PHÂN TÍCH MỘT SỐ MÔ HÌNH TRÍCH XUẤT ĐẶC TRƯNG ẢNH CHO BÀI TOÁN TÌM KIẾM ẢNH DỰA TRÊN NỘI DUNG

Trần Văn Khánh [1, 2*], Nguyễn Ngọc Thuỷ [1], Lê Mạnh Thạnh [1]

[1] Trường Đại học Khoa học, Đại học Huế

[2] Trường Đại học Khánh Hoà

* Email: tvkhanh.dhkh24@hueuni.edu.vn

Ngày nhận bài: 13/12/2024; ngày hoàn thành phản biện: 19/12/2024; ngày duyệt đăng: 20/12/2024

TÓM TẮT

Bài báo tập trung phân tích và so sánh hiệu quả của một số mô hình học sâu cho bài toán CBIR thông qua các tiêu chí đánh giá như độ chính xác, khả năng tổng quát hóa. Các mô hình học sâu được sử dụng trong thực nghiệm như VGG16, Resnet50, EfficientNetB0, Densenet201 và Vision Transformer (ViT), là các mô hình đã được huấn luyện trọng số trên tập dữ liệu lớn Imagenet. Trên cơ sở các mô hình đã đề xuất, thực nghiệm được tiến hành trên bốn tập dữ liệu Corel1K, Oxford17Flowers, Caltech101 và một phần của tập Cifar10. Kết quả thực nghiệm cho thấy rõ hiệu quả vượt trội của các mô hình hiện đại trong quá trình trích xuất đặc trưng ảnh cho bài toán CBIR. ViT luôn đạt độ chính xác cao nhất trên tất cả các tập dữ liệu, gần như tuyệt đối trên hai Corel1K, Oxford17Flowers.

Từ khoá: CBIR, CNN, Học sâu, Transformer.

1. MỞ ĐẦU

Ngày nay, các tập dữ liệu chứa hình ảnh kỹ thuật số ngày càng nhiều và việc tìm kiếm hình ảnh chính xác, nhanh chóng trở thành một nhu cầu cấp thiết trong mọi lĩnh vực của đời sống. Tìm kiếm hình ảnh dựa trên nội dung ra đời như một giải pháp quan trọng, thay thế các phương pháp tìm kiếm hình ảnh dựa trên văn bản vốn còn tồn tại nhiều nhược điểm [1]. CBIR sử dụng các đặc trưng cấp thấp của hình ảnh như màu sắc, kết cấu và hình dạng trong tìm kiếm. Tuy nhiên, các phương pháp CBIR truyền thống vẫn còn nhiều hạn chế. Phương pháp dựa vào màu sắc, kết cấu thường không đủ chính xác khi xử lý hình ảnh có nội dung phức tạp hoặc trong những trường hợp ánh sáng, góc chụp khác nhau. Tương tự, các phương pháp sử dụng đặc trưng hình dạng cũng gặp khó khăn trong việc biểu diễn các chi tiết phức tạp và phụ thuộc nhiều vào quá trình tiền xử lý dữ liệu [2]. Các hạn chế này dẫn đến hiệu quả của hệ thống tìm kiếm chưa đáp

ứng được kỳ vọng, đặc biệt khi khối lượng dữ liệu ngày càng lớn và đa dạng. Điều này đòi hỏi sự cải tiến cả về phương pháp trích xuất, biểu diễn đặc trưng lẫn khả năng tiền xử lý hình ảnh.

Các mô hình học sâu ra đời đã trở thành công cụ mạnh mẽ trong lĩnh vực xử lý hình ảnh, mang lại những cải tiến vượt bậc cho bài toán CBIR. Không giống như các phương pháp truyền thống các đặc trưng được trích xuất thủ công, đối với mô hình học sâu quá trình trích xuất và biểu diễn đặc trưng hoàn toàn tự động từ dữ liệu đầu vào. Mô hình mạng nơ-ron tích chập đã chứng minh khả năng vượt trội trong việc trích xuất các đặc trưng hình ảnh từ nhiều cấp độ, từ thông tin chi tiết đến các cấu trúc trừu tượng [3]. Bên cạnh đó, các kiến trúc hiện đại như Resnet, EfficientNet, DenseNet và gần đây là ViT không chỉ cải thiện độ chính xác mà còn tối ưu hóa tài nguyên tính toán, mở ra khả năng ứng dụng của bài toán CBIR trong thực tế. Điểm nổi bật của các mô hình học sâu là khả năng tổng quát hóa trên nhiều tập dữ liệu, giúp cải thiện hiệu quả tìm kiếm ngay cả khi dữ liệu đầu vào đa dạng về nguồn gốc và nội dung [4]. Ngoài ra, các mô hình học sâu còn hỗ trợ xử lý hình ảnh ở quy mô lớn thông qua việc sử dụng GPU và các nền tảng tính toán đám mây, góp phần tăng tốc quá trình huấn luyện, suy luận và triển khai hệ thống CBIR.

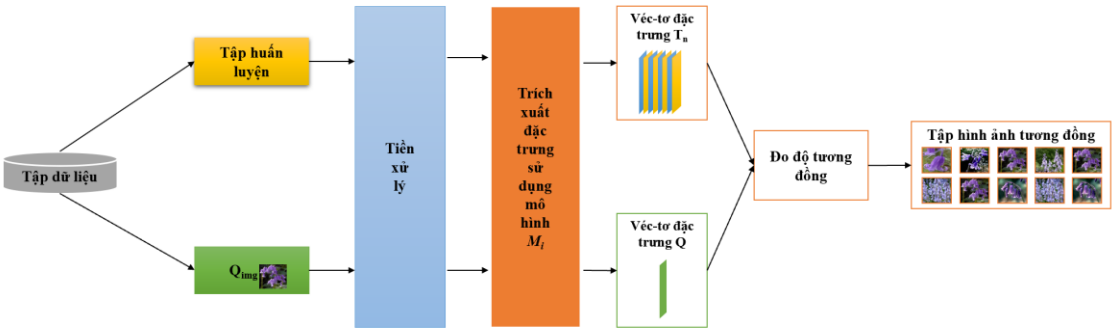
Với bài báo này, chúng tôi tập trung phân tích các mô hình học sâu phổ biến sử dụng cho quá trình trích xuất đặc trưng hình ảnh trong bài toán CBIR. Mục tiêu chính là cung cấp cái nhìn toàn diện về các kiến trúc học sâu, từ các mô hình cổ điển như VGG16 (VGG) đến các mô hình hiện đại như Resnet50 (RES), EfficientNetB0 (EFF), DenseNet201 (DEN) và cuối cùng là ViT. Đồng thời, bài viết đánh giá hiệu quả của các mô hình thông qua các bộ dữ liệu chuẩn, sử dụng các tiêu chí như độ chính xác, khả năng tổng quát hóa.

2. PHƯƠNG PHÁP NGHIÊN CỨU

2.1. CBIR

CBIR là phương pháp tìm kiếm ảnh sử dụng các đặc trưng hình ảnh như màu sắc, kết cấu, hình dạng để thực hiện quá trình tìm kiếm. Phương pháp này được áp dụng rộng rãi trong nhiều lĩnh vực như tìm kiếm sản phẩm trực quan và chẩn đoán y khoa. Qui trình hoạt động của CBIR được trình bày trong hình ***Hình 1***, bao gồm các bước chính sau: ***Tiền xử lý ảnh*** là quá trình xử lý ảnh đầu vào để giảm nhiễu, cân bằng màu sắc và chuẩn hóa kích thước trước khi trích xuất đặc trưng. Các kỹ thuật phổ biến bao gồm chuyển đổi không gian màu và lọc nhiễu bằng Gaussian hoặc Median; ***Trích xuất đặc trưng*** là bước quan trọng trong bài toán CBIR giúp chuyển đổi thông tin hình ảnh từ dạng thô sang dạng số hoá có thể sử dụng để so sánh và tìm kiếm. Trong bài toán CBIR, các đặc trưng thường được sử dụng bao gồm màu sắc, kết cấu và hình dạng. Bên cạnh

đó, các đặc trưng trích xuất từ mô hình học sâu cũng được áp dụng rộng rãi để nâng cao hiệu quả tìm kiếm; ***Đo độ tương đồng và trả về kết quả*** là quá trình so sánh hình ảnh tìm kiếm với các ảnh trong tập dữ liệu bằng các phép đo khoảng cách như khoảng cách Euclidean, Manhattan hoặc độ tương đồng Cosine. Hệ thống sau đó xếp hạng các hình ảnh theo mức độ tương đồng và hiển thị kết quả tìm kiếm phù hợp nhất.



Hình 1. Quy trình hoạt động của một hệ thống CBIR.

2.2. Các mô hình học sâu

Chúng tôi sử dụng các mô hình học sâu được phát triển thông qua các thời kì cho quá trình trích xuất đặc trưng hình ảnh, từ đó so sánh và đánh giá hiệu quả của các mô hình trên các tập dữ liệu cụ thể. Các mô hình VGG, RES, EFF, DEN, ViT mô tả cấu trúc trong ***Bảng 1*** được sử dụng trong thực nghiệm, các mô hình này đã được huấn luyện trên tập dữ liệu lớn ImageNet.

Mô hình VGG: Được đề xuất bởi Simonyan and Zisserman năm 2014 [5], chứa các khối với số lượng lớp tích chập tăng dần. Mô hình nổi bật trong việc sử dụng các lớp tích chập nhỏ 3x3 giúp tăng độ sâu mà không làm tăng số lượng tham số quá lớn. Kích thước của bản đồ đặc trưng giảm đi một nửa do đi qua lớp trộn cực đại nằm giữa các khối. Khối phân loại chứa 2 lớp, mỗi lớp có 4096 nơ ron và lớp đầu ra cuối cùng có 1000 nơ ron.

Mô hình RES: Mô hình được đề xuất bởi He năm 2015 [6] với khái niệm kết nối dư thừa. RES chứa 50 lớp nhưng yêu cầu bộ nhớ ít hơn 5 lần so với mô hình VGG vì sử dụng lớp trộn trung bình để chuyển đổi bản đồ đặc trưng hai chiều của lớp cuối cùng thành véc-tơ n lớp để tính toán xác suất. Với mô hình CNN thông thường, đầu vào X sẽ thu được đầu ra $H(X)$ bằng cách cho X đi qua các lớp khác nhau. Đối với khối dư thừa của mô hình RES thì X thành $F(X) + X$, nó chứa một đường chuyển tiếp được thiết kế để bỏ qua một hoặc nhiều lớp tích chập trong khối. Điều này giúp mô hình tránh hiện tượng biến mất đạo hàm và cho phép mô hình học được các hàm phức tạp hơn.

Mô hình DEN: là một trong những biến thể mở rộng của RES, mỗi lớp được kết nối trực tiếp với mỗi lớp khác nhau theo kiểu chuyển tiếp trong mỗi khối dày đặc. Đối

với mỗi lớp, các bản đồ đặc trưng của tất cả các lớp ở phần trước được coi là các đầu vào riêng biệt và các bản đồ đặc trưng lại tiếp tục làm đầu vào cho tất cả các lớp tiếp theo. DEN có cấu trúc gồm các khối dày đặc và các lớp chuyển tiếp [7].

Mô hình EFF: là mạng nơ ron học sâu được thiết kế nhằm tối ưu hoá sự cân bằng giữa hiệu suất và độ phức tạp tính toán, được giới thiệu bởi Google AI năm 2019 [8]. EFF có số lượng tham số khoảng 5.3 triệu, là phiên bản nhỏ nhất và cơ bản nhất của dòng EfficientNet, được xây dựng dựa trên việc mở rộng kích thước mô hình sử dụng phương pháp Compound Scaling, giúp mô hình phát triển đều theo cả ba hướng là chiều sâu, chiều rộng và độ phân giải.

***Mô hình ViT*:** là một mô hình học sâu hiện đại sử dụng kiến trúc Transformer nổi tiếng trong xử lý ngôn ngữ tự nhiên để giải quyết bài toán thị giác máy tính. Thay vì sử dụng các lớp tích chập như CNN, ViT dựa trên cơ chế self-attention để học các đặc trưng của hình ảnh. ViT được giới thiệu lần đầu trong bài báo [9] của nhóm nghiên cứu tại Google vào năm 2020, là một bước tiến lớn trong thị giác máy tính, đặc biệt là khi dữ liệu lớn trở nên phổ biến.

Bảng 1. Mô tả cấu trúc các mô hình học sâu.

Mô hình	VGG	RES	EFF	DEN	ViT
Số lớp	16 lớp	50 lớp	16 lớp chính	201 lớp	12-24 khối Encoder tuỳ cấu hình
Số kênh	64 đến 512	64 đến 2048	Tối ưu theo scaling	64 đến 1024	Cố định theo cấu hình
Kích thước	224x224	224x224	224x224	224x224	224x224 hoặc 384x384
Tham số	~138 triệu	~25 triệu	~5.3 triệu	~20 triệu	~85-90 triệu (tuỳ cấu hình)

2.3. Học chuyển giao

Học chuyển giao là phương pháp tận dụng các mô hình đã được huấn luyện để giải quyết các bài toán mới nhanh và hiệu quả hơn. Học chuyển giao cho phép một mô hình được huấn luyện cho một công việc này có thể sử dụng lại hoặc tinh chỉnh để phục vụ cho một công việc khác, từ đó cải thiện hiệu suất cũng như tiết kiệm tài nguyên. Riêng đối với phương pháp học chuyển giao sử dụng cho quá trình trích xuất đặc trưng hình ảnh, các mô hình thường chỉ giữ lại các lớp đầu và giữa để tận dụng trọng số đã được huấn luyện để phục vụ cho quá trình trích xuất và các tầng kết nối đầy đủ ở cuối mô hình sẽ loại bỏ. Trong bài báo này, chúng tôi sử dụng các mô hình với trọng số đã được

huấn luyện trên tập dữ liệu ImageNet, đây là một tập dữ liệu lớn bao gồm hơn 14 triệu hình ảnh với hơn 20.000 lớp khác nhau, bao gồm động vật, đồ vật, cảnh vật, v.v.

2.4. Các tập dữ liệu

Việc đánh giá hiệu quả của CBIR yêu cầu sử dụng các tập dữ liệu chuẩn để đảm bảo tính khách quan và khả năng so sánh giữa các mô hình khác nhau. Các tập dữ liệu thực nghiệm thường chứa một số lượng lớn hình ảnh có nhãn giúp kiểm tra độ chính xác, khả năng tổng quát hóa của hệ thống. Trong quá trình thực nghiệm chúng tôi sử dụng một số tập dữ liệu mô tả ở ***Hình 2*** và

***Bảng* 2**, cụ thể:

Corel1K là tập dữ liệu được sử dụng rộng rãi trong các nghiên cứu của bài toán CBIR, gồm 1000 hình ảnh được chia thành 10 nhãn khác nhau, mỗi nhãn chứa 100 hình ảnh [10].

Oxford17Flower (Oxford17) là tập dữ liệu nổi bật được sử dụng phổ biến trong các nghiên cứu tìm kiếm hình ảnh dựa trên nội dung. Tập dữ liệu gồm 1.360 hình ảnh thuộc 17 nhãn, mỗi nhãn chứa 80 hình ảnh [11].

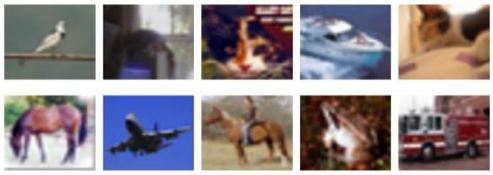
Caltech101 được giới thiệu bởi Fei-Fei năm 2004, là tập dữ liệu đa dạng về đối tượng, gồm 9.144 hình ảnh với 101 nhãn đối tượng khác nhau từ động vật, đồ vật, phương tiện đến các biểu tượng và một nhãn chứa các hình ảnh nền [12].

Cifar10 gồm 60.000 hình ảnh màu 32x32 chứa trong 10 lớp, với 6.000 hình ảnh cho mỗi lớp [13]. Mỗi hình ảnh trong tập dữ liệu có độ phân giải thấp chỉ 32x32. Trong nghiên cứu này, chúng tôi sử dụng một tập con gồm 10.000 hình ảnh, mỗi lớp chứa 1000 hình cho quá trình thực nghiệm và đánh giá.

a) Corel1K

b) Oxford17

c) Caltech101



d) Cifar10

***Hình* 2**. Minh họa hình ảnh trong các tập dữ liệu.

Bảng 2. Mô tả thông tin cơ bản các tập dữ liệu.

Tập dữ liệu	Số ảnh	Số nhãn	Anh mỗi lớp	Kích thước
Corel1K	1,000	10	100	384 × 256 hoặc 256 × 384
Oxford17	1,360	17	80	~500 x 500
Caltech101	9,144	101	40 - 800	Không cố định
Cifar10	10,000	10	1,000	32 x 32

2.5. Phương pháp đánh giá

Độ chính xác (Precision): Là tỷ lệ giữa số lượng ảnh liên quan được truy xuất đúng so với tổng số ảnh được truy xuất, được tính bởi công thức:

$$Precision = \frac{\text{Số ảnh liên quan được tìm kiếm đúng}}{\text{Tổng số ảnh tìm kiếm được}}$$

Độ chính xác trung bình (mAP): Là giá trị trung bình của tất cả các nhãn, dùng để đánh giá hiệu suất tổng thể của hệ thống tìm kiếm ảnh thuộc nhiều nhãn khác nhau được tính bởi công thức bên dưới:

$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i$, với N là số lớp, và AP_i là độ chính xác trung bình của nhãn thứ $\boldsymbol{i}$.

3. THỰC NGHIỆM VÀ KẾT QUẢ

3.1. Thực nghiệm

Chúng tôi tiến hành thực nghiệm lần lượt các mô hình VGG, RES, EFF, DEN, và ViT trên bốn tập dữ liệu Corel1K, Oxford17, Caltech101 và Cifar10. Các tập dữ liệu được chia theo tỷ lệ 80:20 tương ứng với tập huấn luyện và tập kiểm thử. Để so sánh độ tương đồng giữa hai véc-tơ chúng tôi sử dụng độ đo tương đồng Cosine. Cosine thường được sử dụng trong CBIR do khả năng đo lường độ tương đồng về hướng giữa hai véc-tơ đặc trưng thay vì độ lớn tuyệt đối của chúng. Điều này hiệu quả khi các đặc trưng đã được chuẩn hóa hoặc có sự chênh lệch về độ sáng nhưng vẫn giữ nguyên nội dung ảnh. Hơn nữa, Cosine có hiệu suất tính toán tốt trong không gian đa chiều, giúp cải thiện tốc độ và độ chính xác của hệ thống tìm kiếm hình ảnh.

3.2. Kết quả

Kết quả thực nghiệm các mô hình trên tập dữ liệu Corel1K thể hiện hiệu suất khác nhau, với ViT và DEN đạt độ chính xác cao nhất ở hầu hết các nhãn, trong khi VGG

thường có hiệu suất thấp hơn, đặc biệt ở các nhãn như "Beaches" và "People". RES và EFF có hiệu suất ổn định so với các mô hình khác (***Bảng 3).***

Đối với tập dữ liệu Oxford17, ViT vẫn thể hiện khả năng vượt trội so với các mô hình khác, độ chính xác đạt 1.00 trên hầu hết các nhãn. Điều này cho thấy khả năng phân loại mạnh mẽ và chính xác của ViT, đặc biệt là trong các nhãn có hình ảnh gần giống nhau như “Snowdrop”, “Tigerlily” hay “Tulip”. Mô hình DEN cũng thể hiện hiệu suất tương đối cao ở nhiều nhãn, tuy nhiên không vượt qua mô hình ViT. Ngược lại, VGG có độ chính xác thấp hơn hẳn so với các mô hình khác, đặc biệt là ở các nhãn như “Snowdrop” hay “Tulip”. Tóm lại, ViT là mô hình có hiệu suất vượt trội, trong khi các mô hình khác như DEN và EFF cho thấy hiệu quả ổn định (

Bảng 4). Tương tự, kết quả trên tập Cifar10 (***Bảng 5***).

Kết quả so sánh các mô hình trên các tập dữ liệu tại các TopK (***Hình 3***) cho thấy sự khác biệt rõ rệt về hiệu suất giữa các kiến trúc. ViT vượt trội với độ chính xác đạt tối đa trên Corel1K và Oxford17, điều này minh chứng khả năng mô hình hóa tốt các đặc trưng phức tạp của dữ liệu. Tuy nhiên, hiệu suất trên các tập dữ liệu lớn và đa dạng như Caltech101 và Cifar10 thì giảm nhẹ, mặc dù vẫn cao hơn các mô hình khác. DEN đạt hiệu suất cao và ổn định trên tất cả các tập dữ liệu, là lựa chọn mạnh mẽ cho các bài toán CBIR. Ngược lại, VGG đạt kết quả thấp nhất, đặc biệt trên Oxford17 và Cifar10, điều này cho thấy giới hạn trong khả năng học các đặc trưng phức tạp.

Bảng 3. So sánh độ chính xác theo nhãn trên tập Corel1K.

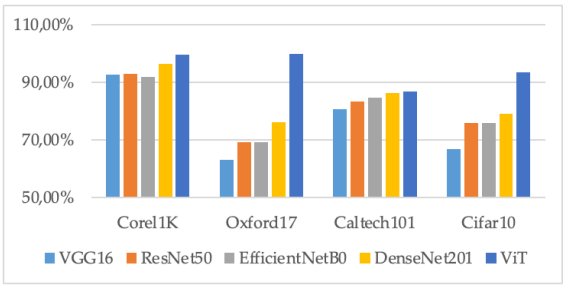
	VGG	**RES**	**EFF**	**DEN**	**ViT**
beaches	0.7333	0.8222	0.7333	0.8667	**1.0000**
bus	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
dinosaurs	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
elephants	0.9900	**1.0000**	**1.0000**	**1.0000**	**1.0000**
flowers	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
foods	0.9308	0.9385	0.8231	**1.0000**	**1.0000**
horses	0.9833	0.9667	0.9833	**1.0000**	**1.0000**
monuments	0.9077	0.8231	0.9154	0.9231	**0.9692**
mountains	0.9571	0.9143	0.9143	**1.0000**	**1.0000**
people	0.8300	0.9200	0.9200	0.9000	**1.0000**

Bảng 4. So sánh độ chính xác theo nhãn của 10 nhãn đầu tiên trên tập Oxford17.

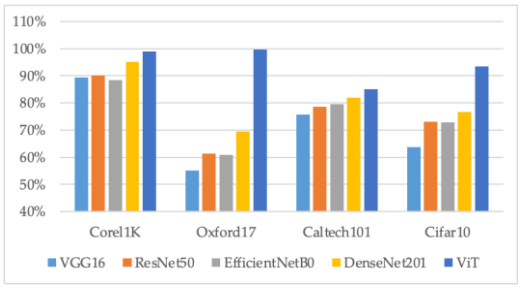
	VGG	**RES**	**EFF**	**DEN**	**ViT**
Bluebell	0.5857	0.6000	0.6286	0.7143	**1.0000**
Buttercup	0.8125	0.7500	0.7625	0.9000	**1.0000**
Coltsfoot	0.4300	0.4200	0.5200	0.6100	**1.0000**
Cowslip	0.6500	0.6250	0.5750	0.7500	**1.0000**
Crocus	0.5833	0.6667	0.5833	0.6833	**1.0000**
Daffodil	0.5000	0.5500	0.6750	0.6500	**1.0000**
Daisy	0.9444	0.9222	0.9000	0.8889	**1.0000**
Dandelion	0.8429	0.8143	0.8286	0.7714	**0.9571**
Fritillary	0.7583	0.7667	0.7750	0.7833	**1.0000**
Iris	0.9167	0.9833	0.9833	0.9333	**1.0000**

Bảng 5. So sánh độ chính xác theo nhãn trên tập Cifar10.

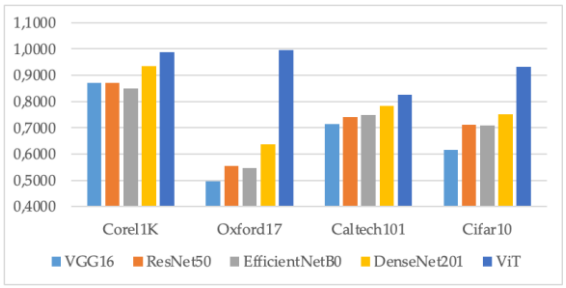
	VGG	**RES**	**EFF**	**DEN**	**ViT**
airplane	0.6582	0.7484	0.7385	0.7703	**0.9066**
automobile	0.5925	0.6613	0.6613	0.6825	**0.9138**
bird	0.6788	0.7697	0.7859	0.8273	**0.9576**
cat	0.7021	0.8206	0.8196	0.8474	**0.9649**
deer	0.7123	0.7889	0.7654	0.8333	**0.9654**
dog	0.5949	0.7342	0.7076	0.7354	**0.9405**
frog	0.700	0.7709	0.7660	0.7748	**0.9184**
horse	0.6724	0.7552	0.7771	0.8086	**0.9152**
ship	0.655	0.7288	0.7413	0.8025	**0.9538**
truck	0.6929	0.7965	0.7847	0.8071	**0.9247**



a) Top10



b) Top20



c) Top30

Hình 3. So sánh độ chính xác trên các tập dữ liệu.

Bảng 6. Tổng hợp kết quả các mô hình của các tập dữ liệu tại Top10.

Tập dữ liệu	VGG	RES	EFF	DEN	ViT
Corel1K	0.9256	0.9300	0.9189	0.9644	**0.9956**
Oxford17	0.6309	0.6927	0.6927	0.7610	**0.9976**
Caltech101	0.8064	0.8337	0.8477	0.8613	**0.8679**
Cifar10	0.6681	0.7593	0.7576	0.7910	**0.9357**

Từ kết quả tổng hợp ở bảng 6 cho thấy, mô hình ViT thể hiện sự vượt trội rõ rệt so với các mô hình CNN truyền thống, đặc biệt trong các trường hợp dữ liệu nhỏ và nhiệm vụ phân loại khó. Trên tập Oxford17 vốn có số lượng lớp vừa phải nhưng các lớp có đặc điểm tương đồng cao (các loài hoa), ViT đạt độ chính xác gần tuyệt đối (0.9976), cao hơn đáng kể so với các mô hình CNN chỉ ở mức 0.6909–0.7610. Điều này cho thấy khả năng nắm bắt mối liên hệ toàn cục và biểu diễn ngữ cảnh của ViT giúp phân biệt tốt các đối tượng có đặc điểm hình ảnh tương tự. Bên cạnh đó, trên tập Cifar10 gồm các ảnh có độ phân giải thấp và dễ nhiễu, ViT vẫn giữ được hiệu năng cao (0.9357), chứng tỏ tính linh hoạt trong môi trường dữ liệu khó xử lý. Ngược lại, ở các tập dữ liệu đơn giản hơn như Corel1K, khoảng cách độ chính xác giữa ViT và các mô hình CNN thu hẹp đáng kể. Kết quả này chỉ ra rằng ViT đặc biệt vượt trội trong các bài toán phân loại khó hoặc dữ liệu nhỏ, nơi yêu cầu khả năng biểu diễn phức tạp và học được quan hệ dài hạn trong ảnh.

4. KẾT LUẬN

Trong bài báo này, chúng tôi đã tiến hành đánh giá hiệu suất của các mô hình học sâu trong bài toán CBIR, với trọng tâm là việc trích xuất đặc trưng ảnh từ các mô hình như VGG, RES, EFF, DEN và ViT. Kết quả phân tích cho thấy rõ sự tiến bộ vượt bậc của các mô hình hiện đại trong bài toán CBIR. ViT luôn đạt độ chính xác cao nhất trên tất cả các tập dữ liệu, đạt độ chính xác gần như tuyệt đối trên Corel1K, Oxford17. Điều này phản ánh khả năng trích xuất, biểu diễn đặc trưng vượt trội của kiến trúc transformer, đặc biệt trong các tập dữ liệu nhỏ và có cấu trúc phức tạp. Trên các tập dữ liệu lớn hơn như Caltech101 và Cifar10, ViT vẫn thể hiện hiệu suất cao nhất, minh chứng cho khả năng xử lý hiệu quả các dữ liệu đa dạng. DEN duy trì vị trí thứ hai với độ chính xác ổn định, đặc biệt trên các tập Corel1K và Oxford17, cho thấy đây là một lựa chọn đáng tin cậy trong các bài toán thị giác máy tính. RES và EFF thể hiện sự cải thiện đáng kể so với VGG, đặc biệt trên các tập dữ liệu lớn như Cifar10 và Caltech101. Tuy nhiên, độ chính xác của chúng vẫn kém hơn DEN và ViT, điều này phản ánh sự hạn chế trong việc học đặc trưng phức tạp. VGG bộc lộ điểm yếu khi đạt kết quả thấp nhất trên mọi tập dữ liệu, cho thấy kiến trúc này không còn phù hợp với các yêu cầu hiện đại.

Bên cạnh các kiến trúc học sâu được khảo sát trong nghiên cứu này như VGG, RES, EFF, DEN và ViT, những tiến bộ gần đây trong lĩnh vực thị giác máy tính đã xuất hiện nhiều mô hình học sâu hiện đại như Swin Transformer hay ConvNeXt. Các mô hình này được kết hợp giữa kiến trúc chú ý và tích chập để cải thiện khả năng biểu diễn đặc trưng. Việc tích hợp và đánh giá các mô hình này trong bài toán tìm kiếm ảnh sẽ là một hướng nghiên cứu tiềm năng trong tương lai, đặc biệt là khi xem xét đến khả năng tổng quát hóa và hiệu quả tính toán trên các tập dữ liệu lớn và đa dạng hơn. Thêm vào đó, chúng tôi tiếp tục khám phá các thành phần của mô hình học sâu cho quá trình trích xuất đặc trưng, kết hợp các thành phần quan trọng, đặc trưng nổi bật từ nhiều mô hình học sâu khác nhau để nâng cao hiệu suất và cải thiện độ chính xác cho bài toán CBIR.

LỜI CẢM ƠN

Chúng tôi xin trân trọng cảm ơn Khoa Công nghệ thông tin, Trường Đại học Khoa học - Đại học Huế; Trường Đại học Khánh Hoà đã hỗ trợ về chuyên môn và tạo điều kiện về cơ sở vất chất giúp chúng tôi hoàn thành bài nghiên cứu này.

TÀI LIỆU THAM KHẢO

[1]. Smeulders, A. W., Worring, M., Santini, S., Gupta, A., and Jain, R. (2000). Content-based image retrieval at the end of the early years, *IEEE Transactions on pattern analysis and machine intelligence*, 22(12), 1349-1380.

[2]. Wang, W., and He, Q. (2008). A survey on emotional semantic image retrieval, *In 2008 15th IEEE International Conference on Image Processing*, pp. 117-120, IEEE.

[3]. Voulodimos, A., Doulamis, N., Doulamis, A., and Protopapadakis, E. (2018). Deep learning for computer vision: A brief review, *Computational intelligence and neuroscience*, 2018(1), 7068349.

[4]. Hou, D., Wang, S., Tian, X., and Xing, H. (2022). An attention-enhanced end-to-end discriminative network with multiscale feature learning for remote sensing image retrieval, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15, 8245-*8255*.

[5]. Simonyan, K., and Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition*, In Proceedings of the International Conference on Learning Representations* (ICLR 2015).

[6]. He, K., Zhang, X., Ren, S., and Sun, J. **(**2016). Deep Residual Learning for Image Recognition, *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016)*.

[7]. Tan, M., and Le, Q. V. (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks*, In Proceedings of the International Conference on Machine Learning (ICML 2019).*

[8]. Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks, *In Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4700-4708).

[9]. Dosovitskiy, A. (2020). An image is worth 16x16 words: Transformers for image recognition at scale, *arXiv preprint arXiv:2010.11929*.

[10]. Kaggle, Corel-1K dataset, https://www.kaggle.com/datasets/elkamel/corel-images/data, visited 02/2024.

[11]. Kaggle, Oxford 17 flowers dataset, https://www.kaggle.com/datasets/cantonioupao/oxford-flower-17categories-labelled, visited 03/2024.

[12]. Li, F.-F., Andreeto, M., Ranzato, M., and Perona, P.. Caltech 101 (1.0). CaltechDATA. https://doi.org/10.22002/D1.20086. 2022.

[13]. Krizhevsky, A., and Hinton, G.. Learning multiple layers of features from tiny images. 2009.

ANALYSIS OF SOME IMAGE FEATURE EXTRACTION MODELS FOR CONTENT-BASED IMAGE RETRIEVAL PROBLEM

Tran Van Khanh[1,2*], Nguyen Ngoc Thuy[1], Le Manh Thanh[1]

[1] University of Sciences, Hue University

[2] University of Khanh Hoa

* Email: tvkhanh.dhkh24@hueuni.edu.vn

ABSTRACT

The article focuses on analyzing and comparing the performance of several deep learning models for the CBIR task through criteria such as accuracy and generalization ability. The deep learning models utilized in the experiments, including VGG16, Resnet50, EfficientNetB0, Densenet201, and Vision Transformer (ViT), were pre-trained on the enormous Imagenet dataset. Experiments using the proposed models were carried out on four datasets: Corel1K, Oxford17Flowers, Caltech101, and a subset of the Cifar10 dataset. The results of the analysis clearly show the remarkable progress of modern models in handling CBIR tasks. ViT always achieves the highest accuracy on all datasets, with nearly perfect accuracy on the Corel1K and Oxford17Flowers.

Keywords: CBIR, CNN, Deep Learning, Transformer.

Trần Văn Khánh sinh năm 1988. Ông tốt nghiệp kỹ sư ngành Công nghệ thông tin năm 2012 tại Trường Đại học Giao thông vận tải TP HCM, Thạc sĩ ngành Khoa học máy tính năm 2016 tại Trường Đại học Qui Nhơn và hiện đang là NCS tại khoa Công nghệ thông tin, trường Đại học Khoa học - Đại học Huế.

Lĩnh vực nghiên cứu: Học máy, Xử lý hình ảnh.

Nguyễn Ngọc Thủy sinh năm 1990. Ông nhận bằng Thạc sĩ ngành Khoa học máy tính tại Trường Đại học Khoa học - Đại học Huế năm 2014 và Tiến sĩ tại Đại học Khon Kaen, Thái Lan năm 2020.

Lĩnh vực nghiên cứu: Khai phá dữ liệu.

Lê Mạnh Thạnh sinh năm 1953 tại Quảng Trị. Ông nhận học vị Tiến sĩ ngành khoa học máy tính tại Đại học Budapest (ELTE), Hungary vào năm 1994 và trở thành Phó giáo sư tại trường Đại học Khoa học, Đại học Huế vào năm 2004. Ông hiện đang là giảng viên Khoa Công nghệ thông tin, Trường Đại học Khoa học, Đại học Huế.

Lĩnh vực nghiên cứu: cơ sở dữ liệu, cơ sở tri thức và lập trình logic.