Kỷ yếu Hội nghị Khoa học công nghệ Quốc gia lần thứ XVI về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR); Đà Nẵng, ngày 28-29/9/2023
DOI: 10.15625/vap.2023.0072

HỌC BIỂU DIỄN ẢNH VỚI MẠNG NƠRON TÍCH CHẬP ĐỒ THỊ CHO TRA CỨU ẢNH

Nguyễn Văn Thanh[1], Nguyễn Hữu Quỳnh[1], Phạm Huy Hoàng[2], Đào Thị Thúy Quỳnh[2], Cù Việt Dũng[1]

[1]Trường Đại học Thủy lợi
[2]Học viện Công nghệ Bưu chính Viễn thông

nguyenvanthanh@hvtc.edu.vn, quynhnh@tlu.edu.vn, hoang.ph303@gmail.com, quynhdtt@ptit.edu.vn, dungcv@tlu.edu.vn.

TÓM TẮT: *Hiệu năng của hệ thống tra cứu ảnh dựa vào nội dung phụ thuộc chủ yếu vào hai giai đoạn: (1) học biểu diễn ảnh hiệu quả và (2) hàm khoảng cách phù hợp cho biểu diễn ảnh được học. Đã có nhiều phương pháp tra cứu ảnh dựa vào nội dung tận dụng học sâu để học biểu diễn ảnh hiệu quả, tuy nhiên các phương pháp này không tận dụng được các quan hệ giữa các ảnh. Gần đây, mạng nơron tích chập đồ thị (GCN - Graph convolutional networks) đã nổi lên là một cách tiếp cận học sâu mạnh mẽ cho dữ liệu được biểu diễn dưới dạng đồ thị. Trong bài báo này, chúng tôi trình bày phương pháp tra cứu ảnh, gọi là IRLIR (Image Representation Learning for Image Retrieval). Phương pháp đề xuất sử dụng mạng nơron tích chập đồ thị để học biểu diễn ảnh hiệu quả và giải pháp sử dụng biểu diễn đặc trưng sang vectơ nhị phân trong không gian Hamming để chọn tập ứng viên nhằm giảm thời gian tính toán, giúp quá trình tra cứu nhanh hơn. Chúng tôi cũng cung cấp các kết quả thực nghiệm trên cơ sở dữ liệu ảnh CIFAR100 để chỉ ra độ chính xác của phương pháp.*

Từ khóa: *Tra cứu ảnh dựa trên nội dung, Học sâu, Mạng nơron tích chập đồ thị, Học biểu diễn ảnh.*

I. GIỚI THIỆU

Tra cứu hình ảnh dựa vào nội dung (CBIR- Content Based Image Retrieval) là quy trình tự động tra cứu ảnh bằng cách trích xuất các đặc trưng ảnh ở mức thấp của chúng, như màu sắc, kết cấu, hình dạng hoặc bất kỳ đặc điểm nào khác hữu ích cho tra cứu được lấy từ chính bức ảnh đó [1]. Hiệu suất của hệ thống CBIR chủ yếu phụ thuộc vào việc trích rút/lựa chọn các đặc trưng để tính toán độ tương tự giữa đặc trưng của ảnh truy vấn và các ảnh trong cơ sở dữ liệu [1]. Vấn đề thách thức lớn nhất của các hệ thống CBIR là giảm khoảng cách ngữ nghĩa, tức là làm giảm thông tin bị mất do biểu diễn một ảnh theo các đặc trưng của nó [2]. Khoảng cách ngữ nghĩa này tồn tại thường do sự sai khác giữa biểu diễn đặc trưng của thông tin ảnh và thông tin ảnh được cảm nhận bởi hệ thống thị giác người (HVS - Human Vision System). Khoảng cách ngữ nghĩa có thể được giảm bớt bằng cách sử dụng một số kỹ thuật học máy để phát triển các hệ thống thông minh có thể được huấn luyện để hoạt động như HVS. Thu được các biểu diễn ảnh (cụ thể là đặc trưng) phân biệt là một yêu cầu quan trọng đối với bất kỳ hệ thống tra cứu ảnh nào [3, 4]. Để làm cho đặc trưng này trở nên quan trọng và có trọng số tốt hơn về mặt biểu diễn kết hợp các đặc trưng trực quan cấp thấp thì chi phí tính toán cao là cần thiết để thu được kết quả đáng tin cậy hơn [5, 6]. Tuy nhiên, việc trích rút/lựa chọn các đặc trưng không phù hợp có thể làm giảm hiệu suất của mô hình tra cứu ảnh [8]. Vectơ đặc trưng ảnh có thể được sử dụng làm đầu vào cho các thuật toán học máy thông qua các mô hình huấn luyện - kiểm tra (train-test), đồng thời có thể cải thiện hiệu suất của CBIR [1, 7].

Các xu hướng tra cứu ảnh gần đây tập trung vào các mạng nơron sâu (DNN), mạng nơron tích chập (CNN) có khả năng tạo ra kết quả tốt hơn với chi phí tính toán cao. Hệ thống học sâu cơ bản sử dụng mạng nơron tích chập CNN để kết hợp quá trình trích rút và phân loại đặc trưng để cải thiện tốc độ và hiệu quả tra cứu hình ảnh nhờ biểu diễn đặc trưng có đầu ra tốt nhất, khả năng tổng quát hóa của các đặc trưng được trích xuất, mối quan hệ giữa giảm kích thước cũng như mất độ chính xác trong CBIR được cải thiện. Tuy nhiên, các phương pháp tra cứu ảnh sử dụng CNN không tận dụng được mối quan hệ giữa các ảnh có cùng chủ đề, cùng lớp nhãn trong cơ sở dữ liệu dẫn tới đặc trưng ảnh được trích rút không có tính tổng quát hóa cao. Để biểu diễn mối liên quan giữa đặc trưng của các ảnh trong cùng lớp, cùng chủ đề chúng tôi sử dụng cấu trúc đồ thị, các cạnh mô tả quan hệ giữa các ảnh và sử dụng phương pháp tích chập đồ thị để tổng hợp các đặc trưng của các nút lân cận thành đặc trưng có tính tổng quát hóa cao biểu diễn ảnh. Trong bài báo này chúng tôi trình bày một nghiên cứu sử dụng mạng nơron tích chập đồ thị (GCN - Graph Convolution Network) biểu diễn, biến đổi, tính toán lại đặc trưng của ảnh để tính độ tương tự cho xếp hạng và trả về kết quả tra cứu trong hệ thống CBIR. Kết quả thực nghiệm của chúng tôi được thực hiện trên bộ dữ liệu CIFAR-100.

Trong các phần tiếp theo của bài báo này, ở Mục II chúng tôi trình bày ngắn gọn về mạng nơron tích chập đồ thị và các nghiên cứu liên quan về CBIR sử dụng GCN. Mục III trình bày phương pháp đề xuất của chúng tôi. Cuối cùng các kết quả thực nghiệm được mô tả trong Mục IV. Kết luận được đưa ra trong Mục V.

II. NGHIÊN CỨU LIÊN QUAN

Với những ưu điểm và tính hiệu quả của các mô hình GCN trong việc thay đổi thuộc tính của một nút trên đồ thị bằng cách kết hợp các thuộc tính của các nút lân cận, trong lĩnh vực thị giác máy tính nói chung và tra cứu ảnh nói riêng đã thu hút nhiều sự quan tâm của các nhà nghiên cứu để ứng dụng mô hình GCN vào giải quyết bài toán của họ. Tiếp theo từ [9], Yoon cùng cộng sự đề xuất phương pháp tra cứu hình ảnh với tính tương tự của đồ thị cảnh trong [10],

sử dụng đồ thị ngữ nghĩa, được gọi là đồ thị cảnh. Với một ảnh truy vấn, mô hình sẽ tạo ra một đồ thị ngữ nghĩa và so sánh sự giống nhau của nó với các đồ thị ảnh trong cơ sở dữ liệu. Việc so sánh đồ thị này đạt được bằng cách lấy tích bên trong của các phần nhúng đồ thị do GNN tạo ra (GCN [11] hoặc GIN [12]). So sánh đồ thị đầu vào và đồ thị đầu ra đằng sau mô hình cho phép công việc này mở rộng quy mô tốt hơn sang cơ sở dữ liệu ảnh lớn khi so sánh với [13].

Sử dụng đồ thị ảnh *K* - lân cận gần nhất được biểu thị dưới dạng nhúng đặc trưng được Liu và cộng sự [14] đề xuất sử dụng GCN cùng với hàm mất mát mới dựa trên độ tương tự của ảnh. Các tính năng nhúng được tăng cường để giải thích cho toàn bộ đồ thị dựa trên toàn bộ cơ sở dữ liệu hình ảnh bằng GCN. Độ tương tự giữa các ảnh được tính bằng cách lấy tích trong của các đặc trưng nhúng. Độ tương tự càng cao, ứng cử viên tra cứu càng tốt. Hàm mất mát mới của tác giả được thiết kế để di chuyển các ảnh tương tự lại gần nhau hơn trong không gian nhúng và các ảnh khác nhau sẽ ra xa hơn. Zhang và cộng sự [15] cũng sử dụng đồ thị K - lân cận gần nhất, nhưng tập trung vào việc cải thiện quy trình xếp hạng lại trong tra cứu ảnh dựa trên nội dung. GNN được áp dụng cho các đặc trưng tổng hợp được tạo ra từ ma trận kề đã điều chỉnh. Việc sử dụng GNN cho phép quá trình xếp hạng lại không nhấn mạnh đến các nút có điểm tin cậy thấp. Các tác giả của [16] chuyển sang áp dụng cách tiếp cận đa phương thức. Họ sử dụng GraphSAGE [17] để tìm một cách học hiệu quả nhằm nhúng các nút đa phương thức chứa thông tin trực quan và khái niệm từ các kết nối trong đồ thị. Khoảng cách giữa các nút được kết nối giảm, trong khi khoảng cách giữa các nút bị ngắt kết nối tăng lên. Bằng cách sử dụng các nút đồ thị đại diện cho ảnh cũng như các nút đại diện cho thẻ siêu dữ liệu, mô hình của họ có thể cung cấp khả năng tra cứu ảnh dựa trên nội dung cũng như dự đoán nhãn. Tại thời điểm suy luận, các ảnh hiển thị cho mô hình có thể được đính kèm vào đồ thị thông qua K ảnh gần nhất của chúng, được đính kèm với các nhãn có liên quan hoặc cả hai.

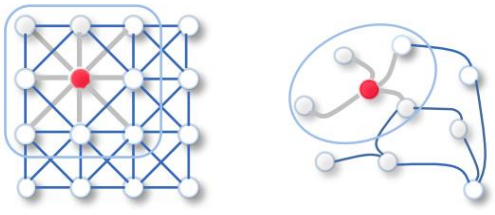
Chaudhuri và cộng sự [18] áp dụng kiến trúc mạng dựa trên Siamese trong đó hai đầu vào tương tự đi vào hai mạng riêng biệt chia sẻ trọng số. Kiến trúc mạng này thường sử dụng suy hao tương phản hoặc suy hao bộ ba để đảm bảo đầu ra của các mạng này giống nhau. Các tác giả sử dụng một Siamese-GCN mới trên đồ thị kề vùng được hình thành bằng cách kết nối các vùng được phân đoạn liền kề và trọng số cạnh chiếm khoảng cách và góc giữa tâm của các vùng. Họ áp dụng kỹ thuật của mình cho các ảnh viễn thám có độ phân giải cao để tra cứu ảnh dựa trên nội dung. Bằng cách sử dụng một Siamese-GCN có độ tương phản giảm, các tác giả có thể tìm hiểu một cách nhúng mang các ảnh tương tự lại với nhau và tách các ảnh khác nhau ra xa nhau. Một nghiên cứu khác để kết hợp một thiết kế mạng Siamese là Zhang cùng các cộng sự [19]. Họ sử dụng thiết kế mạng ba thành phần để thực hiện tra cứu ảnh dựa trên bản phác thảo không ảnh với mạng mã hóa dựa trên Siamese, mạng này tạo ra các đặc điểm của ảnh và bản phác thảo liên quan bằng ResNet50.

Các công trình nghiên cứu liên quan chủ yếu kết hợp giữa văn bản và ảnh để xây dựng hệ thống tra cứu ảnh hoặc phân tách ảnh thành các thành phần riêng biệt dựa trên phương pháp tách biên, đồ thị được xây dựng bởi các nút trong đó các nút là một đối tượng trong ảnh hoặc các sự kết hợp giữa ảnh và văn bản tra cứu liên quan, chưa có công trình nghiên cứu nào về tra cứu hình ảnh sử dụng các ảnh trong cơ sở dữ liệu như các nút. Trong [28], Kipf cùng các cộng sự đề xuất phương pháp sử dụng GCN để phân lớp ảnh với học bán giám sát, tuy nhiên thời gian của mô hình đề xuất chưa được cải thiện.

III. PHƯƠNG PHÁP TRA CỨU ẢNH ĐỀ XUẤT

a. Biến đổi đặc trưng sử dụng mạng nơron tích chập đồ thị

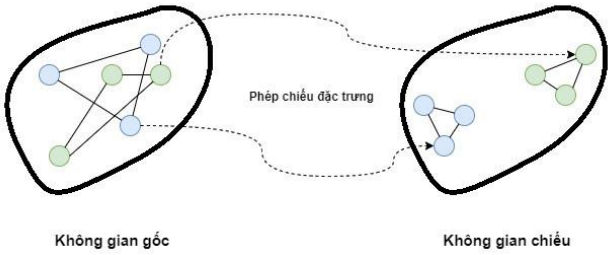
Trong nhiều năm, các mô hình học sâu đã tập trung khai thác, xử lý và huấn luyện trên các tập dữ liệu có cấu trúc. Tuy nhiên, trong thực tế, các đối tượng và các mối quan hệ giữa chúng được thể hiện một cách phi cấu trúc. Đồ thị là một biểu diễn dữ liệu tổng quát và cực kỳ hiệu quả đối với nhiều loại dữ liệu phi cấu trúc như mạng xã hội, các cấu trúc vật lý, hóa học. Các nhà nghiên cứu đã phát triển các mạng nơron hoạt động trên dữ liệu đồ thị (được gọi là mạng nơron đồ thị hoặc GNN - Graph Neural Networks) trong hơn một thập kỷ qua [20] và đã có những ứng dụng thực tế trong các lĩnh vực như khám phá chất kháng khuẩn [21], mô phỏng vật lý [22], phát hiện tin giả [23], dự đoán giao thông [24] và hệ thống khuyến cáo tiên đoán [25].



Hình 1. Tích chập trên ma trận 2 chiều và tích chập trên đồ thị [30]

Nổi bật trong số các mạng GNN là mạng nơron tích chập đồ thị (GCN - Graph Convolutional Networks) được giới thiệu lần đầu vào năm 2014 như một mô hình hiệu quả trong việc học để tổng hợp và biến đổi các đặc trưng của các nút liên quan trong tập dữ liệu đồ thị.

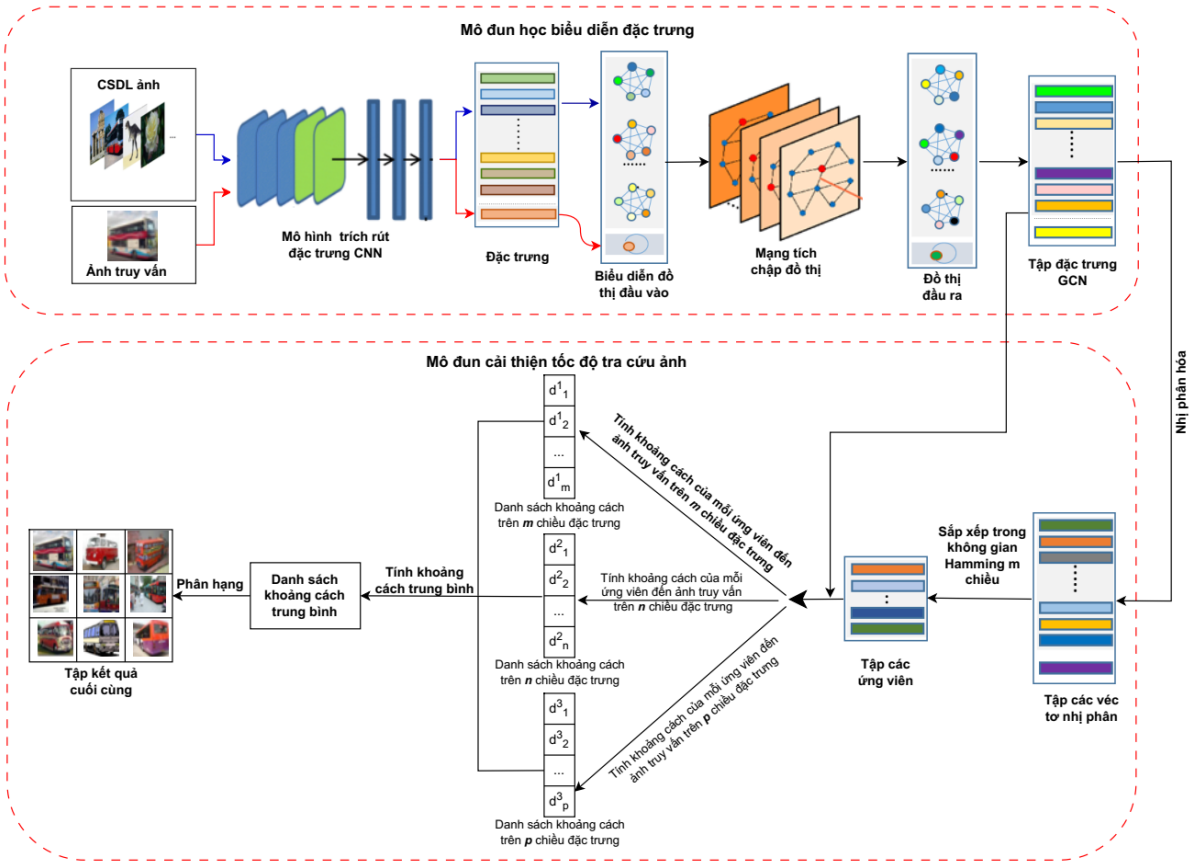
Khi áp dụng vào bài toán tra cứu ảnh, chúng tôi sử dụng GCN là phương pháp chiếu véc-tơ đặc trưng sang không gian mới sao cho ở đây các nút có mối quan hệ với nhau sẽ ở gần nhau và xa các nút không liên quan. Quá trình này được gọi là nhúng đồ thị (Graph Embedding) được biểu diễn trong Hình 2.



Hình 2. Ánh xạ đồ thị sang không gian mới

Kết quả đầu ra của quá trình chiếu rất hữu ích trong việc tìm kiếm các điểm dữ liệu liên quan và không liên quan của một điểm dữ liệu trong đồ thị khi áp dụng vào bài toán CBIR. Quá trình chiếu này gọi là quá trình học biểu diễn ảnh.

b. *Phương pháp đề xuất*



Hình 3. Sơ đồ hệ thống CBIR sử dụng GCN học biểu diễn với kỹ thuật tăng tốc độ tra cứu ảnh

1. *Học biểu diễn đặc trưng hiệu quả*

Mô hình tra cứu ảnh đề xuất của chúng tôi sử dụng các vectơ đặc trưng của ảnh được trích rút bởi một mô hình mạng nơron tích chập CNN như LeNet-5; AlexNet; VGGNe; GoogLeNe; ResNet; DenseNet; MobileNet, sau đó chúng tôi biểu diễn mỗi vectơ đặc trưng của ảnh như một nút trên đồ thị, hai nút ứng với hai ảnh cùng chủ đề sẽ được kết nối với nhau bởi một cạnh. Vectơ của ảnh truy vấn được xem như đồ thị có một nút với một cạnh vòng tới chính nó. Sơ đồ môđun học biểu diễn đặc trưng hiệu quả được trình bày trong khung phía trên của Hình 3.

Sau khi có được đặc trưng ban đầu, chúng tôi xây dựng đồ thị đầu vào $G = (V, A)$ cho quá trình học biểu diễn ảnh sử dụng mạng nơron tích chập đồ thị; trong đó $V = \{v_1, v_2, v_3, \dots, v_N\}, v_i \in R^F$ là tập nút của đồ thị với F là số chiều của vectơ đặc trưng, $A \in R^{N \times N}$ là ma trận kề vô hướng biểu diễn mối quan hệ giữa các nút. $D \in R^{N \times N}$ là ma trận bậc của đồ thị tương ứng với ma trận kề A. Với tập dữ liệu ban đầu $S = (X, L)$; trong đó $X \in R^{N \times F}$ là ma trận đặc trưng và $L \in R^N$ là tập nhãn của tập ảnh. Đồ thị G được xây dựng theo các bước dưới đây:

Bước 1: Khởi tạo

V ← ∅

E ← ∅

Bước 2: Tạo đồ thị

for i = 1…N do:

$v_i = x_i$

$V \leftarrow v_i$

for j = 1…N do:

if L[i] == L[j]:

$E \leftarrow e(i,j)$

end if

end for

end for

Kết quả là đồ thị $G = (V, E)$; $V = \{v_1, v_2, v_3, \dots, v_N\}, v_i \in R^F$ là tập nút của đồ thị với F là số chiều của vector đặc trưng, E là tập cạnh của đồ thị.

Tiếp theo, chúng tôi sẽ trình bày cách thức học biểu diễn ảnh (để thu được các vectơ đặc trưng của ảnh) trong đồ thị sử dụng lớp tích chập đồ thị. Tổng quát, mỗi lớp tích chập đồ thị GCN tiến hành ba thao tác: tổng hợp đặc trưng, biến đổi tuyến tính và kích hoạt tuyến tính. Sự khác nhau chính giữa lớp GCN và các lớp MLP (multi-layer perceptron) nằm ở quá trình tổng hợp đặc trưng từ những nút láng giềng lân cận; biểu diễn của nút đang xét có thể được lấy trung bình từ biểu diễn đặc trưng của các nút lân cận.

Giả định rằng tại lớp tích chập đồ thị thứ k, $H^k \in R^{N \times F_k}$ là vectơ đặc trưng biểu diễn nút và F_k là độ dài vectơ. Biểu diễn khởi tạo của nút là $H^0 = V$. Quá trình tổng hợp đặc trưng diễn ra trong lớp GCN được thực hiện theo công thức sau:

$$\overline{H}^{(k)} = \widehat{D}^{-\frac{1}{2}}\hat{A}\widehat{D}^{-\frac{1}{2}}H^{(k-1)}$$

trong đó $\hat{A} = A + I$ biểu thị ma trận kề với cạnh vòng và $\widehat{D}$ là ma trận bậc tương ứng.

Sau bước tổng hợp đặc trưng, hai bước còn lại của GCN là biến đổi tuyến tính và kích hoạt phi tuyến được thực hiện giống như các lớp MLP cơ bản. Lớp GCN thứ k bao gồm một ma trận trọng số có thể huấn luyện $W^k \in R^{F_{k-1} \times F_k}$ và một hàm kích hoạt phi tuyến $\sigma(x)$, ví dụ như $ReLU(x) = \max(0, x)$. Đặc trưng biểu diễn của nút được cập nhật theo công thức:

$$H^{(k)} = \sigma(\overline{H}^{(k)}W^{(k)})$$

Ma trận trọng số W được tối ưu bằng cách tối thiểu hóa hàm mất mát.

Quá trình thực hiện tích chập trên đồ thị biểu diễn đặc trưng để thực hiện biến đổi và cho kết quả đồ thị đầu ra gồm các đỉnh là các vectơ đặc trưng đã được cập nhật từ quá trình học biểu diễn ảnh theo phương pháp đề xuất của chúng tôi. Từ đồ thị gồm các vectơ đặc trưng đã được cập nhật, chúng tối tiến hành trích xuất và lưu trữ lại tập hợp đặc trưng mới này.

2. Tra cứu nhanh trên tập ảnh cỡ lớn và hàm khoảng cách hiệu quả

Trong Mục III.1 bên trên chúng ta đã thu được một biểu diễn đặc trưng ảnh hiệu quả, mà giúp tăng độ chính xác của phương pháp tra cứu ảnh. Với các hệ thống tra cứu ảnh chung đã được nghiên cứu trước đây thì bước tiếp theo là sử dụng hàm khoảng cách để tính độ tương tự, xếp hạng các vectơ đặc trưng theo thứ tự độ tương tự giảm dần và đưa ra kết quả tra cứu. Một số hàm khoảng cách phổ biến được sử dụng có thể kể đến như: Euclide, Manhatten, Cosine,… Tuy nhiên, với những hệ thống tra cứu như vậy thời gian tra cứu ảnh sử dụng các biểu diễn đặc trưng ảnh sẽ chậm đối với các cơ sở dữ liệu ảnh có kích thước lớn. Do đó, trong phần này chúng tôi đề xuất giải pháp khắc phục để tăng tốc độ tính toán, làm giảm thời gian của quá trình tra cứu. Sơ đồ của giải pháp này được thể hiện trong môđun cải thiện tốc độ tra cứu ảnh ở khung bên dưới của Hình 3, mà chúng tôi sẽ giải thích chi tiết ngay sau đây.

Sau khi thu được tập đặc trưng GCN của cơ sở dữ liệu (CSDL) ảnh và ảnh truy vấn (đầu ra của mô hình biểu diễn đặc trưng trong khung bên trên của Hình 3), chúng tôi tiến hành nhị phân hóa để được tập vectơ nhị phân (trên không giam Hamming) tương ứng. Trên không gian Hamming, chúng tôi tính khoảng cách từ ảnh truy vấn tới mỗi ảnh trong CSDL và sắp xếp các ảnh CSDL theo thứ tự tăng dần của khoảng cách để thu được một tập ứng viên gồm K ảnh. Lý do chúng tôi chuyển các ảnh từ không gian đặc trưng GCN sang không gian Hamming là bởi vì trên không gian này

khoảng cách giữa 2 vectơ được tính một cách nhanh chóng. Bên cạnh đó, chúng ta cũng cần thu một tập ứng viên mà không phải tập kết quả tra cứu cuối cùng bởi vì trên không gian Hamming, khoảng cách từ một ảnh CSDL đến ảnh truy vấn có thể không phản ánh tốt độ tương tự giữa ảnh CSDL và ảnh truy vấn như trên không gian đặc trưng GCN. Chúng tôi cũng sử dụng ba không gian đặc trưng GCN với số chiều của vectơ đặc trưng GCN lần lượt là ba giá trị khác nhau *m, n, p* để gia tăng độ tin cậy của kết quả tìm kiếm trong trường hợp số chiều tổng quát. Trên không gian đặc trưng GCN *m* chiều, chúng tôi tính khoảng cách từ ảnh truy vấn đến mỗi ảnh trong tập ứng viên để thu được danh sách các khoảng cách tương ứng (chúng tôi gọi là DSKC_m). Thực hiện tương tự với không gian đặc trưng GCN *n* chiều và *p* chiều chúng tôi thu được danh sách khoảng cách DSKC_n và DSKC_p. Dựa trên danh sách khoảng cách DSKC_m, DSKC_n, và DSKC_p chúng ta tính khoảng cách trung bình từ ảnh truy vấn đến mỗi ảnh trong tập ứng viên để nhận được danh sách khoảng cách trung bình. Khi tính khoảng cách từ ảnh truy vấn đến mỗi ảnh trong tập ứng viên trên không gian đặc trưng GCN, chúng ta sẽ thu được một tập kết quả chính xác hơn là trên không gian Hamming và kích thước của không gian tìm kiếm sẽ nhỏ hơn. Để tăng tính tổng quát, chúng ta có thể sử dụng nhiều hơn 3 không gian đặc trưng với số chiều cụ thể. Sắp xếp các ảnh trong tập ứng viên theo thứ tự tăng dần của khoảng cách trung bình để thu được tập kết quả tra cứu cuối cùng.

IV. THỰC NGHIỆM

a. Dữ liệu thực nghiệm

Chúng tôi tiến hành thực nghiệm trên tập huấn luyện của bộ dữ liệu ảnh Cifar (Canadian Institute for Advanced Research) 100 là tập dữ liệu bao gồm có 50000 ảnh. Trong đó chúng tôi chia thành 30000 cho huấn luyện và 10000 cho đánh giá và 10000 ảnh cho kiểm thử. Số lượng lớp phân loại là 100 với kích thước 32x32 một ảnh. Chúng tôi sử dụng mô hình Autoencoder để trích rút đặc trưng ban đầu cho các ảnh, với vectơ đặc trưng trích rút bởi Autoencoder là 128 chiều. Các nút của đồ thị là một bộ gồm 2 thành phần u = {nhãn, vectơ đặc trưng}. Hai nút tương ứng với hai hình ảnh cùng nhãn sẽ sẽ có cạnh kết nối.

b. Xây dựng kiến trúc mạng nơron tích chập đồ thị cho thực nghiệm

Trước hết, để có thể có được một mô hình trích rút đặc trưng hiệu quả, chúng tôi xây dựng mô hình cho bài toán phân loại nút trong đồ thị sử dụng lớp tích chập đồ thị GCN. Sau đó, chúng tôi sử dụng phương pháp học chuyển giao, bỏ đi lớp cuối cùng có chức năng phân loại và giữ lại các lớp phía trước có nhiệm vụ học và biến đổi các đặc trưng lên mức cao. Chúng tôi sử dụng mô hình GCN gồm 8 lớp tích chập để biến đổi đặc trưng với số chiều vectơ đặc trưng đầu và và đầu ra cho trong bảng sau:

Bảng 1. Thông số kiến trúc mô hình

Lớp	Conv1	Conv2	Conv3	Conv4	Conv5	Conv6	Conv7	Conv8
Số chiều vectơ đầu vào	128	1024	1024	512	512	256	256	128
Số chiều vectơ đầu ra	1024	1024	512	512	256	256	128	100

Trong đó chúng tôi sử dụng đầu ra của lớp thứ 7 để trích rút đặc trưng cho các nhiệm vụ trong bài toán tra cứu ảnh. Hàm kích hoạt softmax, được định nghĩa là $softmax(x_i) = \frac{1}{Z}e^{x_i}$ với $Z = \sum_i e^{x_i}$ Z, được áp dụng cho lớp cuối cùng làm nhiệm vụ phân loại, có chức năng tính toán tỉ lệ một nút thuộc về một lớp nào đó. Chúng tôi triển khai việc tính toán độ sai lệch trong phân lớp sử dụng hàm mất mát Cross-EntropyLoss, là hàm mất mát đơn giản nhưng hiệu quả và được sử dụng rất phổ biến cho bài toán phân lớp.

c. Chuyển đổi vectơ đặc trưng GCN sang vectơ nhị phân và hàm khoảng cách Hamming

Trong phần nội dung này, chúng tôi sẽ tập trung trình bày phương thức biến đổi vectơ đặc trưng sang không gian nhị phân để có thể sử dụng hàm tính khoảng cách Hamming nhằm tăng tốc độ quá trình tính toán độ tương tự giữa vectơ nhị phân của ảnh truy vấn với các vectơ nhị phân trong CSDL. Trong thực nghiệm này, từ mô hình phân lớp GCN (được trình bày trong Mục IV.B), chúng tôi thực hiện bỏ lớp cuối cùng của mô hình, giữ lại toàn bộ 7 lớp GCN đầu tiên để thu được một mô hình trích rút đặc trưng hiệu quả với đầu ra là vectơ đặc trưng 128 chiều. Lý do chúng tôi chọn số chiều 128 là vì số chiều này thường được các mô hình tra cứu ảnh sử dụng. Sau đó, chúng tôi tiến hành nhị phân hóa một cách hiệu quả các vectơ đặc trưng. Trước hết, chúng tôi tối giản hóa các giá trị trong vectơ về khoảng [0, 1] bằng cách sử dụng một hàm kích hoạt phi tuyến Hard-Sigmoid:

$$f(x) = max\left(0, min\left(1, \frac{(x+1)}{2}\right)\right)$$

Sau đó, để thực hiện nhị phân hóa, chúng tôi thực hiện phân ngưỡng giá trị theo công thức sau:

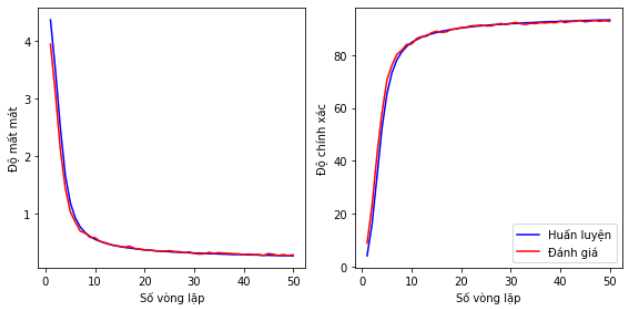
$$f(x) = \begin{cases} 0 \; nếu \; x < 0.5 \\ 1 \; nếu \; x \geq 0.5 \end{cases}$$

Khi đó, việc tính khoảng cách (hay độ tương tự) giữa vectơ đặc trưng đã được nhị phân hóa của ảnh truy vấn và ảnh CSDL sẽ được tính bởi khoảng cách Hamming, là số lượng bit khác nhau trong 2 vectơ nhị phân. Để đơn giản hóa giá trị tính toán, chúng tôi thực hiện chia số lượng bit khác nhau cho độ dài của vectơ để thu được kết quả cuối cùng là độ tương tự giữa 2 vectơ nhị phân.

d. Kết quả thực nghiệm

3. Thực nghiệm đánh giá hiệu quả của mô hình GCN cho bài toán phân lớp

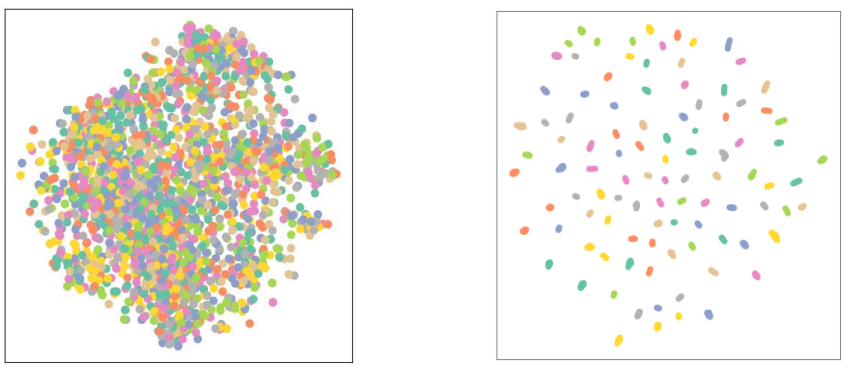
Để đánh giá hiệu quả của mô hình biểu diễn đặc trưng do chúng tôi đề xuất, chúng tôi thực hiện huấn luyện mô hình cho bài toán phân loại nút với 50 vòng lặp huấn luyện sử dụng phương pháp tối ưu Adam với tốc độ học 0.001. Quá trình huấn luyện mô hình phân loại và trích rút đặc trưng mới được thực hiện trên môi trường Python3 với máy tính hệ điều hành Linux sử dụng GPU P100 16 GB được cung cấp bởi Kaggle. Sau đó, quá trình chạy và đánh giá kết quả thực nghiệm cho bài toán tra cứu ảnh được cài đặt trên máy tính hệ điều hành Windows 11, với CPU Intel Core I7-8550U 1.80 GHz và 16 GB RAM. Hình 5 thể hiện kết quả thay đổi của độ chính xác, độ mất mát trong quá trình huấn luyện theo số vòng lặp.



Hình 4. Biểu đồ độ mất mát và độ chính xác trong quá trình huấn luyện và kiểm thử mô hình phân loại

Kết quả cho thấy mô hình hội tụ tại vòng lặp 44 với độ chính xác trên tập đánh giá là ~93,06% và 92,58% trên tập kiểm thử.

Mô hình trích xuất đặc trưng mới được lấy ra từ mô hình phân lớp đã được huấn luyện cho bài toán phân lớp nút trước đó. Việc này cũng được thực hiện tương tự như khi ta xây dựng mô hình trích rút đặc trưng từ một mạng CNN như VGG hay ResNet đã được huấn luyện trước. Chúng tôi cũng sử dụng mô hình hóa trực quan hai chiều để biểu diễn kết quả học biểu diễn sử dụng mô hình đề xuất, từ đó cho thấy hiệu quả trong việc cập nhật đặc trưng. Chúng tôi sử dụng 2000 mẫu ngẫu nhiên trong số 50000 mẫu, kết quả trực quan được thể hiện trong Hình 6.



(a) (b)

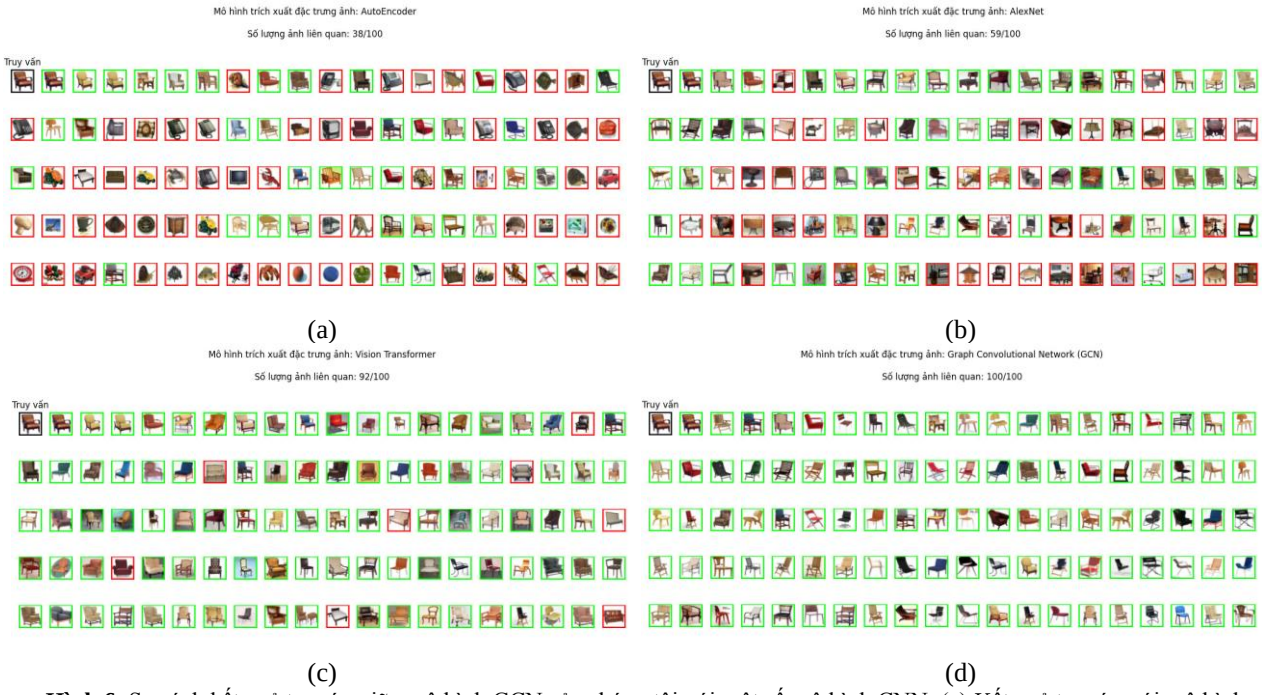
Hình 5. Biểu diễn kết quả phân lớp đặc trưng sau khi áp dụng mô hình GCN của chúng tôi: (a) Biểu diễn tập đặc trưng đầu vào trước khi áp dụng mô hình GCN, (b) Biểu diễn kết quả đầu ra sau khi áp dụng mô hình GCN

Hình 6.a cho thấy rất rõ ràng rằng trước khi được học biểu diễn bởi GCN, đặc trưng của các nút ở mức rất thấp, dẫn tới việc khả năng nhóm dữ liệu kém, vị trí biểu diễn các nút rất lộn xộn, không có sự phân tách rõ ràng; từ đó, rất khó để xác định được nhóm các hình ảnh thuộc cùng một chủ đề. Tuy nhiên, sau khi sử dụng GCN để học biểu diễn, đặc trưng của các nút được cải thiện đáng kể, khả năng nhóm dữ liệu tốt hơn rất nhiều (được biểu diễn trong Hình 6.b), do đó, kết quả của bài toán tra cứu ảnh cũng trở nên tốt hơn nhiều so với lúc ban đầu.

4. Thực nghiệm đánh giá hiệu quả của mô hình biểu diễn đặc trưng mô hình tra cứu ảnh truyền thống

Trong phần này, chúng tôi tập trung thực nghiệm để so sánh hiệu năng học biểu diễn đặc trưng của GCN với một số mô hình học sâu CNN phổ biến như AlexNet, Vision Transformer. Chúng tôi sử dụng đặc trưng trong tập dữ

liệu huấn luyện CIFAR100 được học từ mô hình của chúng tôi và hàm khoảng cách Euclide để tiến hành tra cứu và so sánh kết quả với một số mô hình trích rút đặc trưng khác, minh họa kết quả như sau:



Hình 6. So sánh kết quả tra cứu giữa mô hình GCN của chúng tôi với một số mô hình CNN: (a) Kết quả tra cứu với mô hình trích rút đặc trưng Autoencoder, (b) Kết quả tra cứu với mô hình trích rút đặc trưng AlexNet, (c) Kết quả tra cứu với mô hình trích rút đặc trưng Vision Transformer, (d) Kết quả tra cứu với mô hình trích rút đặc trưng GCN do chúng tôi đề xuất.

Từ kết quả tra cứu trên cho thấy đặc trưng ảnh được học từ mô hình của chúng tôi rất hiệu quả cho việc tra cứu ảnh. Để chứng tỏ hiệu quả của việc học đặc trưng, chúng tôi cũng sử dụng phương pháp đánh giá hiệu năng của mô hình tra cứu bằng cách thực nghiệm tính toán chỉ số mAP. Độ đo mAP là phép đo độ chính xác để so sánh hiệu năng truy vấn trong bài toán tra cứu ảnh. Phương pháp xác định giá trị mAP được trình bày như sau:

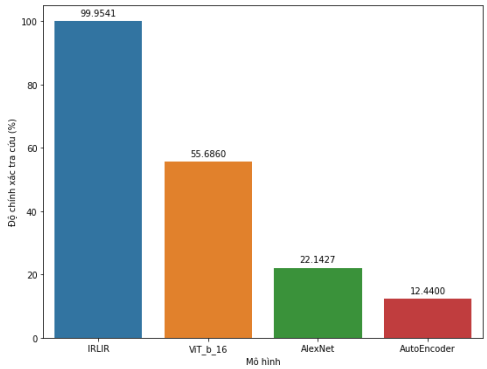
Độ chính xác trung bình (AP): đề cập đến vùng phủ dưới đường cong precision-recall curve. AP lớn ngụ ý đường cong có độ chính xác cao hơn và độ chính xác tra cứu tốt hơn. AP có thể được tính như sau:

$$AP = \frac{\sum_{k=1}^{N} P(k).rel(k)}{R}$$

trong đó, R biểu thị số kết quả có liên quan cho hình ảnh truy vấn từ tổng số N hình ảnh. P(k) là độ chính xác của k ảnh trong kết quả trả về và rel(k) là hàm chỉ báo bằng 1 nếu mục trong hạng k là hình ảnh có liên quan và 0 nếu ngược lại. Độ chính xác trung bình trung bình (mAP) được áp dụng để đánh giá trên tất cả các hình ảnh truy vấn:

$$mAP = \frac{1}{Q}\sum_{q=1}^{Q} AP(q)$$

trong đó Q là số lượng hình ảnh truy vấn.



Hình 7. Kết quả so sánh chỉ số mAP giữa phương pháp IRLIR và các mô hình CNN

Chúng tôi tính toán và so sánh chỉ số mAP trong tra cứu ảnh sử dụng mô hình GCN của chúng tôi với chỉ số mAP của các mô hình CNN đã được đánh giá tốt như: ViT_b_16, AlexNet, Autoencoder để so sánh. Kết quả thể hiện trong Hình 8. Từ kết quả này cho thấy mô hình trích rút đặc trưng của chúng tôi có độ chính xác cao.

5. Thực nghiệm đánh giá hiệu quả của giải pháp tăng tốc độ tính toán

Để đánh giá hiệu quả của việc chuyển đổi vectơ đặc trưng GCN sang vectơ nhị phân trong không gian Hamming cho mục tiêu giảm thời gian tính toán, tăng tốc độ tra cứu của mô hình, chúng tôi chọn các chiều gồm 128, 96, và 32 để khai thác các khía cạnh khác nhau (chúng tôi thực nghiệm với 3 khía cạnh, trong thực tế chúng ta có thể chọn nhiều hơn 3 khía cạnh) của mỗi ảnh. Chúng tôi đánh giá chỉ số mAP và thời gian (tính bằng giây) khi thực nghiệm tra cứu với 128 chiều trên không gian đặc trưng GCN (gọi là IRLIR_GCN_128), trên không gian Hamming (gọi là IRLIR_HAMMING_128), trên không gian kết hợp đặc trưng GCN và Hamming như Mục III.B.2 (gọi là IRLIR_COMBINATION). Kết quả được thể hiện ở Bảng 2.

Bảng 2. Kết quả thực nghiệm trên các không gian 128 chiều khác nhau

	IRLIR_GCN_128	IRLIR_HAMMING_128	IRLIR_COMBINATION
mAP(%)	99.9541	99.945	99.9470
Time(s)	0.4020	0.2050	0.2295

Bảng 2, so sánh kết quả tra cứu và đánh giá thời gian trong 3 trường hợp: Chỉ sử dụng đặc trưng GCN; Chỉ sử dụng đặc trưng được nhị phân hóa; Sử dụng kết hợp đặc trưng GCN và đặc trưng được nhị phân hóa. Từ kết quả trong Bảng 2, chúng tôi đưa ra các kết luận để khẳng định tính hiệu quả của việc sử dụng kết hợp đặc trưng GCN với hàm khoảng cách Hamming như sau:

- Trên không gian GCN 128 chiều: Độ chính xác tra cứu cao nhưng thời gian tra cứu lớn.
- Trên không giam Hamming 128 chiều: Thời gian tra cứu nhanh nhưng độ chính xác tra cứu thấp.
- Trên không gian kết hợp GCN và Hamming: Thời gian tra cứu nhanh và độ chính xác tra cứu vẫn được đảm bảo.

V. KẾT LUẬN

Trong bài báo này, chúng tôi đã trình bày một mô hình đề xuất làm việc hiệu quả để biểu diễn lại đặc trưng của ảnh cho tra cứu ảnh và những nhiệm vụ khác liên quan. Mô hình này tận dụng được hai vấn đề: thứ nhất, kết hợp được đặc trưng các ảnh trong cùng một lớp để có biểu diễn tốt hơn các mô hình sử dụng CNN và thứ hai sử dụng hiệu năng của GCN để học các biểu diễn đặc trưng hiệu quả cho tra cứu ảnh thông qua việc sử dụng nhiều lớp tích chập. Mô hình học này được sử dụng vào việc biểu diễn lại các đặc trưng của các ảnh cơ sở dữ liệu và ảnh truy vấn. Trên cơ sở các biểu diễn đặc trưng này, có thể đưa ra nhiều hướng nghiên cứu mới trong việc cải tiến mô hình tốt hơn, sử dụng mô hình biểu diễn đặc trưng của chúng tôi đã thiết kế để nghiên cứu các nhiệm vụ khác trong thị giác máy tính. Kết quả của mô hình này đã thu được các danh sách phân hạng có chất lượng tốt, vừa khắc phục được sự thiếu hụt đặc trưng liên quan của các mẫu có cùng nhãn vừa tận dụng được ưu điểm của các mạng nơron tích chập đồ thị. Các kết quả thực nghiệm thực hiện trên tập CIFAR-100 đã minh chứng rằng mô hình được đề xuất của chúng tôi sinh ra kết quả có độ chính xác cao.

LỜI CẢM ƠN

Chúng tôi xin chân thành cảm ơn Quỹ phát triển Khoa học và Công nghệ Quốc gia (NAFOSTED) đã tài trợ đề tài "Phát triển các thuật toán học đa tạp ảnh và hàm khoảng cách cho nâng cao độ chính xác và tốc độ tra cứu ảnh" mã số 102.01-2020.10 để chúng tôi hoàn thành nghiên cứu này.

TÀI LIỆU THAM KHẢO

[1] Y. Liu, D. Zhang, G. Lu, and W.-Y. Ma, "A survey of contentbased image retrieval with high-level semantics," *Pattern Recognition*, vol. 40, no. 1, pp. 262-282, 2007.

[2] K. Kranthi Kumar, T. Venu Gopal, "A novel approach to self order feature reweighting in cbir to reduce semantic gap using relevance feedback," *International Conference on Circuits, Power and Computing Technologies*, (2014).

[3] Z. Yu and W. Wang, "Learning DALTS for cross-modal retrieval," *CAAI Transactions on Intelligence Technology*, vol. 4, no. 1, pp. 9-16, 2019.

[4] N. Ali, D. A. Mazhar, Z. Iqbal, R. Ashraf, J. Ahmed, and F. Zeeshan, "Content-based image retrieval based on late fusion of binary and local descriptors," *International Journal of Computer Science and Information Security (IJCSIS)*, vol. 14, no. 11, 2016.

[5] N. Ali, "Image Retrieval Using Visual Image Features and Automatic Image Annotation," University of Engineering and Technology, Taxila, Pakistan, 2016.

[6] B. Zafar, R. Ashraf, N. Ali et al., "Intelligent image classification-based on spatial weighted histograms of concentric circles," *Computer Science and Information Systems*, vol. 15, no. 3, pp. 615-633, 2018.

[7] D. Zhang, M. M. Islam, and G. Lu, "A review on automatic image annotation techniques," *Pattern Recognition*, vol. 45, no. 1, pp. 346-362, 2012.

[8] L. Piras and G. Giacinto, "Information fusion in content based image retrieval: a comprehensive overview," *Information Fusion*, vol. 37, pp. 50-60, 2017.

[9] J. Johnson, R. Krishna, M. Stark, L.-J. Li, D. Shamma, M. Bernstein, and L. Fei-Fei, "Image retrieval using scene graphs," in *CVPR*, 2015, pp. 3668-3678.

[10] S. Yoon, W. Y. Kang, S. Jeon, S. Lee, C. Han, J. Park, and E.-S. Kim, "Image-to-image retrieval by learning similarity between scene graphs," in *AAAI*, vol. 35, no. 12, 2021, pp. 10718-10726.

[11] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *ICLR*, 2017. [Online]. Available: https://openreview.net/forum?id=SJU4ayYgl.

[12] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, "How powerful are graph neural networks?" *arXiv preprint*, 2018.

[13] J. Johnson, R. Krishna, M. Stark, L.-J. Li, D. Shamma, M. Bernstein, and L. Fei-Fei, "Image retrieval using scene graphs," in *CVPR*, 2015, pp. 3668-3678.

[14] C. Liu, G. Yu, M. Volkovs, C. Chang, H. Rai, J. Ma, and S. K. Gorti, "Guided similarity separation for image retrieval," *NeurIPS*, vol. 32, 2019.

[15] X. Zhang, M. Jiang, Z. Zheng, X. Tan, E. Ding, and Y. Yang, "Understanding image retrieval re-ranking: a graph neural network perspective," *arXiv preprint*, 2020.

[16] A. K. Misraa, A. Kale, P. Aggarwal, and A. Aminian, "Multi-modal retrieval using graph neural networks," *arXiv preprint*, 2020

[17] W. L. Hamilton, R. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *NeurIPS*, 2017, pp. 1025-1035.

[18] U. Chaudhuri, B. Banerjee, and A. Bhattacharya, "Siamese graph convolutional network for content based remote sensing image retrieval," *Computer vision and image understanding*, vol. 184, pp. 22-30, 2019.

[19] Z. Zhang, Y. Zhang, R. Feng, T. Zhang, and W. Fan, "Zero-shot sketchbased image retrieval via graph convolution network," in *AAAI*, vol. 34, no. 07, 2020, pp. 12943-12950.

[20] Scarselli, F., Gori, M., Tsoi, A.C., Hagenbuchner, M. and Monfardini, G.; "The Graph Neural Network Model," *IEEE Transactions on Neural Networks*, vol. 20 (1) 2009, pp. 61-80.

[21] Stokes, J.M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N.M., MacNair, C.R., French, S., Carfrae, L.A., Bloom-Ackermann, Z., Tran, V.M., Chiappino-Pepe, A., Badran, A.H., Andrews, I.W., Chory, E.J., Church, G.M., Brown, E.D., Jaakkola, T.S., Barzilay, R. and Collins, J.J, "A Deep Learning Approach to Antibiotic Discovery," *Cell*, vol. 181 (2), pp. 475-483, 2020.

[22] Sanchez-Gonzalez, A., Godwin, J., Pfaff, T., Ying, R., Leskovec, J. and Battaglia, "Learning to simulate complex physics with graph networks," P.W. 2020.

[23] Monti, F., Frasca, F., Eynard, D., Mannion, D. and Bronstein, M.M, "Fake News Detection on Social Media using Geometric Deep Learning," 2019.

[24] O.L. and Perez, "Traffic prediction with advanced Graph Neural Networks," *Proceedings of Machine Learning Research*, vol. 220, 2020.

[25] Eksombatchai, C., Jindal, P., Liu, J.Z., Liu, Y., Sharma, R., Sugnet, C., Ulrich, M. and Leskovec, J, "Pixie: A System for Recommending 3+ Billion Items to 200+ Million Users in {Real-Time}," 2017.

[26] Yuqi Zhang, Qi Qian, Hongsong Wang, Chong Liu, Weihua Chen, Fan Wang, "Graph Convolution Based Efficient Re-Ranking for Visual Retrieval," *IEEE Transactions on Multimedia*, 2023, pp.125-137, 2023.

[27] Zhaolong Zhang, Yuejie Zhang, Rui Feng, Tao Zhang, Weiguo Fan, "Zero-Shot Sketch-Based Image Retrieval via Graph Convolution Network," *The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)*, pp. 12943-12950, 2020.

[28] Kipf, T. N., and Welling, M. 2016, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907Ff.*

[29] Z. Zhang, Y. Zhang, R. Feng, T. Zhang, and W. Fan, "Zero-shot sketchbased image retrieval via graph convolution network," in *AAAI*, vol. 34, no. 07, 2020, pp. 12943-12950.

[30] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, Philip S. Yu, "A Comprehensive Survey on Graph Neural Networks," *Journal of Latex Class Files*, 2019.

[31] Thomas Kipf, "Graph convolutional networks," https://tkipf.github.io/graph-convolutional-networks/, 2016.

IMAGE REPRESENTATION LEARNING WITH GRAPH CONVOLUTIONAL NETWORK FOR CONTENT-BASED IMAGE RETRIEVAL

Nguyen Van Thanh, Nguyen Huu Quynh, Pham Huy Hoang, Dao Thi Thuy Quynh, Cu Viet Dung

ABSTRACT: *The performance of a content-based image retrieval system depends mainly on two stages: (1) learning the effective image representation and (2) the appropriate distance function for the learned image representation. There have been many content-based image retrieval methods that leverage deep learning to learn image representation effectively, but these methods do not take advantage of relationships between images. Recently, Graph convolutional networks (GCNs) have emerged as a powerful deep learning approach for data represented as graphs. In this paper, we present an image retrieval method called IRLIR (Image Representation Learning for Image Retrieval). The proposed mothed uses a graph convolutional neural network to learn effective image representations and a solution that uses feature representation as binary vectors in Hamming space to select candidate set for reducing calculation time, making the retrieval process faster. We also provide experimental results on the CIFAR100 image database to show the accuracy of the method.*