TẠP CHÍ KHOA HỌC
TRƯỜNG ĐẠI HỌC SƯ PHẠM TP HỒ CHÍ MINH
Tập 19, Số 3 (2022): 411-422
Website: http://journal.hcmue.edu.vn
ISSN: 2734-9918

HO CHI MINH CITY UNIVERSITY OF EDUCATION
JOURNAL OF SCIENCE
Vol. 19, No. 3 (2022): 411-422
https://doi.org/10.54607/hcmue.js.19.3.3272(2022)

Bài báo nghiên cứu

MỘT TIẾP CẬN TÌM KIẾM ẢNH THEO NGỮ NGHĨA DỰA TRÊN MẠNG NƠ-RON TÍCH CHẬP VÀ ONTOLOGY

Nguyễn Minh Hải[1*], Trần Văn Lăng[2], Văn Thế Thành[2]

[1]Trường Đại học Sư phạm Thành phố Hồ Chí Minh, Việt Nam

[2]Trường Đại học Ngoại Ngữ - Tin học Thành phố Hồ Chí Minh, Việt Nam

[*]Tác giả liên hệ: Nguyễn Minh Hải – Email: hainm@hcmue.edu.vn

Ngày nhận bài: 13-9-2021; ngày nhận bài sửa: 14-01-2022; ngày duyệt đăng: 13-3-2022

TÓM TẮT

Trích xuất ngữ nghĩa cho hình ảnh là một bài toán mang tính thời sự và được ứng dụng trong nhiều hệ thống tra cứu ngữ nghĩa khác nhau. Trong bài báo này, một tiếp cận tra cứu ngữ nghĩa hình ảnh được đề xuất dựa trên tập ảnh tương tự với ảnh đầu vào; từ đó, ngữ nghĩa của hình ảnh được tra cứu trên ontology qua tập từ vựng thị giác. Các đối tượng trên mỗi hình ảnh được trích xuất và phân lớp dựa trên mạng nơ-ron tích chập nhằm trích xuất ngữ nghĩa cho hình ảnh. Sau đó, câu lệnh SPARQL được tự động tạo ra từ các phân lớp ảnh và thực hiện truy vấn trên ontology đã được xây dựng nhằm truy xuất tập ảnh tương tự và ngữ nghĩa tương ứng. Trên cơ sở phương pháp đã đề xuất, một thực nghiệm được xây dựng và đánh giá trên các bộ ảnh Caltech-256. Kết quả thực nghiệm được so sánh với các công trình công bố gần đây trên cùng một bộ dữ liệu nhằm minh chứng tính hiệu quả của phương pháp đề xuất. Theo kết quả thực nghiệm, phương pháp tra cứu ngữ nghĩa hình ảnh trong bài báo này đã nâng độ chính xác lên 88,7% đối với bộ dữ liệu ảnh Caltech-256.

Từ khóa: phân lớp ảnh; mạng nơ-ron tích chập; truy vấn ảnh dựa trên ngữ nghĩa; ontology

1. Giới thiệu

Ngày nay, với sự phát triển vượt bậc trong việc sử dụng Internet và các thiết bị mobile, số lượng hình ảnh kĩ thuật số đã tăng lên rất nhiều. Do đó, để quản lí và tìm kiếm chính xác hình ảnh trong bộ dữ liệu ảnh khổng lồ này, cần có một hệ thống truy vấn mạnh mẽ. Hiện nay, nhiều lĩnh vực khác nhau ứng dụng hệ thống tìm kiếm ảnh trong thực tế như chẩn đoán bệnh lâm sàn trong lĩnh vực y học, truy vết tội phạm trong lĩnh vực an ninh, hay hệ thống thư viện số… Trong hệ thống truy vấn ảnh theo nội dung CBIR (Content–Based Image Retrival), trích xuất đặc trưng là một tác vụ vô cùng quan trọng. Vì các hình ảnh được so sánh với nhau theo các đặc trưng cấp thấp của chúng, chẳng hạn như màu sắc, kết cấu, hình dạng… để tìm tập ảnh tương tự, nên độ chính xác của hệ thống CBIR chủ yếu phụ thuộc vào

Cite this article as: Nguyen Minh Hai, Tran Van Lang, & Van The Thanh (2022). An approach of semantic-based image retrieval using deep neural network and ontology. *Ho Chi Minh City University of Education Journal of Science, 19*(3), 411-422.

các vectơ đặc trưng được trích xuất từ tập cơ sở dữ liệu hình ảnh. Nhiều kĩ thuật hiện đại được phát triển để nâng cao hiệu suất của CBIR, trong đó mạng nơ-ron tích chập (Convolution neural network – CNN) đã chứng tỏ tính ưu việt của nó trong các lĩnh vực như rút trích đặc trưng ảnh, phân loại và nhận dạng hình ảnh (Hiary, Saadeh, Saadeh, & Yaqub, 2018; Mete & Ensari, 2019; Mohamed, Mohammed, & Brahim, 2017).

Tuy nhiên, hệ thống CBIR chỉ tìm kiếm được các tập ảnh tương tự nhau về nội dung cấp thấp, nhưng có thể có ngữ nghĩa hoàn toàn khác nhau. Vì vậy, luôn tồn tại "khoảng cách ngữ nghĩa" (Sezavar, Farsi, & Mohamadzadeh, 2019) giữa đặc trưng cấp thấp và ngữ nghĩa cấp cao của người dùng. Việc phân tích và tìm kiếm ngữ nghĩa hình ảnh là một trong những thách thức được quan tâm và nghiên cứu trong lĩnh vực thị giác máy tính. Tìm kiếm ảnh dựa trên ontology là phương pháp hiệu quả nhằm truy xuất ngữ nghĩa cấp cao của hình ảnh thông qua các phân cấp miền, mối quan hệ giữa các miền, các khái niệm... Các ngữ nghĩa trên ontology gần gũi với ngữ nghĩa cấp cao của người dùng mà máy tính có thể hiểu và truy xuất được.

Nhiều phương pháp tìm kiếm ảnh theo ngữ nghĩa đã cho thấy độ chính xác cao trong truy vấn và trích xuất ngữ nghĩa cấp cao của hình ảnh. Trong đó, tập trung chủ yếu vào hai vấn đề chính: trích xuất đặc trưng cấp thấp hình ảnh, đồng thời liên kết với ngữ nghĩa cấp cao được truy xuất từ ontology. Mạng học sâu CNNs được sử dụng để trích xuất đặc trưng và phân lớp ảnh được nhiều nhóm nghiên cứu quan tâm. Dingding Cai và cộng sự (2017) đề xuất mạng Nơ-ron tích chập độ phân giải nhận biết (RACNNs) (Cai, Chen, Qian, & Kämäräinen, 2019). Thực nghiệm trên các bộ ảnh Stanford Cars, Caltech-UCSD Birds-200-2011, Oxford 102 Category Flower với độ chính xác của phương pháp đề xuất là 63,8% trên bộ Stanford Cars, 58,1% trên bộ Caltech-UCSD Birds-200-2011. Manjunath Jogin và nhóm cộng sự (2018) (Jogin, Madhulika, Divya, Meghana, & Apoorva, 2018), sử dụng mạng nơ-ron tích chập và kĩ thuật học sâu để sinh ra các đặc trưng một cách tự động và kết hợp nó với bộ phân lớp. Thực nghiệm được tiến hành trên bộ CIFAR-10 với độ chính xác của thuật toán phân lớp đạt 85,97%. Hạn chế của nghiên cứu này là không mã hóa được vị trí và định hướng của đối tượng vào các dự đoán của hệ thống. Busra Rumeysa Mete và cộng sự (2019) biểu diễn một hệ thống phân lớp cho dữ liệu hình ảnh với kĩ thuật Deep CNN và tăng cường dữ liệu (Mete & Ensari, 2019); Nhóm nghiên cứu đã đánh giá hệ thống phân lớp đề xuất trên hai bộ dữ liệu: Oxford-17 Flowers, and Oxford-102 Flowers với độ chính xác cao 99,8% và 98,5% bằng MLP và SVM. Nhóm nghiên cứu của Andres Mafla (2020) đề xuất kết hợp thị giác và các đặc trưng chất liệu được tổng hợp cục bộ trong việc phân lớp và tra cứu ảnh chi tiết (Mafla, Dey, Biten, Gomez, & Karatzas, 2020). Ưu điểm của giải pháp này là tận dụng thông tin dạng văn bản để trích xuất thông tin từ hình ảnh. Khai thác các tín hiệu văn bản có thể mở đường cho các mô hình thị giác máy tính toàn diện hơn (hiểu được ngữ cảnh). Mô hình đã được thử nghiệm trên 2 bộ ảnh Con-Text Dataset; Drink Bottle Dataset với độ chính xác lần lượt là 64,52% và 62,91%.

Một hướng tiếp cận khác là xây dựng hệ thống tìm kiếm ảnh dựa trên ontology. Thông qua tiếp cận này đối tượng ảnh được phân lớp bằng các phương pháp học máy và các quy tắc ngữ nghĩa, sau đó tập ảnh tương tự và ngữ nghĩa của nó sẽ được lấy ra từ Ontology (Filali, Zghal, & Martinet, 2020; Shati, khalid Ibrahim, & Hasan, 2020; Toro Icarte, Baier, Ruz, & Soto, 2017; C. Wang et al., 2020; Xu Wang, Huang, & van Harmelen, 2020). Asim và cộng sự (2019), đã thực hiện khảo sát các phương pháp truy xuất thông tin dựa trên Ontology áp dụng cho truy vấn văn bản, dữ liệu đa phương tiện (hình ảnh, video, audio) và dữ liệu đa ngôn ngữ. Nhóm tác giả đã so sánh hiệu suất với các phương pháp tiếp cận trước đó về truy vấn văn bản, dữ liệu đa phương tiện và dữ liệu đa ngôn ngữ (Asim et al., 2019). Tuy nhiên, nhóm tác giả mới đề xuất mô hình sử dụng Ontology để truy vấn đa đối tượng, chưa đề cập đến kết quả thực nghiệm cụ thể để so sánh với các công trình trước. Chao Wang và cộng sự (2020) đề xuất một khung Ontology tích hợp cho các ảnh viễn thám (Wang et al., 2020). Ontology này được mở rộng dựa trên Ontology mạng cảm biến ngữ nghĩa (SSN) trên ngôn ngữ OWL. Tuy nhiên, trong các ứng dụng mà dữ liệu đa nguồn sẽ gặp phải nhiều trở ngại về ngữ nghĩa. Xu W. và cộng sự (2020) đã cung cấp một hướng tiếp cận tương tự ngữ nghĩa dựa trên Ontology cho bài toán truy xuất tập dữ liệu ảnh y sinh bioCADDIE 2016 (Xu Wang et al., 2020). Với nghiên cứu này, nhóm tác giả đã sử dụng phương pháp MeSH để rút trích các khái niệm từ tập ảnh bioCADDIE. Để truy xuất tập ảnh tương tự này, nhóm tác giả đã sử dụng hai độ đo Wu-Palmer và Resnik để đo độ tương đồng ngữ nghĩa giữa các khái niệm.

Trong bài báo này, chúng tôi đề xuất một phương pháp kết hợp giữa mạng nơ-ron tích chập (CNN) và ontology cho bài toán tìm kiếm ảnh theo ngữ nghĩa. Các đóng góp chính của bài báo bao gồm: (1) sử dụng mạng CNN để rút trích đặc trưng và phân lớp cho cơ sở dữ liệu hình ảnh; (2) xây dựng cấu trúc ontology; (3) tự động tạo câu lệnh SPARQL từ các phân lớp ảnh và thực hiện truy vấn trên ontology đã được xây dựng để truy xuất tập ảnh tương tự và ngữ nghĩa tương ứng.

Phần còn lại của bài báo được tổ chức như sau: trong phần 2, phương pháp truy vấn ảnh theo tiếp cận ngữ nghĩa được trình bày; thực nghiệm và đánh giá kết quả của phương pháp đề xuất được mô tả trong phần 3; Phần 4 là kết luận và hướng phát triển tiếp theo.

2. Phương pháp truy vấn ảnh theo tiếp cận ngữ nghĩa

2.1. Mạng OverFeat

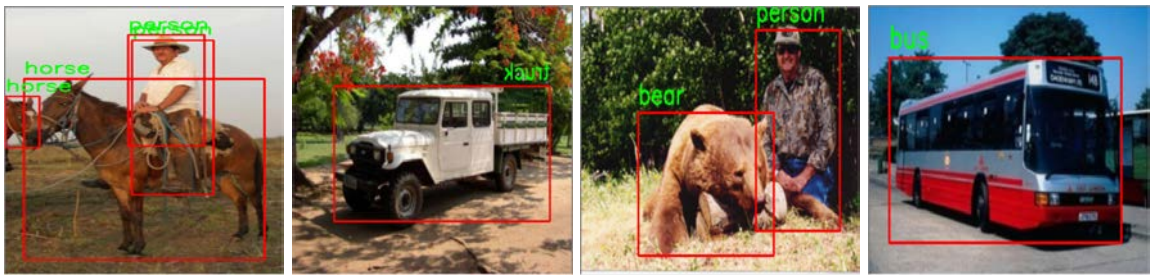
LeNet, AlexNet, GoogLeNet, VGGNet là các kiến trúc CNN phổ biến được sử dụng trong nghiên cứu học sâu hiện đại để giải quyết các vấn đề khác nhau của thị giác máy tính như phân loại hình ảnh, nhận dạng đối tượng, nhận dạng giọng nói… Học sâu được sử dụng trong việc huấn luyện một lượng lớn dữ liệu bằng GPU. Điều này là do số lần lập lại lớn trong quá trình huấn luyện dữ liệu, đặc biệt là đối với dữ liệu hình ảnh. Vì vậy, thay vì huấn luyện CNN từ đầu với số lượng lớn hình ảnh cho mỗi lớp, một phương pháp được gọi là "Học chuyển giao" được sử dụng mà trong đó mạng được huấn luyện trước trên một tập dữ liệu rất lớn (ImageNet challenge) như OverFeat, Inception-v3, Xception được coi như là

công cụ trích xuất đặc trưng bằng cách giữ lại tất cả các lớp được huấn luyện trước ngoại trừ lớp kết nối đầy đủ cuối cùng. Trong nghiên cứu này, mô hình được huấn luyện trước OverFeat được đề xuất cho việc trích xuất và phân lớp cho ảnh đầu vào nhằm tăng độ chính xác cho việc phân lớp ảnh đầu vào.

Bảng 1. *Kiến trúc mạng OverFeat*

Layer	Stage	#filters	Filter size	Conv. Stride	Pooling size	Pooling stride	Spatial input size
1	Conv + max	96	11x11	4	2	2	231x231
2	Conv + max	256	5x5	1	2	2	24x24
3	Conv	512	3x3	1	-	-	12x12
4	Conv	1024	3x3	1	-	-	12x12
5	Conv + max	1024	3x3	1	2	2	12x12
6	Full	3072	-	-	-	-	6x6
7	Full	4096	-	-	-	-	1x1
8	Full	1000	-	-	-	-	1x1

Mạng OverFeat được giới thiệu và huấn luyện bởi (Mathieu et al., 2013) trên tập dữ liệu huấn luyện ImageNet 2012 chứa 1,2 triệu hình ảnh trên 1000 phân lớp. Kiến trúc mạng được biểu thị trong Bảng 1 bao gồm 8 lớp với kích hoạt phi tuyến tính ReLU được áp dụng sau mỗi lớp tích chập và lớp kết nối đầy đủ tương ứng. Trong kiến trúc này, kích thước bộ lọc giảm giảm dần và số lượng bộ lọc bắt đầu nhỏ và sau đó được tăng lên ở các lớp cấp cao hơn của mạng. Hình 1 mô tả các kết quả nhận dạng và phân lớp các đối tượng trên bộ dữ liệu Caltech-256 bằng OverFeat.



Hình 1. *Các kết quả của OverFeat trên các ảnh trong bộ dữ liệu Caltech-256*

Trong bài báo này, mô hình mạng OverFeat được chúng tôi sử dụng nhằm phát hiện các đối tượng trong ảnh; từ đó, xác định tập phân lớp cho tập dữ liệu ảnh Caltech-256. Độ chính xác của việc phân lớp các tập dữ liệu ảnh này được so sánh với các mô hình CNN hiện đại khác được trình bày trong Bảng 2.

Bảng 2. *Độ chính xác trích xuất và phân lớp các tập dữ liệu ảnh sử dụng cấu trúc mạng OverFeat*

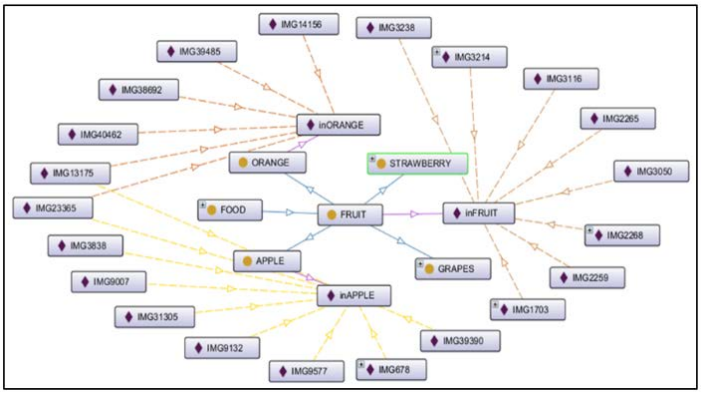
Tập dữ liệu ảnh	Mô hình	Độ chính xác
Caltech-256	Inception-v3	89,68%
	Xception	87,58%
	OverFeat	92,58%

Từ số liệu Bảng 2 cho thấy, việc sử dụng các kiến trúc CNN sâu có thể tăng độ chính xác nhận dạng và phân lớp đối tượng ảnh tăng lên khá tốt. Điều này giúp cho việc truy vấn ảnh tương tự ảnh đầu vào đạt hiệu suất cao hơn.

2.2. Ontology

Nhằm giảm khoảng cách ngữ nghĩa giữa các đặc trưng thị giác cấp thấp và ngữ nghĩa cấp cao của hình ảnh, chúng tôi xây dựng một ontology cho các bộ dữ liệu ảnh dựa trên ngôn ngữ bộ ba RDF/XML và OWL. Kết quả của truy vấn hình ảnh dựa vào cụm dữ liệu là tập các ảnh tương tự được sắp xếp theo độ chính xác. Từ tập hình ảnh này, thuật toán phân lớp k-NN được thực hiện để lấy các phân lớp láng giềng gần nhất và lưu thành các từ vựng là các phân lớp đại diện cho hình ảnh truy vấn. Các phân lớp này được truy vấn dựa trên ontology để xác định khái niệm, ngữ nghĩa cấp cao của hình ảnh. Quá trình ánh xạ ngữ nghĩa được sử dụng để phân tích và tìm khái niệm tốt nhất cho các đối tượng trong hình ảnh.

Các phân lớp của các bộ dữ liệu ảnh được xây dựng theo dạng phân cấp. Một từ điển ngữ nghĩa nhằm định nghĩa cho các phân lớp của hình ảnh được trích xuất từ WordNet. Mỗi hình ảnh là một cá thể/thể hiện (individual/ instance) của một hay nhiều phân lớp trong ontology. Hình 2 là một ví dụ về ontology được xây dựng trên Protégé cho các bộ dữ liệu ảnh Caltech-256.



Hình 2. *Một ví dụ về ontology áp dụng trên bộ dữ liệu ảnh Caltech-256*

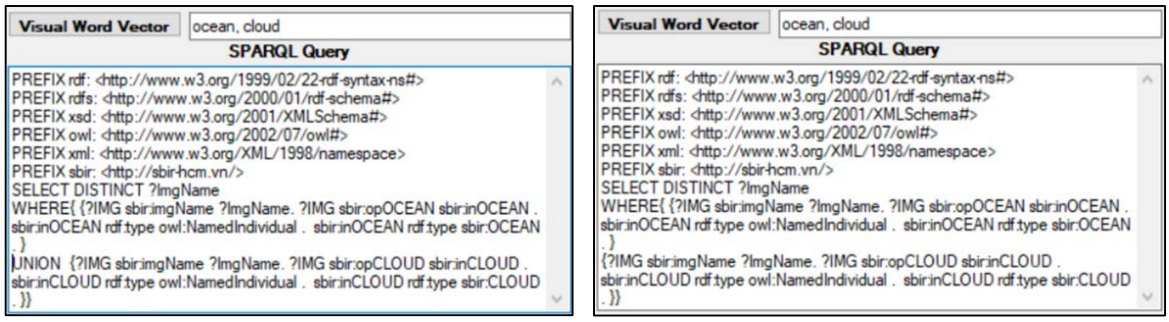
SPARQL là ngôn ngữ truy vấn trên các nguồn dữ liệu được mô tả dưới dạng bộ ba RDF hoặc OWL. Với ảnh truy vấn đầu vào có thể chứa một đối tượng hoặc nhiều đối tượng, dựa vào mạng OverFeat để tìm các phân lớp của ảnh đầu vào; từ đó, tạo vectơ từ vựng thị giác; vec-tơ này chứa một hay nhiều lớp ngữ nghĩa của ảnh truy vấn, và tự động tạo câu lệnh

SPARQL (AND hoặc OR), sau đó truy vấn trên ontology để tìm tập các ảnh tương tự và ngữ nghĩa của ảnh. Thuật toán tự động tạo câu truy vấn SPRQL được thực hiện như sau:

Thuật toán tự động tạo câu truy vấn SPARQL

Đầu vào: Vec-tơ từ thị giác W
Đầu ra: câu lệnh *SPARQL*
Begin
SPARQL=∅;
n=W.count;
SELECT DISTINCT ?Img
WHERE{";
For (i=0..n) **do**
SPARQL+="<subject>: W(i)+ "rdf:type" +
<object>: + ?Img" +"UNION/AND";
End
SPARQL+=}";
Return SPARQL;
End.

Kết quả truy vấn trên ontology là một tập các URIs và các metadata của tập dữ liệu ảnh tương tự và ngữ nghĩa của nó. Hình 3 minh họa cho câu truy vấn SPARQL được tạo ra từ vec-tơ từ thị giác theo hai cách: "UNION Query" hoặc "AND Query"

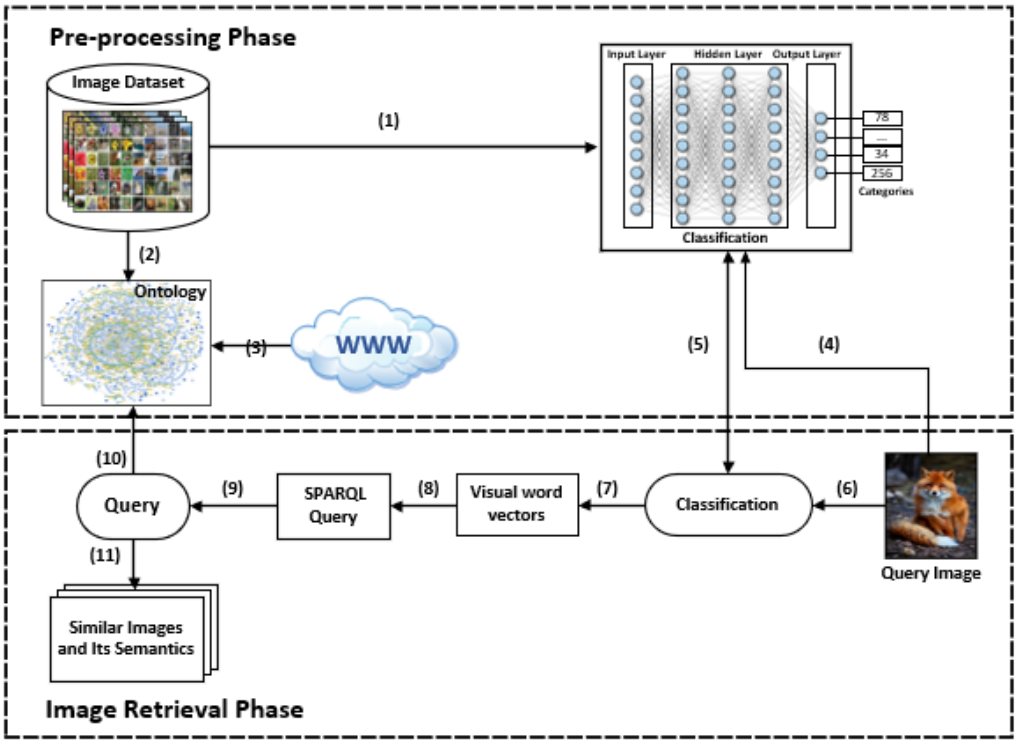


Hình 3. *Một kết quả tạo tự động câu truy vấn SPARQL "UNION" và "AND"*

3. Thực nghiệm và đánh giá

3.1. Kiến trúc hệ thống CNN_SIR

Kiến trúc hệ thống phân lớp và truy vấn tập ảnh tương tự cũng như phân tích ngữ nghĩa hình ảnh CNN_SIR được mô tả tại Hình 4. Hệ thống này thực hiện tìm kiếm tập ảnh tương tự dựa vào kết quả phân lớp ảnh các phân cụm hình ảnh, từ đó phân tích ngữ nghĩa và truy vấn hình ảnh dựa vào ontology.



Hình 4. *Hệ truy vấn CNN_SIR bao gồm hai pha*

- **Pha 1. Tiền xử lí**

Bước 1. Tập dữ liệu ảnh được huấn luyện qua mô hình mạng CNN đã huấn luyện, kết quả thu được tập phân lớp của ảnh đầu vào;

Bước 2. Xây dựng Ontology từ tập dữ liệu ảnh (2) và Word Wide Web (3) dựa vào ngôn ngữ bộ ba RDF/XML.

- **Pha 2. Truy vấn ảnh**

Bước 1. Với mỗi ảnh truy vấn (4), hệ thống thực hiện trích xuất đặc trưng và phân lớp ảnh thông qua mạng CNN đã huấn luyện ở pha tiền xửlí (5). Kết quả thu được là tập phân lớp của ảnh đầu vào (6). Mỗi phân lớp ảnh này sẽ tương ứng với một véc-tơ từ thị giác (7);

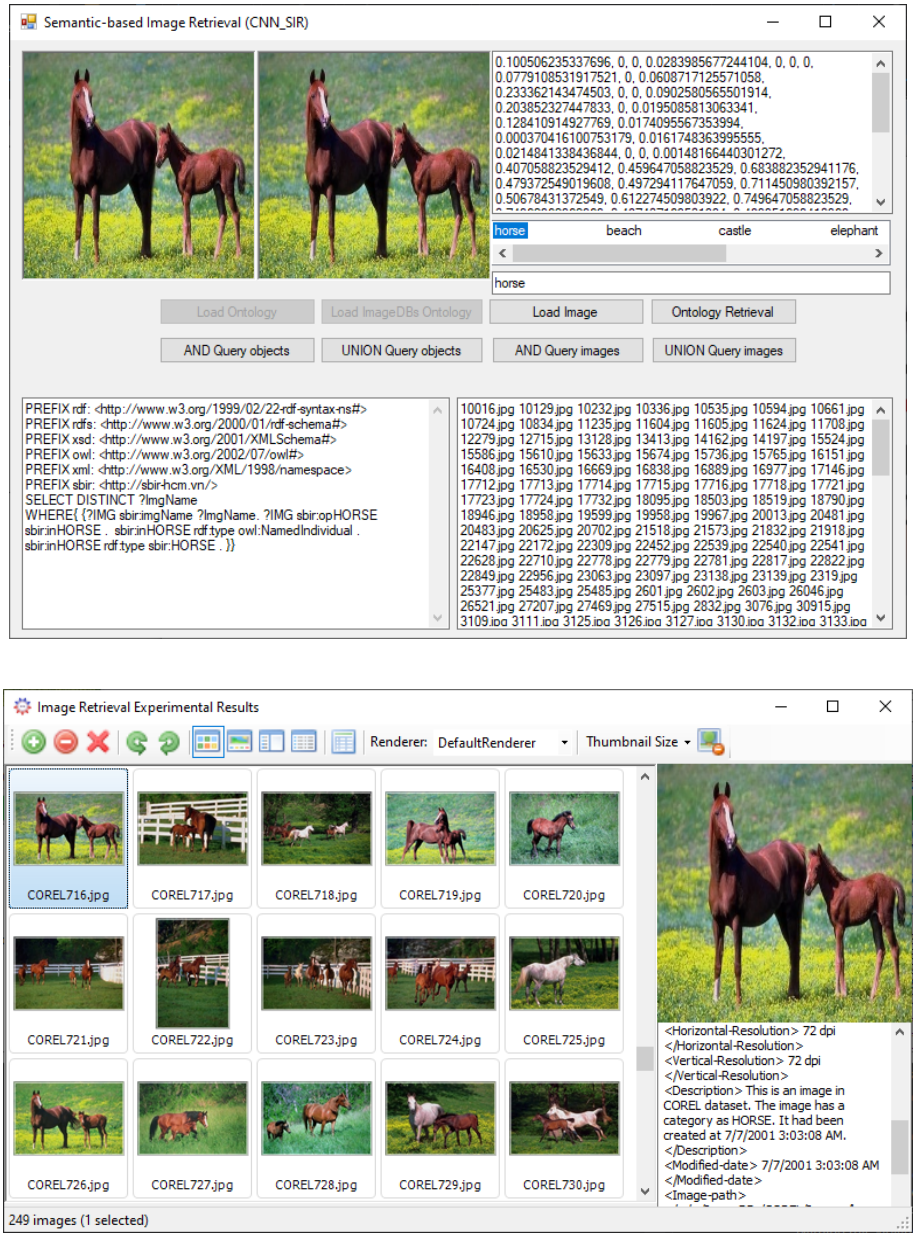
Bước 2. Tự động tạo câu truy vấn SPARQL dựa vào vec-tơ từ thị giác (8) để truy vấn ngữ nghĩa cho hình ảnh trên ontology (9). Kết quả của quá trình truy vấn trên ontology là các URIs, metadata của hình ảnh (10) và tập các hình ảnh tương tự cùng ngữ nghĩa của nó (11).

3.2. Môi trường thử nghiệm

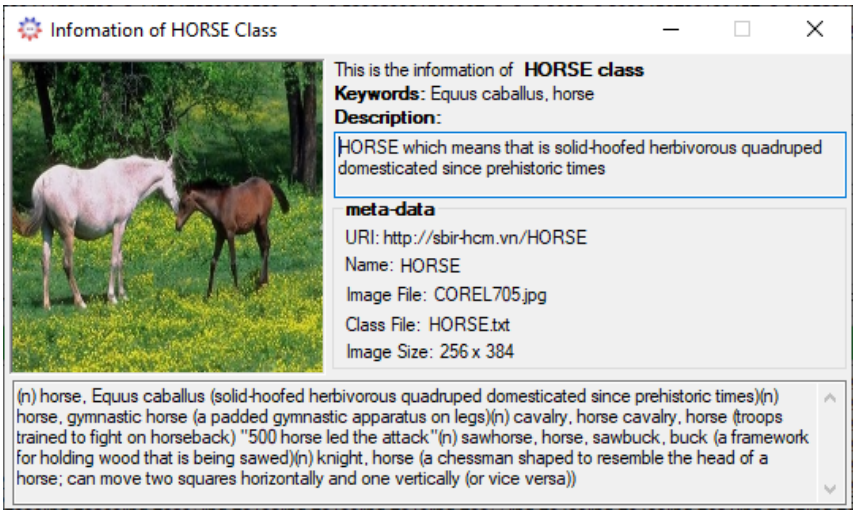
Hệ truy vấn CNN_SIR được xây dựng nhằm truy vấn hình ảnh dựa trên cây CNN và ontology, dựa trên nền tảng dotNET Framework 4.8, ngôn ngữ lập trình C#. Các đồ thị được xây dựng trên Mathlab 2015. Cấu hình máy tính của thực nghiệm: Intel(R) CoreTM i9-9200H, CPU 4,20GHz, RAM 16GB và hệ điều hành Windows 10 Professional. Tập dữ liệu được sử dụng trong thực nghiệm là bộ dữ liệu ảnh Caltech-256 với 30,000 ảnh và 256 phân lớp.

3.3. Ứng dụng

Với một ảnh đầu vào, hệ thống CNN_SIR trích xuất đặc trưng và phân lớp ảnh bằng CNN. Hình 5 một kết quả tìm kiếm hình ảnh theo ngữ nghĩa của hệ CNN_SIR. Từ phân lớp của ảnh đầu vào, câu truy vấn SPARQL (UNION hoặc AND) cũng được tự động tạo ra để truy vấn trên ontology. Hình 6 là một kết quả của sự phân lớp và truy vấn theo SPARQL của hệ truy vấn CNN_SIR. Với mỗi hình ảnh trong tập ảnh tương tự sẽ được mô tả ngữ nghĩa với các meta-data cho chú thích hình ảnh, định danh URI của hình ảnh.



Hình 5. *Một kết quả tìm kiếm hình ảnh theo ngữ nghĩa của hệ truy vấn CNN_SIR*



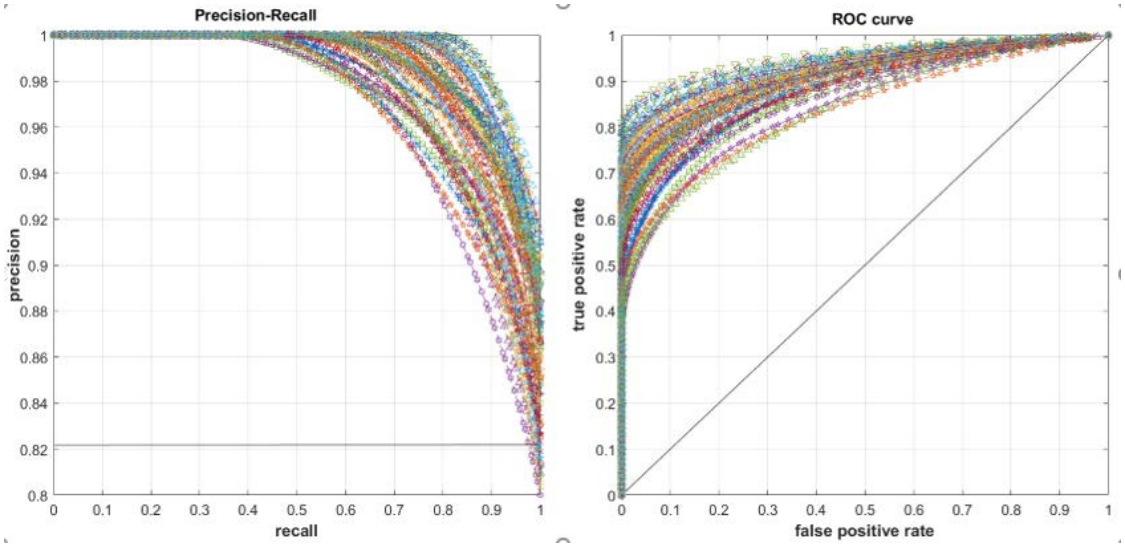
Hình 6. *Một kết quả tìm kiếm hình ảnh theo ngữ nghĩa của hệ truy vấn CNN_SIR*

3.4. Đánh giá thực nghiệm

Để đánh giá hiệu quả tìm kiếm hình ảnh, bài viết sử dụng các yếu tố để đánh giá, bao gồm: precision, recall và F-measure, thời gian truy vấn (milli seconds).

Trên cơ sở giá trị hiệu suất đã có, các giá trị hiệu suất và thời gian tìm kiếm trung bình của tập dữ liệu Caltech-256 của hệ CNN_SIR với độ chính xác 88,7%, độ phủ 84,98% và thời gian truy vấn trung bình là 966,884 ms.

Dựa trên các số liệu thực nghiệm, Hình 7 mô tả các đồ thị Precision-Recall curve và ROC curve được thực hiện để đánh giá độ chính xác của hệ truy vấn ảnh CNN_SIR, mỗi đường cong mô tả một bộ dữ liệu được truy vấn; diện tích dưới các đường cong này cho thấy độ chính xác của việc truy vấn ảnh. Hiệu suất truy vấn hình ảnh của CNN_SIR trên tập ảnh Caltech-256 cho thấy phương pháp cải tiến được đề xuất trong bài báo là hiệu quả.



Hình 7. *Hiệu suất truy vấn ảnh trên tập dữ liệu ảnh Caltech-256 của hệ CNN_SIR*

Để đánh giá độ chính xác và hiệu quả của hệ truy vấn ảnh CNN_SIR, kết quả thực nghiệm được so sánh với các công trình nghiên cứu khác trên cùng một bộ dữ liệu ảnh. Giá trị trung bình độ chính xác của hệ CNN_SIR được so sánh với các phương pháp khác trên cùng mỗi bộ dữ liệu được mô tả trong Bảng 3 cho thấy kết quả truy vấn của phương pháp đề xuất tương đối chính xác so với các hệ truy vấn ảnh theo ngữ nghĩa.

Bảng 3. *So sánh độ chính xác giữa các phương pháp trên bộ dữ liệu Caltech-256*

Phương pháp	Mean Average Precision (MAP)
Feature fusion + BoW (Xinzhi Wang, Zou, Bakker, & Wu, 2020)	74,22%
LDA_DCT with Scaling 32x32 DWT (0.5) (Sivakumar & Soundar, 2021)	81,8%
Deep Feature Selection Network (DFS-Net) with InceptionV3 (Kumar, Tripathi, & Pant, 2020)	83,91%
CNN_SIR	**88,7%**

Qua số liệu của các Bảng 3 cho thấy, phương pháp đề xuất của có độ chính xác cao hơn khi so sánh với các phương pháp truy vấn khác trên cùng tập ảnh. Kết quả này chứng minh rằng, phương pháp đề xuất của chúng tôi là hiệu quả.

4. Kết luận

Trong bài báo này, một phương pháp tìm kiếm ảnh theo ngữ nghĩa được đề xuất với sự kết hợp của mạng nơ-ron tích chập (CNN) và ontology. Với mỗi hình ảnh đầu vào, đặc trưng được trích xuất và tìm kiếm, sau đó phân lớp trên mạng CNN để tạo thành tập từ vựng thị giác. Từ đó, câu lệnh SPARQL được tự động tạo ra từ các từ vựng thị giác và thực hiện truy vấn trên ontology nhằm truy xuất tập ảnh tương tự và ngữ nghĩa tương ứng. Một mô hình truy vấn ảnh dựa trên mạng CNN và ontology (CNN_SIR) được đề xuất và thực nghiệm trên bộ ảnh Caltech-256 với độ chính xác là 88,7%. Kết quả thực nghiệm được so sánh với các nghiên cứu khác trên cùng một tập ảnh, cho thấy, phương pháp của đề xuất của chúng tôi có độ chính xác cao hơn. Trong định hướng nghiên cứu tương lai, chúng tôi tiếp tục cải tiến các phương pháp trích xuất đặc trưng, bổ sung, làm giàu cho ontology và xây dựng một hệ tìm kiếm ảnh dựa trên ontology trên WWW.

❖ ***Tuyên bố về quyền lợi:*** *Các tác giả xác nhận hoàn toàn không có xung đột về quyền lợi.*

❖ ***Lời cảm ơn:*** *Nhóm tác giả chân thành cảm ơn Trường Đại học Sư phạm Thành phố Hồ Chí Minh đã bảo trợ cho nghiên cứu này. Trân trọng cảm ơn nhóm nghiên cứu SBIR-HCM đã hỗ trợ về chuyên môn để nhóm tác giả hoàn thành nghiên cứu này.*

TÀI LIỆU THAM KHẢO

Asim, M. N., Wasim, M., Khan, M. U. G., Mahmood, N., & Mahmood, W. (2019). The use of ontology in retrieval: a study on textual, multilingual, and multimedia retrieval. *IEEE Access, 7*, 21662-21686.

Cai, D., Chen, K., Qian, Y., & Kämäräinen, J. K. (2019). Convolutional low-resolution fine-grained classification. *Pattern Recognition Letters, 119*, 166-171.

Filali, J., Zghal, H. B., & Martinet, J. (2020). Ontology-based image classification and annotation. *International Journal of Pattern Recognition and Artificial Intelligence, 34*(11), 2040002.

Hiary, H., Saadeh, H., Saadeh, M., & Yaqub, M. (2018). Flower classification using deep convolutional neural networks. *IET Computer Vision, 12*(6), 855-862.

Jogin, M., Madhulika, M., Divya, G., Meghana, R., & Apoorva, S. (2018). *Feature extraction using convolution neural networks (CNN) and deep learning.* Paper presented at the 2018 3rd IEEE international conference on recent trends in electronics, information & communication technology (RTEICT).

Kumar, V., Tripathi, V., & Pant, B. (2020). *Content based fine-grained image retrieval using convolutional neural network.* Paper presented at the 2020 7th International Conference on Signal Processing and Integrated Networks (SPIN).

Mafla, A., Dey, S., Biten, A. F., Gomez, L., & Karatzas, D. (2020). *Fine-grained image classification and retrieval by combining visual and locally pooled textual features.* Paper presented at the Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision.

Mathieu, M., LeCun, Y., Fergus, R., Eigen, D., Sermanet, P., & Zhang, X. (2013). Overfeat: Integrated recognition, localization and detection using convolutional networks.

Mete, B. R., & Ensari, T. (2019). *Flower Classification with Deep CNN and Machine Learning Algorithms.* Paper presented at the 2019 3rd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT).

Mohamed, O., Mohammed, O., & Brahim, A. (2017). *Content-based image retrieval using convolutional neural networks.* Paper presented at the First International Conference on Real Time Intelligent Systems.

Sezavar, A., Farsi, H., & Mohamadzadeh, S. (2019). Content-based image retrieval by combining convolutional neural networks and sparse representation. *Multimedia Tools and Applications, 78*(15), 20895-20912.

Shati, N. M., khalid Ibrahim, N., & Hasan, T. M. (2020). A review of image retrieval based on ontology model. *Journal of Al-Qadisiyah for computer science and mathematics, 12*(1), 10-14, 10-14.

Sivakumar, R., & Soundar, K. R. (2021). A novel generative adversarial block truncation coding schemes for high rated image compression on E-learning resource environment. *Materials Today: Proceedings*.

Toro Icarte, R., Baier, J. A., Ruz, C., & Soto, A. (2017). How a General-Purpose Commonsense Ontology can Improve Performance of Learning-Based Image Retrieval. *arXiv e-prints*, arXiv: 1705.08844.

Wang, C., Zhuo, X., Li, P., Chen, N., Wang, W., & Chen, Z. (2020). An Ontology-Based Framework for Integrating Remote Sensing Imagery, Image Products, and In Situ Observations. *Journal of Sensors, 2020*.

Wang, X., Huang, Z., & van Harmelen, F. (2020). *Ontology-Based Semantic Similarity Approach for Biomedical Dataset Retrieval*. Paper presented at the International Conference on Health Information Science.

Wang, X., Zou, X., Bakker, E. M., & Wu, S. (2020). Self-constraining and attention-based hashing network for bit-scalable cross-modal retrieval. *Neurocomputing, 400*, 255-271.

AN APPROACH OF SEMANTIC-BASED IMAGE RETRIEVAL USING DEEP NEURAL NETWORK AND ONTOLOGY

Nguyen Minh Hai[1*], Tran Van Lang[2] Van The Thanh[2]

[1]Ho Chi Minh City University of Education, Vietnam

[2] Ho Chi Minh University of Food Industry, Vietnam

[*]Corresponding author: Nguyen Minh Hai – Email: hainm@hcmue.edu.vn

Received: September 13, 2021; Revised: January 14, 2022; Accepted: March 03, 2022

ABSTRACT

Semantic extraction for images is a computational problem and is applied in many different semantic retrieval systems. In this paper, a semantic-based image retrieval approach is proposed based on images similar to the input image; since then, the semantic of the image is retrieved on the ontology through the set of visual words. The objects on each image are extracted and classified based on the CNN network to extract semantics for the image. Then, the command of SPARQL is automatically generated from the visual words of the image and executes the query on the built ontology for extracting corresponding semantics. The proposed base method, an experiment was built and evaluated on the Caltech-256 dataset. Experimental results are compared with recently published work on the same dataset results to demonstrate the effectiveness of the proposed method. According to the experimental results, the image semantic lookup method in this paper has increased the accuracy to 0.88712 for the Caltech-256 dataset.

Keywords: classification; CNN; Semantic-based Image Retrieval; ontology