on Graph Computing (GC), IEEE, Sep. 2019, pp. 45–50. doi: 10.1109/GC46384.2019.00015.

[3] J. Jia et al., "Image captioning based on scene graphs: A survey," Expert Syst Appl, vol. 231, p. 120698, Nov. 2023, doi: 10.1016/j.eswa.2023.120698.

[4] J. Johnson, A. Gupta, and L. Fei-Fei, "Image Generation from Scene Graphs," Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1219–1228, Dec. 2018, doi: 10.1109/CVPR.2018.00133.

[5] R. Krishna et al., "Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations," Int J Comput Vis, vol. 123, no. 1, pp. 32–73, Feb. 2016, Accessed: Dec. 01, 2023. [Online]. Available: http://arxiv.org/abs/1602.07332

[6] J. Johnson et al., "Image retrieval using scene graphs," in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Jun. 2015, pp. 3668–3678. doi: 10.1109/CVPR.2015.7298990.

[7] R. Zellers, M. Yatskar, S. Thomson, and Y. Choi, "Neural Motifs: Scene Graph Parsing with Global Context," in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, Jun. 2018, pp. 5831–5840. doi: 10.1109/CVPR.2018.00611.

[8] K. Tang, H. Zhang, B. Wu, W. Luo, and W. Liu, "Learning to Compose Dynamic Tree Structures for Visual Contexts," in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Jun. 2019, pp. 6612–6621. doi: 10.1109/CVPR.2019.00678.

[9] P. Tian, H. Mo, and L. Jiang, "Scene graph generation by multi-level semantic tasks," Applied Intelligence, vol. 51, no. 11, pp. 7781–7793, Nov. 2021, doi: 10.1007/s10489-020-02115-2.

[10] B. Jiang and C. J. Taylor, "Hierarchical Relationships: A New Perspective to Enhance Scene Graph Generation," Mar. 2023, Accessed: Nov. 12, 2023. [Online]. Available: http://arxiv.org/abs/2303.06842.

[11] H. Li et al., "Scene Graph Generation: A comprehensive survey," Neurocomputing, vol. 566, p. 127052, Jan. 2024, doi: 10.1016/j.neucom.2023.127052.

[12] B. Dai, Y. Zhang, and D. Lin, "Detecting Visual Relationships with Deep Relational Networks," Apr. 2017.

[13] W. Cong, W. Wang, and W.-C. Lee, "Scene Graph Generation via Conditional Random Fields," Nov. 2018, [Online]. Available: http://arxiv.org/abs/1811.08075

[14] Y. Li, W. Ouyang, X. Wang, and X. Tang, "ViP-CNN: Visual Phrase Guided Convolutional Neural Network," in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Jul. 2017, pp. 7244–7253. doi: 10.1109/CVPR.2017.766.

[15] Y. Li, W. Ouyang, B. Zhou, J. Shi, C. Zhang, and X. Wang, "Factorizable Net: An Efficient Subgraph-Based Framework for Scene Graph Generation," 2018, pp. 346–363. doi: 10.1007/978-3-030-01246-5_21.

[16] W. Liao, B. Rosenhahn, L. Shuai, and M. Y. Yang, "Natural Language Guided Visual Relationship Detection," in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE, Jun. 2019, pp. 444–453. doi: 10.1109/CVPRW.2019.00058.

[17] Y. Li, W. Ouyang, B. Zhou, K. Wang, and X. Wang, "Scene Graph Generation from Objects, Phrases and Region Captions," in 2017 IEEE International Conference on Computer Vision (ICCV), IEEE, Oct. 2017, pp. 1270–1279. doi: 10.1109/ICCV.2017.142.

[18] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei, "Scene Graph Generation by Iterative Message Passing," in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Jul. 2017, pp. 3097–3106. doi: 10.1109/CVPR.2017.330.

[19] N. T. U. Nhi, T. M. Le, and Thanh The Van, "A Model of Semantic-Based Image Retrieval Using C-Tree and Neighbor Graph," Int J Semant Web Inf Syst, vol. 18, no. 1, pp. 1–23, Feb. 2022, doi: 10.4018/IJSWIS.295551.

[20] N. T. Dinh, T. T. Van, and T. M. Le, "Semantic Relationship-Based Image Retrieval Using KD-Tree Structure," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 13757 LNAI, Springer Science and Business Media Deutschland GmbH, 2022, pp. 455–468. doi: 10.1007/978-3-031-21743-2_36.

[21] L. T. V. Thanh, T. T. Van, and T. M. Le, "Semantic-Based Image Retrieval Using RS-Tree and Knowledge Graph," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 13757 LNAI, Springer Science and Business Media Deutschland GmbH, 2022, pp. 481–495. doi: 10.1007/978-3-031-21743-2_38.

A SCENE GRAPH-BASED MODEL FOR IMAGE RETRIEVAL

Nguyen Phuong Nam[1,2*], Van The Thanh[2], Le Manh Thanh[1]

[1] Faculty of Information Technology, University of Sciences, Hue University

[2] Faculty of Information Technology, Ho Chi Minh City University of Education

*Email: npnam@hueuni.edu.vn

ABSTRACT

Scene graphs describe the objects, attributes, and their relationships in a certain context of the image. Using scene graphs can improve the performance and capability of image retrieval systems in search tasks that require the understanding of semantic relationships and spatial arrangements among objects in the image. This research will provide an overview of different scene graph-based image search approaches, as well as the pros and cons of object relationship determination methods for creating scene graphs for images. The DPTree model was proposed, which predicts the relationships between objects using a set of decision tree rules based on these evaluations. Afterwards, this model is applied to extract triple sets from images, creating an image search task model based on scene graphs.

Keywords: Scene graph, image retrieval, semantic-based image retrieval.

Nguyễn Phương Nam sinh năm 1981, tốt nghiệp kỹ sư chuyên ngành Công nghệ thông tin tại trường Khoa học tổ chức (FOS), Đại học Belgrade, Cộng hòa Serbia, tốt nghiệp thạc sĩ chuyên ngành Khoa học máy tính năm 2020 tại Trường Đại học Sư phạm TP.HCM. Hiện là nghiên cứu sinh chuyên ngành Khoa học máy tính tại Trường Đại học Khoa học, Đại học Huế.

Lĩnh vực nghiên cứu: công nghệ phần mềm, xử lý ảnh và tìm kiếm ảnh.

Văn Thế Thành sinh năm 1979, tốt nghiệp cử nhân chuyên ngành Toán tin năm 2001 tại trường Đại học Khoa học Tự nhiên, Đại học Quốc gia TP.HCM, tốt nghiệp thạc sĩ chuyên ngành Khoa học máy tính năm 2008 tại Đại học Quốc gia TP.HCM; nhận học vị tiến sĩ Khoa học máy tính năm 2016 tại trường Đại học Khoa học, Đại học Huế.

Lĩnh vực nghiên cứu: xử lý ảnh, khai thác dữ liệu ảnh và tìm kiếm ảnh.

Lê Mạnh Thạnh sinh năm 1953, nhận học vị tiến sĩ Khoa học máy tính năm 1994 tại Đại học Budapest (ELTE), Hungary; nhận học hàm Phó giáo sư năm 2004 tại trường Đại học Khoa học, Đại học Huế.

Lĩnh vực nghiên cứu: cơ sở dữ liệu, cơ sở tri thức và lập trình logic.