AN IMAGE RETRIEVAL MODEL USING KNOWLEDGE GRAPH AND BAG OF VISUAL WORDS

Tran Duc Tai, Nguyen Ngoc Sang, To Thanh Tuan, Nguyen Do Thai Nguyen*
Ho Chi Minh City University of Education

ARTICLE INFO

Received: 17/4/2025
Revised: 29/6/2025
Published: 30/6/2025

KEYWORDS

Image retrieval
Bag of visual words
Knowledge graph
YOLOv8
Object detection

ABSTRACT

In the context of the growing demand for image retrieval based on content and semantic understanding, traditional techniques that rely solely on visual features are increasingly revealing limitations, especially in representing semantic relationships among entities within an image. This study proposes an integrated model comprising three key components: entity detection using YOLOv8, visual feature representation through the bag of visual words model, and information organization via a knowledge graph. Detected entities are encoded into bag of visual words, from which relational triples are constructed and mapped into the knowledge graph. During querying, the system generates triples from the input image to perform semantic retrieval within the knowledge graph. The model was evaluated on two widely used image datasets OpenImagesV7 and MS-COCO, achieving accuracies of 84.1% and 89.6%, respectively. These results outperform many traditional approaches, reflecting the reliability and feasibility of the proposed model.

MỘT MÔ HÌNH TRUY VẤN ẢNH SỬ DỤNG ĐỒ THỊ TRI THỨC VÀ TÚI TỪ THỊ GIÁC

Trần Đức Tài, Nguyễn Ngọc Sang, Tô Thanh Tuấn, Nguyễn Đỗ Thái Nguyên*
Trường Đại học Sư phạm Thành phố Hồ Chí Minh

THÔNG TIN BÀI BÁO

Ngày nhận bài: 17/4/2025
Ngày hoàn thiện: 29/6/2025
Ngày đăng: 30/6/2025

TỪ KHÓA

Truy vấn ảnh
Túi từ thị giác
Đồ thị tri thức
YOLOv8
Trích xuất đối tượng

TÓM TẮT

Trong bối cảnh nhu cầu tra cứu ảnh theo hướng hiểu nội dung và ngữ nghĩa ngày một trở nên phổ biến, những kỹ thuật truyền thống vốn chỉ dựa vào đặc trưng thị giác đang dần bộc lộ nhiều hạn chế, đặc biệt khi phải diễn tả được các quan hệ ngữ nghĩa giữa những thực thể có trong ảnh. Nghiên cứu này đề xuất một mô hình tích hợp gồm ba yếu tố chính: phát hiện thực thể bằng YOLOv8, biểu diễn đặc trưng thị giác với túi từ thị giác, và tổ chức thông tin bằng đồ thị tri thức. Các thực thể được phát hiện sẽ được tổ chức vào túi từ thị giác từ đó tạo các bộ ba quan hệ để ánh xạ vào đồ thị tri thức. Khi truy vấn, hệ thống sinh các bộ ba từ ảnh đầu vào để thực hiện tra cứu trong đồ thị tri thức. Mô hình được triển khai đánh giá trên hai tập ảnh dữ liệu phổ biến là OpenImagesV7 và MS-COCO với độ chính xác đạt được ở mức 84,1% và 89,6%, vượt qua nhiều mô hình truyền thống, phản ánh độ tin cậy và khả thi của mô hình đề xuất.

DOI: https://doi.org/10.34238/tnu-jst.12608

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

1. Giới thiệu

Hệ thống tìm kiếm hình ảnh dựa vào đặc điểm nội dung (Content-Based Image Retrieval – CBIR) hiện nay được xem là một trong những hướng nghiên cứu trọng tâm của thị giác máy tính, với phạm vi ứng dụng rộng rãi trong các lĩnh vực như chẩn đoán hình ảnh y khoa, giám sát an ninh và quản lý kho dữ liệu đa phương tiện. Các phương pháp truyền thống thường sử dụng các đặc trưng thị giác để tính toán sự tương đồng giữa ảnh truy vấn và ảnh trong cơ sở dữ liệu. Tuy nhiên, các đặc trưng này chủ yếu phản ánh thông tin cục bộ, thiếu khả năng biểu diễn mối quan hệ ngữ nghĩa giữa các đối tượng, dẫn đến việc hệ thống khó hiểu được nội dung ảnh ở mức độ khái niệm cao [1], [2]. Ví dụ, một bức ảnh chứa "người cầm cốc" và "cốc đặt trên bàn" có thể được xem là tương đồng về màu sắc nhưng khác biệt hoàn toàn về ngữ cảnh, điều mà các phương pháp truyền thống không phân biệt được.

Để giải quyết hạn chế này, nhiều nghiên cứu gần đây đã hướng đến việc tích hợp tri thức bên ngoài vào quá trình truy vấn. Trong đó, đồ thị ngữ cảnh (scene graph) nổi lên như một công cụ hiệu quả để mô tả các thực thể và mối liên hệ giữa chúng thông qua cấu trúc đồ thị [3], [4]. Đồ thị này không chỉ liệt kê các thực thể trong ảnh (ví dụ: "người", "ghế", "chó") mà còn mã hóa tương tác giữa chúng (ví dụ: "người →ngồi → ghế", "chó → chạy → cạnh xe"). Cách tiếp cận này nâng cao khả năng diễn giải ngữ nghĩa, từ đó cải thiện hiệu quả tra cứu ảnh dựa trên các mối liên kết ngữ nghĩa giữa các thực thể. Song song với đó, đồ thị tri thức (Knowledge Graph – KG) được tạo từ các bộ ba quan hệ (subject-predicate-object), cho phép nhúng thông tin vào không gian vector để hỗ trợ suy luận logic và học máy [5], [6]. Sự kết hợp giữa hai đồ thị này đã chứng minh khả năng nâng cao chất lượng truy vấn thông qua việc kết nối thông tin thị giác với tri thức nền [3], [6], mở đường cho các hệ thống CBIR thông minh và có khả năng giải thích.

Bên cạnh các phương pháp dựa trên tri thức, việc biểu diễn đặc trưng hình ảnh cũng không ngừng được cải tiến. Mô hình túi từ thị giác (Bag of Visual Words – BoVW) vẫn giữ vị trí quan trọng nhờ khả năng ánh xạ các đối tượng thành các "từ" thị giác có tính tổ chức cao, tương tự cách biểu diễn văn bản trong xử lý ngôn ngữ tự nhiên. Các nghiên cứu gần đây [1], [2] chỉ ra rằng, dù các mô hình học sâu như mạng nơ-ron tích chập (Convolutional Neural Networks - CNN) đã cải thiện đáng kể hiệu suất trích xuất đặc trưng, việc tổ chức chúng thành dạng có thể truy vấn vẫn là thách thức. BoVW đóng vai trò cầu nối giữa thông tin thị giác và tri thức, đặc biệt hiệu quả khi được kết hợp với những phương pháp phát hiện đối tượng tiên tiến như YOLOv8. YOLOv8, một phiên bản của họ YOLO, nổi bật nhờ tốc độ xử lý nhanh và độ chính xác cao trong việc nhận diện đối tượng, là cơ sở cho việc xây dựng BoVW và đồ thị tri thức [7].

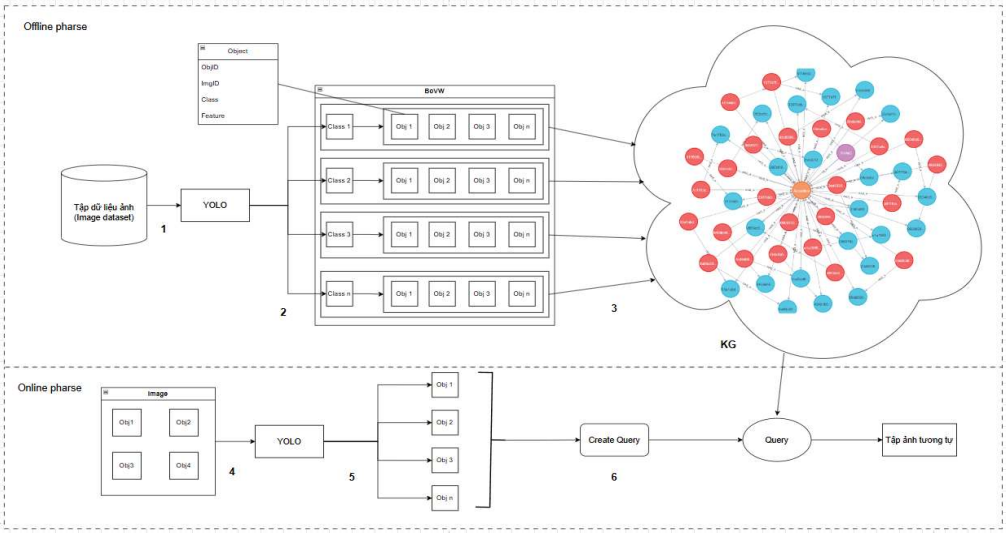
Tại Việt Nam, một số nghiên cứu đã tiếp cận hướng tích hợp tri thức vào CBIR. Chẳng hạn, nhóm nghiên cứu của Lê Thị Vĩnh Thanh [8] đưa ra mô hình kết hợp đồ thị láng giềng và đồ thị ngữ nghĩa giúp cải thiện hiệu suất trong tìm kiếm ảnh, trong khi nhóm của Phan Minh Tiến [9] sử dụng thống kê và biểu diễn tri thức để tối ưu hóa quá trình tìm kiếm. Tuy nhiên, các phương pháp này vẫn tồn tại hạn chế như thiếu khả năng tự động hóa, phụ thuộc vào dữ liệu huấn luyện cục bộ và chưa kết nối hiệu quả với nguồn tri thức đa dạng bên ngoài.

Trong bối cảnh đó, nghiên cứu này đề xuất một mô hình tra cứu hình ảnh tích hợp ba thành phần chính: YOLOv8 để nhận diện đối tượng chính xác, BoVW để tổ chức đặc trưng thị giác có cấu trúc, và đồ thị tri thức đóng vai trò trong việc để biểu diễn và tổ chức các liên hệ ngữ nghĩa. Khác với cách tiếp cận truyền thống, phương pháp đề xuất khai thác không chỉ dựa trên sự tương đồng về hình ảnh mà còn xem xét logic giữa các đối tượng để thiết kế một hệ thống đáp ứng yêu cầu diễn giải kết quả và mở rộng tri thức. Chúng tôi đã triển khai thực nghiệm trên các tập ảnh chuẩn như MSCOCO và OpenImagesV7 nhằm kiểm chứng hiệu quả ngữ nghĩa và khả năng tổng quát hóa của phương pháp đề xuất.

2. Phương pháp nghiên cứu

2.1. Cấu trúc tổng quan của mô hình truy vấn ảnh

Nghiên cứu này giới thiệu một mô hình tra cứu ảnh tích hợp ba thành phần cốt lõi: phát hiện các thực thể bằng YOLOv8, biểu diễn đặc trưng với mô hình túi từ thị giác (BoVW), và tổ chức thông tin thông qua đồ thị tri thức (KG). Mô hình gồm hai pha: (1) Pha ngoại tuyến để rút trích và xây dựng cơ sở tri thức từ tập huấn luyện; và (2) Pha trực tuyến để tra cứu ảnh và cho các ảnh tương đồng. Hình 1 minh họa mô hình đề xuất.



Hình 1. *Mô hình đề xuất*

Trong Hình 1, hệ thống gồm có các bước sau:

Ở pha ngoại tuyến:

(1) YOLOv8 nhận diện và gắn nhãn các thực thể trong ảnh đầu vào.

(2) Các thực thể được tổ chức vào BoVW cùng với đặc trưng histogram, phân loại theo từng nhóm đối tượng.

(3) Từ BoVW, hệ thống xây dựng KG để biểu diễn các mối liên hệ giữa các thực thể.

Ở pha trực tuyến:

(4) Ảnh truy vấn được phân tích bằng YOLOv8 để xác định các thực thể và trích xuất đặc trưng.

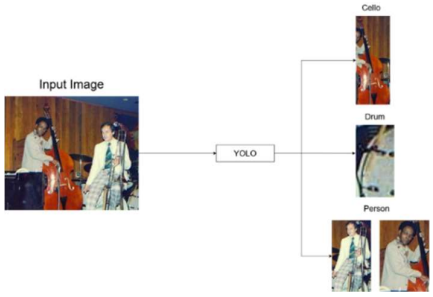
(5) Hệ thống sử dụng các bộ ba quan hệ từ ảnh đầu vào để tra cứu ngữ nghĩa trong KG.

Tập ảnh thu được thể hiện sự tương đồng ngữ nghĩa với ảnh đầu vào theo các bộ ba quan hệ được truy xuất. Hệ thống này không chỉ so khớp đặc trưng thị giác mà còn sử dụng các mối liên kết giữa các đối tượng, mang lại khả năng truy vấn chính xác và dễ diễn giải hơn.

2.2. Các thành phần của mô hình đề xuất

2.2.1. Phát hiện đối tượng với YOLOv8

Trong mô hình đề xuất, YOLOv8 đóng vai trò nhận diện và gán nhãn các thực thể xuất hiện trong ảnh, làm nền tảng cho quá trình trích xuất và tổ chức thông tin thị giác. Mỗi đối tượng được xác định kèm theo điểm tin cậy (confidence score), sau đó được chuyển vào mô hình túi từ thị giác (BoVW) để tổ chức và lưu trữ. Kết quả này là nền tảng cho việc xây dựng đồ thị tri thức, thể hiện các mối liên kết giữa các đối tượng. Trong giai đoạn truy vấn, YOLOv8 tiếp tục phân tích ảnh đầu vào để tạo ra các bộ ba quan hệ, làm cơ sở cho việc tra cứu các ảnh tương tự trong KG. Quá trình nhận diện đối tượng bằng YOLOv8 được minh họa cụ thể trong Hình 2: từ ảnh đầu vào qua YOLOv8 sẽ nhận diện được các đối tượng thuộc các lớp Cello, Drum, Person.



Hình 2. *Quá trình nhận diện đối tượng sử dụng YOLOv8*

Tuy nhiên, chất lượng của các đối tượng phát hiện bởi YOLOv8 quyết định trực tiếp đến độ chính xác của BoVW và KG. Nếu YOLOv8 bỏ sót hoặc nhận diện sai nhãn, các bộ ba quan hệ sinh ra sẽ không đầy đủ hoặc sai lệch, ảnh hưởng đến khả năng truy vấn chính xác. Do đó, việc huấn luyện YOLOv8 với tập dữ liệu phù hợp và tối ưu hóa tham số nhận diện là rất quan trọng để đảm bảo tính nhất quán của hệ thống.

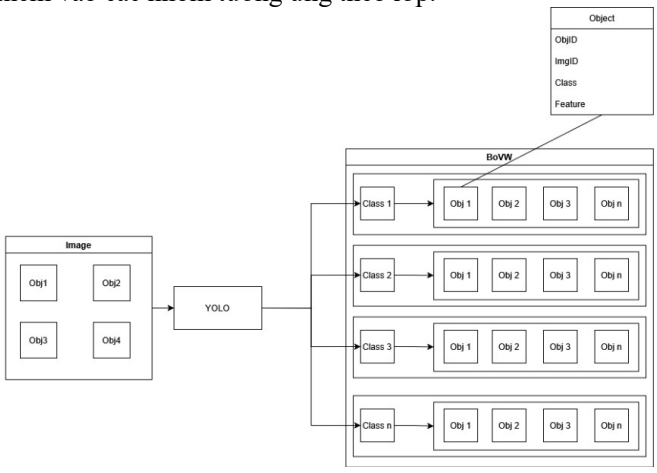
2.2.2. Túi từ thị giác

Mô hình BoVW biểu diễn nội dung hình ảnh dựa trên việc tổ chức các đặc trưng cục bộ thành các từ thị giác. Trong hệ thống này, BoVW đóng vai trò cầu nối giữa thông tin từ YOLOv8 và việc ánh xạ vào KG.

Cụ thể, mỗi ảnh đầu vào được chuyển đổi thành một tập các từ thị giác, phản ánh các đối tượng được nhận dạng và gán nhãn bởi YOLOv8. Các đối tượng này được tổ chức theo hai cấu trúc chính:

- **Object**: Đại diện cho một thực thể riêng lẻ trong ảnh, chứa các thông tin như ObjID (ID của đối tượng), ImgID (ID của ảnh chứa đối tượng), Class (nhãn phân loại), Feature (đặc trưng), và Conf (điểm tin cậy).
- **Class**: Đại diện cho nhóm đối tượng cùng loại, chứa các thông tin như ClassName, ObjCount (số lượng đối tượng), và ObjList (danh sách các đối tượng thuộc nhóm).

Tập hợp các nhóm Class tạo nên BoVW, hình thành một không gian phân loại có cấu trúc. Quá trình tổ chức đối tượng vào BoVW được thực hiện bằng cách duyệt qua từng ảnh, trích xuất danh sách đối tượng và thêm vào các nhóm tương ứng theo lớp.



Hình 3. *Minh họa cấu trúc túi từ thị giác*

Hình 3 minh họa cấu trúc túi từ thị giác: từ ảnh đầu vào qua YOLOv8 sẽ nhận dạng các đối tượng Object, sau đó các đối tượng sẽ được phân bổ vào túi từ tương ứng với từng Class.

2.2.3. Khung đồ thị tri thức (KGF)

Đồ thị tri thức đóng vai trò là khung lưu trữ, tổ chức thông tin về thực thể, liên hệ và thuộc tính trong phương pháp đề xuất. Mỗi thực thể là một nút trong KG, còn mối quan hệ được thể hiện bằng các cạnh nối giữa các nút. Các quan hệ chính bao gồm: *IsA* (xác định phân cấp giữa lớp và nút gốc), *HasA* (liên kết đối tượng với lớp hoặc ảnh), *HasSubcategory* (mô tả quan hệ phân lớp con) và *IsAttributeOf* (gắn đặc trưng histogram với ảnh). Trong đó, *IsA* và *HasA* là quan hệ cốt lõi để hỗ trợ truy vấn. Các từ thị giác trong BoVW được ánh xạ thành các thực thể và quan hệ dựa trên cấu trúc Class–Object đã tổ chức trước đó. Cụ thể, mỗi thực thể sẽ tạo ít nhất hai quan hệ: (ClassName, HasA, Object) và (ImgID, HasA, Object). Các quan hệ này đóng vai trò tạo nên mạng lưới tri thức để suy luận khi tra cứu. Ngoài các quan hệ cơ bản, mô hình có thể bổ sung các thuộc tính mở rộng như vị trí, màu sắc để tăng độ phong phú cho KG.

Để chuyển đổi BoVW thành KG, hệ thống sử dụng thuật toán CKGF (Create Knowledge Graph Framework), gồm các bước:

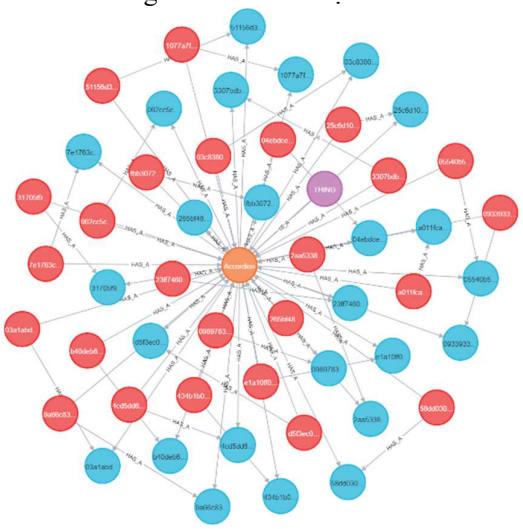
Bước 1: Với mỗi lớp trong BoVW, tạo quan hệ Thing – IsA → ClassName

Bước 2: Với mỗi đối tượng thuộc lớp đó, tạo các bộ ba:

- ClassName – HasA → Object
- ImgID – HasA → Object
- ImgID – HasA → ClassName (nếu chưa tồn tại)

Bước 3: Thêm các bộ ba vào KG nếu chưa xuất hiện trong KG

Nhờ cấu trúc bộ ba, đồ thị tri thức cho phép thực hiện truy vấn ngữ nghĩa hiệu quả và dễ dàng mở rộng với các nguồn tri thức bổ sung. Hình 4 minh họa cho cấu trúc logic của KGF.



Hình 4. *Minh họa cấu trúc của KGF*

Trong Hình 4, node màu cam là ClassName, màu đỏ là ImgID, màu xanh là Object, bộ ba ClassName – HasA → Object được biểu diễn bằng mũi tên HAS_A từ node cam sang node xanh, bộ ba ImgID – HasA → Object được biểu diễn bằng mũi tên HAS_A từ node đỏ sang node xanh và bộ ba ImgID – HasA → ClassName được biểu diễn bằng mũi tên HAS_A từ node đỏ sang node cam.

2.3. Thuật toán truy vấn

Sau khi KG được xây dựng từ tập ảnh huấn luyện, hệ thống triển khai thuật toán truy vấn để tìm kiếm các ảnh trong cơ sở dữ liệu dựa trên mức độ tương đồng ngữ nghĩa với ảnh đầu vào. Thuật toán RIKGF (Retrieval in Knowledge Graph Framework) như sau:

Input: Ảnh truy vấn

```
Output: Tập các ảnh tương đồng
BEGIN
    image_profile = get_item_from_image(YOLO, I)
    Triples = []
    Foreach obj ∈ image_profile do
    Triples.add(CreateTriple(obj[“ImgID”],HasA,obj[“ClassName”]))
    End Foreach
    Query_Imgs = query_images(Triples)
    listTopK = sort_by_cosine(I, Query_Imgs)
    Return listTopK[:K]
END
```

Ảnh đầu vào được phân tích bởi YOLOv8 để nhận diện các thực thể, sau đó chuyển thành các bộ ba. Các bộ ba này dùng để tìm kiếm trong KG, nhằm thu được tập ảnh tương đồng. Ví dụ: một ảnh đầu vào có ID là Img_001 và YOLOv8 phát hiện được một đối tượng thuộc lớp “Dog” với ID là Obj_10. Khi đó, hệ thống sẽ sinh ra các bộ ba quan hệ cụ thể: (1) Dog – HasA → Obj_10, nghĩa là lớp “Dog” sở hữu một đối tượng cụ thể là Obj_10; (2) Img_001 – HasA → Obj_10, thể hiện ảnh Img_001 chứa đối tượng Obj_10. Kế tiếp, thuật toán sẽ truy vấn toàn bộ KG để tìm các ảnh khác cũng chứa các quan hệ “HasA” với cùng lớp “Dog” hoặc đối tượng cùng loại. Những ảnh có nhiều quan hệ trùng khớp sẽ được ưu tiên trong danh sách kết quả. Cuối cùng, hệ thống sẽ tính khoảng cách cosine của ảnh đầu vào với kết quả được tìm thấy, công thức như sau:

$$d(Q, O) = 1 - \cos\theta = 1 - \frac{Q.O}{\|Q\|.\|O\|} \quad (1)$$

Các ảnh có khoảng cách cosine càng thấp thì sẽ càng tương đồng với ảnh đầu vào. Vì thế, hàm xếp hạng đã áp dụng điều này để sắp xếp lại theo độ ưu tiên các ảnh có khoảng cách cosine thấp nhất lên trước, từ đó lấy ra một top K ảnh bất kì (với K là một hằng số thể hiện số lượng ảnh muốn truy vấn).

3. Kết quả và thảo luận

3.1. Kết quả thực nghiệm

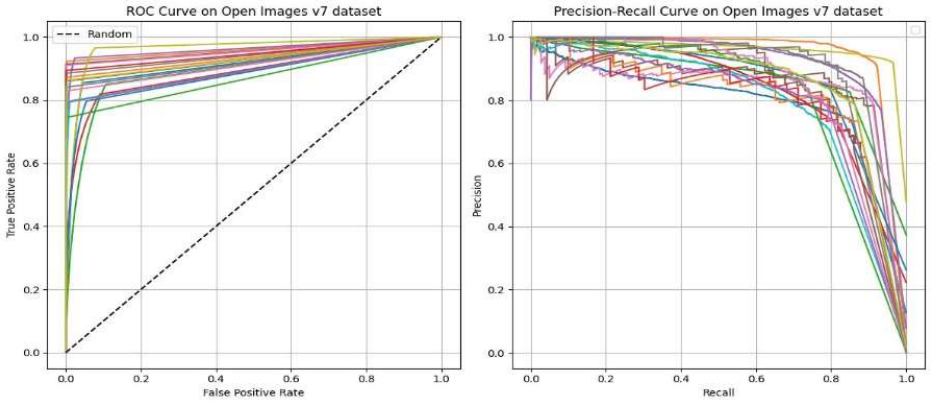
Nhằm kiểm tra tính hiệu quả của hệ thống truy vấn ảnh, nhóm nghiên cứu đã triển khai thực nghiệm trên hai tập ảnh dữ liệu tiêu chuẩn là MS COCO và OpenImagesV7. Các chỉ số đánh giá chính bao gồm Precision (độ chính xác), Recall (độ phủ), và F1-score (chỉ số trung bình điều hòa giữa Precision và Recall), phản ánh khả năng truy vấn chính xác và đầy đủ của mô hình.

Dữ liệu trong Bảng 1 cho thấy mô hình đạt Precision 89,6%, Recall 80,2%, và F1-score 85,7% trên tập MS COCO. Trong khi đó, trên tập OpenImagesV7, mô hình đạt Precision 84,1%, Recall 77,4%, và F1-score 80,3%. Các chỉ số này cho thấy mô hình đều có độ chính xác và độ phủ tốt trên cả hai tập dữ liệu có đặc điểm khác biệt về nội dung và phân phối hình ảnh.

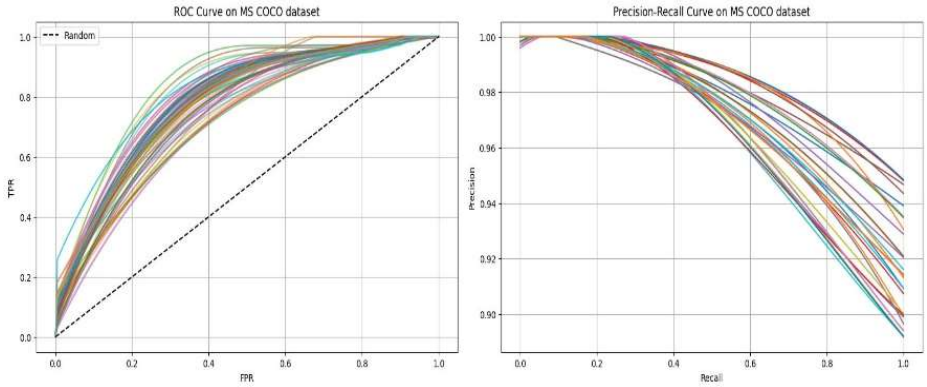
Bảng 1. *Kết quả thử nghiệm trên các tập dữ liệu MSCOCO và OpenImagesV7*

Tập ảnh dữ liệu	Precision	Recall	F1-score
MS COCO	89,6%	80,2%	85,7%
OpenImagesV7	84,1%	77,4%	80,3%

Ngoài ra, Hình 5 và Hình 6 minh họa đường cong ROC và biểu đồ Precision–Recall trên tập OpenImagesV7 và MSCOCO, cho thấy khả năng phân tách và xếp hạng ảnh tương đồng rõ rệt của hệ thống. Các đường cong này phản ánh mối quan hệ giữa tỷ lệ tìm kiếm đúng và tỷ lệ tìm kiếm sai, đồng thời cho thấy độ chính xác duy trì ổn định ở nhiều ngưỡng truy vấn khác nhau.

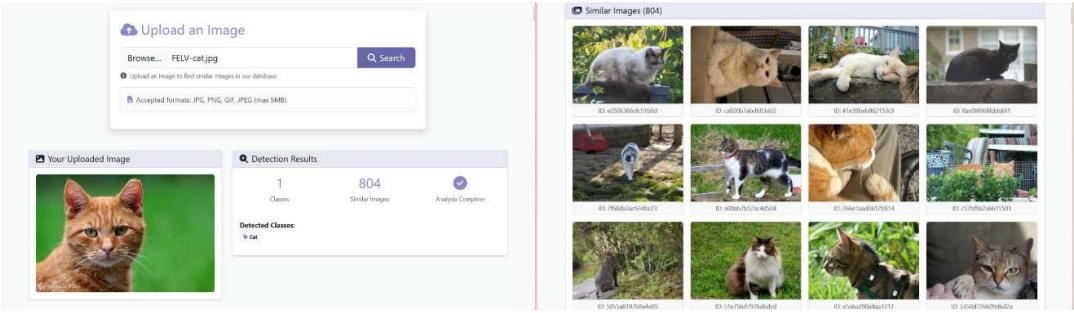


Hình 5. *Đường cong ROC, Precision – Recall của bộ dữ liệu OpenImagesV7*



Hình 6. *Đường cong ROC, Precision – Recall của bộ dữ liệu MS COCO*

Hình 7 cũng cung cấp một ví dụ thực tế về kết quả truy vấn ảnh, trong đó hệ thống đã xác định được các ảnh chứa thực thể và bối cảnh tương đồng với ảnh đầu vào.



Hình 7. *Ví dụ tra cứu ảnh*

Ảnh FELV-cat được nhận diện có một đối tượng thuộc lớp "Cat" với ID là Obj1, hệ thống sẽ sinh ra các bộ ba quan hệ cụ thể: (1) Cat – HasA → Obj1; (2) FELV-cat – HasA → Obj_1, sau đó thuật toán sẽ truy vấn toàn bộ KG để tìm các ảnh khác cũng chứa các quan hệ "HasA" với cùng lớp "Cat" hoặc đối tượng cùng loại.

Để làm rõ tính vượt trội của mô hình, Bảng 2 trình bày so sánh với một số phương pháp truy vấn ảnh nổi bật trên tập MS COCO, bao gồm CLIP [10], Faster R-CNN kết hợp VQA [11], IronKG [8], và NeiKG [9]. Mô hình đề xuất vượt qua tất cả các phương pháp kể trên với độ chính xác

89,6%, cho thấy tiềm năng mạnh mẽ khi kết hợp giữa biểu diễn thị giác có cấu trúc và suy luận ngữ nghĩa trong đồ thị.

Bảng 2. *So sánh với các mô hình khác trên tập MSCOCO*

Phương pháp	Độ chính xác
CLIP với encoder ResNet-50/ViT [10]	72,8%
Faster R-CNN kết hợp VQA [11]	77,8%
IronKG [8]	79,3%
NeiKG [9]	88,32%
Mô hình đề xuất	89,6%

3.2. Thảo luận

Dữ liệu thực nghiệm cho thấy phương pháp đề xuất đạt kết quả tốt trên cả hai tập dữ liệu, đặc biệt hiệu quả hơn so với các phương pháp như CLIP hay Faster R-CNN kết hợp VQA. Việc tích hợp YOLOv8, BoVW và KG giúp hệ thống không chỉ có thể tìm kiếm ảnh dựa trên đặc trưng thị giác mà còn xét đến mối liên kết ngữ nghĩa giữa các thực thể, cải thiện mức độ chính xác và tăng cường khả năng hiểu thông tin hình ảnh.

Khác với các phương pháp truyền thống chỉ dựa trên đặc trưng hình ảnh, mô hình đề xuất tận dụng các quan hệ bộ ba từ KG để lấp đầy khoảng trống thông tin khi ảnh bị phân mảnh hoặc thiếu chi tiết. Nhờ khả năng suy luận trên KG, hệ thống vẫn có thể tìm kiếm và truy xuất các ảnh tương đồng về mặt ngữ nghĩa ngay cả khi một số đối tượng không được nhận diện đầy đủ.

Tuy nhiên, hiệu quả hệ thống vẫn phụ thuộc vào hiệu quả của YOLOv8. Nếu nhận diện sai hoặc thiếu thực thể quan trọng, KG sẽ bị thiếu thông tin, ảnh hưởng đến kết quả truy vấn. Ngoài ra, việc khai thác các quan hệ ngữ nghĩa ẩn vẫn là một thách thức cần nghiên cứu thêm.

Đáng chú ý, kết quả trên tập MS COCO cao hơn OpenImagesV7. Nguyên nhân có thể do MS COCO có chú thích đối tượng chi tiết và ít nhiễu hơn, giúp YOLOv8 phát hiện chính xác. Trong khi đó, OpenImagesV7 chứa nhiều ảnh phức tạp, nhiều đối tượng nhỏ và nhãn không đồng nhất, ảnh hưởng đến chất lượng bộ ba quan hệ.

Khác biệt nổi bật của mô hình so với các nghiên cứu trước đó là sự tích hợp toàn diện giữa phát hiện đối tượng, biểu diễn đặc trưng (BoVW) và tổ chức tri thức (KG) trong một quy trình thống nhất. Các nghiên cứu trước thường chỉ áp dụng BoVW hoặc KG riêng lẻ, trong khi mô hình này cho phép tự động hóa hoàn toàn việc sinh bộ ba từ ảnh đầu vào, tăng cường khả năng truy vấn ngữ nghĩa mà không cần mô tả thủ công. Đây chính là điểm mới mẻ và giá trị thực tiễn mà nghiên cứu mang lại.

Mặc dù mô hình đã cho thấy hiệu quả cao trên các tập dữ liệu chuẩn, việc triển khai trong các môi trường thực tế đòi hỏi khả năng thích ứng với dữ liệu đa dạng, nhiễu hoặc không đầy đủ. Trong những tình huống như vậy, chất lượng phát hiện đối tượng và độ hoàn thiện của đồ thị tri thức có thể bị ảnh hưởng, làm giảm hiệu quả truy vấn. Do đó, các nghiên cứu tiếp theo sẽ tập trung vào việc nâng cao tính linh hoạt của mô hình, tối ưu hóa YOLOv8 cho các nguồn dữ liệu phức tạp và phát triển các cơ chế suy luận bù đắp thông tin thiếu hụt.

4. Kết luận

Nghiên cứu này đã giới thiệu một mô hình truy vấn ảnh kết hợp giữa YOLOv8, túi từ thị giác và đồ thị tri thức. Hướng tiếp cận này khai thác khả năng phát hiện thực thể mạnh mẽ của YOLOv8, biểu diễn hình ảnh hiệu quả của túi từ thị giác và khả năng suy luận ngữ nghĩa từ KG. Việc đánh giá mô hình trên các tập ảnh chuẩn như OpenImagesV7 và MS-COCO đã cho các kết quả khả quan, với độ chính xác đạt được là 84,1% và 89,6%, phản ánh tính hiệu quả của phương pháp đề xuất trong việc nhận diện và truy vấn ảnh có liên quan. Trong các nghiên cứu tiếp theo, mô hình có thể được mở rộng để xử lý các trường hợp dữ liệu không đầy đủ hoặc thiếu nhãn, cũng như tích hợp thêm dữ liệu đa phương thức như kết hợp hình ảnh với mô tả văn bản hoặc âm thanh, nhằm tăng khả năng hiểu và tra cứu ngữ nghĩa phức tạp.

Lời cám ơn

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

TÀI LIỆU THAM KHẢO/ REFERENCES

[1] I. M. Hameed, S. H. Abdulhussain, and B. M. Mahmmod, "Content-based image retrieval: A review of recent trends," *Cogent Engineering*, vol. 8, no. 1, 2021, Art. no. 1927469, doi: 10.1080/23311916.2021.1927469.

[2] X. Li, J. Yang, and J. Ma, "Recent developments of content-based image retrieval (CBIR)," *Neurocomputing*, vol. 452, no. 10, pp. 675-689, 2021, doi: 10.1016/j.neucom.2020.07.139.

[3] A. Zareian, S. Karaman, and S. F. Chang, "Bridging Knowledge Graphs to Generate Scene Graphs," July 18, 2020. [Online]. Available: https://arxiv.org/pdf/2001.02314. [Accessed March 30, 2025].

[4] L. Giacomo, "Semantic Aware Image Search with Scene," October 17, 2022. [Online]. Available: https://thesis.unipd.it/bitstream/20.500.12608/36544/1/Loreggia_Giacomo.pdf. [Accessed March 30, 2025].

[5] X. Chang, P. Ren, P. Xu, Z. Li, X. Chen, and A. Hauptmann, "A Comprehensive Survey of Scene Graphs: Generation and Application," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 1-26, 2023, doi: 10.1109/TPAMI.2021.3137605.

[6] Z. Mohamad, H. Reza, and M. Rabiei, "A Review of Knowledge Graph Completion," *Information,* vol. 13, no. 8, 2022, doi: 10.3390/info13080396.

[7] W. H. Li, S. Yang, Y. Wang, D. Song, and X. Li, "Multi-level similarity learning for image-text retrieval," *Information Processing & Management,* vol. 58, no. 1, 2021, doi: 10.1016/j.ipm.2020.102432.

[8] T. V. T. Le and T. T. Van, "An Image Retrieval Model combining Neighbor Graph and Semantic Graph," (in Vietnamese), *Proceedings of the 15th National Conference on Fundamental and Applied Information Technology Research (FAIR)*, 2022, pp. 400-412, doi: 10.15625/vap.2022.0249.

[9] M. T. Phan and T. T. Van, "An image retrieval model combining statistical methods and knowledge graphs," (in Vietnamese), *Proceedings of the 17th National Conference on Fundamental and Applied Information Technology Research (FAIR)*, 2024, pp. 531-539.

[10] A. Radford, J. W. Kim, C. Hallacy, A.Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," *Proceedings of the 38th International Conference on Machine Learning*, 2021, pp. 8748-8763.

[11] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, Y. Choi, and J. Gao, "Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks," *Computer Vision – ECCV 2020: 16th European Conference*, 2020, pp. 121-137, doi: 10.1007/978-3-030-58577-8_8.