TẠP CHÍ KHOA HỌC
TRƯỜNG ĐẠI HỌC SƯ PHẠM TP HỒ CHÍ MINH
Tập 20, Số 5 (2023): 831-841
Website: https://journal.hcmue.edu.vn
ISSN: 2734-9918

HO CHI MINH CITY UNIVERSITY OF EDUCATION
JOURNAL OF SCIENCE
Vol. 20, No. 5 (2023): 831-841
https://doi.org/10.54607/hcmue.js.20.5.3644(2023)

Bài báo nghiên cứu

MỘT PHƯƠNG PHÁP PHÂN LỚP ẢNH ĐA NHÃN DỰA TRÊN MẠNG TÍCH CHẬP ĐỒ THỊ

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

[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 Văn Thịnh – Email: thinhnv@hcmue.edu.vn

Ngày nhận bài: 18-10-2022; ngày nhận bài sửa: 21-02-2023; ngày duyệt đăng: 27-02-2023

TÓM TẮT

Phân lớp ảnh đa nhãn là một trong những tác vụ quan trọng và thách thức trong thị giác máy tính. Trong bài báo này, một phương pháp phân lớp ảnh đa nhãn được đề xuất dựa trên mạng tích chập đồ thị hướng đến việc khai thác mối quan hệ giữa các nhãn lớp trong tập dữ liệu và giữa các đối tượng trong ảnh nhằm nâng cao độ chính xác. Đầu tiên, nội dung hình ảnh được học biểu diễn bằng mạng nơ-ron tích chập (CNN – Convolutional Neural Network) và mạng tích chập đồ thị (GCN – Graph Convolutional Network) dựa trên đồ thị ngữ cảnh (scene graph) của ảnh. Sau đó, đồ thị mô tả sự phụ thuộc giữa các nhãn đối tượng trong tập dữ liệu được xây dựng làm cơ sở cho việc học các bộ phân lớp cho các nhãn bằng cách sử dụng GCN, từ đó, áp dụng các bộ phân lớp này cho đặc trưng ảnh để tạo ra các giá trị nhãn lớp dự đoán. Cuối cùng, toàn bộ mạng được huấn luyện sử dụng cách phân lớp đa nhãn truyền thống. Thực nghiệm được xây dựng và đánh giá trên tập dữ liệu là phần giao giữa tập Visual Genome và MS COCO. Kết quả thực nghiệm cho thấy phương pháp đề xuất là hiệu quả và vượt trội hơn một số công trình đã công bố gần đây.

Từ khóa: convolutional neural network; graph convolutional network; label graph; multi-label image classification; scene graph

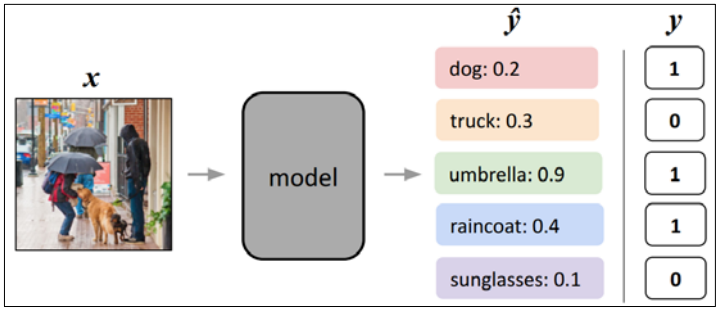
1. Giới thiệu

Hình ảnh trong các ứng dụng thực thế thường miêu tả nhiều đối tượng và ngữ cảnh phức tạp, phân lớp ảnh đa nhãn (*multi-lable image classification*) là việc dự đoán tập nhãn tương ứng với các đối tượng, thuộc tính hoặc các thực thể khác có trong ảnh đầu vào (Lanchantin, Wang, Ordonez, & Qi, 2021). Một ví dụ về bài toán phân lớp ảnh đa nhãn như ở Hình 1, đầu vào là ảnh x, cần dự đoán tập các nhãn $\{y_1, y_2, \dots, y_L\}, y_i \in \{0,1\}$. Bài toán này có nhiều ứng dụng trong các lĩnh vực như: xác nhận các chẩn đoán y khoa (Ge, Mahapatra, Sedai, Garnavi, & Chakravorty, 2018), nhận dạng thanh toán bán lẻ (Wei, Cui, Yang, Wang, & Liu, 2019)…

Cite this article as: Nguyen Van Thinh, Tran Van Lang, & Van The Thanh (2023). A multi-label image classification method based on convolutional neural network. *Ho Chi Minh City University of Education Journal of Science, 20*(5), 831-841.

Tập nhãn đầu ra của một hình ảnh thường có cấu trúc phản ánh mối liên hệ tự nhiên trong thế giới thực. Ví dụ, cá heo hiếm khi cùng xuất hiện với cỏ trong một hình ảnh, trong khi con dao sẽ xuất hiện bên cạnh một cái nĩa với khả năng rất cao (Lanchantin et al., 2021). Do đó, để có mô hình phân lớp hiệu quả ngoài việc trích xuất đặc trưng tốt để dự đoán các nhãn lớp của ảnh, cần phải khai thác được các mối quan hệ phức tạp và phụ thuộc lẫn nhau giữa đặc trưng ảnh với nhãn lớp, và giữa các nhãn lớp với nhau.

Ngày nay, đồ thị ngữ cảnh (*scene graph*) (Lu, Xiong, Parikh, & Socher, 2017) đã trở nên phổ biến trong việc biểu diễn ngữ nghĩa về ngữ cảnh và mối quan hệ giữa các đối tượng một cách chính xác và hiệu quả. Bên cạnh đó, GCN là sự mở rộng của CNN để tổng hợp thông tin từ dữ liệu cấu trúc đồ thị đã cho thấy tính hiệu quả trong việc khai thác mối quan hệ giữa các đối tượng và nhận được sự quan tâm ngày càng tăng của các nhóm nghiên cứu (Zhang, Tong, Xu, & Maciejewski, 2019). Do đó, trong bài báo này, một phương pháp phân lớp ảnh đa nhãn được đề xuất dựa trên đồ thị và mạng GCN để biểu diễn giàu ngữ nghĩa về mối quan hệ giữa các đối tượng trong ảnh, giữa các nhãn lớp trong tập dữ liệu cũng như sự phụ thuộc giữa chúng nhằm nâng cao độ chính xác. Đóng góp chính của bài báo gồm: (1) trích xuất đồ thị ngữ cảnh của hình ảnh và học biểu diễn đồ thị bằng mạng tích chập đồ thị nhằm khai thác đặc điểm và mối quan hệ ngữ nghĩa của các đối tượng trong ảnh; (2) đề xuất mô hình phân lớp ảnh đa nhãn dựa trên mạng tích chập đồ thị; (3) xây dựng ứng dụng thực nghiệm dựa trên mô hình và thuật toán đề xuất.



Hình 1. *Minh họa mô mình phân lớp ảnh đa nhãn*

Các công trình liên quan

Có nhiều phương pháp phân lớp ảnh đa nhãn đã được nhiều nhóm tác giả quan tâm nghiên cứu và công bố trong những năm gần đây như: phân lớp ảnh đa nhãn sử dụng khung CNN-RNN (Wang et al., 2016), phân lớp ảnh sử dụng đồ thị tri thức (Marino, Salakhutdinov, & Gupta, 2016), các phương pháp phân lớp ảnh đa nhãn sử dụng thuật giải di truyền (Gonçalves, Freitas, & Plastino, 2018), phân lớp ảnh đa nhãn sử dụng mạng học sâu bán giám sát (Cevikalp, Benligiray, & Gerek, 2020), học sự đồng xuất hiện của nhãn với mạng tích chập đồ thị cho phân loại ảnh X-quang phổi nhiều nhãn (Chen, Li, Lu, Yu, & Zhang, 2020)…

Feng Zhu và cộng sự (2017) (Zhu, Li, Ouyang, Yu, & Wang, 2017) đã giới thiệu phương pháp phân lớp ảnh đa nhãn dựa trên mạng nơ-ron học sâu bằng cách đề xuất mạng SRN (*Spatial Regularization Network*) phát sinh ra các bản đồ đặc trưng trội (*attention maps*)

cho các nhãn lớp trong ảnh, từ đó kết hợp với đặc trưng ảnh trích xuất từ mạng tích chập ResNet-101 để làm đầu vào cho việc huấn luyện mô hình phân lớp, toàn bộ mạng có thể huấn luyện từ đầu đến cuối (*end-to-end*). Tuy nhiên, công trình này có hạn chế là chỉ khai thác mối quan hệ giữa các đối tượng trong ảnh mà bỏ qua mối quan hệ giữa các nhãn lớp trong tập dữ liệu ảnh. Bên cạnh đó, việc sử dụng đặc trưng mức ảnh thay vì mức vùng cũng làm giảm tốc độ và độ chính xác.

Zhao-Min Chen và cộng sự (2019) (Chen, Wei, Wang, & Guo, 2019) đã xây dựng một mô hình phân lớp ảnh đa nhãn dựa trên đồ thị có hướng biểu diễn sự phụ thuộc giữa các nhãn trong tập dữ liệu, sau đó, một mạng tích chập đồ thị được sử dụng để học các bộ phân lớp cho các nhãn lớp của ảnh. Các bộ phân lớp này áp dụng cho đặc trưng ảnh được trích xuất bằng mạng CNN nhằm huấn luyện mạng theo phương pháp học có giám sát để dự đoán tập phân lớp ảnh đầu vào. Mô hình này có nhược điểm là chưa tận dụng được mối quan hệ ngữ nghĩa giữa các đối tượng nội tại trong từng hình ảnh để biểu diễn đặc trưng hình ảnh đầy đủ ngữ nghĩa nhằm tăng độ chính xác.

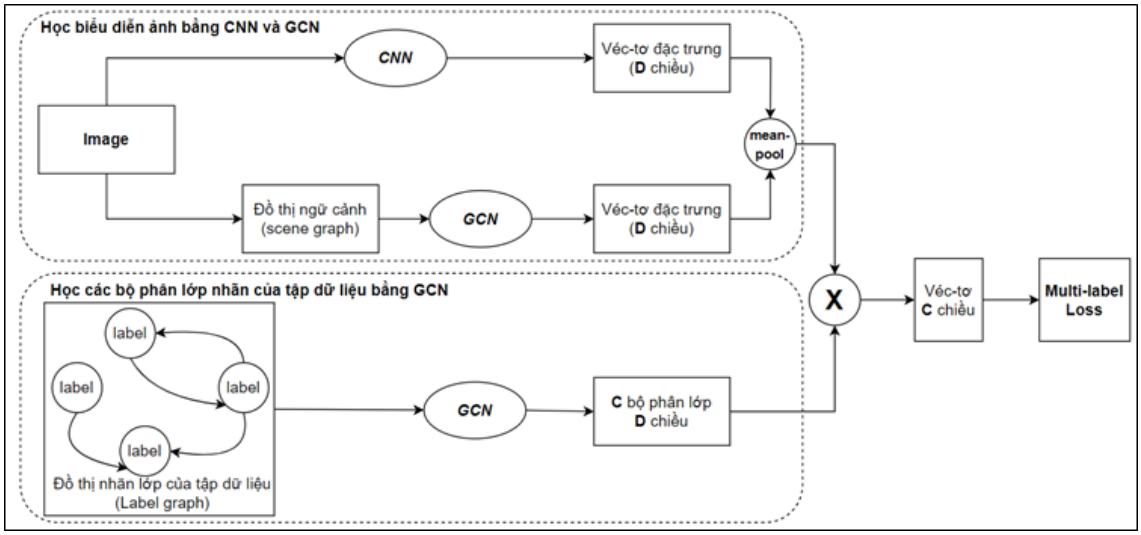
Ya Wang và cộng sự (2020) đã đề xuất phương pháp phân loại ảnh nhiều nhãn bằng cách xếp chồng đồ thị nhãn (*label graph*). Mối tương quan giữa các nhãn được mô tả bằng cách xếp chồng đồ thị nhãn (*được xây dựng từ thông tin thống kê đồng xuất hiện nhãn đã biết*), từ đó các lớp tích chập đồ thị được sử dụng để tạo ra các embedding cho nhãn. Bài báo này có nhược điểm là chưa khai thác mối quan hệ tương tác giữa các đối tượng trong ảnh để biểu diễn đặc trưng ảnh đầy đủ và chính xác hơn về ngữ nghĩa.

Từ các công trình đã công bố cho thấy bài toán phân lớp ảnh đa nhãn có nhiều sự quan tâm của các nhóm tác giả. Hơn nữa, việc áp dụng đồ thị để biểu diễn hình ảnh, sử dụng mạng tích chập đồ thị (GCN) để học biểu diễn đồ thị giúp nâng cao hiệu quả cho phân lớp ảnh (Milewski, Moens, & Calixto, 2020; Yang, Tang, Zhang, & Cai, 2019). Trên cơ sở kế thừa từ các công trình đã có và khắc phục những hạn chế của các phương pháp liên quan đã công bố, đồng thời tạo ra một hệ phân lớp ảnh hiệu quả, một phương pháp phân lớp ảnh đa nhãn dựa trên đồ thị và mạng tích chập đồ thị được đề xuất nhằm nâng cao độ chính xác.

2. Phương pháp nghiên cứu

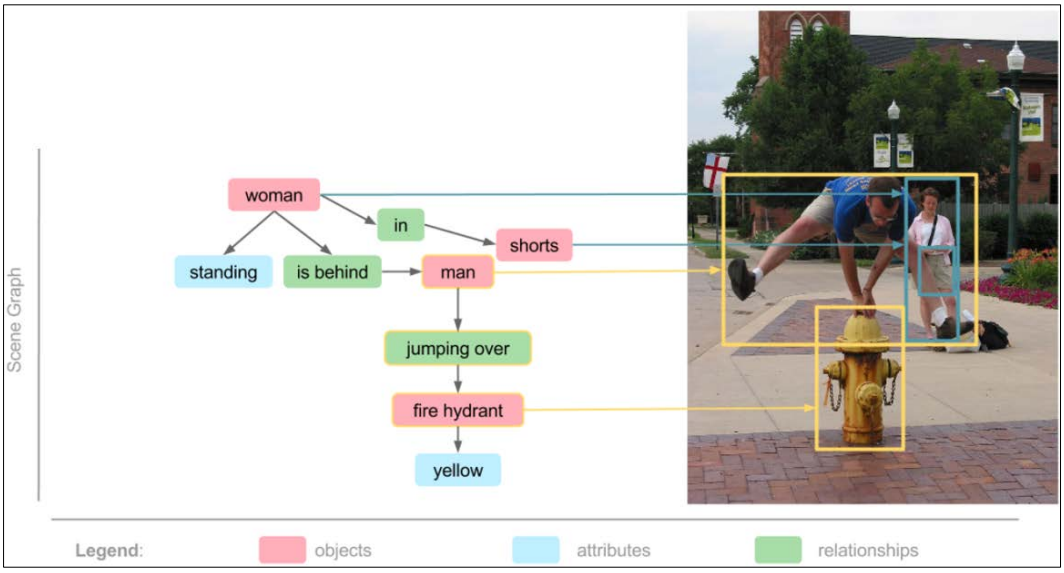
2.1. Phương pháp đề xuất

Mô hình tổng quát của phương pháp phân lớp ảnh đa nhãn đề xuất như ở Hình 2, gồm 3 phần chính. Một là học biểu diễn hình ảnh bằng mạng nơ-ron tích chập và mạng tích chập đồ thị dựa trên đồ thị ngữ cảnh nhằm tạo ra véc-tơ đặc trưng mô tả nội dung hình ảnh. Hai là xây dựng đồ thị từ các nhãn đối tượng trong tập dữ liệu ảnh để biểu diễn sự phụ thuộc giữa chúng, từ đó học các bộ phân lớp cho các nhãn tương ứng bằng mạng tích chập đồ thị. Ba là huấn luyện mạng sử dụng phương pháp phân lớp đa nhãn truyền thống. Các bước xử lý chính trong mô hình đề xuất gồm: học biểu diễn hình ảnh, xây dựng đồ thị nhãn và học bộ phân lớp cho các nhãn, huấn luyện mạng phân lớp đa nhãn.



Hình 2. *Mô hình tổng quát của phương pháp đề xuất*

Trong Hình 2, các nhãn đối tượng (*label*) được biểu diễn bởi các word embeding $Z \in \mathbb{R}^{Cxd}$ (C là số phân lớp của tập dữ liệu và d là số chiều của véc-tơ word embedding). Đồ thị có hướng (*label graph*) được xây dựng từ các biểu diễn của các nhãn này, mỗi đỉnh biểu diễn cho một nhãn. GCN được học để ánh xạ đồ thị nhãn này thành tập các bộ phân lớp đối tượng phụ thuộc lẫn nhau, nghĩa là $W \in \mathbb{R}^{CxD}$.



Hình 3. *Một ví dụ về đồ thị ngữ cảnh trong tập Visual Genome*

2.2.1. Giới thiệu đồ thị ngữ cảnh và mạng tích chập đồ thị

- *Đồ thị ngữ cảnh*

Đồ thị ngữ cảnh là một cấu trúc dữ liệu mô tả nội dung của một ngữ cảnh. Nó là sự biểu diễn dưới dạng cấu trúc của một ngữ cảnh mà có thể mô tả một cách tường minh các đối tượng, thuộc tính và mối quan hệ giữa các đối tượng trong ngữ cảnh đó. Một đồ thị ngữ cảnh bao gồm các đỉnh là các đối tượng trong ảnh, và các cạnh biểu diễn mối quan hệ giữa

các đối tượng này (Johnson et al., 2015). Đồ thị ngữ cảnh được ứng dụng nhiều trong các tác vụ thuộc lĩnh vực thị giác máy tính như phân lớp ảnh (*Image Classification*), tìm kiếm ảnh (*Image Retrieval*), chú thích ảnh (*Image Captioning*)… (Maheshwari, Chaudhry, & Vinay, 2021). Hình 3 minh họa một đồ thị ngữ cảnh tương ứng một ngữ cảnh trong tập dữ liệu Visual Genome. Trong đó objects là các đối tượng, attributes là các thuộc tính của đối tượng, relationships là các mối quan hệ giữa các đối tượng.

Trong bài báo này, cấu trúc đồ thị ngữ cảnh của hình ảnh được đề xuất là một đồ thị có hướng như sau:

Mỗi ảnh I có một đồ thị ngữ cảnh là một bộ $G = (V, E)$, trong đó V và E lần lượt là tập đỉnh và tập cạnh. Có 3 loại đỉnh gồm: đỉnh chứa đối tượng o, đỉnh chứa thuộc tính a và đỉnh chứa mối quan hệ r. Kí hiệu o_i là đối tượng thứ i, r_{ij} là mối quan hệ giữa đối tượng o_i và đối tượng o_j, $a_{i,k}$ là thuộc tính thứ k của đối tượng o_i. Mỗi đỉnh được biểu diễn bằng một véc-tơ D chiều. Tập cạnh E được xây dựng theo quy tắc:

- Nếu đối tượng o_i có thuộc tính $a_{i,k}$ thì tạo một cạnh có hướng từ $a_{i,k}$ đến o_i;
- Nếu có một mối quan hệ bộ ba $< o_i - r_{ij} - o_j >$ thì tạo 2 cạnh có hướng lần lượt từ o_i đến r_{ij} và r_{ij} đến o_j.

Ví dụ như đồ thị ngữ cảnh ở Hình 3 có các đối tượng: woman, man, shorts, fire hydrant; các mối quan hệ giữa đối tượng với thuộc tính: $<$ woman, standing $>$, $<$ fire hydrant, yellow $>$; các mối quan hệ giữa các đối tượng: $<$ woman, in, shorts $>$, $<$ woman, behind, man $>$, $<$ man, jumping over, fire hydrant $>$. Là một đồ thị có hướng gồm 9 đỉnh và 8 cạnh.

Mặc dù đồ thị ngữ cảnh diễn đạt được toàn bộ thông tin ngữ nghĩa của ngữ cảnh một cách chính xác và hiệu quả. Tuy nhiên, nó không phù hợp để làm đầu vào cho hầu hết các thuật toán được xây dựng để sử dụng thông tin ngữ nghĩa bởi bản chất không đồng nhất của nó và về sự phụ thuộc bắc cầu (Kumar, Aggarwal, Bathwal, & Singh, 2021). Do đó, cần phải chuyển đổi biểu diễn đồ thị này thành dạng tuyến tính sao cho bảo toàn được thông tin của đồ thị và có thể làm đầu vào cho các kiến trúc học.

- *Mạng tích chập đồ thị*

Mạng nơ-ron đồ thị (*GNN - Graph Neural Network*) (Scarselli, Gori, Tsoi, Hagenbuchner, & Monfardini, 2008) là sự kết hợp giữa dữ liệu cấu trúc đồ thị và mạng nơ-ron. Gần đây, GCN là sự mở rộng của CNN để tổng hợp thông tin từ dữ liệu cấu trúc đồ thị đã nhận được sự quan tâm ngày càng tăng của các nhóm nghiên cứu, đồng thời cũng là một kiến trúc GNN phổ biến nhất dựa trên khái niệm tích chập đồ thị (*graph convolutions*). Có 2 phương pháp thực hiện phép tích chập trên đồ thị là tích chập quang phổ (*spectral convolution*) (Defferrard, Bresson, & Vandergheynst, 2016) và tích chập không gian (*spatial convolution*) (Chen, Li, Fei-Fei, & Gupta, 2018; Monti et al., 2017). Tích chập quang phổ thực hiện bằng cách chuyển biểu diễn của node sang miền quang phổ (*spectral domain*) sử dụng phép biến đổi fourier đồ thị. Tích chập không gian thực hiện phép tích chập bằng cách

kết hợp các đỉnh láng giềng. Trong bài báo này, chúng tôi thực hiện học biểu diễn đặc trưng của đồ thị ngữ cảnh bằng mạng tích chập đồ thị sử dụng phép tích chập không gian trên đồ thị. Nó thực hiện một chuỗi các phép toán tích chập trên đồ thị, sau đó là lớp tổng hợp để gộp ngữ cảnh từ các đối tượng khác nhau có trong ảnh.

2.2. Học biểu diễn nội dung ảnh

Để biểu diễn đầy đủ các đối tượng và mối quan hệ ngữ nghĩa giữa các đối tượng và thuộc tính của chúng trong ảnh, chúng tôi đề xuất phương pháp học biểu diễn ảnh gồm 2 phần: rút trích đặc trưng bằng mạng CNN và học biểu diễn đồ thị ngữ cảnh của ảnh bằng mạng GCN.

- *Học biểu diễn ảnh bằng mạng CNN*

Đặc trưng của ảnh có thể học được bằng cách dùng bất kì mô hình mạng CNN nào. Tuy nhiên, trong thực nghiệm của bài báo này, mạng *ResNet-101* (He, Zhang, Ren, & Sun, 2016) được sử dụng. Do đó, nếu một ảnh I có kích thước 448x448, các bản đồ đặc trưng (*feature maps*) tại tầng (*layer*) conv5_x sẽ là 2048x14x14. Sau đó, thực hiện glogal max − pooling để đạt được đặc trưng của ảnh là e:

$$e = f_{GMP}(f_{cnn}(I;\theta_{cnn})) \in \mathbb{R}^D, \quad (1)$$

trong đó, θ_{cnn} cho biết các tham số của mô hình và $D = 2048$.

- *Học biểu diễn ảnh từ đồ thị ngữ cảnh dựa trên GCN*

Mỗi đỉnh v được biểu diễn như là một véc-tơ $x_v \epsilon R^D$, với D là số chiều. Các véc-tơ này được cập nhật bởi phép toán tích chập từ các láng giềng tương ứng của chúng. GCN nguyên thủy hoạt động trên đồ thị vô hướng, mã hóa thông tin về các láng giềng của đỉnh v dưới dạng một véc-tơ giá trị thực, được tính bằng công thức sau:

$$h_v^0 = x_v$$
$$h_v^l = \sigma\left(W^l \sum_{u \in N(v)} \frac{h_u^{(l-1)}}{|N(v)|}\right), \forall l \in \{0,1...,L-1\} \quad (2)$$
$$z_v = h_v^{(L)}$$

trong đó: h_v^0 là emdedding khởi tạo của đỉnh v, x_v là đặc trưng của đỉnh v, h_v^l là embedding của đỉnh v tại tầng (*layer*) l, N(v) là các đỉnh láng giềng của v (bao gồm v), L là tổng số tầng của mạng, σ là hàm kích hoạt (*ví dụ như ReLU*), z_v là embedding của đỉnh v sau L tầng của các kết hợp láng giềng, W^l là ma trận trọng số tại tầng L.

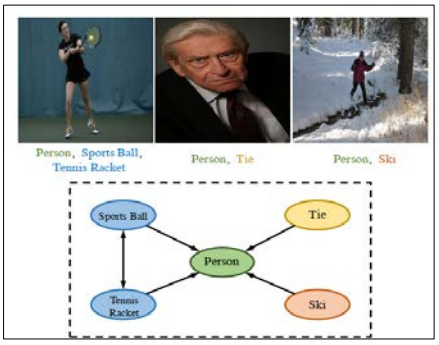
Với trường hợp đồ thị ngữ cảnh có hướng $G = (V, E)$ trong bài báo này, mỗi đỉnh v được mã hóa thông qua GCN như sau:

$$h_v^l = \sigma\left(W_{dir(u,v)}^l \sum_{u \in N(v)} \frac{h_u^{(l-1)}}{|N(v)|}\right), \forall l \in \{0,1...,L-1\} \quad (3)$$

Kết hợp với đặc trưng học từ mạng CNN ở công thức (1), đặc trưng của ảnh x là trung bình của e và đặc trưng tổng hợp trung bình của các đỉnh của đồ thị ngữ cảnh.

2.3. Học các bộ phân lớp nhãn dựa trên GCN

Trong phần này, phương pháp (Chen et al., 2019) được sử dụng để học các bộ phân lớp phụ thuộc lẫn nhau từ các biểu diễn nhãn cho trước trong tập dữ liệu bằng cách tạo đồ thị có hướng (*label graph*) mô tả sự phụ thuộc giữa các nhãn lớp trong tập dữ liệu, từ đó sử dụng GCN với cơ chế lan truyền thông điệp giữa các nhãn, và do đó học được các bộ phân lớp cho từng nhãn của hình ảnh. Một ví dụ về đồ thị phụ thuộc nhãn như trong Hình 4, trong đó "$\text{Label}_A \rightarrow \text{Label}_B$" nghĩa là khi có Label_A thì có khả năng có Label_B, nhưng điều ngược lại có thể thì không đúng.



Hình 4. *Một ví dụ về đồ thị mô tả sự phụ thuộc giữa các nhãn lớp*

Các bộ phân lớp đối tượng $W = \{w_i\}_{i=1}^{C}$ học từ các biểu diễn nhãn đối tượng thông qua một hàm ánh xạ dựa trên GCN, trong đó C là số phân lớp của tập dữ liệu. Sử dụng mạng GCN như ở phần 2.2, công thức (2) và (3), đầu vào của tầng thứ nhất là $Z \in R^{Cxd}$, với d là chiều của word embedding mức nhãn, đầu ra của tầng cuối cùng là $W \in R^{CxD}$ với D là số chiều biểu diễn ảnh.

2.4. Huấn luyện mạng phân lớp ảnh đa nhãn

Áp dụng các bộ phân lớp đã học cho đặc trưng biểu diễn ảnh, giá trị dự đoán có được là:

$$\hat{y} = Wx \tag{4}$$

Giả sử nhãn tập nhãn cho trước (*ground truth*) của ảnh là $y \in R^C$, trong đó $y^i = \{0,1\}$ cho biết nhãn i có trong ảnh hay không. Khi đó, toàn bộ mạng được huấn luyện sử dụng phương pháp mất mát phân lớp đa nhãn truyền thống như sau:

$$L = \sum_{c=1}^{C} y^c \log\left(\sigma(\hat{y}^c)\right) + (1 - y^c) \log\left(1 - \sigma(\hat{y}^c)\right), \tag{5}$$

trong đó, $\sigma(.)$ là hàm sigmoid.

3. Kết quả và thảo luận

3.1. Dữ liệu thực nghiệm

Phương pháp đề xuất được thực nghiệm trên tập dữ liệu ảnh Visual Genome (Krishna et al., 2017) gồm 108,077 ảnh và đồ thị ngữ cảnh của chúng. Để đánh giá hiệu suất phân lớp ảnh, chúng tôi sử dụng tập con gồm 51,498 ảnh, đây là phần giao của tập Visual Gennome và MS COCO (Lin et al., 2014). Tập ảnh này cũng được chọn lọc các loại đối tượng và mối quan hệ xuất hiện từ 25 lần trở lên, kết quả có 2416 đối tượng và 478 loại mối quan hệ. Đồng thời chỉ sử dụng các ảnh có từ 3 đến 40 đối tượng và có ít nhất một mối quan

hệ. Kết quả còn lại 45,358 ảnh, trong đó trung bình có 21 đối tượng và 15 mối quan hệ trên một ảnh. Tập dữ liệu cũng được chia thành các tập huấn luyện (train), kiểm định (validation) và kiểm tra (test) với tỉ lệ lần lượt là 70%, 20% và 10%.

3.2. Chi tiết cài đặt

Tập đồ thị ngữ cảnh của 45,358 ảnh thực nghiệm được trích xuất từ tập dữ liệu Visual Genome. Mô hình GCN gồm 2 tầng với số chiều đầu ra lần lượt là 1024 và 2048. Về phần biểu diễn nhãn đối tượng, GloVe (Pennington, Socher, & Manning, 2014) 300 chiều được chọn. Mạng ResNet-101 được sử dụng để trích xuất đặc trưng ảnh. Về huấn luyện mạng phân lớp, sử dụng SGD, momentum được thiết lập là 0.9, hệ số học là 0.001, mạng được huấn luyện thông qua 100 epochs.

3.3. Các độ đo đánh giá hiệu suất

Độ đo đánh giá sử dụng trong bài báo gồm độ chính xác trung bình theo từng phân lớp (CP), độ phủ (CR), độ dung hòa F1 (CF1) và độ chính xác trung bình trên tất cả phân lớp (OP), độ phủ (OR), độ dung hòa F1 (OF1).

3.4. Kết quả thực nghiệm

Thực nghiệm được thực thi trên máy PC CPU Intel Core i7-7200U 11800H @ 2.30GHz, 16.0GB RAM, hệ điều hành Windows 11 Pro 64 bit, sử dụng ngôn ngữ lập trình python và các thư viện Pytorch, Stellar graph. Kết quả thực nghiệm được đánh giá trên bộ dữ liệu là phần giao giữa tập dữ liệu Visual Genome và MS COCO như mô tả ở trên.

Hiệu suất của phương pháp được trình bày trong Bảng 1 với các độ đo mAP, CP, CR, CF1, OP, OR và OF có giá trị lần lượt là 84.7, 86.1, 72.8, 78.9, 88.6, 76.5 và 82.1. Các giá trị hiệu suất của phương pháp đề xuất được so sánh với các phương pháp khác trên cùng bộ dữ liệu được mô tả trong Bảng 2. Kết quả trong Bảng 2 cho thấy phương pháp đề xuất của chúng tôi tương đối chính xác, hầu hết các độ đo đều cao hơn các công trình khác. Từ kết quả này cho thấy phương pháp được đề xuất là khả thi và hiệu quả.

Bảng 1. *Hiệu suất phân lớp ảnh của phương pháp đề xuất trên tập dữ liệu thực nghiệm*

Tổng số ảnh	Số ảnh huấn luyện	Số ảnh kiểm định	Số ảnh kiểm tra	mAP	CP	CR	CF1	OP	OR	OF
45,358	31750	9071	4537	84.7	86.1	72.8	78.9	88.6	76.5	82.1

Bảng 2. *So sánh hiệu suất giữa các phương pháp trên tập dữ liệu thực nghiệm*

Phương pháp	mAP	CP	CR	CF1	OP	OR	OF
SRN (Zhu et al., 2017)	77.1	81.6	65.4	71.2	82.7	69.9	75.8
ML-GCN (Z.-M. Chen et al., 2019)	83.0	85.1	72.0	78.0	85.8	75.4	80.3
KSSNet (Y. Wang et al., 2020)	83.7	84.6	**73.2**	77.2	87.8	76.2	81.5
Đề xuất của bài báo	**84.7**	**86.1**	72.8	**78.9**	**88.6**	**76.5**	**82.1**

4. Kết luận

Trong bài báo này, một phương pháp phân lớp ảnh đa nhãn dựa trên mạng tích chập đồ thị đã được xây dựng. Một mô hình GCN để học biểu diễn đặc trưng của đồ thị ngữ cảnh và học bộ phân lớp phụ thuộc lẫn nhau của các nhãn đối tượng dựa trên đồ thị nhãn đối tượng được đưa ra làm cơ sở cho việc huấn luyện mạng phân lớp ảnh. Dựa trên phương pháp và mô hình đã đưa ra, thực nghiệm được xây dựng và đánh giá thông qua các độ đo mAP, CP, CR, CF1, OP, OR, OF. Kết quả thực nghiệm được so sánh với các phương pháp khác trên cùng một tập dữ liệu ảnh đã cho thấy phương pháp đề xuất tương đối hiệu quả. Thực nghiệm cũng cho thấy tính đúng đắn của phương pháp và các thuật toán đã giới thiệu, do đó phương pháp này có thể làm cơ sở cho việc phát triển các hệ thống phân lớp ảnh đa nhãn và ứng dụng trong thực tế. Hướng phát triển tiếp theo là xây dựng mạng học sâu để tạo tự động đồ thị ngữ cảnh từ ảnh đầu vào nhằm tạo thành một mạng huấn luyện từ đầu đến cuối, đồng thời kết hợp thêm tri thức về mối quan hệ giữa các đối tượng từ các cơ sở tri thức bên ngoài nhằm tăng độ chính xác.

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

TÀI LIỆU THAM KHẢO

Cevikalp, H., Benligiray, B., & Gerek, O. N. (2020). Semi-supervised robust deep neural networks for multi-label image classification. *Pattern Recognition, 100*, 107164.

Chen, B., Li, J., Lu, G., Yu, H., & Zhang, D. (2020). Label co-occurrence learning with graph convolutional networks for multi-label chest x-ray image classification. *IEEE journal of biomedical and health informatics, 24*(8), 2292-2302.

Chen, X., Li, L.-J., Fei-Fei, L., & Gupta, A. (2018). *Iterative visual reasoning beyond convolutions.* Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition.

Chen, Z. M., Wei, X. S., Wang, P., & Guo, Y. (2019). *Multi-label image recognition with graph convolutional networks.* Paper presented at the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.

Defferrard, M., Bresson, X., & Vandergheynst, P. (2016). Convolutional neural networks on graphs with fast localized spectral filtering. *Advances in neural information processing systems, 29*.

Ge, Z., Mahapatra, D., Sedai, S., Garnavi, R., & Chakravorty, R. (2018). Chest x-rays classification: A multi-label and fine-grained problem. *arXiv preprint arXiv:1807.07247*.

Gonçalves, E. C., Freitas, A. A., & Plastino, A. (2018). A survey of genetic algorithms for multi-label classification. Paper presented at the 2018 IEEE Congress on Evolutionary Computation (CEC).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition.

Johnson, J., Krishna, R., Stark, M., Li, L.-J., Shamma, D., Bernstein, M., & Fei-Fei, L. (2015). *Image retrieval using scene graphs.* Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., . . . Shamma, D. A. (2017). Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision, 123*(1), 32-73.

Kumar, V., Aggarwal, D., Bathwal, V., & Singh, S. (2021). *A Novel Approach to Scene Graph Vectorization.* Paper presented at the 2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS).

Lanchantin, J., Wang, T., Ordonez, V., & Qi, Y. (2021). *General multi-label image classification with transformers.* Paper presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Li, Y., Huang, C., Loy, C. C., & Tang, X. (2016). *Human attribute recognition by deep hierarchical contexts.* Paper presented at the European conference on computer vision.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., . . . Zitnick, C. L. (2014). *Microsoft coco: Common objects in context.* Paper presented at the European conference on computer vision.

Lu, J., Xiong, C., Parikh, D., & Socher, R. (2017). *Knowing when to look: Adaptive attention via a visual sentinel for image captioning.* Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition.

Maheshwari, P., Chaudhry, R., & Vinay, V. (2021). *Scene graph embeddings using relative similarity supervision.* Paper presented at the Proceedings of the AAAI Conference on Artificial Intelligence.

Marino, K., Salakhutdinov, R., & Gupta, A. (2016). The more you know: Using knowledge graphs for image classification. *arXiv preprint arXiv:1612.04844*.

Milewski, V., Moens, M.-F., & Calixto, I. (2020). Are scene graphs good enough to improve image captioning? *arXiv preprint arXiv:2009.12313*.

Monti, F., Boscaini, D., Masci, J., Rodola, E., Svoboda, J., & Bronstein, M. M. (2017). *Geometric deep learning on graphs and manifolds using mixture model cnns.* Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition.

Pennington, J., Socher, R., & Manning, C. D. (2014). *Glove: Global vectors for word representation.* Paper presented at the Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP).

Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., & Monfardini, G. (2008). The graph neural network model. *IEEE transactions on neural networks, 20*(1), 61-80.

Wang, J., Yang, Y., Mao, J., Huang, Z., Huang, C., & Xu, W. (2016). *Cnn-rnn: A unified framework for multi-label image classification.* Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition.

Wang, Y., He, D., Li, F., Long, X., Zhou, Z., Ma, J., & Wen, S. (2020). *Multi-label classification with label graph superimposing.* Paper presented at the Proceedings of the AAAI Conference on Artificial Intelligence.

Wei, X.-S., Cui, Q., Yang, L., Wang, P., & Liu, L. (2019). RPC: A large-scale retail product checkout dataset. *arXiv preprint arXiv:1901.07249*.

Yang, X., Tang, K., Zhang, H., & Cai, J. (2019). *Auto-encoding scene graphs for image captioning.* Paper presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Zhang, S., Tong, H., Xu, J., & Maciejewski, R. (2019). Graph convolutional networks: a comprehensive review. *Computational Social Networks, 6*(1), 1-23.

Zhu, F., Li, H., Ouyang, W., Yu, N., & Wang, X. (2017). *Learning spatial regularization with image-level supervisions for multi-label image classification.* Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition.

A MULTI-LABEL IMAGE CLASSIFICATION METHOD BASED ON GRAPH CONVOLUTIONAL NETWORK

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

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

[2]HCM City University of Foreign Languages – Information Technology, Vietnam

[*]Corresponding author: Nguyen Van Thinh – Email: thinhnv@hcmue.edu.vn

Received: October 18, 2022; Revised: February 21, 2023; Accepted: February 27, 2023

ABSTRACT

Multi-label image classification is one of the critical and challenging tasks in computer vision. In this paper, a multi-label image classification method is proposed based on the Graph Convolutional Network (GCN) to exploit the relationship between object labels in the dataset and between objects in the image to improve accuracy. First, the image content is representation learning by a convolutional neural network (CNN), and GCN relies on the scene graph of the image. Then, the graph describing the dependency between object labels in the dataset is built as the basis for learning classifiers for the labels using GCN and applying these classifiers to the image feature to generate predicted scores. Finally, the entire network is trained using the traditional multi-label classification loss. Experiments are built and evaluated on the dataset, which is the intersection between Visual Genome and MS COCO. The results show that the proposed method is effective and superior to some recently published works.

Keywords: convolutional neural network; graph convolutional network; label graph; multi-label image classification; scene graph