Kỷ yếu Hội nghị Khoa học công nghệ Quốc gia lần thứ XVII về Nghiên cứu cơ bản và ứng dụng Công Nghệ thông tin (FAIR); Hà Nội, ngày 08-09/8/2024
DOI: 10.15625/vap.2024.0253

NHẬN DẠNG MẶT HÀNG BẰNG HỌC SÂU VỚI TRI THỨC NGỮ NGHĨA

Lê Phước Lộc[1], Trần Công Hùng[2], Đào Văn Tuyết[2], Ngô Hoàng Huy[3], Chung Tấn Lâm[2], Phạm Đình Phong[4], Nguyễn Văn Quyền[5], Nguyễn Thành Ý[6]

[1]Viện Cơ học và Tin học ứng dụng, Viện Hàn lâm Khoa học và Công nghệ Việt Nam

[2] Trường Đại học Quốc tế Sài Gòn

[3]Trường Đại học CMC

[4]Trường Đại học Giao thông vận tải Hà Nội

[5]Trường Đại học Hải Phòng

[6]Trường Đại học FPT, Hòa Lạc Campus

phuocloc@iami.vast.vn, tranconghung@siu.edu.vn, daovantuyet@siu.edu.vn, nhhuy@cmc-edu,vn
chungtanlam@siu.edu.vn, dinhphongpham@gmail.com, quyennv@dhhp.edu.vn, ynt4@fe.edu.vn

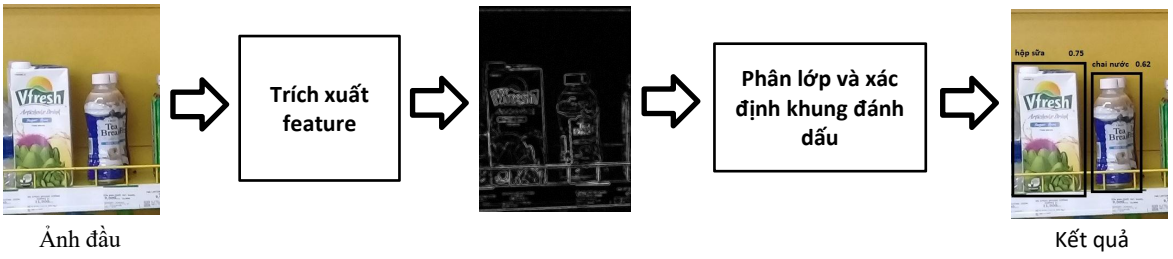
TÓM TẮT: *Thông qua bài toán nhận dạng các mặt hàng trong siêu thị, bài báo này trình bày kết quả việc áp dụng tri thức ngữ nghĩa vào phương pháp phát hiện đối tượng dựa trên kỹ thuật học sâu. Nghiên cứu được tiến hành trên một tập dữ liệu nhỏ gồm hình ảnh các mặt hàng nước giải khát trên các kệ hàng trong một số siêu thị. Bài báo sử dụng phương pháp phát hiện mối tương quan về ngữ nghĩa của các đối tượng và nhóm đối tượng mặt hàng thông qua mô hình thống kê tần suất xuất hiện và áp dụng nó vào hàm mục tiêu của mạng học sâu Faster R-CNN.*

Từ khóa: *Object detection, semantic consistency, frequency-based knowledge.*

I. GIỚI THIỆU

Bài toán phát hiện đối tượng (object detection) là một ứng dụng của kỹ thuật thị giác máy tính, trong đó dữ liệu hình ảnh đầu vào sẽ được xử lý để dự đoán xem có sự xuất hiện của các đối tượng đã biết hay không, đồng thời cho biết vị trí của chúng thông qua khung đánh dấu (bounding box). Bên trên mỗi khung đánh dấu của đối tượng phát hiện được có thể ghi chú lớp đối tượng kèm confidence score của phán đoán đó.

Bài toán nhận dạng mặt hàng áp dụng phương pháp phát hiện đối tượng nói chung được thể hiện ở Hình 1. Đầu vào của hệ thống là ảnh đơn hoặc chuỗi ảnh lấy từ camera, tùy vào mỗi hệ thống ảnh đầu vào có thể là ảnh màu, ảnh mức xám hay các loại ảnh đa phổ (multi-spectral) như ảnh kèm độ sâu, ảnh hồng ngoại,… Sau quá trình pre-processing, hệ thống sẽ tiến hành rút trích các feature của ảnh đầu vào, sau đó đưa vào bộ phân lớp để xác định loại đối tượng cũng như tìm ra vị trí của nó trong ảnh. Sau quá trình hậu xử lý, kết quả đầu ra được thể hiện bằng cách đánh dấu trên ảnh gốc vị trí và gán nhãn đối tượng đã phát hiện được.



Hình 1. Sơ đồ bài toán nhận dạng mặt hàng thông qua phương pháp phát hiện đối tượng

Hiện nay các phương pháp phát hiện đối tượng phổ biến đều sử dụng kỹ thuật học sâu (deep learning) thông qua mạng nơ-ron tích chập CNN (Convolutional Neural Network) nhờ hiệu năng vượt trội so với các phương pháp máy học cũ, tiêu biểu như Faster R-CNN [1], YOLO [2], DETR [3]. Tuy nhiên các phương pháp hiện nay đều thực hiện thuật toán dựa trên phân tích feature của ảnh mà không xét đến các thông tin ngữ nghĩa. Những thông tin này rất có ích để giúp phát hiện các đối tượng liên quan trong ảnh. Cụ thể bằng mắt, ta có thể quan sát thấy một số loại đối tượng thường xuyên xuất hiện cùng nhau nhiều hơn các loại khác. Chẳng hạn xe hai bánh chạy trên đường phải luôn có người điều khiển nó, bởi thế khi phát hiện đối tượng xe hai bánh chạy trên đường thì đối tượng kèm theo khả năng cao sẽ là người.

Bài báo này trình bày một phương pháp nhận dạng một số nhãn hàng nước giải khát được bày bán trên các kệ hàng trong một số siêu thị bằng cách áp dụng mô hình ngữ nghĩa ở đầu ra của mạng học sâu trên tập dữ liệu nhỏ gồm hình ảnh các mặt hàng trên kệ siêu thị. Mô hình ngữ nghĩa được chọn là thống kê tần suất xuất hiện của các đối tượng hay nhóm đối tượng liên quan trong tập dữ liệu. Phương pháp có thể áp dụng với đầu ra của một mạng học sâu bất kỳ, cụ thể chúng tôi thử nghiệm với mạng Faster R-CNN. Kết quả đạt được sẽ được so sánh, đánh giá với giải pháp chưa áp dụng mô hình ngữ nghĩa.

II. CÔNG TRÌNH LIÊN QUAN

Thành phần cơ bản của tất cả các phương pháp học sâu phát hiện đối tượng là mạng xương sống (backbone) CNN, được dùng để rút trích feature của ảnh để tạo ra ma trận các tham số biểu diễn mô hình tính toán. Trên cơ sở đó, mỗi phương pháp phát hiện đối tượng sẽ bổ sung thêm hay sửa đổi các tầng thích hợp, áp dụng các mô hình tính toán khác nhau để tăng hiệu năng hay cải thiện độ chính xác của phương pháp. Chẳng hạn mô hình phổ biến hiện nay là Faster R-CNN [1] bổ sung mạng đề xuất vùng RPN (Region Proposal Network) để rà soát và chọn lọc ra các khung đối tượng tiềm năng trên ảnh, sau đó mới ra soát ảnh theo các khung đó để đưa vào mạng CNN phân lớp. Một họ phương pháp khác đang được dùng rộng rãi là YOLO [2], phương pháp này trải qua nhiều phiên bản khác nhau, gồm cả phiên bản chính thức và nhiều bản không chính thức. YOLO không tiến hành rà soát ảnh theo 2 giai đoạn như Faster R-CNN mà chỉ tiến hành rà soát một lượt theo nhiều khung mẫu tại cùng vị trí (anchor box). Gần đây đang phổ biến mô hình biến đổi (transformer) có xuất xứ từ phương pháp học sâu trên mô hình ngôn ngữ, được áp dụng cho bài toán phát hiện đối tượng với đại diện là phương pháp DETR [3]. Phương pháp này tạo dựng mạng xương sống CNN có tích hợp bộ biến đổi mã hóa và giải mã giúp giảm số chiều của CNN.

Mô hình ngữ nghĩa áp dụng cho bài toán phát hiện đối tượng cũng mới xuất hiện gần đây, chẳng hạn như của nhóm tác giả Liu (2021) [4] về việc dùng mạng tương tự (similarity network) để tính mức tương đồng ngữ nghĩa giữa các cặp đối tượng cho trước. Một công trình khác của Nies (2023) [5] thì tính độ nhất quán ngữ nghĩa (semantic consistency) giữa hai đối tượng bất kỳ dựa vào việc thống kê tần suất xuất hiện đồng thời của chúng trong tập dữ liệu huấn luyện. Bài báo của chúng tôi sẽ phát triển dựa theo ý tưởng này.

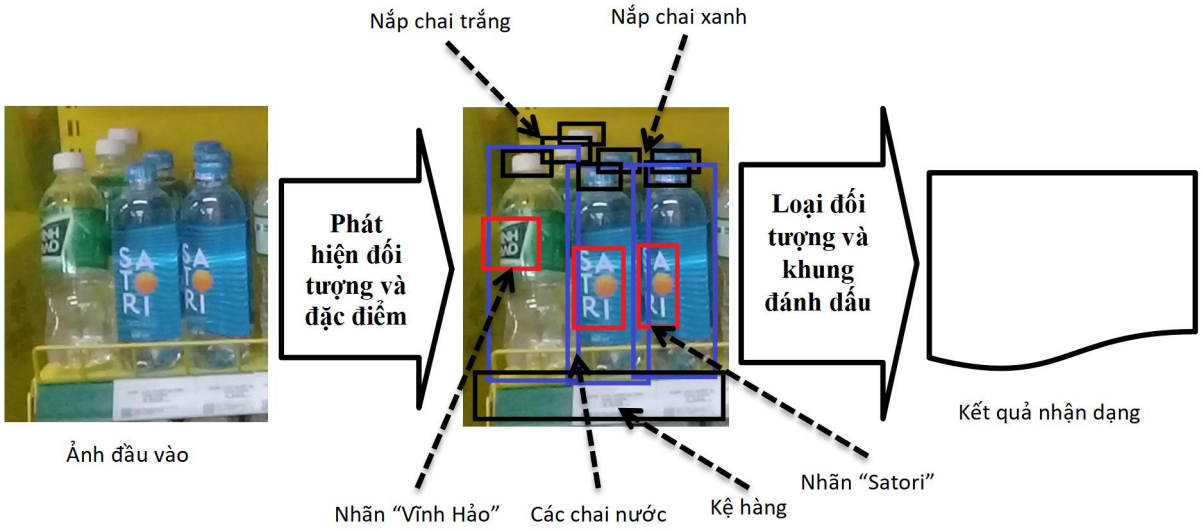
Ngoài ra còn có các phương pháp như xây dựng knowledge graph mô tả quan hệ ngữ nghĩa để thiết lập trọng số quan hệ giữa các đối tượng [6], phương pháp xây dựng knowledge graph đa mô hình gồm hình ảnh và văn bản được đưa vào chung một cấu trúc [7], phương pháp tích hợp thông tin ngữ nghĩa và vị trí không gian của đối tượng để tăng độ chính xác [8], phương pháp vận dụng thông tin metadata trong ảnh gồm vị trí và thời điểm chụp ảnh để đánh giá xác suất xuất hiện đồng thời của các đối tượng [9].

Các thử nghiệm trong bài báo này được áp dụng trên tập dữ liệu chúng tôi đã tạo ra ở bài báo trước [10].

III. BÀI TOÁN NHẬN DẠNG MẶT HÀNG

A. Mô tả

Bài toán nhận dạng nhãn hàng trên kệ hàng trong các siêu thị phục vụ cho nhu cầu kiểm tra, thống kê các mặt hàng nước giải khát đang được bày bán trên kệ. Dữ liệu đầu vào là các ảnh chụp từ những kệ hàng, sau khi xử lý sẽ trả về mỗi loại sản phẩm kèm số lượng ước tính trên mỗi tầng của kệ hàng. Bài toán sẽ được xử lý qua hai giai đoạn bao gồm: phát hiện các đối tượng và đặc tính liên quan; nhận dạng sản phẩm và ước đoán số lượng trên mỗi kệ. Hình 2 trình bày ví dụ minh họa quy trình này. Ảnh đầu vào được đưa vào bộ phát hiện đối tượng và đặc tính sẽ trả về các khung đánh dấu những loại sản phẩm, nhãn hàng, hình ảnh trên đó.



Hình 2. Quy trình nhận dạng mặt hàng trên kệ

B. Tập dữ liệu

Tập dữ liệu *sieuthi* bao gồm các ảnh chụp các kệ hàng nước giải khát từ nhiều siêu thị khác nhau. Sau đó chọn lọc lại và xử lý phân tách thành các ảnh nhỏ hơn kích thước khoảng 350x350 đến 500x500 nhằm tăng số lượng mẫu huấn luyện và kiểm tra. Ngoài ra, kích thước này cũng phù hợp khi huấn luyện trên mạng xương sống CNN.

Số lượng ảnh trong tập dữ liệu *sieuthi* được phân bổ như sau:

- Tập huấn luyện (training set): 1103 ảnh + chú giải (annotation).
- Tập kiểm tra (validation set): 271 ảnh + chú giải (annotation).
- Tập thử nghiệm (test set): Thực tế tập kiểm tra sẽ bị bỏ bớt ¼ ngẫu nhiên để làm dùng làm tập thử nghiệm và đánh giá độ chính xác của phương pháp.

Bảng 1. Danh mục các lớp đối tượng và số lượng ảnh có chứa đối tượng tương ứng

STT	Tên lớp	Mô tả	Số ảnh tập huấn luyện	Số ảnh tập kiểm tra
1	Kehang	Một tầng của kệ trưng bày hàng	797	194
2	Hop	Nước giải khát loại hộp giấy	38	7
3	Thung	Thùng các-tong đựng sản phẩm	57	14
4	Chai	Nước giải khát đóng chai	194	46
5	Chai-cam-nho	Chai nước cam loại nhỏ	16	4
6	Chai-cam-lon	Chai nước cam loại lớn	22	4
7	Chai-coca-nho	Chai coca loại nhỏ	40	7
8	Chai-coca-lon	Chai coca loại lớn	51	15
9	Chai-nuoc-nho	Chai nước lọc loại nhỏ	40	13
10	Chai-nuoc-lon	Chai nước lọc loại lớn	25	4
11	Chai-soda-nho	Chai soda loại nhỏ	16	9
12	Chai-soda-lon	Chai soda loại lớn	18	11
13	Chai-sua-nho	Chai sữa loại nhỏ	13	7
14	Chai-sua-lon	Chai sữa loại lớn	21	5
15	Binh-5L-nuoc	Bình nước lọc 5L	2	1
16	Binh-5L-dau	Bình dầu 5L	4	1
17	Ly	Ly thủy tinh	2	1
18	Loc-hopsua	Lốc các hộp sữa	15	5
19	Loc-lon	Lốc các lon (bia/nước ngọt)	95	20
20	Loc-chai	Lốc các chai	107	27
21	Loc-voboc	Lốc các sản phẩm được bọc kín	70	21
22	Lon	Lon (bia/nước ngọt)	261	57
23	Lon-lon	Lon (bia/nước ngọt) loại lớn	11	4
24	Nap-chai-den	Nắp chai màu đen	27	9
25	Nap-chai-do	Nắp chai màu đỏ	75	18
26	Nap-chai-trang	Nắp chai màu trắng	275	71
27	Nap-chai-vang	Nắp chai màu vàng	59	10
28	Nap-chai-xanh	Nắp chai màu xanh dương	33	8
29	Nap-chai-xanhla	Nắp chai màu xanh lá	83	28
30	Nap-lon	Nắp chai lớn	127	28
31	Label-do	Nhãn dán màu đỏ	57	15
32	Label-xanh	Nhãn dán màu xanh dương	40	8
33	Label-xanhla	Nhãn dán màu xanh lá	41	17
34	Brand-333	Nhãn hàng bia 333	15	3
35	Brand-7up	Nhãn hàng 7up	44	16
36	Brand-Cocacola	Nhãn hàng Coca Cola	73	18
37	Brand-Dasani	Nhãn hàng Dasani	2	2
38	Brand-Fanta	Nhãn hàng Fanta	35	9
39	Brand-Heineken	Nhãn hàng bia Heineken	22	7
40	Brand-Larue	Nhãn hàng bia Larue	16	4
41	Brand-Mirinda	Nhãn hàng Mirinda	45	10
42	Brand-NutriBoost	Nhãn hàng NutriBoost	25	5
43	Brand-Pepsi	Nhãn hàng Pepsi	73	19
44	Brand-Sprite	Nhãn hàng Sprite	24	12
45	Brand-Tiger	Nhãn hàng bia Tiger	46	15
46	Brand-Crystal	Nhãn hàng Crystal	25	8
47	Brand-Vfresh	Nhãn hàng Vfresh	13	5
48	Brand-Vinhhao	Nhãn hàng Vĩnh Hảo	9	2
49	Image-atiso	Ảnh atiso trên sản phẩm	1	1
50	Image-cachua	Ảnh cà chua trên sản phẩm	5	1
51	Image-cam	Ảnh cam trên sản phẩm	18	6
52	Image-dao	Ảnh đào trên sản phẩm	17	5
53	Image-dau	Ảnh dâu trên sản phẩm	4	1
54	Image-nho	Ảnh nho trên sản phẩm	6	2
55	Image-oi	Ảnh ổi trên sản phẩm	6	4
56	Image-tao	Ảnh táo trên sản phẩm	14	1

Thông tin chú giải đi kèm mỗi ảnh được biên soạn thủ công, bao gồm tên và tọa độ, kích thước của mỗi đối tượng mà mắt người có thể chủ quan phân biệt được. Sau đó các chú giải này được chuyển sang định dạng COCO annotation (file .json) để tương thích với chương trình Faster R-CNN dựa trên bản Cafe mà chúng tôi sử dụng. Bảng 1 thống kê tất cả các lớp đối tượng trong tập dữ liệu *sieuthi* bao gồm 56 loại.

IV. PHƯƠNG PHÁP TRI THỨC NGỮ NGHĨA

Bài báo này tham khảo ý tưởng của phương pháp nhất quán ngữ nghĩa (semantic consistency) [11] mà các tác giả đã áp dụng trong lĩnh vực xe tự lái để tăng tính chính xác phát hiện đối tượng dựa vào tri thức ngữ cảnh khi tham gia giao thông. Khái niệm tri thức ngữ nghĩa gồm nhiều khía cạnh như ngữ cảnh, chủ đề,… nhưng ở đây chúng tôi chỉ khai thác ở khía cạnh mối quan hệ giữa các đối tượng trong ảnh, cụ thể là thể hiện các loại đối tượng nào thường xuất hiện đồng thời. Với hai lớp đối tượng bất kỳ, sẽ có một giá trị nhất quán ngữ nghĩa thể hiện khả năng chúng sẽ xuất hiện đồng thời trong khung cảnh. Giá trị này cũng dùng để biểu diễn khả năng xuất hiện lặp lại của chính lớp đối tượng đó trong ảnh nhiều hơn so với các lớp đối tượng khác. Phương pháp này còn gọi là quá trình tái tối ưu (re-optimization) từ kết quả của mạng học sâu để cải thiện độ chính xác dựa vào tri thức đã biết, mà cụ thể là một ma trận nhất quán ngữ nghĩa được xây dựng sẵn.

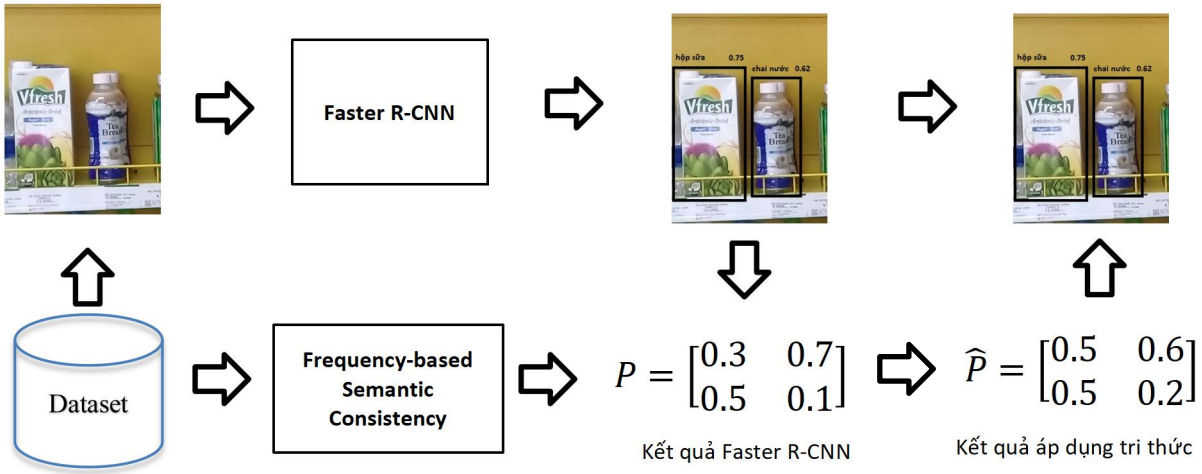
Kết quả phát hiện đối tượng từ các mạng học sâu bất kỳ sẽ kết hợp với ma trận nhất quán ngữ nghĩa để tăng lên hay giảm đi confidence score của các kết quả phát hiện được từ mạng học sâu. Để phát hiện mối quan hiện đồng xuất hiện của các đối tượng trong ảnh, có hai mô hình: thứ nhất là mô hình dựa trên tần suất xuất hiện (frequency), thứ hai là mô hình dựa trên knowledge graph. Dựa vào kết quả nghiên cứu từ tác giả [5], phương pháp knowledge graph lấy từ ConceptNet thực tế chưa phát huy hiệu quả với tập dữ liệu COCO và VOC hơn so với phương pháp dựa trên tần suất (xem Bảng 2). Nguyên nhân có thể do các tập dữ liệu này không phù hợp với knowledge graph quá khái quát như ConceptNet. Vì vậy trước mắt trong phạm vi nghiên cứu này, chúng tôi chỉ lựa chọn tiến hành thử nghiệm với mô hình nhất quán ngữ nghĩa dựa trên tần suất xuất hiện.

Bảng 2. Hiệu năng Faster R-CNN (Baseline) so với dùng Frequency (Freq), Knowledge Graph (KG), kết hợp (Hybrid) trên COCO

Phương pháp	AP @ 100				Recall @ 100			
	Baseline	KG	Hybrid	Freq	Baseline	KG	Hybrid	Freq
Person	25,29	24,92	24,42	23,32	33,51	33,78	33,94	33,48
Rider	29,76	29,40	29,73	29,75	41,31	40,26	40,18	41,25
Car	44,28	43,74	42,73	44,26	49,87	50,67	50,76	49,82
Truck	19,36	19,45	19,46	19,49	31,51	32,58	32,37	32,69
Bus	36,83	36,86	36,90	36,86	46,37	46,84	46,63	46,73
Train	12,14	12,15	11,82	12,15	23,48	23,91	23,91	23,91
Motorcycle	14,87	15,00	15,11	15,04	27,65	27,99	27,99	27,72
Bicycle	20,28	20,26	19,95	20,33	31,35	31,19	31,19	31,14

A. Mô hình tri thức ngữ nghĩa dựa trên tần suất

Đầu tiên ta tiến hành khảo sát tập dữ liệu huấn luyện (đã đánh dấu) để thống kê số lần xuất hiện của các lớp đối tượng (hay còn gọi là concept) trong mỗi ảnh và trên toàn tập dữ liệu, kết quả thu được ma trận semantic consistency S.



Hình 3. Quy trình phát hiện mặt hàng trên kệ siêu thị dùng Faster R-CNN kết hợp tri thức ngữ nghĩa dựa trên tần suất

Sau đó tiến hành nhận dạng ảnh từ tập dữ liệu thông qua mô hình Faster R-CNN thu được danh mục các lớp đối tượng đã dự đoán được L và tập các khung vị trí B của chúng trong ảnh. Xác suất (chính là confidence score) của mỗi khung đối tượng b thuộc về lớp dự đoán l được thể hiện trong ma trận P. Hình 3 thể hiện quá trình tích hợp ma trận nhất quán ngữ nghĩa theo tần suất S với ma trận confidence score P từ đầu ra của Faster R-CNN để thu được ma trận đã hiệu

chỉnh $\widehat{P}$ là kết quả phân lớp sau cùng. Quy trình này còn gọi là tái tối ưu (re-optimization) kết quả nhận dạng để ưu tiên chọn các lớp có tần suất xuất hiện chung cao với các lớp khác trong tập dữ liệu ảnh huấn luyện. Cụ thể mỗi phần tử của ma trận S thể hiện tần suất xuất hiện của lớp *l* so với lớp *l'* mô tả như công thức (1):

$$S_{l,l'} = \max\left(log\frac{n(l,l')N}{n(l)n(l')}, 0\right) \quad (1)$$

trong đó: *n(l)* và *n(l')* tương ứng là số lần xuất hiện của đối tượng lớp *l* và lớp *l'* trong tập dữ liệu ảnh, *n(l, l')* kí hiệu số lần đồng xuất hiện của đối tượng thuộc cả hai lớp đó, *N* là tổng số tất cả các đối tượng. Nếu các đối tượng lớp *l* và *l'* xuất hiện độc lập hay ít khi cùng lúc thì $S_{l,l'}$ sẽ bằng 0, nếu đồng xuất hiện càng nhiều thì $S_{l,l'}$ là số dương càng lớn.

Trường hợp nhiều đối tượng của cùng một lớp đồng thời xuất hiện thì giá trị *n(l, l)* sẽ được tính theo kiểu "bắt tay" như sau:

$$n(l,l) = n(l)\frac{n(l)-1}{2} \quad (2)$$

B. Công thức tái tối ưu theo tri thức ngữ nghĩa

Sau khi thiết lập ma trận nhất quán ngữ nghĩa *S*, ta sẽ thực hiện bước tái tối ưu (re-optimization) bằng cách tìm cực tiểu cho hàm mục tiêu (hàm loss) gồm 2 số hạng dương ở công thức (3). Số hạng đầu tiên nhằm cực tiểu phần nhất quán ngữ nghĩa nếu hai đối tượng đồng xuất hiện và số hạng thứ hai nhằm cực tiểu sai số kết quả dự đoán.

$$E(\widehat{P}) = (1-\epsilon)\sum_{b=1}^{B}\sum_{\substack{b'=1\\b'\neq b}}^{B}\sum_{l=1}^{L}\sum_{l'=1}^{L}S_{l,l'}\left(\widehat{P}_{b,l} - \widehat{P}_{b',l'}\right)^2 + \epsilon\sum_{b=1}^{B}\sum_{l=1}^{L}B\left\|S_{l,*}\right\|_1\left(\widehat{P}_{b,l} - P_{b,l}\right)^2 \quad (3)$$

trong đó: $\{\widehat{P}_{b,l}: b \in B, l \in L\}$ là xác suất kết quả từ Faster R-CNN dự đoán khung đánh dấu *b* thuộc lớp *l*. Còn $\{\widehat{P}_{b,l}: b \in B, l \in L\}$ và $\{\widehat{P}_{b',l'}: b \in B, l \in L\}$ là hai xác suất tái tối ưu thể hiện khung đánh dấu *b* có nhãn lớp là *l*, *b'* có nhãn lớp là *l'*. Nếu *l* và *l'* thường xuyên đồng xuất hiện thì $S_{l,l'}$ sẽ có giá trị lớn hơn, khi đó hai xác suất này buộc phải gần bằng nhau để cực tiểu số hạng thứ nhất.

Số hạng thứ hai ràng buộc kết quả tái tối ưu không được quá sai khác so với kết quả dự đoán từ Faster R-CNN. Hệ số $B\left\|S_{l,*}\right\|_1$ = tích số lượng khung đánh dấu B với độ lớn tổng nhất quán ngữ nghĩa của lớp *l* đối với các lớp còn lại nhằm mục đích cân bằng số phép tính tổng ở số hạng thứ hai so với số hạng thứ nhất. Tỉ lệ đánh đổi (trade-off) giữa số hạng thứ hai và số hạng thứ nhất được điều chỉnh bởi hệ số $\epsilon \in (0,1)$.

C. Phần đóng góp của bài báo

Mô hình tri thức ngữ nghĩa dựa theo tần suất được biểu diễn bởi ma trận *S* thể hiện tính đồng xuất hiện của mỗi cặp lớp đối tượng trong tập dữ liệu. Ma trận *S* được tạo ra thông qua việc thống kê số lần xuất hiện của các lớp đối tượng trên tập dữ liệu đã có. Vì vậy kết quả của phương pháp này phụ thuộc vào từng tập dữ liệu được chọn mà ít mang tính tổng quát với các dữ liệu mới. Để cải thiện vấn đề trên, chúng tôi nhận thấy rằng các lớp đối tượng cùng loại thường có giá trị S tương tự nhau, ví dụ mỗi lớp hình ảnh trái cây (cam, táo, ổi, …) trên các hộp nước ép tương ứng đều xuất hiện đồng thời cùng với lớp đối tượng hộp nước ép. Do vậy nếu ta gom các lớp hình ảnh trái cây vào một nhóm và đếm tần suất xuất hiện chung cho cả nhóm đó thì có thể dễ đoán hơn một hình ảnh khác trên hộp nước ép dù nó ít xuất hiện. Dựa trên tiêu chí này, chúng tôi gom nhóm một số lớp đối tượng trong tập dữ liệu và định nghĩa thêm ma trận *G* để thể hiện tần suất xuất hiện của mỗi nhóm đối tượng so với các nhóm đối tượng khác. Công thức (3) thành:

$$E(\widehat{P}) = (1-\epsilon)\beta\sum_{b=1}^{B}\sum_{\substack{b'=1\\b'\neq b}}^{B}\sum_{l=1}^{L}\sum_{l'=1}^{L}S_{l,l'}\left(\widehat{P}_{b,l} - \widehat{P}_{b',l'}\right)^2 + (1-\epsilon)(1-\beta)\sum_{b=1}^{B}\sum_{\substack{b'=1\\b'\neq b}}^{B}\sum_{l=1}^{L}\sum_{l'=1}^{L}G_{M_l,M_{l'}}\left(\widehat{P}_{b,l} - \widehat{P}_{b',l'}\right)^2$$
$$+ \epsilon\sum_{b=1}^{B}\sum_{l=1}^{L}B(\beta\left\|S_{l,*}\right\|_1 + (1-\beta)\left\|G_{M_l,*}\right\|_1)(\widehat{P}_{b,l} - P_{b,l})^2 \quad (4)$$

trong đó: vector *M* có kích thước *L* + 1 thể hiện chỉ số nhóm của các lớp đối tượng, với quy ước giá trị 0 khi lớp đối tượng không được gom nhóm, giá trị dương là chỉ số nhóm của lớp đối tượng tương ứng. Như vậy M_l là chỉ số nhóm của lớp đối tượng *l* và $M_{l'}$ là chỉ số nhóm của lớp đối tượng *l'*. Hệ số $\beta \in [0,1]$ là tỉ lệ đánh đổi (trade-off) giữa số hạng thứ nhất, thể hiện nhất quán ngữ nghĩa theo lớp đối tượng và số hạng thứ hai, thể hiện nhất quán ngữ nghĩa theo nhóm lớp đối tượng trong công thức (4).

D. Thực hiện tối ưu hóa

Để cực tiểu hàm mục tiêu (4), ta tính đạo hàm của nó theo $\widehat{P}_{b,l}$ như sau:

$$\frac{\partial E(\widehat{P})}{\partial \widehat{P}_{b,l}} \propto (1-\epsilon)\sum_{\substack{b'=1\\b'\neq b}}^{B}\sum_{l'=1}^{L}\left[\beta S_{l,l'} + (1-\beta)G_{M_l,M_{l'}}\right]\left(\widehat{P}_{b,l} - \widehat{P}_{b',l'}\right) + \epsilon B\left[\beta\left\|S_{l,*}\right\|_1 + (1-\beta)\left\|G_{M_l,*}\right\|_1\right]\left(\widehat{P}_{b,l} - P_{b,l}\right) \quad (5)$$

Ta tìm điểm cực trị bằng cách giải phương trình từ biểu thức (5) bằng 0, $\forall b \in B, l \in L$. Từ đó tính được $\widehat{P}_{b,l}$:

$$\widehat{P}_{b,l} = (1-\epsilon)\frac{\sum_{b'=1,\ b'\neq b}^{B'}\sum_{l'=1}^{L'}\left[\beta S_{l,l'}+(1-\beta)G_{M_l,M_{l'}}\right]\widehat{P}_{b',l'}}{\sum_{b'=1,\ b'\neq b}^{B'}\sum_{l'=1}^{L'}\left[\beta S_{l,l'}+(1-\beta)G_{M_l,M_{l'}}\right]} + \epsilon P_{b,l} \quad (6)$$

Nhận xét rằng lời giải cho (6) là tìm giới hạn của chuỗi (7) với i ∈ {1, 2,…}. Đặc biệt, với $\widehat{P_{b,l}^{(0)}}$ được khởi gán bất kỳ thì $\widehat{P_{b,l}^{(i)}}$ vẫn luôn hội tụ về cùng một lời giải khi i → ∞.

$$\widehat{P_{b,l}^{(i)}} = (1-\epsilon)\frac{\sum_{b'=1,\ b'\neq b}^{B'}\sum_{l'=1}^{L'}\left[\beta S_{l,l'}+(1-\beta)G_{M_l,M_{l'}}\right]\widehat{P_{b',l'}^{(i-1)}}}{\sum_{b'=1,\ b'\neq b}^{B'}\sum_{l'=1}^{L'}\left[\beta S_{l,l'}+(1-\beta)G_{M_l,M_{l'}}\right]} + \epsilon P_{b,l} \quad (7)$$

Mặc dù độ phức tạp của giải thuật về lý thuyết là $O(B^2L^2I)$ với I là số lần lặp, tuy nhiên ta có thể điều chỉnh về thời gian đa thức như sau. Quá trình hội tụ thường diễn ra nhanh, trong khoảng 50 đến 100 lần lặp đối với tập dữ liệu *sieuthi*. Để tiết kiệm thời gian tính toán số hạng nhất quán ngữ nghĩa, ta có thể giới hạn số lượng khung đánh dấu lại còn B_k khung gần đối tượng đang xét nhất thay vì phải tính toán tất cả các khung đánh dấu có trong ảnh. Tương tự ta cũng giới hạn số lớp đối tượng cần xét chỉ còn L_k lớp có giá trị nhất quán ngữ nghĩa cao nhất với đối tượng đang xét. Khi đó độ phức tạp thực tế chỉ còn $O(BLI)$ với giả sử B_k, L_k là các hằng số nhỏ.

V. THỬ NGHIỆM VÀ ĐÁNH GIÁ

A. Thử nghiệm

1. Tập dữ liệu

Chúng tôi tiến hành thử nghiệm trên tập dữ liệu *sieuthi* được như đã giới thiệu và trình bày cách phân hoạch ở phần trước. Tập dữ liệu được định dạng theo tiêu chuẩn của dữ liệu MS COCO để thuận tiện cho chương trình truy xuất. Trước khi áp dụng các phương pháp tri thức ngữ nghĩa, tập dữ liệu sẽ được thống kê tính toán tần suất xuất hiện của các cặp lớp đối tượng và các cặp nhóm lớp đối tượng. Sau đó lưu các ma trận nhất quán ngữ nghĩa S và G thu được vào file để bước đánh giá sử dụng.

2. Hiện thực chương trình

Chúng tôi sử dụng và hiệu chỉnh và cải tiến dựa trên mô hình học sâu Faster R-CNN. Mô hình này được cài đặt dựa trên bản Python Caffe sửa đổi để tích hợp tri thức ngữ nghĩa và được công bố từ bài báo [11][5].

Tập dữ liệu *sieuthi* được huấn luyện ban đầu từ bộ tham số pre-trained của COCO trên mạng ResNet50 để có bộ tham số kết quả đưa vào thử nghiệm. Huấn luyện sử dụng thuật toán SGD (Stochastic Gradient Descent) với momentum = 0,9, kích thước mini-batch = 1, learning rate = 0,001, learning rate decay = 0,0001, weight decay = 0,00005. Quá trình huấn luyện diễn ra trong 100 epoch với epoch decay = 6.

Quá trình thử nghiệm hiện thực công thức tái tối ưu (7) để tích hợp tri thức ngữ nghĩa từ ma trận S và G vào kết quả thử nghiệm của Faster R-CNN. Thử nghiệm diễn ra trong 50 lần lặp với các thông số cố định sau:

- Số đối tượng tối đa / ảnh = 100.
- Điểm số tối thiểu (threshold) của khung đánh dấu = 0,00005.
- B_k = 40, L_k = 15.
- Số nhận dạng tối đa (top k) = 100.

Thử nghiệm tiến hành và đánh giá 4 phương pháp sau đây theo tiêu chí mAP (độ chính xác trung bình) và Recall (độ bao phủ của các dự đoán đúng) trên cùng tập dữ liệu *sieuthi*:

- **Baseline**: Chỉ sử dụng Faster R-CNN.
- **Freq**: Faster R-CNN áp dụng tri thức ngữ nghĩa dựa trên tần suất xuất hiện các lớp đối tượng.
- **GroupFreq**: Faster R-CNN áp dụng tri thức ngữ nghĩa dựa trên tần suất xuất hiện các nhóm lớp đối tượng.
- **Hybrid**: Kết hợp 2 phương pháp Freq và GroupFreq.

Ở phương pháp nhóm lớp đối tượng (GroupFreq), chúng tôi định nghĩa các nhóm lớp sau đây:

- ORANGE = {Chai-cam-nho, Chai-cam-lon}.
- COCA = {Chai-coca-nho, Chai-coca-lon}.
- WATER = {Chai-nuoc-nho, Chai-nuoc-lon}.
- SODA = {Chai-soda-nho, Chai-soda-lon}.
- MILK = {Chai-sua-nho, Chai-sua-lon}.

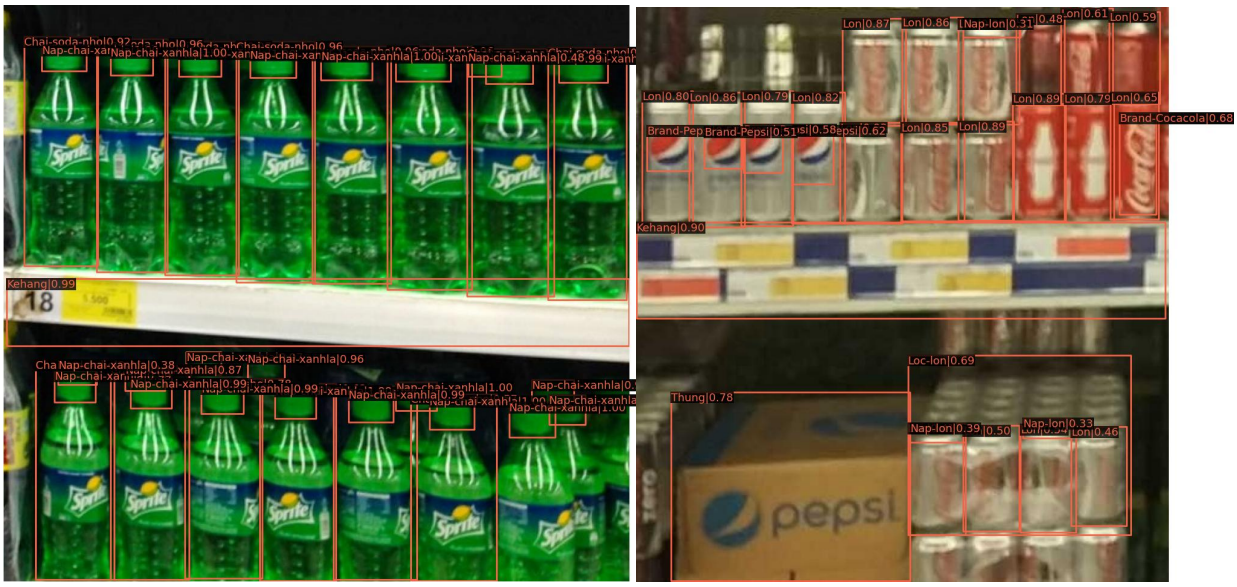
- IMAGE = {Image-atiso, Image-cachua, Image-cam, Image-dao, Image-dau, Image-nho, Image-oi, Image-tao}.
- Các nhóm còn lại là nhóm đơn, mỗi nhóm chứa duy nhất lớp chính nó.

Các trường hợp đánh giá tùy theo 2 hệ số:

- **epsilon** $\epsilon = \{0.8, 0.5, 0.2\}$.
- **beta** $\beta = \{1.0, 0.5, 0\}$.

B. Kết quả

1. Một số hình ảnh



Hình 4: Một số hình ảnh nhận dạng mặt hàng nước giải khát trên kệ siêu thị

2. Số liệu thử nghiệm

Kết quả so sánh đánh giá các chỉ số mAP và Recall của 3 phương pháp Baseline, Freq và GroupFreq theo các hệ số epsilon và beta khác nhau được trình bày ở Bảng 3. Trong đó, khi gán *epsilon* = 1,0 nghĩa là chọn giải pháp Faster R-CNN nguyên bản (Baseline). Nếu gán *beta* = 1,0 nghĩa là chọn phương pháp tích hợp tri thức ngữ nghĩa dựa trên tần suất các lớp đối tượng (Freq) vào Faster R-CNN. Nếu chọn *beta* = 0,0 nghĩa là chọn phương pháp tích hợp tri thức ngữ nghĩa dựa trên tần suất các nhóm lớp đối tượng (GroupFreq) vào Faster R-CNN. Nếu chọn *beta* = 0,5 nghĩa là chọn phương pháp kết hợp (Hybrid) cả Freq và GroupFreq cho Faster R-CNN.

Bảng 3. Kết quả so sánh Faster R-CNN trước (Baseline) và sau khi áp dụng tri thức ngữ nghĩa loại Freq, GroupFreq và kết hợp

epsilon	beta	mAP @ 100	Recall @ 100
1,0 (Baseline)		44,80	54,27
0,8	1,0 (**Freq**)	44,86	54,57
	0,5 (**Hybrid**)	44,89	54,55
	0 (GroupFreq)	44,89	54,55
0,5	1,0 (**Freq**)	44,82	54,98
	0,5 (**Hybrid**)	44,84	54,92
	0 (GroupFreq)	44,84	54,92
0,2	1,0 (**Freq**)	44,58	54,98
	0,5 (**Hybrid**)	44,70	54,93
	0 (GroupFreq)	44,70	54,93

VI. BÀN LUẬN

Căn cứ vào số liệu Bảng 3, trước hết ta nhận thấy các giải pháp tích hợp tri thức ngữ nghĩa đều cho kết quả tốt hơn khi chưa tích hợp. Tuy nhiên điều này chỉ hiệu quả với hệ số epsilon không dưới 0,5, điều này có nghĩa số hạng duy trì kết quả nhận dạng của Faster R-CNN vẫn phải là yếu tố chủ đạo để phân lớp nên cần có trọng số cao hơn số hạng nhận dạng theo tri thức ngữ nghĩa.

Trong 3 phương pháp tích hợp tri thức ngữ nghĩa - Freq, GroupFreq và Hybrid thì Freq cho kết quả Recall nhỉnh hơn, trong khi GroupFreq và phương pháp kết hợp cho kết quả mAP nhỉnh hơn. Điều này cho thấy việc tăng cường nhận

dạng theo nhóm lớp đối tượng giúp tăng khả năng nhận dạng chính xác như mong đợi. Tuy nhiên vẫn cần phải cải tiến để tránh bỏ sót các đối tượng, tăng chỉ số Recall.

Do hạn chế về kích thước tập dữ liệu nên khả năng tổng quát của giải pháp còn hạn chế, dẫn đến kết quả chưa cải thiện nhiều như mong đợi. Việc thống kê và thử nghiệm với nhiều tập dữ liệu khác sẽ hứa hẹn các kết quả thú vị.

VII. KẾT LUẬN

Thông qua bài toán nhận dạng nhãn hàng trên kệ hàng trong siêu thị, chúng tôi đã thử nghiệm và đánh giá việc áp dụng tri thức ngữ nghĩa loại thống kê theo tần suất vào một mạng nhận dạng đối tượng phổ biến hiện nay là Faster R-CNN. Kết quả cho thấy việc áp dụng giải pháp này trên tập dữ liệu do chúng tôi tạo ra hay tập dữ liệu chuẩn như MS COCO hay VOC đều cho thấy luôn tăng hiệu quả nhận dạng. Đối với phương pháp thống kê tần suất theo nhóm do chúng tôi bổ sung, mặc dù có tăng độ nhận dạng chính xác hơn nhưng vẫn cần cải tiến thêm để tăng chỉ số Recall. Việc thử nghiệm và cải tiến phương pháp nhận dạng sử dụng đồ thị tri thức vẫn còn nhiều hứa hẹn.

Bài toán nhận dạng với hỗ trợ của tri thức ngữ nghĩa còn đặt ra một số bài toán cần thiết khác như vấn đề nhận dạng kết hợp với các thông tin ghi chú, thông tin định vị. Ngoài ra vấn đề đếm hay ước tính số lượng đối tượng thông qua tri thức ngữ nghĩa cũng có nhiều tiềm năng.

TÀI LIỆU THAM KHẢO

[1] S. Ren, K. He, R. Girshick, and J. Sun., “Faster R-CNN:Towards Real-Time Object Detection with Region Proposal Networks,” in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137-1149, 2017.

[2] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," *arXiv:1506.02640* [cs], 2016.

[3] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, “End-to-End Object Detection with Transformers,” *ECCV*, pp. 213-229, Springer, 2020.

[4] J. Liu, W. Liu, Y. Cheng, S. Ge, and X. Wang, “Similarity network fusion based on random walk and relative entropy for cancer subtype prediction of multigenomic data,” *Scientific Programming*, 2021.

[5] J. F. Nies, “Semantic Information for Object Detection,” *CoRR abs/2308.08990*, 2023.

[6] D. P. Fan, L. Zheng, J. Zhao, Y. Liu, Z. Zhang, Q. Hou, M. Zhu, and M. M. Cheng, “Rethinking RGB-D Salient Object Detection: Models, Data Sets, and Large-Scale Benchmarks,” in *IEEE Transactions on Neural Networks and Learning Systems*, pp. 2075-2089, 2019.

[7] N. Kumari, B. Zhang, R. Zhang, E. Shechtman, J. Y. Zhu, “Multi-Concept Customization of Text-to-Image Diffusion,” *CVPR*, 2023.

[8] Y. Zhu, B. Sun, X. Lu, and S. Jia, “Geographic Semantic Network for Cross-View Image Geo-Localization,” in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1-15, 2022.

[9] O. M. Aodha, “Presence-Only Geographical Priors for Fine-Grained Image Classification,” *ICCV*, 2019.

[10] L. P. Lộc, Đ. V. Tuyết, N. N. M. Thúy, và N. T. Thành, “Nhận dạng một số nhãn hàng trên kệ hàng siêu thị sử dụng kỹ thuật học sâu,” *Tạp chí KH&CN Trường ĐH Bình Dương*, vol. 3, no. 3, pp. 347-364, 2021.

[11] Y. Fang, K. Kuan, J. Lin, C. Tan, and V. Chandrasekhar, “Object detection meets knowledge graphs,” in *Proceedings of the Twenty-Sixth IJCAI*, pp. 1661-1667, 2017.

IDENTIFYING ITEMS USING DEEP LEARNING WITH SEMANTIC KNOWLEDGE

Loc Le Phuoc, Hung Tran Cong, Tuyet Dao Van, Huy Ngo Hoang,
Lam Chung Tan, Phong Pham Dinh, Quyen Nguyen Van, Y Nguyen Thanh

ABSTRACT: *Through the problem of identifying items in a supermarket, this paper presents the results of applying semantic knowledge to object detection methods using deep learning. The research was conducted on a small dataset consisting of images of beverage items on shelves in several supermarkets. We use methods of discovering semantic relations of item classes and groups of item classes through the statistical model of co-occurrence frequency and applies them to optimization function of Faster R-CNN network.*