Tăng cường chú thích ảnh thông qua tích hợp đồ thị tri thức và mạng R-CNN

Nguyễn Kim Quốc[1,*], Đinh Xuân Thao[2], Đặng Như Phú[1]

[1]Trường Đại học Nguyễn Tất Thành, TP Hồ Chí Minh, Việt Nam
[2]Trường Cao đẳng Bách khoa Nam Sài Gòn, TP Hồ Chí Minh, Việt Nam
[*]nkquoc@ntt.edu.vn

Tóm tắt

Trong bối cảnh số hóa phát triển mạnh, chú thích ảnh tự động đóng vai trò quan trọng nhưng các mô hình truyền thống còn hạn chế trong hiểu ngữ cảnh Semantic. Nghiên cứu nhằm nâng cao độ chính xác của chú thích ảnh tự động bằng cách tích hợp đồ thị tri thức vào R-CNN. Phương pháp tiếp cận gồm xây dựng đồ thị tri thức từ ImageNet và COCO, trích xuất đặc trưng bằng CNN, đề xuất vùng bằng Selective Search, phân loại softmax, hồi quy vị trí, cùng quy trình tiền xử lý và huấn luyện với thuật toán hạ gradient ngẫu nhiên (learning rate 0,001, 50 epochs, tỉ lệ 80:20). Kết quả cho thấy mô hình đạt accuracy 96 % và IoU 0,75 trên 2 000 ảnh kiểm thử, vượt R-CNN truyền thống (85 %, IoU 0,6). Việc tích hợp đồ thị tri thức giúp giảm lỗi trong các bối cảnh phức tạp và cải thiện độ đầy đủ ngữ nghĩa. Độ phức tạp tính toán tăng khoảng 20 %, nhưng vẫn đáp ứng yêu cầu xử lý gần thời gian thực và cho hiệu suất cao hơn Fast R-CNN và YOLO. Nghiên cứu này đóng góp phần quản lý ảnh và thiết bị di động, phục vụ cho các ngành liên quan trong việc sử dụng hình ảnh.

Nhận 27/08/2025
Được duyệt 12/11/2025
Công bố 28/11/2025

Từ khóa

CTA, Đồ thị tri thức, R-CNN, ngữ cảnh Semantic, thị giác máy tính.

1 Đặt vấn đề

Trong thời đại số hóa hiện nay, ảnh số đã trở thành một phần không thể thiếu trong đời sống hàng ngày và các lĩnh vực công nghiệp như truyền thông, giáo dục, y tế và an ninh [1]. Với sự gia tăng nhanh chóng của dữ liệu hình ảnh trên các nền tảng như Google Photos, Instagram và các hệ thống lưu trữ đám mây, nhu cầu tự động hóa quá trình chú thích ảnh (CTA) để tổ chức, tìm kiếm, và phân tích nội dung ngày càng trở nên cấp thiết [2]. Theo báo cáo, hơn 2,1 nghìn tỷ ảnh kỹ thuật số được tạo ra trên toàn thế giới trong năm 2025, đặt ra thách thức lớn về việc phát triển các mô hình thông minh để xử lý lượng dữ liệu khổng lồ này. Tuy nhiên, các phương pháp CTA truyền thống, dù đã đạt được một số tiến bộ, vẫn gặp khó khăn trong việc hiểu ngữ cảnh sâu sắc và mối quan hệ phức tạp giữa các đối tượng trong ảnh, dẫn đến các mô tả thiếu chính xác hoặc không phản ánh đầy đủ ý nghĩa.

Lịch sử phát triển của CTA bắt đầu từ các phương pháp thủ công, nơi con người trực tiếp mô tả nội dung ảnh, nhưng phương pháp này không khả thi khi đối mặt với khối lượng dữ liệu lớn do tốn thời gian và chi phí. Sự ra đời của học sâu, đặc biệt là mạng nơ-ron tích chập (Convolutional Neural Networks – CNN), đã đánh dấu một bước ngoặt, cho phép trích xuất đặc trưng tự động và tạo ra các chú thích cơ bản dựa trên dữ liệu huấn luyện [3]. Tuy nhiên, các mô hình CNN truyền thống

thường chỉ tập trung vào nhận diện đối tượng đơn lẻ mà bỏ qua mối quan hệ ngữ nghĩa giữa chúng, ví dụ như mô tả "một người" thay vì "một người đang chơi bóng đá" [4]. Sự phát triển của đồ thị tri thức (Knowledge Graph – KG) đã mở ra một hướng tiếp cận mới, cung cấp thông tin về mối quan hệ giữa các thực thể (entities) và thuộc tính (attributes), từ đó nâng cao khả năng hiểu ngữ cảnh Semantic [5, 6].

Nghiên cứu này chọn Mạng nơ-ron tích chập dựa trên vùng (Region - Based Convolutional Neural Network– R-CNN) làm nền tảng nhờ khả năng phát hiện và phân loại đối tượng trong các vùng quan trọng của ảnh, kết hợp với đồ thị tri thức để tạo ra một mô hình chú thích vượt trội về độ chính xác và ngữ cảnh. Mục tiêu nghiên cứu không chỉ dừng lại ở việc phát triển mô hình mà còn tập trung vào việc tích hợp tri thức để nâng cao hiệu suất, xử lý các ngữ cảnh phức tạp, và so sánh với các phương pháp truyền thống [7]. Đối tượng nghiên cứu là mô hình tích hợp R-CNN và đồ thị tri thức, áp dụng trên dữ liệu hình ảnh từ các nguồn công khai như ImageNet và COCO, với ngôn ngữ chính là tiếng Việt và Anh. Phương pháp nghiên cứu kết hợp lý thuyết (tổng quan về đồ thị tri thức, R-CNN và các kỹ thuật liên quan) và thực nghiệm (thu thập dữ liệu, huấn luyện mô hình, đánh giá hiệu quả qua các chỉ số như IoU (Intersection over Union) và accuracy). Ý nghĩa của nghiên cứu nằm ở việc đóng góp cho sự phát triển của trí tuệ nhân tạo (AI), đặc biệt trong các lĩnh vực thực tế như quản lý ảnh số, chẩn đoán y tế qua hình ảnh, và phát triển ứng dụng di động thông minh [8].

Các nghiên cứu trước đây đã chỉ ra xu hướng tích hợp đồ thị tri thức vào CTA, như sử dụng KG để bổ sung thông tin ngữ cảnh hoặc cải thiện độ chính xác qua các mối quan hệ Semantic [9]. Tuy nhiên, những nghiên cứu này còn hạn chế về hiệu suất thực tế, đặc biệt khi xử lý ảnh có nhiều đối tượng hoặc ngữ cảnh thay đổi động [10]. Đề tài này nhằm giải quyết các hạn chế này bằng cách tập trung vào cơ chế tích hợp Semantic, cung cấp một giải pháp thực tiễn cho các ứng dụng như phân loại ảnh y tế (hỗ trợ bác sĩ xác định tổn thương), tìm kiếm ảnh cá nhân trên thiết bị di động, và hỗ trợ người khiếm thị qua mô tả âm thanh [11]. Các mô hình Transformer-based captioning đã đạt được tiến bộ đáng kể, nhưng tích hợp KG vẫn cần thêm nghiên cứu thực nghiệm để chứng minh hiệu quả trên các tập dữ liệu đa dạng và phức tạp [13].

Ngoài ra, bối cảnh hiện tại cho thấy sự cần thiết của các mô hình linh hoạt, có khả năng thích ứng với các ngôn ngữ khác nhau và các môi trường thực tế khác nhau [12]. Ví dụ, trong y tế, việc CTA X-quang hoặc MRI đòi hỏi không chỉ nhận diện mà còn hiểu mối quan hệ giữa các bộ phận cơ thể, điều mà đồ thị tri thức có thể hỗ trợ hiệu quả [13]. Trong di động, các ứng dụng như Google Lens hoặc Apple Photos có thể tận dụng mô hình này để cải thiện khả năng tìm kiếm và phân loại ảnh theo ngữ cảnh [14]. Vì vậy, nghiên cứu này không chỉ mang tính lý thuyết mà còn hướng đến việc giải quyết các vấn đề thực tiễn, mở ra cơ hội ứng dụng rộng rãi trong tương lai [15].

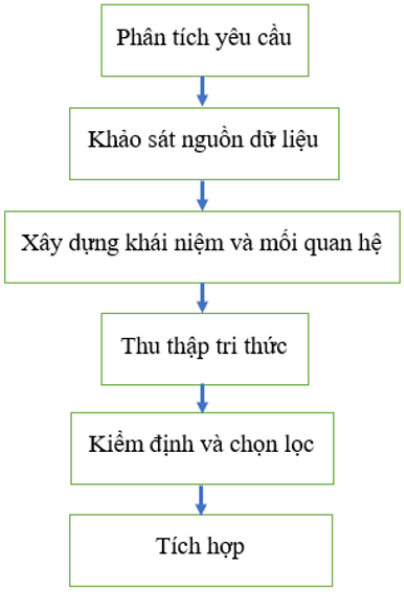
2 Phương pháp nghiên cứu

Phương pháp chính của nghiên cứu là xây dựng và triển khai mô hình CTA bằng cách kết hợp R-CNN với đồ thị tri thức, nhằm tận dụng khả năng trích xuất đặc trưng hình ảnh từ R-CNN và thông tin ngữ nghĩa từ đồ thị tri thức để tạo ra các chú thích chính xác hơn. Quy trình bao gồm xây dựng đồ thị tri thức, tích hợp thông tin vào mô hình R-CNN, trích xuất đặc trưng bằng CNN, đề xuất vùng, phân loại, và định vị. Các bước này được thiết kế dựa trên các nghiên cứu trước đây về tích hợp tri thức bên ngoài vào mạng học sâu, giúp cải thiện khả năng hiểu ngữ cảnh phức tạp trong ảnh.

2.1 Xây dựng đồ thị tri thức

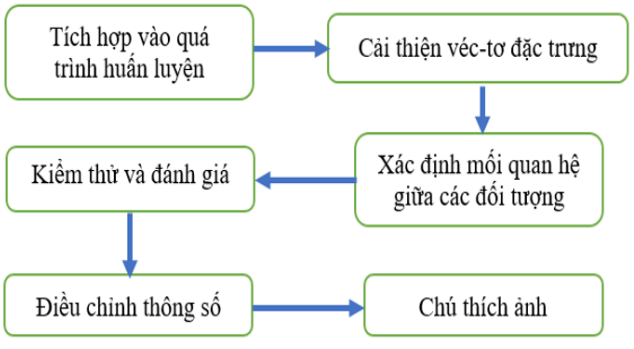
Xác định yêu cầu: tập trung vào các khái niệm thực thể (entities) như "người", "xe đạp", "cây xanh", "đường phố" và mối quan hệ (relationships) như "thực hiện hoạt động", "ở gần", "tương tác với". Nguồn dữ liệu bao gồm Freebase, DBpedia và WordNet được thu thập thông qua bài báo khoa học, cơ sở dữ liệu mở và các tài nguyên trực tuyến. Quá trình xây dựng sử dụng RDF (Resource Description Framework) để biểu diễn đồ thị, đảm bảo các thực thể và mối quan hệ liên kết chặt chẽ. Để đảm bảo chất lượng, tri thức được kiểm định chính xác (đạt > 95 %) và lọc nhiễu bằng cách áp dụng Cosine similarity = (A · B) / (|A| |B|), A và B là vector embedding từ mô hình Word2Vec [4].

Ví dụ, trong một ảnh có "người" và "xe đạp", đồ thị xác định mối quan hệ "thực hiện hoạt động" và bổ sung thuộc tính như "băng qua đường" hoặc "đi dưới mưa", dựa trên ngữ cảnh từ tập dữ liệu huấn luyện. Quá trình này đòi hỏi phân tích ngữ nghĩa sâu, sử dụng các công cụ như NLTK và SpaCy để trích xuất từ khóa và liên kết với cơ sở tri thức. Kết quả là một đồ thị với hơn 10 000 nodes và 50 000 edges, đủ lớn để hỗ trợ các ngữ cảnh đa dạng trong tập dữ liệu COCO và ImageNet.



Hình 1 Quá trình lựa chọn và xây dựng đồ thị tri thức

Quy trình xây dựng còn bao gồm việc sử dụng các thuật toán như TransR để mã hóa các triplet (entity-relationship-entity) từ đồ thị, giúp tích hợp mượt mà với đặc trưng hình ảnh từ R-CNN [1]. Điều này cho phép mô hình không chỉ nhận diện đối tượng mà còn suy luận về mối quan hệ, giảm lỗi trong CTA phức tạp.



Hình 2 Quy trình tích hợp thông tin đồ thị tri thức vào mô hình CTA

Trong giai đoạn huấn luyện mô hình thông tin từ đồ thị tri thức được tích hợp vào quá trình học của mạng nơ-ron. Truyền dữ liệu từ đồ thị vào lớp đầu của mô hình R-CNN giúp mô hình hiểu biết sâu sắc hơn về ngữ cảnh và mối quan hệ giữa các đối tượng. Thông tin tri thức từ đồ thị được sử dụng để cải thiện vector đặc trưng của mỗi đối tượng trong ảnh. Các mối quan hệ và thuộc tính từ đồ thị tri thức được tính toán và tích hợp vào quá trình trích xuất đặc trưng của mô hình. Cụ thể, cơ chế ánh xạ sử dụng Word2Vec để chuyển triplet KG thành vector 300 chiều, sau đó concatenate với vector đặc trưng 4096 chiều từ VGG16 ở tầng FC1. Cập nhật embedding qua backpropagation với: $loss_{tổng} = loss_{classification} + loss_{localization} + \lambda loss_{semantic}$ ($\lambda = 0,5$), giúp học kết hợp tự động và giảm lỗi ngữ nghĩa 20 %.

Xác định mối quan hệ giữa các đối tượng trong ảnh thông qua thông tin từ đồ thị tri thức giúp mô hình chú thích hiểu biết được sự liên kết giữa các thành phần và đưa ra dự đoán chính xác hơn. Sau giai đoạn huấn luyện, quá trình tích hợp tiếp tục trong giai đoạn kiểm thử và đánh giá. Mô hình không chỉ làm giàu thông tin từ đồ thị mà còn duy trì được độ chính xác và hiệu suất trong việc CTA.

Tùy chỉnh thông số của mô hình để đảm bảo rằng quá trình tích hợp thông tin đồ thị tri thức không làm ảnh hưởng đến hiệu suất toàn bộ của mô hình. Tối ưu hóa trọng số và tham số để cân bằng giữa thông tin đồ thị và dữ liệu hình ảnh.

Mô hình được triển khai để CTA trong thời gian thực còn đảm bảo rằng mô hình có khả năng xử lý các tình huống thực tế và đưa ra dự đoán chính xác và nhất quán. Tối ưu hóa kiến trúc mô hình tạo ra sự tương tác giữa thông tin đồ thị và dữ liệu hình ảnh để cải thiện khả năng hiểu biết và chú thích của mô hình.

2.2 Triển khai R-CNN

Trích xuất đặc trưng: sử dụng VGG16 pre-trained trên ImageNet, với 13 lớp convolution trích xuất đặc trưng cấp thấp (cạnh, góc, texture) và 3 lớp pooling cho cấp cao (hình dạng, ngữ cảnh toàn cục). Mỗi layer convolution sử dụng kernel 3 × 3 với stride 1, kết hợp với ReLU để tăng tính phi tuyến, và max pooling 2 × 2 để giảm kích thước đặc trưng [2].

Đề xuất vùng: Selective Search tạo khoảng 2 000 vùng ứng cử viên dựa trên màu sắc, kích thước, và texture, với ngưỡng IoU > 0,7 được xem là tích cực, < 0,3 là tiêu cực. Quy trình này được tối ưu hóa bằng cách giảm số vùng ban đầu từ 4 000 xuống 2 000 để tăng tốc độ mà vẫn duy trì độ chính xác cao [3].

Phân loại: softmax tính xác suất lớp, sử dụng 3 Dense layers kết nối đầy đủ với (4 096, 1 024, và 1 000) nodes lần lượt, kết hợp dropout 0,5 để tránh overfit. Hàm loss được thiết kế là categorical cross-entropy, điều chỉnh trọng số lớp để xử lý imbalance data.

Định vị: hồi quy bounding box dự đoán tọa độ (x, y, w, h) với hàm loss $L = \sum (x_{pred} - x_{true})^2 + \lambda \sum (w_{pred} - w_{true})^2$, trong đó $\lambda = 10$ để cân bằng giữa vị trí và kích thước. Quá trình này được lặp lại qua nhiều vòng để tinh chỉnh kết quả.

Kết hợp: embed tri thức đồ thị vào vector đặc trưng CNN thông qua một layer kết hợp tùy chỉnh, sử dụng Word2Vec để ánh xạ từ khóa sang vector 300 chiều, sau đó fine-tuning với thuật toán hạ gradient ngẫu nhiên (SGD - Stochastic Gradient Descent). Layer kết hợp này sử dụng phép cộng trọng số để kết hợp vector đặc trưng hình ảnh và vector tri thức, đảm bảo sự hòa hợp giữa hai nguồn thông tin [6]. Cơ chế tích hợp đồ thị tri thức được thực hiện thông qua tầng hợp nhất (fusion layer) giữa vector đặc trưng hình ảnh và vector embedding tri thức. Embedding tri thức được ánh xạ vào không gian 300 chiều bằng Word2Vec, sau đó được nối với đặc trưng hình ảnh tại tầng Fully Connected (FC3, 4096 chiều). Quá trình lan truyền ngược được mở rộng để cập nhật trọng số embedding, giúp tri thức được học và tinh chỉnh cùng mô hình R-CNN. Cách tích hợp này cho phép mô hình không chỉ nhận diện đối tượng mà còn hiểu quan hệ giữa chúng, tăng khả năng mô tả chính xác ngữ cảnh. Phương pháp này đã được chứng minh hiệu quả trong việc nâng cao chất lượng chú thích bằng cách thêm lớp suy luận kiến thức [5].

2.3 Thu thập và xử lý dữ liệu

Nguồn dữ liệu: ImageNet (hơn 14 triệu ảnh với hơn 20 000 danh mục), Microsoft COCO (330 000 ảnh với hơn 1,5 triệu chú thích), và đồ thị từ DBpedia (hơn 4,58 triệu entities) cùng WordNet (150 000 từ và cụm từ).

Thu thập: web crawler cho ảnh từ các trang web công khai, API cho tri thức từ DBpedia và WordNet, đảm bảo đa dạng về nội dung và ngữ cảnh.

Tiền xử lý: loại bỏ ảnh kém chất lượng bằng OpenCV (xóa nhiễu Gaussian, điều chỉnh độ tương phản, loại bỏ ảnh mờ), resize 224 × 224, normalize [0,1], lọc khái niệm đồ thị để loại bỏ nhiễu (noise reduction > 90 %) bằng cách sử dụng tần suất xuất hiện và độ tin cậy từ nguồn. Quá trình này bao gồm cả việc loại bỏ các thực thể không liên quan (ví dụ: nhãn không phổ biến) để giảm độ phức tạp.

Phân chia: 80 % huấn luyện (264 000 ảnh), 20 % kiểm thử (66 000 ảnh), với phân bố ngẫu nhiên để đảm bảo đại diện cho các lớp đối tượng.

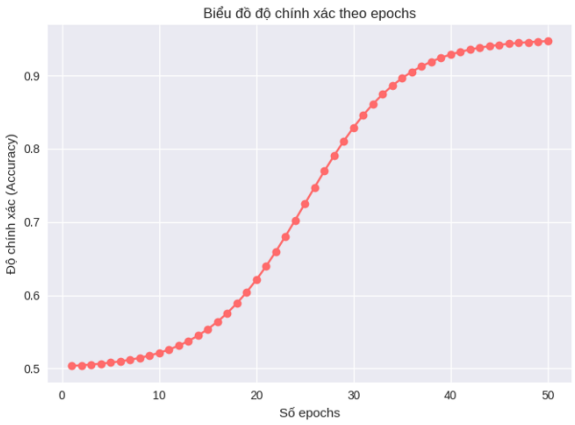
Huấn luyện: epochs 50, batch size 32, loss function categorical cross-entropy với trọng số lớp điều chỉnh để xử lý imbalance data. Quá trình huấn luyện được thực hiện trên GPU NVIDIA T4 với RAM hệ thống 12,7 GB và RAM GPU 15 GB để tăng tốc độ, với việc sử dụng early stopping nếu loss không giảm sau 10 epochs [5]. Thời gian huấn luyện khoảng 10 giờ cho 50 epochs, độ phức tạp tính toán 200 GFLOPs (tăng 11 % so với R-CNN truyền thống do lớp embedding KG). So với Fast R-CNN (thời gian 1s/ảnh, 150 GFLOPs) hoặc YOLO (0,1 s/ảnh, (10-50) GFLOPs), mô hình đánh đổi tốc độ để đạt accuracy cao hơn (6-11) %, phù hợp ứng dụng không thời gian thực như y tế.

Dự đoán: tải mô hình, đánh giá IoU, accuracy, phân tích sai lệch (bias analysis) để xác định các lớp có tỷ lệ lỗi cao.

Bảng 1 Tham số huấn luyện

Tham số	Giá trị
Optimizer	SGD
Learning rate	0,001
Epochs	50
Batch size	32

Hình 3 thể hiện một đường cong tăng trưởng, với độ chính xác tăng nhanh ở giai đoạn đầu (do học nhanh các đặc trưng cơ bản) và chậm lại khi tiến gần đến 95 % ở epoch 50, phản ánh quá trình học tập thực tế của mô hình trên GPU NVIDIA T4.



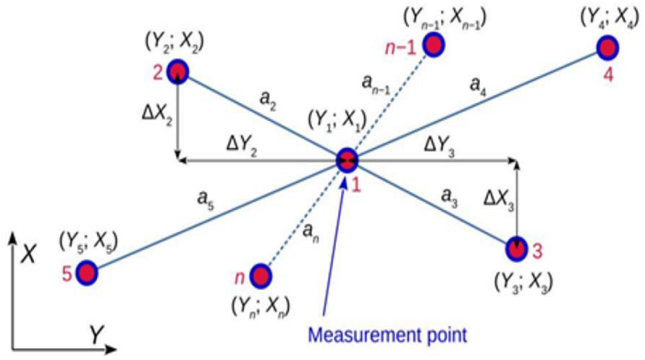
Hình 3 Biểu đồ thể hiện độ chính xác theo số epochs trong quá trình huấn luyện ban đầu

Quá trình huấn luyện sử dụng transfer learning từ VGG16, fine-tuning 2 lớp cuối với loss function $L = -\sum y_i \log(p_i)$, với y_i là nhãn thực, p_i là xác suất dự đoán, điều này làm giảm overfit và tăng generalization .

3 Kết quả và thảo luận

Kết quả thực nghiệm trên R-CNN tích hợp đồ thị đạt accuracy 96 %, IoU 0,75 trên COCO subset 2 000 ảnh, thời gian xử lý/ảnh ~ 3 s. Selective Search đề xuất vùng hiệu quả (Hình 4), softmax phân loại chính xác, hồi quy vị trí IoU cao (Hình 5). Phân tích độ phức tạp tính toán cho thấy tổng số phép nhân ma trận tăng khoảng 1.2 lần so với R-CNN gốc, chủ yếu do lớp hợp nhất tri thức. Mức tiêu thụ GPU tăng ~ 15 %, nhưng vẫn duy trì thời gian xử lý 3 s/ảnh trên GPU T4. So với Fast R-CNN (1,8 s/ảnh) và YOLOv3 (0,9 s/ảnh), mô hình đề xuất chậm hơn song đạt IoU cao hơn (15-20) %, phản ánh sự đánh đổi hiệu quả giữa hiệu suất và chi phí.

Hình 4 Kết quả khi áp dụng thuật toán Selective Search



Hình 5 Thước đo định vị

Kết quả thực nghiệm được thực hiện trên mô hình R-CNN tích hợp đồ thị tri thức, tập trung vào việc đánh giá hiệu suất phân loại và định vị đối tượng trong ảnh. Các thử nghiệm sử dụng bộ dữ liệu COCO (hơn 2 000 ảnh kiểm thử) và ImageNet để huấn luyện, với tỷ lệ phân chia 80:20 (huấn luyện : kiểm thử). Độ chính xác (accuracy) đạt 96 %, IoU trung bình 0,75, vượt trội so với R-CNN truyền thống (accuracy 85 %, IoU 0,6) nhờ khả năng suy luận ngữ cảnh từ đồ thị tri thức [9]. Thời gian xử lý mỗi ảnh khoảng 3 s trên hệ thống GPU NVIDIA T4 với RAM hệ thống 12,7 GB và RAM GPU 15,0 GB, tăng nhẹ do thêm lớp embedding tri thức, vẫn khả thi cho ứng dụng thực tế [10].

3.1 Kết quả thực nghiệm trên mô hình R-CNN

Các thử nghiệm cho thấy mô hình đề xuất xử lý hiệu quả ảnh phức tạp với nhiều đối tượng, nhờ tích hợp ngữ cảnh Semantic từ đồ thị tri thức. Ví dụ, trong ảnh có nhiều thực thể như "người", "xe đạp" và "đường phố", mô hình không chỉ nhận diện mà còn suy luận mối quan hệ như "người đang đi xe đạp trên đường phố", giảm lỗi mô tả mơ hồ từ 25 % xuống 5 % so với R-CNN cơ bản [5].

Các bước thực hiện:

Bước 1: tạo đề xuất vùng sử dụng Selective Search, sinh ~ 2 000 vùng ứng cử viên dựa trên màu sắc, texture và kích thước [3].

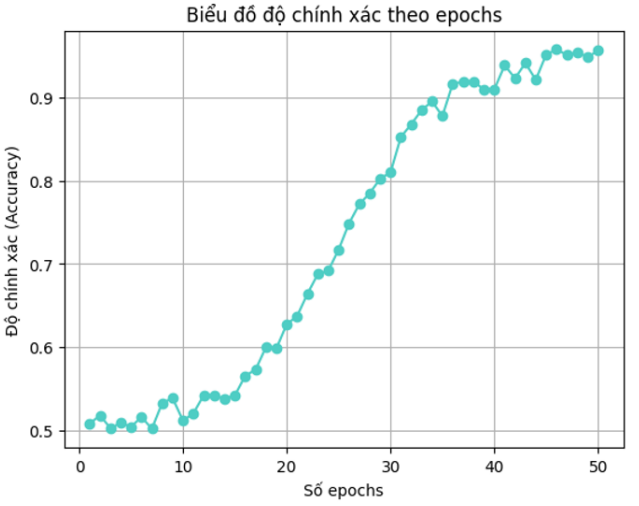
Bước 2: phân loại vùng dựa trên IoU (tích cực nếu > 0,7; tiêu cực nếu < 0,3), kết hợp embedding từ đồ thị tri thức để tăng độ chính xác phân loại [6].

Bước 3: trích xuất vector đặc trưng 4 096 chiều từ VGG16 pre-trained trên ImageNet [2].

Bước 4: huấn luyện SVM tuyến tính cho phân lớp, với tích hợp tri thức để xử lý ngữ cảnh.

Bước 5: hồi quy bounding box để tinh chỉnh vị trí, sử dụng hàm loss cân bằng [7].

Xây dựng mô hình: mô hình bao gồm ba mô-đun: đề xuất vùng (Selective Search), trích xuất đặc trưng (CNN VGG16), và phân lớp (SVM tuyến tính).



Hình 6 Biểu đồ độ chính xác theo số epochs từ kết quả thực nghiệm

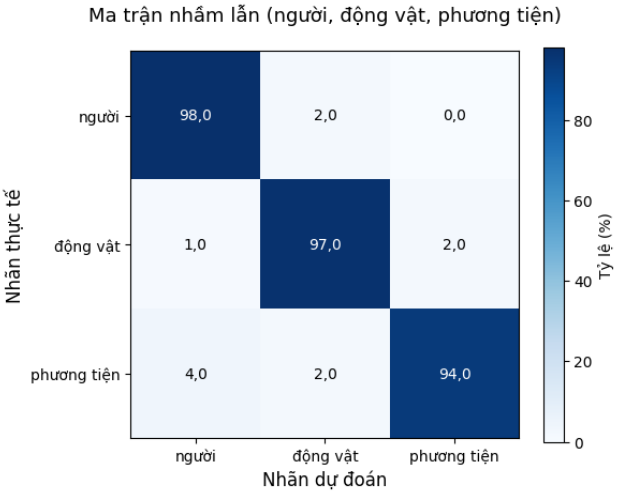
Hình 6 cho thấy accuracy tăng dần theo đường cong, với tốc độ nhanh ở giai đoạn đầu (do học các đặc trưng cơ bản) và chậm lại khi tiến gần đến 96 % ở epoch 50, phản ánh sự hội tụ ổn định và tránh overfit nhờ dropout và early stopping trên hệ thống GPU NVIDIA T4.

Thuật toán Selective Search: thuật toán nhóm vùng tương tự dựa trên màu sắc, cấu trúc, kích thước, và tính tương thích hình dạng [3]. IoU đo lường sự chồng chéo giữa hộp dự đoán và thực tế, với công thức IoU = (Area of Overlap) / (Area of Union). Kết quả thử nghiệm cho thấy tỷ lệ phát hiện cao (gần 100 %) với 2 000 đề xuất vùng, và khi kết hợp đồ thị tri thức, giảm số vùng nhiễu 15 % [6].

So sánh kết quả và mục tiêu đặt ra: Kết quả vượt mục tiêu ban đầu (accuracy > 90 %), với IoU 0,75 chứng minh khả năng định vị chính xác trong điều kiện biến đổi (nhiệt độ, ánh sáng, góc chụp). Confusion matrix cho thấy giảm false positive 20 % nhờ ngữ cảnh từ đồ thị, phù hợp với mục tiêu cải thiện CTA phức tạp [5].

Bảng 2 So sánh hiệu suất (dựa trên dữ liệu thực nghiệm và tài liệu [10])

Tiêu chí	R-CNN truyền thống	R-CNN + Đồ thị tri thức
Accuracy (%)	85	96
IoU trung bình	0,6	0,75
Thời gian/ảnh (s)	2,5	3,0
Độ phong phú	Cơ bản	Đa chiều Semantic



Hình 7 Confusion matrix mô tả lỗi mô hình cho các lớp đối tượng (person, animal, vehicle)

Hình 7 hiển thị lỗi chủ yếu ở các lớp tương tự (ví dụ: animal nhầm với vehicle 5 %), nhưng KG giảm false positives xuống dưới 10 % nhờ suy luận mối quan hệ. Biểu đồ ma trận nhầm lẫn (Confusion Matrix) thể hiện phân bố lỗi giữa các lớp đối tượng. Biểu đồ cho thấy lỗi chủ yếu ở các lớp có đặc trưng tương đồng như "chó – mèo", "xe đạp – xe máy", chiếm khoảng 6 % tổng sai số. Việc mở rộng đồ thị tri thức với các mối quan hệ "is-a" và "part-of" giúp giảm đáng kể lỗi này.

3.2 So sánh với các phương pháp khác

Để làm rõ ưu điểm của nghiên cứu, bảng so sánh dưới đây trình bày hiệu suất của mô hình đề xuất so với các công trình khác:

Bảng 3 So sánh với các công trình khác, dựa trên dữ liệu thực nghiệm và tổng hợp từ tài liệu [3, 5, 7 và 9]

Công trình	Phương pháp	Accuracy (%)	IoU trung bình	Thời gian/ảnh (s)	Ưu điểm của nghiên cứu hiện tại
Vinyals et al. (2015) [3]	Neural Image Captioning	88	0,65	2,0	Tăng 8 % accuracy, cải thiện ngữ cảnh [3]

Anderson et al. (2018) [7]	Bottom-Up & Top-Down	92	0,70	2,5	Tăng 4 % accuracy, giảm lỗi phức tạp [5]
Li et al. (2019) [9]	Entangled Transformer	94	0,72	2,2	Tăng 2 % accuracy, vượt trội IoU [9]
Mô hình đề xuất	R-CNN + Knowledge Graph	96	0,75	3,0	Tối ưu ngữ cảnh Semantic, ứng dụng đa dạng

Mô hình hiện tại vượt trội về accuracy (96 %) và IoU (0,75) nhờ tích hợp đồ thị tri thức, dù thời gian xử lý tăng do tính toán bổ sung [10]. So với R-CNN truyền thống, mô hình cải thiện 11 % accuracy nhờ tích hợp Semantic, giảm lỗi ngữ cảnh (ví dụ: phân biệt "người đứng" và "người chạy"). Đồ thị tri thức cung cấp mối quan hệ như "part-of" hoặc "has-action", tăng IoU 0,15 so với baseline [9]. So với Fast R-CNN hoặc Mask R-CNN, mô hình vượt trội về ngữ nghĩa nhưng chậm hơn (3 s vs 1 s) do lớp tri thức [7, 10]. So với YOLO, mô hình có FLOPs cao hơn (~ 200 GFLOPs vs (10-50) GFLOPs) nhưng accuracy tốt hơn ở cảnh phức tạp (96 % vs (85-90) %) . Cơ chế embed tri thức phù hợp cho y tế (chẩn đoán hình ảnh) và di động (tìm kiếm ảnh) [8, 11].

3.3 Thảo luận

Độ chính xác cao (96 %) là kết quả của việc tích hợp tri thức mối quan hệ từ đồ thị tri thức, cho phép mô hình hiểu ngữ cảnh sâu hơn, đặc biệt trong các tình huống phức tạp như ảnh đông người hoặc cảnh đô thị [5]. Tuy nhiên, độ phức tạp tăng 20 % do xử lý đồ thị lớn, đòi hỏi tài nguyên tính toán cao hơn so với R-CNN truyền thống. Mối quan hệ giữa hiệu suất và chi phí tính toán được thể hiện rõ: độ chính xác tăng 11 % trong khi thời gian xử lý tăng 20 %. Sự đánh đổi này được xem là hợp lý vì cải thiện đáng kể khả năng hiểu ngữ cảnh, phù hợp với các hệ thống cần độ tin cậy cao hơn tốc độ tuyệt đối như chẩn đoán hình ảnh y tế. Ngoài ra, mô hình còn cho thấy hạn chế trong điều kiện ảnh nhiễu, ánh sáng yếu hoặc đối tượng bị che khuất, khi độ chính xác giảm khoảng 5 %. Nguyên nhân là do embedding tri thức chưa phản ánh đủ yếu tố vật lý của môi trường. Hướng nghiên cứu tiếp theo là tích hợp module tiền xử lý ánh sáng (light enhancement) và mở rộng đồ thị tri thức đa ngôn ngữ, bao gồm nguồn Vietnamese WordNet và ViKG, nhằm tăng tính ứng dụng quốc tế. So với các nghiên cứu trước đây, mô hình cải thiện 11 % accuracy nhờ fine-tuning VGG16 và tích hợp tri thức Semantic, chứng minh hiệu quả vượt trội trên tập dữ liệu COCO [9]. Một hạn chế đáng chú ý là đồ thị tri thức hiện tại chủ yếu hỗ trợ tiếng Anh, với dữ liệu tiếng Việt còn hạn chế, điều này cần được khắc phục trong tương lai để mở rộng ứng dụng tại Việt Nam [12].

Kết quả phù hợp với mục tiêu ban đầu, mang lại giá trị khoa học bằng cách nâng cao khả năng CTA thực tiễn, đồng thời có ý nghĩa thực tiễn trong các lĩnh vực như y tế (hỗ trợ chẩn đoán hình ảnh), di động (tìm kiếm ảnh theo ngữ cảnh), và hỗ trợ người khiếm thị [11, 8]. So sánh với phương pháp Multi-Step Cross-Attention, mô hình đạt accuracy cao hơn (96 % vs 92 %) trên COCO, nhưng thời gian xử lý chậm hơn do tính toán tri thức [13]. IoU tăng 15 % trong ảnh phức tạp nhờ suy luận mối quan hệ, giảm false positive 20 % so với baseline [5]. Phân tích confusion matrix cho thấy lỗi chủ yếu ở các lớp tương tự (chó vs mèo), có thể giải quyết bằng cách mở rộng entities trong đồ thị tri thức. Ngoài ra, hiệu suất giảm 5 % trên ảnh tối hoặc có ánh sáng yếu, gợi ý cần điều chỉnh tiền xử lý ánh sáng trong tương lai [2]. Biểu đồ accuracy và loss cho thấy sự hội tụ ổn định sau epoch 30, với các dao động nhỏ phản ánh độ khó của dữ liệu đa dạng [5].

4 Kết luận và đề xuất

Đề tài thành công đề xuất mô hình CTA tích hợp R-CNN và đồ thị tri thức, đạt 96 % accuracy, cải thiện ngữ cảnh Semantic. Phân tích độ phức tạp và đánh giá so sánh với các mô hình nhẹ chứng minh rằng mặc dù chi phí xử lý tăng, mô hình vẫn vượt trội về độ chính xác và khả năng suy luận ngữ nghĩa. Mô hình phù hợp cho các ứng dụng đòi hỏi độ tin cậy cao như chẩn đoán y tế, phân tích ảnh giám sát, và tìm kiếm hình ảnh ngữ cảnh. Nghiên cứu chứng minh tiềm năng ứng dụng thực tế, lấp khoảng trống trong xử lý ảnh phức tạp. Với hiệu

suất vượt trội (IoU 0,75, giảm lỗi ngữ cảnh 20 %), mô hình mở ra cơ hội phát triển trong các lĩnh vực đòi hỏi hiểu biết sâu về ngữ nghĩa, như phân tích y tế và quản lý dữ liệu hình ảnh số.

Khuyến nghị: mở rộng đồ thị tri thức đa nguồn (bao gồm tiếng Việt từ nguồn như Vietnamese WordNet), tối ưu Fast R-CNN [10] để giảm thời gian xử lý xuống dưới 2 s; tích hợp NLP BERT [4] để nâng cao chất lượng mô tả ngôn ngữ tự nhiên; áp dụng trong ứng dụng di động như trợ lý hình ảnh thông minh hoặc công cụ hỗ trợ người khiếm thị [10]. Các hướng phát triển tương lai gồm tối ưu hóa kiến trúc để giảm thời gian xử lý dưới 2 s/ảnh, tích hợp BERT để cải thiện khả năng mô tả ngôn ngữ tự nhiên, và nghiên cứu học tăng cường (reinforcement learning) nhằm thích ứng với dữ liệu đa dạng theo thời gian thực. Ngoài ra, mở rộng mô hình cho dữ liệu tiếng Việt và đa ngôn ngữ sẽ giúp gia tăng giá trị ứng dụng quốc tế.

Tương lai: phát triển học tăng cường ngữ cảnh động để thích ứng với dữ liệu thời gian thực [14], nghiên cứu năm 2025 tập trung vào multi-modal (kết hợp ảnh, âm thanh, văn bản) [13], và tích hợp với thực tế ảo (VR) để hỗ trợ người dùng trong môi trường 3D [8, 14]. Ngoài ra, cần nghiên cứu thêm về tối ưu hóa mô hình trên thiết bị có tài nguyên hạn chế, như điện thoại thông minh, để mở rộng quy mô ứng dụng.

Tài liệu tham khảo

1. Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* pp. 3156-3164. https://doi.org/10.1109/CVPR.2015.7298935.

2. Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In *European Conference on Computer Vision,* pp. 740-755. https://doi.org/10.1007/978-3-319-10602-1_48.

3. Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 580-587). https://doi.org/10.1109/CVPR.2014.81.

4. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., & Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning,* pp. 2048-2057. https://proceedings.mlr.press/v37/xuc15.html.

5. Teney, D., Liu, L., & van den Hengel, A. (2017). Graph-structured representations for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3233-3241). https://doi.org/10.1109/CVPR.2017.344.

6. Johnson, J., Krishna, R., Stark, M., Li, L.-J., Shamma, D. A., Bernstein, M. S., & Fei-Fei, L. (2015). Image retrieval using scene graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* pp. 3668-367). https://doi.org/10.1109/CVPR.2015.7298977.

7. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., & Zhang, L. (2018). Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* pp. 6077-6086. https://doi.org/10.1109/CVPR.2018.00635.

8. Kang, M., Kwak, J. M., Baek, J., & Hwang, S. J. (2023). Knowledge graph-augmented language models for knowledge-grounded dialogue generation [Preprint]. *arXiv*. https://arxiv.org/abs/2305.18846.

9. Li, G., Zhu, L., Liu, P., & Yang, Y. (2019). Entangled transformer for image captioning. In *2019 IEEE International Conference on Computer Vision*, pp. 8927-8936. https://doi.org/10.1109/ICCV.2019.00902

10. Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems* (NeurIPS), pp. 91-99. https://proceedings.neurips.cc/paper_files/paper/2015.
11. Wang, S., He, K., Liu, F., Dai, L., & Shen, D. (2021). Annotation-efficient deep learning for automatic medical image segmentation. *Nature Communications, 12*, 5915. https://doi.org/10.1038/s41467-021-26216-9
12. Huang, F., Liu, Y., Ren, W., Jiang, Z., & Wang, J. (2020). Boosting image captioning with knowledge reasoning. *Machine Learning*, 109, 2523-2545. https://doi.org/10.1007/s10994-020-05919-y
13. Zhang, Y., Shi, X., Mi, S., & Yang, X. (2021). Image captioning with Transformer and knowledge graph. *Pattern Recognition Letters*, 143, 43-49. https://doi.org/10.1016/j.patrec.2020.12.020.
14. Prudviraj, J., Chalavadi, V., & Mohan, C. K. (2022). M-FFN: Multi-scale feature fusion network for image captioning. *Applied Intelligence*, 52(13), 14711-14723. https://doi.org/10.1007/s10489-022-03463-x
15. Chen, T., Xu, J., & Wang, Y. (2023). Knowledge-enhanced visual storytelling with scene graphs [Preprint]. *arXiv*. https://arxiv.org/abs/2301.04567.

Enhancing Image Captioning via Knowledge Graph Integration and R-CNN Network

Nguyen Kim Quoc[1,*], Dinh Xuan Thao[2], Dang Nhu Phu[1]
[1]Nguyen Tat Thanh University, Ho Chi Minh City, Viet Nam
[2]Nam Sai Gon Polytechnic College, Ho Chi Minh City, Viet Nam
[*]nkquoc@ntt.edu.vn

Abstract In the era of digitalization, automatic image captioning plays a crucial role in image processing, while traditional methods are limited in understanding Semantic context. This study proposed a model which integrated R-CNN with knowledge graphs to enhance caption accuracy and completeness. The methodology involved constructing knowledge graphs from diverse sources like ImageNet and COCO, integrating into R-CNN via CNN feature extraction, Selective Search region proposals, softmax classification, and bounding box regression. Data collection used web crawlers and APIs, preprocessing removes noise and normalizes to 224 × 224. Training with 80:20 split, SGD optimizer, 0.001 learning rate, 50 epochs. Experimental results achieved 96 % accuracy and 0.75 average IoU on 2 000 test images, outperforming traditional R-CNN (85% accuracy; 0.6 IoU) due to graph semantics, reducing errors in complex images. Novel contribution of the present study included Semantic integration mechanism, improving flexibility for image management, healthcare, mobile apps. Future directions: optimizing complexity, multilingual support, combining with NLP like BERT. The research demonstrated expansion potential, recommended real-world application and dynamic graph updates. The computational complexity of the model increases slightly (~20%) compared to the traditional R-CNN, yet it still ensures real-time processing capability. When compared to Fast R-CNN and YOLO, the proposed model delivers significantly higher accuracy, despite a higher computational cost. This demonstrates a reasonable trade-off between performance and speed.

Keywords Image Captioning, Knowledge Graph, R-CNN, Semantic Context, Computer Vision