TRUY XUẤT HÌNH ẢNH SỬ DỤNG PHƯƠNG PHÁP KẾT HỢP XẾP HẠNG VÀ VISION TRANSFORMER

IMAGE RETRIEVAL USING COMBINATION METHOD OF RATINGS AND VISION TRANSFORMER

PHAN THƯỢNG CANG[1], ĐỖ THỊ NGỌC HIỀN[2,a], PHAN ANH CANG[2]

[1]Khoa Công nghệ thông tin, Trường Đại học Cần Thơ

[2] Trường Đại học Sư phạm Kỹ thuật Vĩnh Long

[a]Tác giả liên hệ: ngochiendtnh.h@gmail.com,

Nhận bài(Received): 17/7/2023; Phản biện (Reviewed):24/7/2023; Chấp nhận (Accepted):9/10/2023

Tóm tắt

Truy xuất hình ảnh là một nhiệm vụ quan trọng trong thị giác máy tính liên quan đến việc truy xuất các hình ảnh có liên quan từ cơ sở dữ liệu hình ảnh dựa trên nội dung trực quan của chúng. Phương pháp truyền thống có thể không nắm bắt được ý nghĩa ngữ nghĩa của hình ảnh một cách hiệu quả, dẫn đến kết quả truy xuất dưới mức tối ưu. Trong bài báo này, chúng tôi đề xuất phương pháp truy xuất hình ảnh theo nội dung (Content Based Images Retrieval - CBIR) dựa trên kiến trúc mạng Vision Transformer kết hợp phương pháp VisualRank để xếp hạng các hình ảnh dựa trên sự tương đồng của chúng với hình ảnh truy vấn. Qua kết quả đào tạo cho thấy mô hình đề xuất đạt độ chính xác lên đến 97%.

Từ khóa: Truy vấn ảnh theo nội dung, xếp hàng truy vấn ảnh, học chuyển tiền vision.

Abstract

Image retrieval is an important task in computer vision that involves retrieving relevant images from image databases based on their visual content. Traditional methods may not capture the semantic meaning of images effectively, leading to suboptimal retrieval results. In this paper, we propose a Content-Based Images Retrieval (CBIR) method based on Vision Transformer network architecture combined with the VisualRank method to rank images based on their similarity of images with query images. Through the training results, the proposed model has an accuracy of up to 97%.

Keywords: *Content-Based Images Retrieval, VisualRank, Vision Transformer..*

1. MỞ ĐẦU

1.1. Giới thiệu bài toán

Dữ liệu đa phương tiện, đặc biệt là ảnh số đã trở nên thân thuộc với cuộc sống hàng ngày và được sử dụng trên nhiều thiết bị khác nhau như camera, mobile, smartphone, tablet,… Theo báo cáo của IDC (International Data Corporation) năm 2015, thế giới đã tạo và chia sẻ hơn 1,6 nghìn tỷ hình ảnh, trong đó 70% hình ảnh được tạo ra từ thiết bị mobile [3]. Theo báo cáo mới nhất của Cisco vào năm 2021, dự kiến dung lượng dữ liệu gia tăng trong năm 2025 là 180 zettabyte [11]....Tìm kiếm hình ảnh tương tự từ các tập dữ liệu ảnh lớn là một bài toán quan trọng trong lĩnh vực thị giác máy tính [15], [2]. Việc thiết kế chỉ mục, xếp hạng hình ảnh, xây dựng cấu

trúc dữ liệu và đưa ra thuật toán tìm kiếm là trọng tâm của bài toán tìm kiếm dữ liệu ảnh [16]. Vấn đề đặt ra là xây dựng phương pháp tìm kiếm ảnh hiệu quả, nghĩa là tìm kiếm nhanh các hình ảnh tương tự trong một tập dữ liệu ảnh lớn với độ chính xác cao. Các kiến trúc mạng học sâu dựa trên mạng nơ-ron tích chập (CNN - Convolutional Nural Network) [17] đã liên tục được nghiên cứu cải tiến mang đến kết quả thử nghiệm ngày càng cao như Long Short-Term Memory [7], Vision Transformer [1], các kỹ thuật xếp hạng VisualRank [5], Multiclass VisualRank [9] góp phần không nhỏ trong các kết quả thử nghiệm. Các kết quả thử nghiệm được so sánh đánh giá để đề xuất một kiến trúc mạng học sâu phù hợp có chất lượng tốt nhất làm tiền đề cho việc xây dựng ứng dụng hỗ trợ truy vấn hình ảnh theo nội dung.

1.2. Những nghiên cứu liên quan

Rohit Raja và cộng sự [14] đã thực hiện một phương pháp tìm kiếm ảnh tương tự dựa vào đa đặc trưng của hình ảnh sử dụng vùng đặc trưng ROI (**Region of Interest**) theo màu sắc bằng phép lọc Sobel và Canny. Kết quả đầu ra của giai đoạn này được tiếp tục phân vùng trong không gian màu HSV. Kết quả thực nghiệm trên bộ ảnh COREL-1k, COREL- 5k đã cho độ chính xác là 87,33%. Công trình nghiên cứu của Nguyễn Hoài Nam [6] dựa trên cơ sở một số phương pháp tìm kiếm và xếp hạng trang cơ bản, từ đó đưa ra những đề xuất cải tiến cho thuật toán PageRank theo chủ đề. Phương pháp này đã được áp dụng thử nghiệm cho máy tìm kiếm Vietseek và bước đầu đã mang lại hiệu quả. Một nghiên cứu khác cũng về vấn đề xếp hạng là nghiên cứu về học xếp hạng trong tính hạng đối tượng và tạo nhãn cụm tài liệu của Nguyễn Thu Trang [12]. Các kết quả thu được đã chứng minh vai trò và hiệu quả của học xếp hạng áp dụng vào máy tìm kiếm. Nguyễn Hoàng Trung [10] đã tiến hành xây dựng thử nghiệm một thành phần tìm kiếm MP3 cho tiếng Việt cho máy tìm kiếm Socbay. Phần mềm tìm kiếm này cho kết quả tương đối chính xác đối với cả những tìm kiếm tiếng Việt không dấu và có dấu trong thời gian cho phép.

1.3. Đặc điểm của mô hình xếp hạng

Bộ xếp hạng chịu trách nhiệm tìm ra tài liệu thích hợp nhất từ truy vấn của người dùng và các tài liệu được đánh chỉ mục. Đặc điểm của mô hình xếp hạng là tạo ra một danh sách các tài liệu được xếp hạng theo mức độ liên quan giữa tài liệu và truy vấn. Sau đó sắp xếp tất cả các tài liệu theo thứ tự giảm dần theo mức độ liên quan của chúng.

1.4. Các độ đo đánh giá mô hình

Accuracy [11] được đánh giá dựa trên các giá trị như true positives (TP), false positives (FP), true negatives (TN) và false negatives (FN) được mô tả theo công thức (1). Hàm tính toán độ đo Loss (L) được xác định bởi công thức (2) với i là nhãn dữ liệu, c là giá trị thực tế và p là giá trị dự đoán. Giá trị chính xác là tỷ lệ trung bình giữa số lượng hình ảnh có liên quan trong hình ảnh N được truy xuất bởi sự giống nhau của từng hình ảnh q. Gọi tập hợp các phần tử có liên quan đến truy vấn $q \in Q$ là {d1, d2, …, dmj}, độ chính xác cho tất cả các truy vấn được tính bởi công thức (3).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$L(C, P) = \sum C_i \log(P_i) \quad (2)$$

$$mAP = \left(\frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{m_j}{N} \right) * 100 \quad (3)$$

2. KẾT QUẢ NGHIÊN CỨU

2.1. Mô hình mạng học sâu dùng trong huấn luyện và xếp hạng

a) Mạng Long Short-Term Memory (LSTM): Long Short-Term Memory [7] là một dạng đặc biệt của RNN, nó có khả năng học được các phụ thuộc xa. LSTM được giới thiệu bởi Hochreiter & Schmidhuber (1997), và sau đó đã được cải tiến. Nó được sử dụng để xử lý, dự đoán và phân loại trên cơ sở dữ liệu chuỗi thời gian. Chúng hoạt động cực kì hiệu quả trên nhiều bài toán khác nhau nên dần đã trở nên phổ biến như hiện nay.

b) Mạng Vision Transformer (ViT): Mô hình Vision Transformer (ViT) [1] là một mạng nơ-ron được sử dụng để xử lý và phân loại hình ảnh dựa trên cơ chế tự chú ý; cho phép xử lý hình ảnh bằng cách chia ảnh thành các patch nhỏ và xử lý chúng dưới dạng chuỗi mã thông báo. Điều này giúp mô hình nhận biết các đặc trưng không gian trong ảnh và mối quan hệ giữa các patch; quá trình xử lý thông tin được thực hiện dựa trên các phép chuyển đổi tuyến tính và cơ chế tự chú ý mà không yêu cầu các phép tích chập như các mạng CNN. Ngoài ra, Vision Transformer có thể mở rộng để xử lý ảnh có kích thước lớn hơn bằng cách tăng số lượng patch [5].

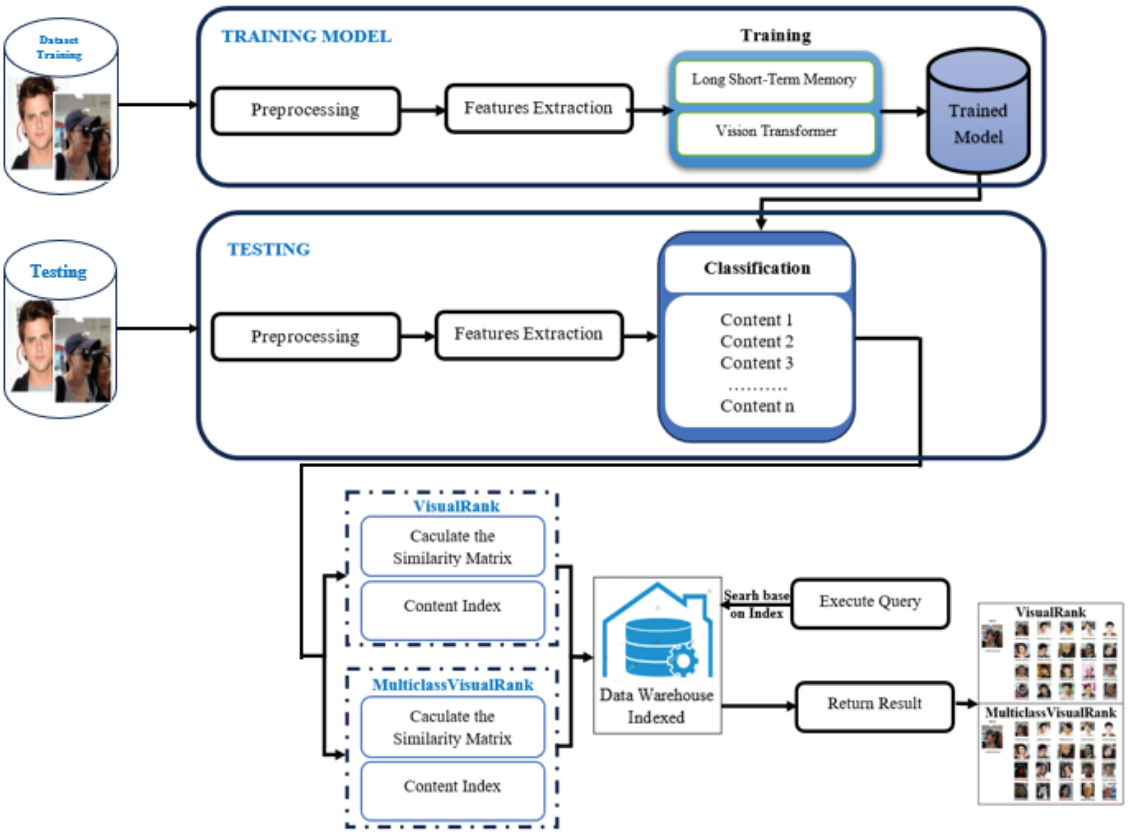
c) Phương pháp VisualRank: Phương pháp xếp hạng VisualRank là thuật toán tính hạng ảnh dựa vào việc phân tích độ tương đồng về nội dung giữa các bức ảnh do Yushi Jing và Shumeet Baluja [5][9] đề xuất. Phương pháp mà Jing và Baluja đưa ra lấy tư tưởng cơ bản từ thuật toán phân tích liên kết PageRank. Cũng giống như PageRank, thuật toán VisualRank sử dụng lý thuyết đồ thị để xây dựng đồ thị ảnh và dùng vector đặc trưng trung tâm để tính hạng cho các ảnh. Thuật toán đã được các tác giả thử nghiệm và cho kết quả tốt hơn kết quả xếp hạng của máy tìm kiếm ảnh Google trong phần lớn các truy vấn trong khi vẫn duy trì được hiệu quả tính toán hợp lý cho việc triển khai quy mô lớn.

d) Phương pháp Multiclass VisualRank: Phương pháp xếp hạng Multiclass VisualRank là thuật toán xếp hạng ảnh mở rộng ý tưởng từ phương pháp VisualRank của Jing và Baluja [5][9] để xếp hạng ảnh cho nhiều phân loại ảnh, do Misur Ambai và Yuichi Yoshida [14] đề xuất. Thuật toán Multiclass VisualRank chia các ảnh được trả về từ máy tìm kiếm thành những phân loại khác nhau dựa vào các đặc trưng nội dung của ảnh và tiến hành xếp hạng trong từng phân loại đó. Multiclass VisualRank gồm ba bước sau: Tính độ tương đồng về nội dung ảnh, phân cụm, tính hạng.

2.2. Phương pháp đề xuất

Hệ thống tìm kiếm hình ảnh dựa vào nội dung mà chúng tôi đề xuất có quy trình thực hiện như sau: Từ file hình ảnh đầu vào thực hiện cắt ra thành các frame hình, tiếp theo chúng tôi sẽ phát hiện đối tượng (nội dung) trên hình ảnh đó. Kết thúc quá trình xử lý dữ liệu nguồn, kết quả thu được là các đặc trưng tương ứng đối với nội dung của hình ảnh đã được trích xuất. Các thông tin rút trích được sẽ được lập chỉ mục và lưu vào cơ sở dữ liệu của công cụ tìm kiếm để phục vụ cho quá trình tiếp theo là truy vấn. Bộ xử lý truy vấn nhận các truy vấn của người dùng và bộ xếp hạng sẽ tìm ra đặc trưng thích hợp nhất từ truy vấn của người dùng và các đặc trưng được đánh chỉ mục. Bộ xếp hạng có thể lấy trực tiếp các truy vấn và các đặc trưng để tính toán một điểm số (score) hoặc cũng

có thể trích xuất những đặc điểm giữa các cặp đặc trưng và truy vấn để tạo ra điểm số được kết hợp từ những đặc điểm đó, và kết quả trả về cho người dùng là những hình ảnh liên quan được xếp hạng dựa theo thuật toán mà công cụ tìm kiếm sử dụng. Chi tiết các giai đoạn thực hiện của hệ thống tìm kiếm hình ảnh dựa vào nội dung mà chúng tôi đề xuất được miêu tả như hình 1.

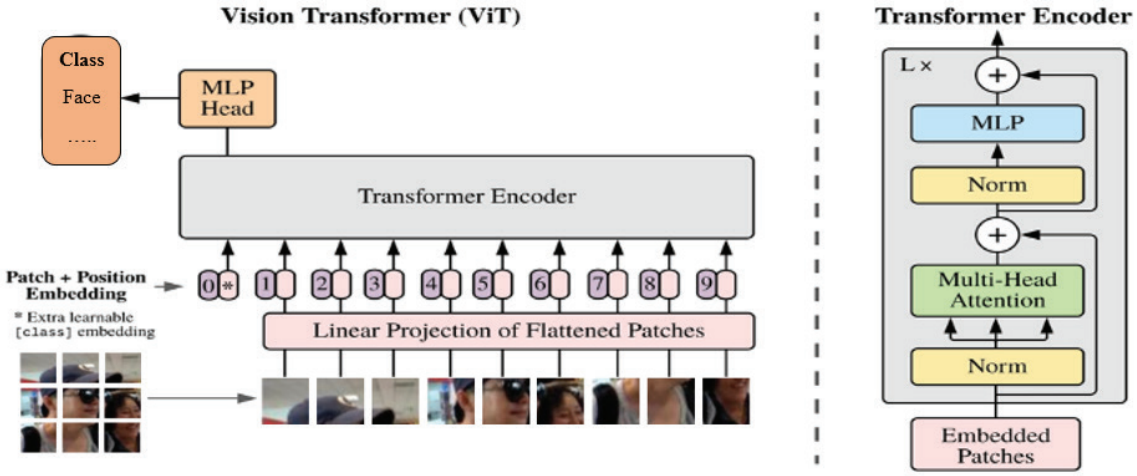


Hình 1. Mô hình tổng quát phương pháp đề xuất

2.2.1. Đào tạo mô hình

a) Tiền xử lý và chuẩn bị dữ liệu: Đầu vào của hệ thống là tập dữ liệu hình ảnh có sẵn, chúng tôi thực hiện chuẩn hóa tất cả hình ảnh thành kích thước 224 x 224 pixel, đây là kích thước phù hợp cho quá trình trích xuất đặc trưng tiếp theo. Chúng tôi chia dữ liệu thành các tệp train/test và đưa vào mô hình để đánh giá.

b) Trích chọn đặc trưng, huấn luyện: Để có thể phát hiện và phân loại được đối tượng, chúng tôi tiến hành rút trích đặc trưng trên tập dữ liệu đã được tiền xử lý. Chúng tôi đề xuất phương pháp rút trích đặc trưng với 2 mô hình mạng: Long Short-Term Memory và Vision Transformer. Tập dữ liệu sau khi rút trích đặc trưng sẽ được huấn luyện trên hai mô hình mạng Long Short-Term Memory(a) và Vision Transformer(b). Đặc biệt với mô hình Vision Transfomer [1] chúng tôi thực hiện bằng cách sau khi hình ảnh đầu vào được tiền xử lý sẽ cắt thành các patch đưa qua các tầng để tiến hành mã hóa, rút trích đặc trưng, phân loại và xếp hạng. Kết quả của quá trình này sẽ là ảnh được phân loại, lưu trữ và xếp hạng.



Hình 2. Mô hình đề xuất Vision Tranformer

2.2.2. Kiểm thử

Ở giai đoạn kiểm thử chúng tôi cũng tiến hành trích xuất đặc trưng với 2 mô hình mạng: Long Short-Term Memory và Vision Transformer. Dựa vào các thuật toán phân loại và xếp hạng như đã trình bày ở phần 1 và CSDL huấn luyện, ta tiến hành đưa ảnh đầu vào để xác nhận đối tượng và đưa ra kết quả. Đặt một ngưỡng để quy định độ chính xác khi nhận dạng, nếu lớn hơn ngưỡng này tức là đối tượng này tồn tại trong CSDL huấn luyện kết quả trả về sẽ là một id của đối tượng đó.

2.3. Kết quả thực nghiệm

a) Môi trường cài đặt: Phương pháp đề xuất được cài đặt trên môi trường Google Colab, cấu hình RAM 12GB và dùng GPU Nvidia Geforce. Thư viện hỗ trợ huấn luyện mô hình mạng sử dụng là Tensorflow, Keras và OpenCV.

b) Tập dữ liệu thực nghiệm: Để đánh giá phương pháp đề xuất chúng tôi sử dụng tập hình ảnh lấy từ bộ dữ liệu VGGFace2. Tập dữ liệu gồm 1,539 ảnh với 38 bộ phân lớp được chia theo tỉ lệ 80% (1231 ảnh) cho tập Training và 20% (308 ảnh) cho tập Test.

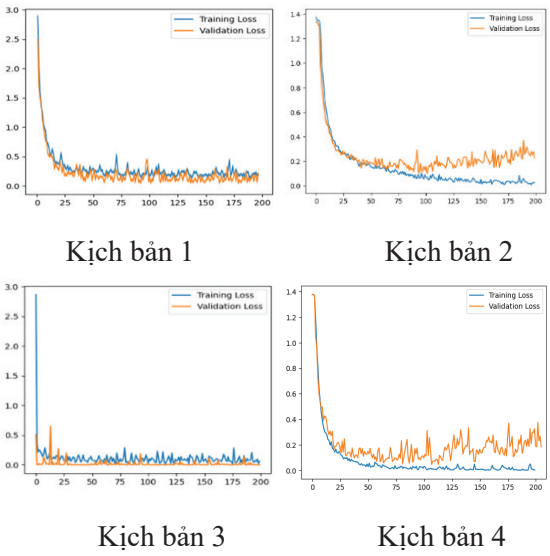
c) Các kịch bản áp dụng: Để so sánh và đánh giá các mô hình, chúng tôi thực hiện 4 kịch bản với các tham số trong Bảng 1.

Bảng 1. *Các kịch bản được đề xuất và các tham số huấn luyện*

Kịch bản	Kiến trúc mạng rút trích và huấn luyện	Phương pháp xếp hạng	Epochs	Learning rate	Patch_size	Image size
1	Vision Tranformer	VisualRank	200	1e-3	10	224 x 224
2	Vision Tranformer	MultiClass VisualRank	200	1e-3	10	224 x 224
3	Long Short-Term Memory	VisualRank	200	1e-3	10	224 x 224
4	Long Short-Term Memory	MultiClass VisualRank	200	1e-3	10	224 x 224

2.4. Kết quả huấn luyện

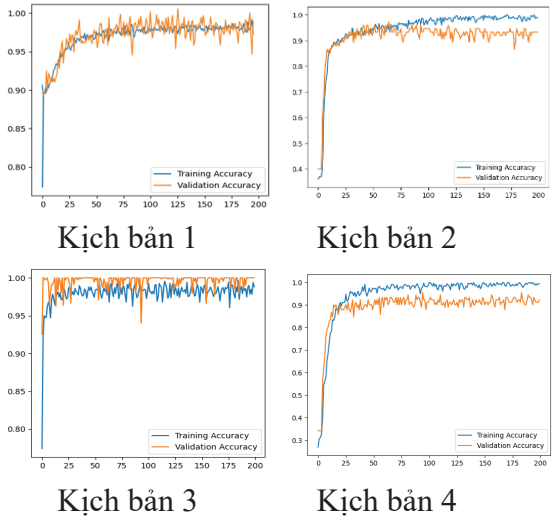
a) Độ đo Loss của các kịch bản



Hình 3 - ***Độ đo Loss của các kịch bản trong giai đoạn huấn luyện***

Hình 3 biểu diễn giá trị mất mác (loss) của các kịch bản trong giai đoạn huấn luyện. Giá trị val loss của các kịch bản lần lượt là kịch bản 1 là 0.0058, kịch bản 2 là 0.4836, kịch bản 3 là 0.0062, kịch bản 4 là 0.3662. Ở kịch bản 2 và 4, giá trị val_loss cho kết quả cao hơn giá trị train loss và không ổn định, đều này cho thấy mô hình huấn luyện chưa tối ưu và có thể dẫn đến sai lệch trong quá trình dự đoán. Ngược lại, kịch bản 1 và 3 lại cho kết quả tối ưu hơn so với các kịch bản trên, khi giá trị val loss gần bằng với giá trị train loss đối với kịch bản 3, và giá trị val loss thấp hơn so với giá trị train loss đối với kịch bản 1. Tuy nhiên, giá trị val loss trùng với train loss cũng có thể là do vấn đề underfitting, trong khi đó kịch bản 1 lại có giá trị val loss thấp hơn so với giá trị train loss. Điều này cho thấy mô hình trong kịch bản 1 cho kết quả tốt hơn các mô hình trong các kịch bản còn lại.

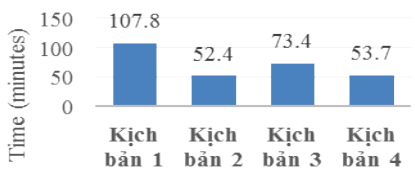
b) Độ đo Accuacy của các kịch bản



Hình 4 - Độ đo Accuracy của các kịch bản trong giai đoạn huấn luyện.

Hình 4 biểu diễn độ chính xác (accuracy) của các kịch bản trong giai đoạn huấn luyện. Kịch bản 1 đạt độ chính xác 98.3%, kịch bản 2 có độ chính xác là 88.5%. Kịch bản 3 có độ chính xác là 94.8%, kịch bản 4 có độ chính xác là 85.8%. Mặt khác, khi quan sát kịch bản 2 và 4 ta thấy rằng, giá trị val accuracy cho kết quả thấp hơn giá trị train accuracy và không ổn định, đều này cho thấy mô hình huấn luyện chưa thật sự hiệu quả và có thể dẫn đến sai lệch trong quá trình dự đoán. Ngược lại, kịch bản 1 và 3 lại cho kết quả tối ưu hơn so với các kịch bản trên khi giá trị val accuracy gần bằng với giá trị train accuracy đối với kịch bản 1 và giá trị val accuracy cao hơn so với giá trị train accuracy đối với kịch bản 3. Điều này cho thấy mô hình trong kịch bản 1 có khả năng dự đoán tốt hơn các mô hình trong ba kịch bản còn lại.

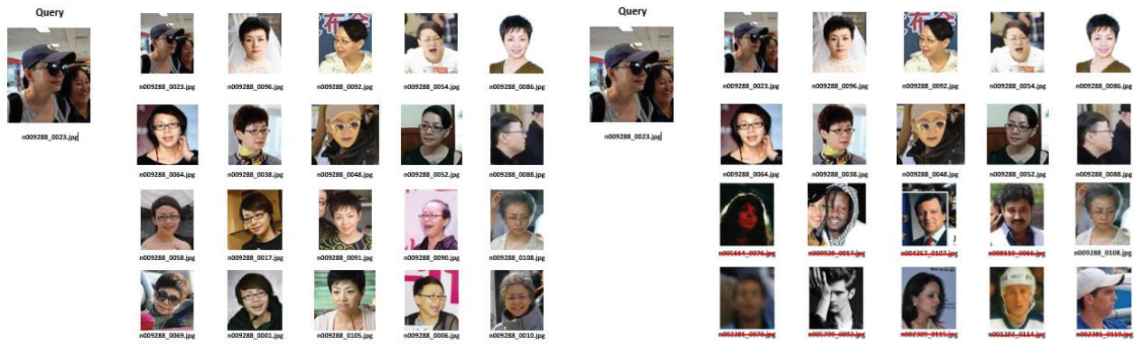
c) Thời gian huấn luyện



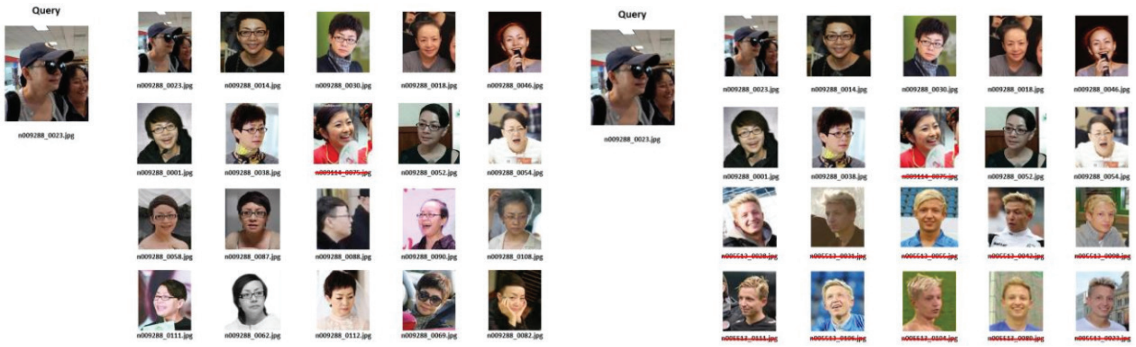
Hình 5. Thời gian huấn luyện của các kịch bản

Hình 5 biểu diễn thời gian huấn luyện của các kịch bản đề xuất lần lượt với kịch bản 1 là 108.7 phút, kịch bản 2 và 4 lần lượt là 52.4 phút và 53.7 phút, kịch bản 3 là 73.4 phút. Qua kết quả trên, có thể thấy kịch bản 2 và 4 cho kết quả huấn luyện nhanh hơn so với hai mô hình còn lại, mô hình kịch bản 1 cho kết quả huấn luyện lâu nhất, với thời gian gấp gần 2 lần so với kịch bản 2 và 4, gần 1.5 lần so với kịch bản 3.

d) Một số hình ảnh kết quả phân loại



a) Kịch bản 1 b) Kịch bản 2

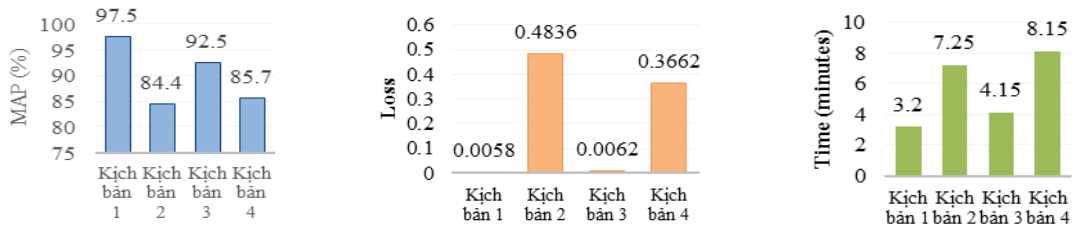


c) Kịch bản 3 d) Kịch bản 4

Hình 6 - Một số kết quả thực nghiệm của các kịch bản

Trong phần thực nghiệm này, chúng tôi sử dụng mô hình ViT kết hợp VisualRank (kịch bản 1), mô hình ViT kết hợp MulticlassVisualRank (kịch bản 2), mô hình LSTM kết hợp VisualRank (kịch bản 3), mô hình LSTM kết hợp MulticlassVisualRank (kịch bản 4) để xếp hạng tương ứng và hình 6 là một số hình ảnh minh họa. Kịch bản 1 và 3 cho kết quả truy vấn với độ chính xác cao, các kịch bản còn lại tuy nhận diện được hình ảnh truy vấn nhưng với tỉ lệ tương đối thấp, các hình ảnh được trả lại có tỷ lệ lỗi rất cao, thậm chí lên tới 100% kịch bản 2 và 4 là một ví dụ.

2.5. So sánh, đánh giá các mô hình



Hình 7. Biểu đồ mAP, Loss, thời gian kiểm thử của các kịch bản trên tập dữ liệu thực nghiệm

Kết quả thực nghiệm được tổng hợp qua biểu đồ hình 7 cho thấy giá trị loss của mô hình kịch bản 1 là 0.0058, thấp nhất trong bốn kịch bản. Thời gian kiểm thử ngắn nhất với 3.2 phút, nhanh hơn gấp 2 lần so với kịch bản 2 và gần 2,5 lần so với kịch bản 4. Ngoài ra mAP của 3 kịch bản cho thấy độ chính xác trung bình tương đối cao, tuy nhiên ở kịch bản 1 cho thấy tỉ lệ độ chính xác trung bình là cao nhất, hơn 97% so với kịch bản 2 là 84.4%, kịch bản 3 là 92.5% và kịch bản 4 là 85.7%. Từ kết quả trên có thể thấy mô hình kịch bản 1 sử dụng VisualRankVision Transformer nhận diện, phân loại và xếp hạng truy vấn hiệu quả hơn.

Bảng 2. *So sánh kết quả của ViT và các phương pháp khác*

Kịch bản	Kiến trúc mạng rút trích và huấn luyện	Phương pháp xếp hạng	Độ chính xác
1	Vision Tranformer	VisualRank	98.3%
2	Vision Tranformer	MultiClass VisualRank	88.5%
3	Long Short-Term Memory	VisualRank	94.8%
4	Long Short-Term Memory	MultiClass VisualRank	85.8%

3. KẾT LUẬN

Trong nghiên cứu này, chúng tôi đã trình bày phương pháp sử dụng các mô hình học sâu kết hợp thuật toán tính hạng để biểu diễn các thuộc tính nội dung có trong hình ảnh trên tập các ảnh lấy ngẫu nhiên từ tập dữ liệu có sẵn. Khi thực hiện và huấn luyện bằng các mô hình học sâu cụ thể là kiến trúc mạng VissualRank ViT, MulticlassVissualRank ViT, VissualRank LSTM, MulticlassVissualRank LSTM ta thấy được là phương pháp này đều đạt độ chính xác cao từ 87% đến 98% cho các mô hình kiến trúc. Đồng thời phương pháp này cũng chứng minh được sự thay đổi của độ chính xác tùy theo bộ tham số sử dụng. Các kết quả thực nghiệm cũng thể hiện tính khả thi của phương pháp khi áp dụng trên các công cụ tìm kiếm, cả về độ chính xác và thời gian thực hiện. Phương pháp có ưu điểm là tận dụng được khả năng xử lý của mạng nơ ron sâu cho cả thao tác trích xuất đặc trưng và phân loại đối tượng. Tuy nhiên, nhược điểm của phương pháp là cần được thực hiện tối ưu nhằm tìm ra bộ tham số tốt nhất do việc huấn luyện trên mạng nơ ron sâu là một hoạt động tiêu tốn tài nguyên và thời gian, nghiên cứu này chưa thực hiện tối ưu một cách triệt để các tham số của thuật toán. Hướng phát triển trong tương lai có thể tập trung vào việc giải quyết những hạn chế này bằng cách kết hợp các tính năng bổ sung, khám phá các kiến trúc mới và kết hợp dữ liệu từ nhiều nguồn, chẳng hạn như kết hợp truy vấn và xếp hạng cho video nhằm giảm việc tiêu tốn tài nguyên và thời gian.

TÀI LIỆU THAM KHẢO

[1] Abbas A. H., Mirza N. M., Qassir S. A., & Abbas L. H. - Maize leaf images segmentation using color threshold and K-means clustering methods to identify the percentage of the affected areas, In IOP Conference Series: Materials Science and Engineering **745** (1) (2020, February), 012048, IOP Publishing.

[2] ACI. http://www.aci.aero/ (2015)

[3] C. Chute - Worldwide Digital Image 2015–2019 Forecast: The Image Capture and Share Bible, International Data Corporation. (2015) p.13.

[4] Cyril Goutte, Eric Gaussier, "A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation", *In: European Conference on Information Retrieval (ECIR)*, 2005

[5] Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv 2020, arXiv:2010.11929.

[6] Gao Huang, Zhuang Liu, Laurens van der Maaten, Kilian Q. Weinberger, "Densely Connected Convolutional Networks," arXiv:1608.06993, 2016.

[7] Han L., Tian Y., Qi Q. - Research on edge detection algorithm based on improved sobel operator, In MATEC Web of Conferences: EDP Sciences (309) (2020) 03031.

[8] Hochreiter, S. and Schmidhuber, J., 1997. Long short-term memory. Neural computation, 9(8), pp.1735-1780

[9] Hong Hui Tan, King Hann Lim, "Vanishing Gradient Mitigation with Deep Learning Neural Network Optimization," In: 2019 7th International Conference on Smart Computing & Communications (ICSCC), 2019.

[10] https://arxiv.org/pdf/2010.11929.pdf

[11]https://cisco.com/c/en/us/solutions/collateral/executive-perspectives/annualinternetreport/white-paper-c11-741490.html

[12] https://www.pluralsight.com/guides/introduction-to-lstm-units-in-rnn

[13] https://paperswithcode.com/datasets

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, "Deep Residual Learning for Image Recognition," **arXiv:1512.03385,** 2015.