p-ISSN 3030 - 4628
e-ISSN 3030 - 4636

PHÁT TRIỂN PHẦN MỀM GÁN NHÃN VÀ CHÚ THÍCH ẢNH BÁN TỰ ĐỘNG ỨNG DỤNG TRÍ TUỆ NHÂN TẠO

Nguyễn Đình Công[1], Hoàng Anh Công[2], Nguyễn Hoàng Long[1] , Phạm Thế Anh[1], Lê Văn Sâm[3]

TÓM TẮT

Bài báo giới thiệu một hệ thống phần mềm gán nhãn và chú thích ảnh bán tự động ứng dụng trí tuệ nhân tạo, hỗ trợ xây dựng tập dữ liệu huấn luyện cho các mô hình thị giác máy tính. Hệ thống tích hợp YOLOv8 để tự động phát hiện đối tượng, cho phép người dùng hiệu chỉnh hộp giới hạn (bounding box), viết và chỉnh sửa chú thích, truy xuất và tải dữ liệu. Phần mềm còn hỗ trợ phân quyền người dùng, quản lý danh mục dữ liệu, tải ảnh theo lô và truy vấn theo yêu cầu. Giao diện thân thiện, đa nền tảng, dễ triển khai trong giáo dục và nghiên cứu. Kết quả thực nghiệm cho thấy hệ thống giảm đáng kể thời gian gán nhãn mà vẫn duy trì độ chính xác cao, đồng thời mở ra hướng tích hợp mô hình sinh chú thích (caption) tự động trong tương lai.

Từ khoá: *Gán nhãn ảnh, chú thích ngữ nghĩa, học sâu.*

DOI: *https://doi.org/10.70117/hdujs.79.09.2025.977*

1. ĐẶT VẤN ĐỀ

Trong những năm gần đây, sự phát triển của học sâu đã thúc đẩy mạnh mẽ các bài toán thị giác máy tính như nhận dạng, phát hiện, phân loại và mô tả ảnh. Tuy nhiên, hiệu quả của các mô hình này phụ thuộc lớn vào tập dữ liệu huấn luyện được gán nhãn chính xác và đầy đủ [1]. Việc xây dựng tập dữ liệu quy mô lớn gặp nhiều khó khăn do gán nhãn thủ công tốn kém thời gian, nhân lực và dễ sai lệch [2]. Các công cụ gán nhãn hiện nay, kể cả nền tảng thương mại như Labelbox hay Amazon Rekognition, chủ yếu tập trung vào khoanh vùng đối tượng, thiếu khả năng chú thích ngữ nghĩa đa ngôn ngữ, đặc biệt là tiếng Việt, đồng thời đòi hỏi hạ tầng phức tạp [3]. Điều này tạo rào cản cho môi trường học thuật và đào tạo trong nước. Do đó, nhu cầu cấp thiết là phát triển một phần mềm gán nhãn bán tự động, dễ triển khai, hỗ trợ đa tác vụ (gán nhãn, chú thích, truy xuất dữ liệu), quản lý người dùng và mở rộng linh hoạt.

Nghiên cứu này đề xuất một hệ thống phần mềm với kiến trúc ba lớp: giao diện người dùng, dịch vụ AI, và cơ sở dữ liệu. Hệ thống tích hợp YOLOv8 để phát hiện đối tượng, sinh nhãn bán tự động, đồng thời hỗ trợ viết chú thích, phân quyền và đóng gói dữ liệu huấn luyện. Ngoài ra, kiến trúc còn cho phép tích hợp các mô hình chú thích ảnh dựa trên Transformer như BLIP [4] hoặc GIT [5]. Bài báo trình bày toàn diện quá trình thiết kế, phát triển, thử nghiệm và đánh giá hệ thống, cho thấy giải pháp không chỉ giảm chi phí, thời gian xây dựng dữ liệu mà còn có tiềm năng ứng dụng rộng rãi trong đào tạo, nghiên cứu và phát triển AI tại Việt Nam. Để làm rõ hơn vai trò quan trọng và những thách thức hiện tại trong lĩnh vực gán nhãn ảnh và chú thích ảnh, chúng tôi sẽ trình bày tổng quan một số vấn đề có liên quan.

[1] *Khoa Kỹ thuật, Công nghệ và Truyền thông, Trường Đại học Hồng Đức; Email: nguyendinhcong@hdu.edu.vn*
[2] *Trường Đại học Văn hoá, Thể thao và Du lịch Thanh Hoá*
[3] *Phân hiệu Trường Đại học Y Hà Nội tại Thanh Hoá*

p-ISSN 3030 - 4628
e-ISSN 3030 - 4636

1.1. Gán nhãn ảnh và vai trò trong thị giác máy tính

Gán nhãn ảnh là bước tiền xử lý quan trọng trong thị giác máy tính, giúp cung cấp dữ liệu có cấu trúc cho huấn luyện mô hình học máy và học sâu. Các dạng nhãn phổ biến gồm: phân loại, phát hiện đối tượng, phân vùng và chú thích ngữ nghĩa. Trong đó, phát hiện và phân vùng đặc biệt quan trọng trong các ứng dụng như xe tự hành, giám sát thông minh hay y học ảnh học [6].

Tuy gán nhãn thủ công có độ chính xác cao nhưng tốn nhiều thời gian và chi phí, nên các hệ thống bán tự động được phát triển nhằm kết hợp sức mạnh AI với phản hồi của người dùng [7]. Một số công cụ như LabelImg, Label Studio, CVAT đã hỗ trợ quá trình này, nhưng chủ yếu tập trung vào khoanh vùng đối tượng, chưa tích hợp tốt chức năng chú thích ngữ nghĩa và hỗ trợ đa ngôn ngữ, đặc biệt là tiếng Việt.

1.2. Mô hình học sâu cho nhận diện đối tượng và sinh mô tả ảnh

Về kỹ thuật, các mô hình phát hiện đối tượng tiêu biểu như YOLO, Faster R-CNN và EfficientDet đã chứng minh hiệu năng vượt trội, với YOLOv8 đặc biệt nổi bật nhờ cân bằng tốt giữa tốc độ và độ chính xác, đồng thời dễ triển khai trên web và thiết bị biên. Trong chú thích ảnh, nghiên cứu ban đầu thường kết hợp CNN và RNN [11], sau đó phát triển thành các mô hình sử dụng attention như "Show, Attend and Tell" [12]. Gần đây, Transformer và các mô hình tiền huấn luyện như BLIP, GIT hay Flamingo đã nâng cao khả năng sinh mô tả ngữ cảnh tự nhiên, song chủ yếu chỉ hỗ trợ tiếng Anh và yêu cầu hạ tầng tính toán mạnh.

Một số hệ thống gán nhãn tích hợp AI đã xuất hiện, chẳng hạn Labelbox autoML (tự động phát hiện đối tượng sau vài vòng gán nhãn), Amazon SageMaker Ground Truth (ứng dụng Active Learning để gợi ý nhãn), hay Facebook Detectron2 (kết hợp Visual Genome để sinh nhãn phong phú). Tuy nhiên, các nền tảng này chủ yếu hướng tới doanh nghiệp lớn, chi phí cao và khó tiếp cận với các nhóm nghiên cứu độc lập hoặc cơ sở giáo dục đại học ở Việt Nam. Hơn nữa, việc hỗ trợ tiếng Việt gần như chưa được chú trọng trong các mô hình chú thích hiện tại [14].

1.3. Khoảng trống và định hướng tiếp cận

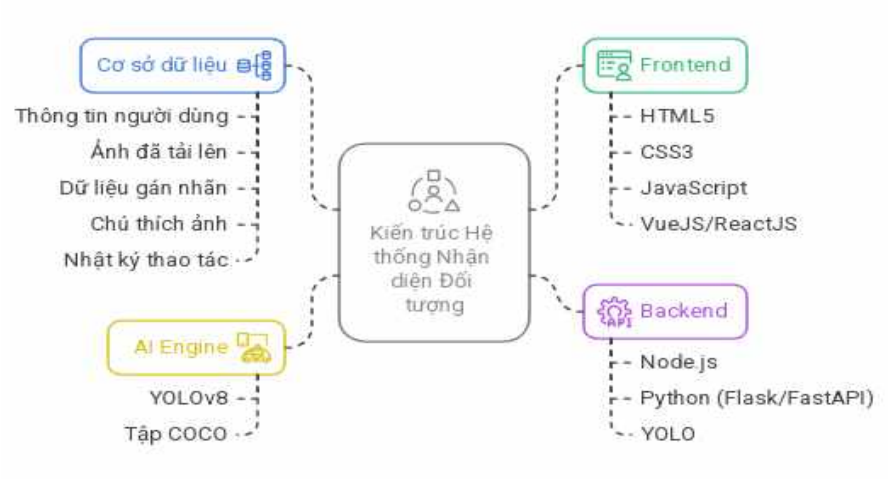
Từ các khảo sát trên có thể thấy: (i) các công cụ hiện hành chưa tích hợp chặt chẽ mô hình phát hiện đối tượng với giao diện chú thích trực quan; (ii) gần như chưa có giải pháp hỗ trợ sinh mô tả ngữ nghĩa tiếng Việt trên nền web; (iii) nhiều hệ thống còn thiếu thân thiện, khó triển khai trong môi trường giáo dục hạn chế tài nguyên. Do đó, việc phát triển một nền tảng web tích hợp YOLOv8 [15], cho phép hiệu chỉnh trực tiếp, viết chú thích tiếng Việt, hỗ trợ tải ảnh theo lô, phân quyền người dùng và truy xuất tập dữ liệu đầu ra là hướng đi thiết thực, vừa phục vụ nghiên cứu vừa có tiềm năng ứng dụng rộng rãi. Đây chính là trọng tâm của hệ thống được trình bày trong bài báo này.

2. PHƯƠNG PHÁP NGHIÊN CỨU VÀ THIẾT KẾ HỆ THỐNG

Hệ thống gán nhãn và chú thích ảnh bán tự động được thiết kế theo mô hình ba lớp tiêu chuẩn gồm: (1) lớp giao diện người dùng, (2) lớp xử lý nghiệp vụ và trí tuệ nhân tạo, và (3) lớp quản lý dữ liệu. Kiến trúc này nhằm đảm bảo tính tách biệt chức năng, dễ bảo trì, mở rộng và thích ứng với nhiều môi trường triển khai khác nhau. Phương pháp nghiên cứu bao gồm khảo sát yêu cầu nghiệp vụ, lựa chọn công nghệ phù hợp và mô hình hoá kiến trúc phần mềm. Phương tiện nghiên cứu sử dụng các framework và công cụ hiện đại như ReactJS cho giao diện người dùng, Node.js và Python (Flask/FastAPI) cho xử lý nghiệp vụ và tích hợp AI.

p-ISSN 3030 - 4628
e-ISSN 3030 - 4636

2.1. Thiết kế kiến trúc hệ thống



Hình 1. Kiến trúc tổng thể của hệ thống đề xuất

Kiến trúc hệ thống được trình bày ở Hình 1, bao gồm các thành phần chính như sau.

Frontend: sử dụng framework ReactJS, cho phép người dùng tải ảnh, xem kết quả nhận diện, vẽ/chỉnh sửa hộp bao quanh đối tượng và viết chú thích.

Backend: phát triển bằng Node.js kết hợp với Python (Flask/FastAPI) để điều phối xử lý nghiệp vụ, định tuyến API, quản lý truy cập và tích hợp mô hình học sâu (YOLOv8).

AI Engine: sử dụng mô hình YOLOv8 được huấn luyện trước trên tập Open Image V7 và tinh chỉnh trên dữ liệu thực tế để phát hiện đối tượng trong ảnh đầu vào, trả về danh sách hộp bao quanh đối tượng cùng nhãn dự đoán.

Cơ sở dữ liệu (MongoDB): lưu trữ thông tin người dùng, ảnh đã tải lên, dữ liệu gán nhãn, chú thích ảnh, và nhật ký thao tác.

Hệ thống sử dụng RESTful API để kết nối giữa frontend và backend, đồng thời áp dụng JWT để xác thực và phân quyền truy cập.

2.2. Luồng xử lý chức năng

Phương pháp nghiên cứu tập trung vào phân tích nghiệp vụ và thiết kế luồng xử lý tuần tự nhằm tối ưu trải nghiệm người dùng và đảm bảo độ chính xác của quá trình gán nhãn bán tự động. Phương tiện nghiên cứu bao gồm các công cụ phát triển phần mềm (ReactJS, Python) và tích hợp mô hình YOLOv8 phục vụ nhận diện đối tượng. Luồng xử lý chính của hệ thống được tổ chức theo các bước sau.

Tải ảnh: Người dùng (vai trò người gán nhãn hoặc người thu thập dữ liệu) đăng nhập và tải ảnh lên hệ thống (đơn lẻ hoặc theo lô).

Nhận diện đối tượng: Backend gửi ảnh đến mô hình YOLOv8, mô hình trả về danh sách hộp bao quanh đối tượng và loại đối tượng.

Hiển thị và hiệu chỉnh: Kết quả nhận diện được hiển thị trên giao diện; người dùng có thể vẽ mới, chỉnh sửa hoặc xóa hộp bao quanh đối tượng.

Chú thích ảnh: Người dùng viết mô tả bằng tiếng Việt cho từng ảnh hoặc từng vùng đối tượng theo ngữ cảnh.

Lưu và truy xuất: Thông tin ảnh, hộp bao và chú thích ảnh được lưu trữ đồng bộ vào MongoDB, và có thể truy xuất theo truy vấn từ phía người dùng.

p-ISSN 3030 - 4628
e-ISSN 3030 - 4636

Trong trường hợp mô hình bỏ sót một đối tượng, người dùng có thể trực tiếp vẽ mới hộp bao quanh đối tượng và gán nhãn thủ công. Các mẫu này được lưu lại trong cơ sở dữ liệu và có thể sử dụng cho việc tinh chỉnh hoặc huấn luyện lại mô hình trong các vòng cải tiến tiếp theo, giúp giảm dần tỷ lệ bỏ sót theo thời gian.

2.3. Các phân hệ chức năng chính

Phương pháp nghiên cứu tập trung vào phân tích yêu cầu người dùng và xây dựng mô hình ca sử dụng (use cases) để đảm bảo hệ thống đáp ứng đầy đủ nghiệp vụ gán nhãn và quản lý dữ liệu. Phương tiện nghiên cứu gồm kỹ thuật thiết kế phần mềm hướng đối tượng, công cụ mô hình hóa ca sử dụng, và môi trường phát triển ReactJS.

Dưới đây mô tả hệ thống bao gồm tám nhóm chức năng chính được hiện thực hóa thông qua các ca sử dụng như sau:

UC-1: Phân quyền người dùng, quản trị viên cấp quyền truy cập chỉnh sửa.

UC-2: Nhận diện đối tượng tích hợp AI để nhận dạng các đối tượng trong ảnh đầu vào.

UC-3: Tải dữ liệu ảnh - hỗ trợ tải theo file đơn và dạng nén (.zip/.rar).

UC-4: Tạo và hiệu chỉnh hộp bao quanh, vẽ, điều chỉnh, xóa vùng khoanh đối tượng.

UC-5: Viết và chỉnh sửa chú thích, tạo chú thích ảnh ngắn hoặc mô tả ngữ nghĩa cho từng ảnh/vùng ảnh.

UC-6: Truy xuất ảnh và chú thích - truy vấn dữ liệu đã gán nhãn theo nhiều tiêu chí và tải về tập dữ liệu đầu ra.

UC-7: Quản lý danh mục - chỉnh sửa danh sách lớp nhãn, nhóm dữ liệu, bộ ảnh.

UC-8: Quản lý tài khoản - tạo, sửa, xóa và theo dõi nhật ký người dùng.

2.4. Yêu cầu phi chức năng và triển khai hệ thống

Phương pháp nghiên cứu được xây dựng dựa trên phân tích yêu cầu phi chức năng của người dùng và đặc điểm hạ tầng triển khai phổ biến, nhằm đảm bảo hệ thống hoạt động ổn định, an toàn và dễ mở rộng. Phương tiện nghiên cứu bao gồm các tiêu chuẩn thiết kế phần mềm web hiện đại, công nghệ xác thực bảo mật, và cơ sở dữ liệu NoSQL.

Hệ thống được thiết kế đảm bảo các yêu cầu phi chức năng sau:

Phản hồi thời gian thực: mỗi truy vấn xử lý ảnh phải hoàn thành trong vòng ≤10 giây.

Khả năng đồng thời: hỗ trợ ít nhất 50 người dùng truy cập đồng thời.

Bảo mật: xác thực người dùng bằng JWT, phân quyền theo vai trò, lưu vết thao tác.

Tương thích trình duyệt: hoạt động ổn định trên Chrome, Firefox, Edge; hỗ trợ hiển thị trên desktop và thiết bị di động.

Phần mềm có thể triển khai tại chỗ (on-premise) hoặc trên máy chủ đám mây với cấu hình tối thiểu: CPU ≥ 4 core, RAM ≥ 8GB, GPU hỗ trợ CUDA (nếu muốn suy diễn mô hình tại backend).

3. KẾT QUẢ NGHIÊN CỨU VÀ THẢO LUẬN

3.1. Thiết lập môi trường thử nghiệm

Để đánh giá hiệu quả và tính khả dụng của hệ thống, nhóm tác giả triển khai phần mềm trên máy chủ cục bộ với cấu hình như sau: CPU: Intel Core i5-12400 (6 cores, 12 threads); RAM: 16 GB DDR4; Hệ điều hành: Ubuntu 22.04 LTS; Backend: Node.js 20 và

p-ISSN 3030 - 4628
e-ISSN 3030 - 4636

Python 3.10 (FastAPI); Frontend: ReactJS 18; Cơ sở dữ liệu: MongoDB 5.4; trình duyệt thử nghiệm: Google Chrome, Firefox. Quá trình thử nghiệm được triển khai trong môi trường Internet, với sự tham gia của 08 thành viên thực hiện các tác vụ gán nhãn, viết chú thích và truy xuất dữ liệu.

3.2. Tập dữ liệu thử nghiệm

Để đánh giá hiệu quả trong điều kiện thực tế, nhóm tác giả đã xây dựng hai tập dữ liệu chuyên biệt phản ánh bối cảnh giáo dục và sinh hoạt thường ngày tại Việt Nam:

Tập dữ liệu học đường: 5000 ảnh chụp tại các trường tiểu học, trung học và đại học ở Thanh Hóa, ghi lại cảnh lớp học, sân trường, giáo viên, học sinh, sách vở và thiết bị giảng dạy.

Tập dữ liệu giao thông: 5000 ảnh chụp tại các tuyến phố, nút giao và phương tiện (ô tô, xe máy, xe đạp), người đi bộ, biển báo, vạch kẻ đường, thu thập ở nhiều thời điểm để tăng tính đa dạng ánh sáng và bối cảnh.

Ảnh được thu thập bằng điện thoại thông minh với độ phân giải tối thiểu 1280×720 pixels, tuân thủ nguyên tắc đạo đức (không ghi rõ mặt người, không chứa thông tin nhạy cảm, không vi phạm quyền riêng tư). Dữ liệu sau đó được tiền xử lý gồm loại bỏ trùng lặp, điều chỉnh kích thước và khử nhiễu để đảm bảo định dạng đầu vào thống nhất.

3.3. Kết quả định lượng

Hệ thống sử dụng YOLOv8-s, được huấn luyện trên Open Images V7, để nhận diện và hiển thị đối tượng cho người dùng. Kết quả (Bảng 1) cho thấy YOLOv8-s đạt mAP@0.5 cao nhất (89,5% trên dữ liệu giao thông) và tỷ lệ hộp cần hiệu chỉnh thấp nhất (10,2% trên dữ liệu học đường). Tỷ lệ nhãn đúng ngay đạt 88,7%, khẳng định tính ứng dụng thực tế, dù dữ liệu giao thông vẫn đòi hỏi mức hiệu chỉnh cao hơn (15,5%) do đặc thù phức tạp.

Cần lưu ý, mAP@0.5 chỉ yêu cầu IoU ≥ 0,5, nên nhiều khung dự đoán vẫn phải chỉnh thủ công về vị trí/kích thước. Vì vậy, việc báo cáo song song mAP và tỷ lệ hiệu chỉnh phản ánh đầy đủ cả hiệu năng học thuật và mức độ hữu dụng triển khai thực tế.

Nhìn chung, YOLOv8-s là lựa chọn phù hợp nhất để giảm thiểu nỗ lực thủ công mà vẫn đảm bảo độ chính xác cao trong nhiều bối cảnh khác nhau.

Bảng 1. Đánh giá độ tin cậy và mức độ hiệu chỉnh của hệ thống gán nhãn

Mô hình	Tập dữ liệu	mAP@0.5 (%)	Tỷ lệ nhãn đúng ngay	Bounding box cần chỉnh (%)
YOLOv5s	Học đường	77,7	85,7	15,2
	Giao thông	75,1	84,5	15,5
YOLOv7	Học đường	85,2	88,7	11,3
	Giao thông	82,5	84,5	15,5
YOLOv8-s	Học đường	85,2	88,7	10,2
	Giao thông	89,5	86,5	15,5

p-ISSN 3030 - 4628
e-ISSN 3030 - 4636

3.4. Tốc độ và hiệu suất

Để đánh giá hiệu năng của hệ thống trong môi trường triển khai thực tế qua Internet, nhóm tác giả tiến hành đo thời gian xử lý trung bình cho từng giai đoạn chính trong quy trình gán nhãn và chú thích ảnh. Kết quả được ghi nhận trong điều kiện người dùng truy cập từ xa thông qua trình duyệt web, kết nối Internet ổn định (tốc độ tải ≥ 20 Mbps).

Bảng 2. Thời gian xử lý trung bình của hệ thống trên môi trường Internet

Giai đoạn	Thời gian trung bình (giây)	Mô tả
Tải ảnh lên	1,2 giây	Ảnh JPEG/PNG dung lượng <=3MB
Nhận diện và đưa ra kết quả	2,5 giây	Gồm xử lý YOLOv8 và trả về hộp bao quanh đối tượng
Chỉnh sửa và chú thích	4,2 giây	Thao tác trung bình của người dùng
Lưu dữ liệu	0,9 giây	Bao gồm xác thực và ghi log

Bảng 2 trình bày thời gian xử lý trung bình cho từng giai đoạn chính trong quy trình gán nhãn khi triển khai hệ thống trên môi trường Internet. Kết quả cho thấy toàn bộ quy trình hoàn tất trong khoảng dưới 10 giây, đáp ứng tốt yêu cầu phản hồi thời gian thực. Cụ thể, thời gian nhận diện và hiển thị kết quả từ mô hình YOLOv8 chỉ mất 2,5 giây, cho thấy khả năng tích hợp AI nhẹ hiệu quả. Thao tác chỉnh sửa và viết chú thích trung bình mất 4,2 giây, phản ánh sự tối ưu trong giao diện người dùng và luồng tương tác. Việc lưu dữ liệu vào cơ sở dữ liệu chỉ mất 0.9 giây, nhờ sử dụng kiến trúc RESTful API kết hợp MongoDB tối ưu. Tổng thể, các chỉ số này khẳng định hệ thống hoạt động ổn định ngay cả trong môi trường phân tán và có thể mở rộng để phục vụ đồng thời nhiều người dùng mà không gây nghẽn.

Bảng 3. So sánh năng suất gán nhãn ảnh trước và sau khi sử dụng hệ thống

Tiêu chí	Trước hệ thống (thủ công)	Với hệ thống
Số ảnh gán nhãn/giờ	~22	~56
Thời gian trung bình/ảnh (phút)	2,7	1,0
Chú thích trung bình mỗi ảnh	5 câu	5 câu

Kết quả ở Bảng 2 và Bảng 3 cho thấy hệ thống đáp ứng tốt yêu cầu phản hồi thời gian thực, ngay cả trong môi trường phân tán qua Internet. Độ trễ mạng trung bình không ảnh hưởng đáng kể đến tốc độ xử lý tổng thể nhờ việc tối ưu kiến trúc API (RESTful), sử dụng mô hình AI nhẹ (YOLOv8-small) và kết nối cơ sở dữ liệu nội bộ. Trong quá trình thử nghiệm, để đánh giá khả năng xử lý đồng thời của hệ thống, 08 thành viên đã thực hiện gán nhãn cùng lúc, nhưng hệ thống vẫn duy trì hiệu suất ổn định, không xảy ra tình trạng nghẽn API hoặc treo giao diện. Ngoài ra, hệ thống cũng được ghi nhận về mức độ tiêu tốn tài nguyên hợp lý, đảm bảo khả năng vận hành lâu dài.

p-ISSN 3030 - 4628
e-ISSN 3030 - 4636

4. KẾT LUẬN

Kết quả thực nghiệm cho thấy, hệ thống gán nhãn bán tự động tích hợp YOLOv8-s đáp ứng tốt yêu cầu về độ chính xác, tốc độ xử lý và tính khả dụng trong môi trường Internet. Với thời gian phản hồi ngắn, tỷ lệ nhãn đúng cao và mức hiệu chỉnh thấp, hệ thống góp phần tối ưu hóa quy trình xây dựng tập dữ liệu huấn luyện. Trong tương lai, nhóm tác giả dự kiến khai thác dữ liệu từ các trường hợp người dùng chỉnh sửa hoặc bổ sung thủ công để huấn luyện lại mô hình theo hướng học chủ động (active learning) hoặc bán tự động. Điều này không chỉ cải thiện hiệu quả nhận diện mà còn giảm dần tỷ lệ hộp bao quanh đối tượng cần chỉnh, góp phần nâng cao chất lượng và độ tin cậy của hệ thống gán nhãn trong môi trường triển khai thực tế.

TÀI LIỆU THAM KHẢO

[1] Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... & Zitnick, C. L. (2014), *Microsoft coco: Common objects in context*, In Computer vision-ECCV, 740-755.

[2] Moon, Y. B., & Oh, T. H. (2024), *Label-efficient learning methods for computer vision applications*, IEIE Transactions on Smart Processing & Computing, 13(2), 120-128.

[3] Osaid, M., & Memon, Z. A. (2022), *A Survey On Image Captioning*, In 2022 (ICETST), IEEE, 1-6.

[4] Li, J., Li, D., Xiong, C., & Hoi, S. (2022), *Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation*, PMLR, 12888-12900.

[5] Wang, J., Yang, Z., Hu, X., Li, L., Lin, K., Gan, Z., ... & Wang, L. (2022), *Git: A generative image-to-text transformer for vision and language*, arXiv preprint arXiv:2205.14100.

[6] Liu, L., Ouyang, W., Wang, X., Fieguth, P. (2020), *Deep learning for generic object detection: A survey*, International Journal of Computer Vision, 128, 261-318.

[7] Adnan, M. M., Rahim, M. S. (2021), *A review of methods for the image automatic annotation*, In Journal of Physics: Conference Series, 1892(1), pp.012002.

[8] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016), *You only look once: Unified, real-time object detection*, In Proceedings of the IEEE CVPR, 779-788.

[9] Ren, S., He, K., Girshick, R., (2015), *Faster r-cnn: Towards real-time object detection with region proposal networks*, Advances in neural information processing systems, 28.

[10] Tan, M., Pang, R., & Le, Q. V. (2020), *Efficientdet: Scalable and efficient object detection*, In Proceedings of the IEEE/CVF, 10781-10790.

[11] Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015), *Show and tell: A neural image caption generator*, In Proceedings of the IEEE CVPR, 3156-3164.

[12] Xu, K., Ba, J., Kiros, R., Cho, K.,(2015), *Show, attend and tell: Neural image caption generation with visual attention*, In International Conference on Machine.

p-ISSN 3030 - 4628
e-ISSN 3030 - 4636

[13] Alayrac, J. B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., ... & Simonyan, K. (2022), *Flamingo: a visual language model for few-shot learning*, Advances in Neural Information Processing Systems, 35, 23716-23736.

[14] Bui, D. C., Nguyen, N. H., & Nguyen, K. (2023), *UIT-OpenViIC: A Novel Benchmark for Evaluating Image Captioning in Vietnamese*, arXiv preprint arXiv:2305.04166.

[15] Sohan, M., Sai Ram, T., & Rami Reddy, C. V. (2024), *A review on yolov8 and its advancements*, In International Conference on Data Intelligence and Cognitive Informatics, Springer, Singapore, 529-545.

DEVELOPMENT OF A SEMI-AUTOMATIC IMAGE LABELING AND ANNOTATION SOFTWARE USING ARTIFICIAL INTELLIGENCE

Nguyen Dinh Cong, Hoang Anh Cong, Nguyen Hoang Long, Pham The Anh, Le Van Sam

ABSTRACT

This paper presents the development of a semi-automatic software system for image labeling and annotation to support training datasets for computer vision models. The system integrates YOLOv8 for automatic object detection, while allowing users to adjust bounding boxes and write semantic captions. It includes user role management, dataset organization, and flexible data retrieval. Experimental results show that the system significantly reduces labeling time while maintaining high accuracy, paving the way for future integration of automatic caption generation models.

Keywords: *Image labeling, semantic annotation, deep learning.*

** Ngày nộp bài: 01/7/2025; Ngày gửi phản biện: 18/7/2025; Ngày duyệt đăng: 25/9/2025*

** Bài báo là kết quả nghiên cứu từ đề tài NCKH cấp cơ sở (mã số ĐT-2024-19) của Trường Đại học Hồng Đức.*