TÌM KIẾM ẢNH THEO NỘI DUNG DỰA TRÊN MẠNG NƠRON TÍCH CHẬP VÀ PHƯƠNG PHÁP SINH MÃ NHỊ PHÂN

Nguyễn Thị Huyền*, Trần Thị Thu Huyền, Vũ Thị Lưu

Khoa Công nghệ thông tin, Học viện Nông nghiệp Việt Nam

Tác giả liên hệ: nthuyen@vnua.edu.vn

Ngày nhận bài: 20.07.2020 Ngày chấp nhận đăng: 02.09.2020

TÓM TẮT

Tìm kiếm ảnh theo nội dung là hướng nghiên cứu đang được quan tâm trong những năm gần đây vì phương pháp tìm kiếm này có thể khắc phục nhược điểm của phương pháp tìm kiếm dựa trên văn bản mô tả là không bị ảnh hưởng bởi sự thiếu hoặc sai của văn bản kèm theo ảnh. Bên cạnh đó, các phương pháp học sâu như mạng nơron tích chập đã chứng minh được khả năng xử lý dữ liệu lớn đặc biệt trong lĩnh vực thị giác máy tính và xử lý ảnh. Mục tiêu của nghiên cứu này là giải bài toán tìm kiếm ảnh theo nội dung và phương pháp để giảm thời gian truy vấn ảnh sử dụng mạng nơtron tích chập. Đồng thời, chúng tôi kết hợp phương pháp này với phương pháp sinh mã nhị phân để cải thiện thời gian truy vấn ảnh. Kết quả thực nghiệm trên hai bộ dữ liệu cifar-10 và mnist cho thấy việc sử dụng mạng nơron tích chập kết hợp phương pháp sinh mã nhị phân trong tìm kiếm ảnh đạt độ chính xác xấp xỉ 89% và 98% và cải thiện đáng kể thời gian truy vấn ảnh.

Từ khóa: Tìm kiếm ảnh theo nội dung, mạng nơron tích chập, sinh mã nhị phân.

Content-based Image Retrieval with Convolutional Neural Networks and Binary Hashing Method

ABSTRACT

Content-based image retrieval has received great attention in recent years because this method overcomes the disadvantages of the text-based image retrieval that is not affected by the lack of or wrong of the text attached to the image. In addition, deep learning methods such as convolutional neural networks have demonstrated their ability to process large-sized data, especially computer vision and image processing. The aims of this study was develop a content-based image retrieval program and method to reduce image query time using the convolutional neural network (CNN). Also, we combined CNN with a binary hashing method to improve image retrieval time. The experimental results on CIFAR-10 and MNIST data sets showed that combining CNN with the binary hashing method for content-based image retrieval achieved an accuracy of approximately 89% on CIFAR-10, 98% on MNIST and significantly improved retrieval time.

Keywords: Content-based image retrieval, CBIR, convolutional neural networks, CNN, binary hashing.

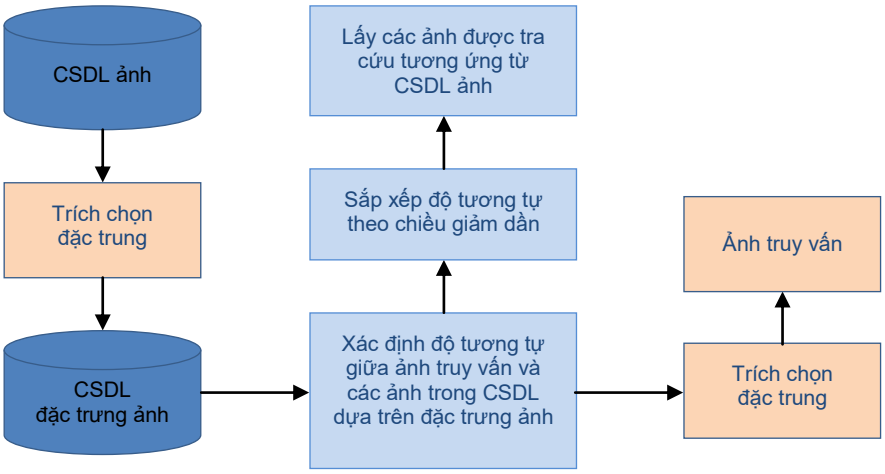
1. ĐẶT VẤN ĐỀ

Ngày nay, với sự phát triển vượt trội của công nghệ kỹ thuật số và sự phổ biến rộng rãi các thiết bị quay phim, chụp ảnh dẫn đến kho dữ liệu hình ảnh về nhiều lĩnh vực khác nhau như: y khoa, hệ thống thông tin địa lý, thư viện số, giáo dục đào tạo, giải trí, mạng xã hội… cũng tăng theo một cách nhanh chóng. Theo báo cáo của Tập đoàn dữ liệu thế giới IDC năm 2016, thế giới đã tạo ra 1.138 nghìn tỷ hình ảnh, gấp hơn 700 lần so với năm 2015 (Photoindustrie-Verband e.V, 2016). Theo báo cáo về chia sẻ ảnh trên toàn cầu, Brandwatch đã tính toán rằng mỗi ngày có 350 triệu hình ảnh được chia sẻ qua Facebook, 95 triệu hình ảnh được chia sẻ qua Instagram, 400 triệu trên Snapchat và 1,6 tỷ hình ảnh trên WhatsApp (Văn Thế Thành & Lê Mạnh Thạnh, 2016).

Vì vậy, nhu cầu tìm kiếm ảnh hay truy xuất dữ liệu ảnh là một nhu cầu tất yếu, và là một trong những lĩnh vực nghiên cứu thu hút sự quan tâm nhất hiện nay. Tìm kiếm ảnh hiểu một cách cơ bản là tìm những ảnh trong cơ sở dữ liệu ảnh có liên quan đến một ảnh truy vấn (query) cụ thể. Hình 1 mô tả sơ lược quá trình tìm kiếm ảnh. Bài toán tìm kiếm ảnh được chia thành hai lớp chính (Văn Thế Thành, 2017): Thứ nhất là tìm kiếm ảnh dựa trên văn bản TBIR (Text-Based Image Retrieval). Phương pháp này mất nhiều thời gian để mô tả chỉ mục của hình ảnh dưới dạng văn bản, có nhiều hạn chế vì tính chủ quan của con người và kết quả tìm kiếm sẽ không chính xác khi các mô tả này bị sai sót hoặc không tồn tại. Ví dụ, Google Images Search là một trong các công cụ tìm kiếm ảnh được sử dụng phổ biến nhất hiện nay. Công cụ này cho phép người sử dụng nhập các từ khóa liên quan đến ảnh cần tìm và thực hiện việc tìm kiếm thông qua việc phân tích các meta-data và văn bản đi kèm với ảnh. Phương pháp này cho kết quả tương đối tốt, đáp ứng nhu cầu cơ bản của người sử dụng. Tuy nhiên, các kết quả trả về sẽ không đúng với yêu cầu đặt ra khi các meta-data đi kèm với ảnh bị thiếu hoặc sai sót và khi những từ khóa truy vấn mang ý nghĩa nhập nhằng (Lê Minh Phúc & Trần Công Án, 2017). Thứ hai là tìm kiếm ảnh dựa trên nội dung CBIR (Content-Based Image Retrieval), tức là tìm tập hình ảnh có nội dung tương tự với hình ảnh cho trước. Phương pháp CBIR thực hiện tìm kiếm dựa trên đặc trưng thị giác của hình ảnh, do đó vượt qua được hạn chế của phương pháp tìm kiếm TBIR. Với phương pháp CBIR, cả hai vấn đề trích xuất tự động các đặc trưng thị giác và phương pháp đánh giá độ tương tự giữa hai ảnh đều đóng vai trò quan trọng, quyết định hiệu quả tìm kiếm. Về vấn đề thứ nhất, phương pháp tìm kiếm ảnh theo nội dung "truyền thống" thường dựa vào các đặc trưng trực quan như màu sắc, kết cấu, hình dạng, đặc trưng cục bộ được rút trích từ ảnh; do đó rất nhiều hệ thống truy vấn ảnh dựa trên nội dung đã ra đời như: QBIC, VisualSeek, WebSeek và BlobWorld... (Văn Thế Thành, 2017). Phương pháp này có hạn chế là khó xác định và chọn ra được những đặc trưng đại diện cho ảnh để việc tìm kiếm đạt kết quả tốt.

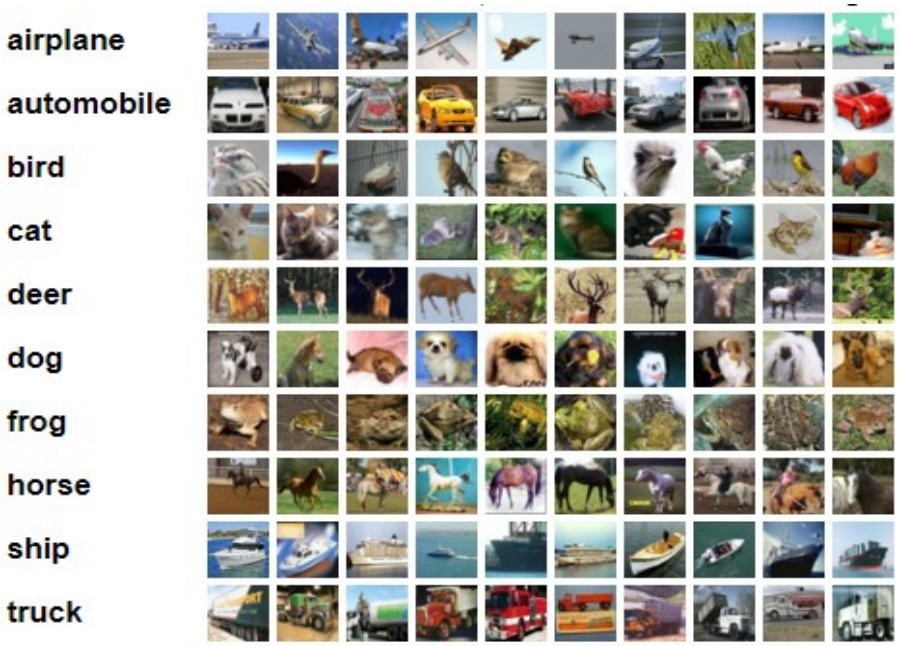
Những năm gần đây, các phương pháp học sâu (Deep Learning) trong đó có mạng nơron tích chập (CNN) đã đạt được thành công to lớn trong xử lý dữ liệu kích thước lớn. Nó đã được chứng minh là rất hiệu quả trong lĩnh vực thị giác máy tính và xử lý ảnh như: phát hiện người đi bộ (Luo và cộng sự, 2014), phát hiện khuôn mặt (Li & cs., 2015), phân loại hình ảnh (Ciressan & cs., 2012), tự động tô màu hình ảnh (Cheng, 2015)… và gần đây các phương pháp dựa trên Deep Learning như CNN đã được áp dụng vào bài toán tìm kiếm ảnh. Lecun & cs. (1998) đã đề xuất mạng nơron tích chập LeNet-5 sử dụng để nhận biết các chữ cái viết tay, và đã cho thấy đó là một thuật toán rất thành công. Sau đó, Krizhevsky & cs. (2012) đã cải thiện thuật toán mạng nơron tích chập, sử dụng đầu ra của lớp thứ 7 làm đặc trưng để truy vấn hình ảnh và đã đạt được kết quả tốt trên tập dữ liệu ImageNet. Babenko & cs. (2014) đã nghiên cứu phương pháp cải thiện hiệu suất truy vấn ảnh bằng cách sử dụng PCA để nén các đặc trưng được trích chọn nhờ CNN. Mặc dù các công trình của Krizhevsky & cs. (2012) và Babenko & cs. (2014) cho thấy khi sử dụng CNN vào việc truy vấn ảnh cho độ chính xác cao nhưng việc đối sánh các ảnh được thực hiện trong không gian Euclide dẫn đến chưa hiệu quả về thời gian tính toán, trong khi yêu cầu đặt ra với một hệ thống tìm kiếm ảnh là phải đưa ra kết quả nhanh chóng.

Xuất phát từ những vấn đề nêu trên, trong bài báo này, chúng tôi sẽ giới thiệu về mạng nơron tích chập (CNN) và phương pháp dựa trên mạng nơron tích chập để trích chọn đặc trưng ảnh (gọi là đặc trưng CNN) sau đó tiếp tục thực hiện việc sinh mã nhị phân (binary hashing) để biến các đặc trưng này thành 1 véctơ nhị phân có độ dài nhỏ, véctơ này được gọi là mã nhị phân (hash code). Sau khi có được mã nhị phân cho từng bức ảnh, việc tính toán sự tương đồng giữa các bức ảnh sẽ trở nên đơn giản hơn vì số chiều thấp hơn và chỉ phải làm việc với các toán tử nhị phân đơn giản, từ đó cải thiện được tốc độ tìm kiếm.



Ghi chú: CSDL: Cơ sở dữ liệu.

Hình 1. Quá trình tìm kiếm ảnh



Hình 2. Hình ảnh minh họa bộ dữ liệu CIFAR-10

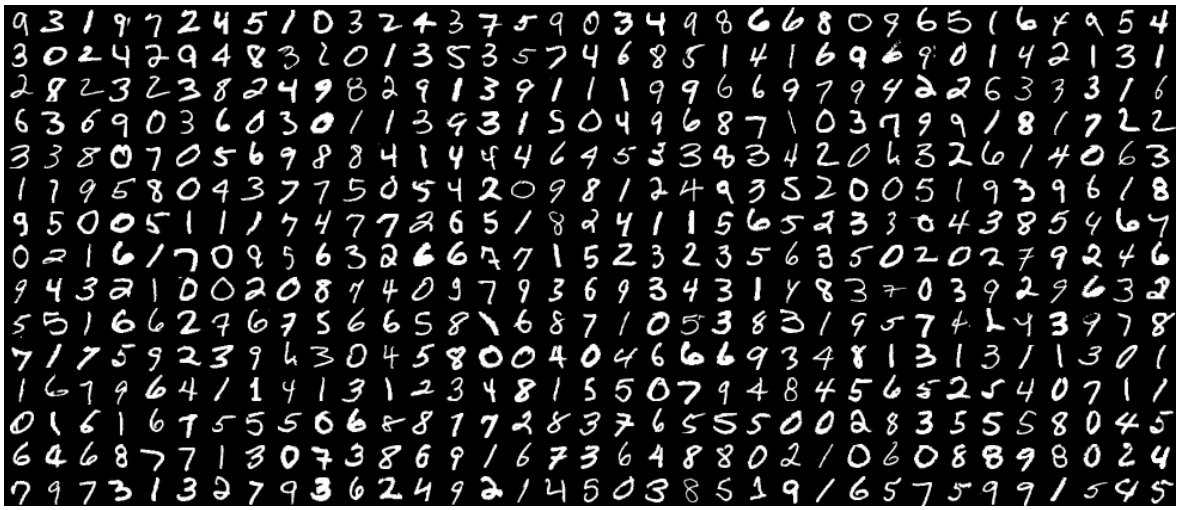
2. PHƯƠNG PHÁP NGHIÊN CỨU

2.1. Vật liệu

2.1.1. Dữ liệu ảnh

Bộ dữ liệu CIFAR-10 do Krizhevsky & cs. (2009) thu thập gồm 10 lớp đối tượng: plane, car, bird, cat, deer, dog, frog, horse, ship và truck, mỗi lớp gồm 6.000 ảnh màu có kích thước 32 × 32 pixel. Tổng số có 60.000 ảnh trong đó 50.000 ảnh được sử dụng cho huấn luyện (train), 10.000 ảnh còn lại được dùng cho kiểm tra (test). Đây là một bộ cơ sở dữ liệu tương đối khó vì ảnh nhỏ và đối tượng trong cùng một lớp cũng biến đổi rất nhiều về màu sắc, hình dáng, kích thước.

Bộ dữ liệu MNIST do Lecun & cs. (1998) xây dựng là cơ sở dữ liệu bao gồm các ảnh đa mức xám của 10 chữ số viết tay từ 0 đến 9 đã được chuẩn hóa về kích thước 28 × 28 pixel. Bộ ảnh gồm 60.000 ảnh dùng để huấn luyện (train), và 10.000 ảnh còn lại được dùng cho kiểm tra (test).



Hình 3. Hình ảnh minh họa bộ dữ liệu MNIST

2.1.2. Công cụ

Chúng tôi sử dụng máy tính cài hệ điều hành Windows 64-bit, Intel, Core™ i5-5200U, CPU@2.20GHz, ngôn ngữ lập trình Matlab2016a, thư viện Caffe CNN (Jia, 2014).

2.2. Phương pháp nghiên cứu

2.2.1. Mạng nơron tích chập

Với mạng nơron truyền thẳng (ANN) thông thường: nhận đầu vào là một véctơ và chuyển đổi nó thông qua một loạt các lớp ẩn. Mỗi lớp ẩn bao gồm một tập các nơ-ron, trong đó mỗi nơron được kết nối đầy đủ với tất cả các nơron trong lớp trước và các nơron trong một lớp không có bất kỳ kết nối nào với nhau. Lớp được kết nối đầy đủ cuối cùng được gọi là lớp đầu ra. Như vậy, với tập dữ liệu gồm các hình ảnh có kích thước [200 × 200 × 3], mỗi nơron trong lớp ẩn đầu tiên của mạng sẽ có 200 × 200 × 3 = 120.000 trọng số kết nối. Điều này gây khó khăn cho việc huấn luyện ANN trên cả hai yếu tố: chi phí để xây dựng dữ liệu huấn luyện lớn và thời gian huấn luyện lâu.

Từ thực tế đó, mạng CNN ra đời với ý tưởng chính là mỗi nơron chỉ cần kết nối tới một vùng cục bộ của ảnh thay vì trên toàn bộ ảnh. Về cơ bản CNN là một kiểu mạng ANN truyền thẳng, trong đó kiến trúc chính gồm nhiều thành phần được ghép nối với nhau theo cấu trúc nhiều lớp đó là: Convolution, ReLU, Pooling và liên kết đầy đủ (Fully connected). Sự sắp xếp về số lượng và thứ tự giữa các lớp này sẽ tạo ra những mô hình khác nhau phù hợp cho các bài toán khác nhau.

a. Lớp Convolution

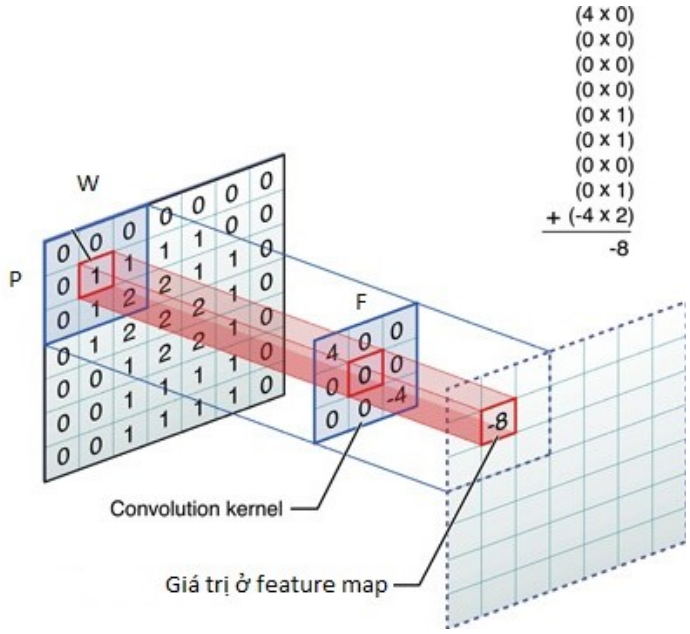
Lớp Convolution (Conv) là lớp quan trọng nhất trong cấu trúc của CNN. Hình 4 mô tả lý thuyết và cách thức Conv hoạt động trên một dữ liệu đầu vào được biểu diễn bằng một ma trận hai chiều. Phép tính này được thực hiện bằng cách dịch chuyển một cửa sổ mà ta gọi là bộ lọc (hay kernel) trên ma trận đầu vào, trong đó kết quả mỗi lần dịch chuyển được tính bằng tổng tích chập (tích của các giá trị giữa 2 ma trận tại vị trí tương ứng), trong hình 4 là giá trị đầu ra khi dịch chuyển bộ lọc có kích thước [3 × 3] trên toàn bộ ma trận đầu vào có kích thước [7 × 7].

Trong trường hợp tổng quát, hình ảnh có kích thước [W1 × H1 × D1], sử dụng K bộ lọc có kích thước [F × F], trong quá trình xử lý sẽ dịch chuyển các bộ lọc trên toàn bộ ảnh với bước dịch chuyển (stride) S (S được tính bằng pixel). Trong một số trường hợp để cân bằng giữa số bước dịch chuyển và kích thước của ảnh người ta có thể chèn thêm P pixel với một giá trị cho trước (thường là 0) xung quanh viền của ảnh khi đó ta được ma trận đầu ra (feature map) là [W2 × H2 × D2] trong đó:

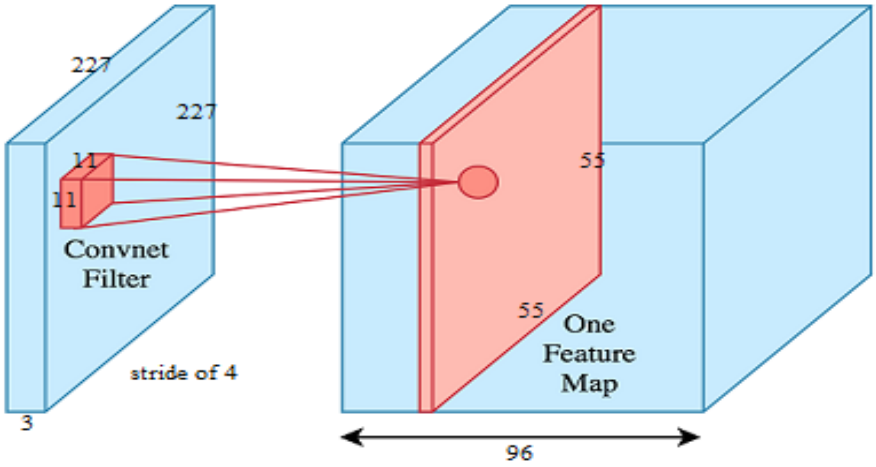
$$W2 = (W1 - F + 2P)/S + 1$$

$$H2 = (H1 - F + 2P)/S + 1$$

$$D2 = K$$



Hình 4. Minh hoa phép nhân chập



Hình 5. Lớp nhân chập thực hiện nhân chập ảnh đầu vào có kích thước [227×227×3] với 96 bộ lọc có kích thước [11×11×3], bước dịch chuyển S = 4 pixel và P = 0. Tương ứng với mỗi bộ lọc sẽ cho một feature map có kích thước là W2 = H2 = (227 – 11)/4 +1 = 55 ở kết quả đầu ra

Lược đồ chia sẻ tham số được sử dụng trong các lớp tích chập để kiểm soát số lượng tham số. Ví dụ trong lớp tích chập ở hình 3, có 55 × 55 × 96 = 290.400 nơron, mỗi nơron có 11 × 11 × 3 = 363 trọng số kết nối và 1 bias. Như vậy có 290.400 × 364 = 105.705.600 tham số. Rõ ràng, con số này rất lớn. Chúng ta có thể giảm đáng kể số lượng tham số bằng cách sử dụng cùng bộ trọng số và bias cho các nơron trong cùng feature map. Với lược đồ chia sẻ tham số này, lớp Conv trong ví dụ của chúng ta bây giờ chỉ có 96 bộ trọng số, với tổng số 96 × 11 × 11 × 3 = 34,848 hoặc 34.944 tham số (96 bias).

b. Lớp ReLU

Lớp ReLU thường được cài đặt ngay sau lớp Conv. Lớp này sử dụng hàm kích hoạt f(x) = max(0,x). Nói một cách đơn giản, lớp này có nhiệm vụ chuyển toàn bộ giá trị âm trong kết quả lấy từ lớp Conv thành giá trị 0. Ý nghĩa của

cách cài đặt này chính là tạo nên tính phi tuyến cho mô hình. Có rất nhiều cách để khiến mô hình trở nên phi tuyến như sử dụng các hàm kích hoạt sigmoid, tanh,... nhưng hàm f(x) = max(0,x) dễ cài đặt, tính toán nhanh mà vẫn hiệu quả (Krizhevsky & cs., 2012).

c. Lớp Pooling

Lớp này sử dụng một cửa sổ trượt quét qua toàn bộ ảnh, mỗi lần trượt theo một bước dịch chuyển cho trước. Khác với lớp Conv, lớp Pooling không tính tích chập mà thực hiện lấy mẫu. Khi cửa sổ trượt trên ảnh, chỉ có một giá trị được xem là giá trị đại diện cho thông tin ảnh tại vùng đó (giá trị mẫu) được giữ lại. Các phương thức lấy mẫu phổ biến trong lớp Pooling là MaxPooling (lấy giá trị lớn nhất), MinPooling (lấy giá trị nhỏ nhất) và AveragePooling (lấy giá trị trung bình).

Với ma trận đầu vào có kích thước [W1 × H1 × D1], thực hiện toán tử pooling trên cửa sổ có kích thước [F×F] với bước dịch chuyển S pixel ta được ma trận đầu ra [W2 × H2 × D2] trong đó:

W2 = (W1– F)/S + 1

H2 = (H1 – F)/S + 1

D2 = D1

Hình 6 là ví dụ minh họa về sử dụng toán tử pooling. Trong đó, hình 6(a) phía bên trái là cách thức lớp pooling xử lý đối với một đầu vào có kích [224 × 224 × 64], cửa sổ có kích thước [2 × 2], bước dịch chuyển S = 2 khi đó đầu ra thu được có kích thước [112 × 112 × 64]. Hình 6(b) phía bên phải mô tả chi tiết cách thức hoạt động của max-pooling với F = 2, S = 2.

d. Lớp liên kết đầy đủ

Liên kết đầy đủ là cách kết nối các nơron ở hai lớp với nhau trong đó lớp phía sau kết nối đầy đủ với các nơron ở lớp phía trước nó. Đây cũng là dạng kết nối thường thấy ở ANN, trong CNN lớp này thường được sử dụng ở các lớp phía cuối của kiến trúc mạng.

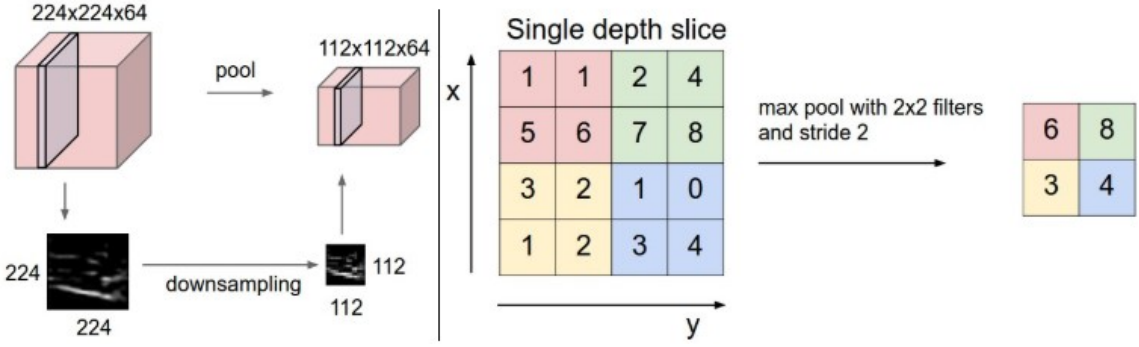
2.2.2. Ứng dụng mạng nơron tích chập vào bài toán tìm kiếm ảnh

Như đã đề cập ở trên, phương pháp tìm kiếm ảnh theo nội dung "truyền thống" thường dựa vào các đặc trưng trực quan như màu sắc, kết cấu, hình dạng, đặc trưng cục bộ được rút trích từ ảnh và thường không hiệu quả trong một số trường hợp. Ví dụ, lược đồ màu có thể được dùng để miêu tả đặc trưng màu của một ảnh. Tuy nhiên, hạn chế chính của lược đồ màu là chưa tận dụng được thông tin không gian của các vùng ảnh. Điều này có thể dẫn đến các sai số không mong muốn như minh họa trong hình 7, hai ảnh có cấu trúc khác nhau nhưng có lược đồ màu giống nhau, hoặc trong trường hợp ảnh bị lệch như minh họa trong hình 8.

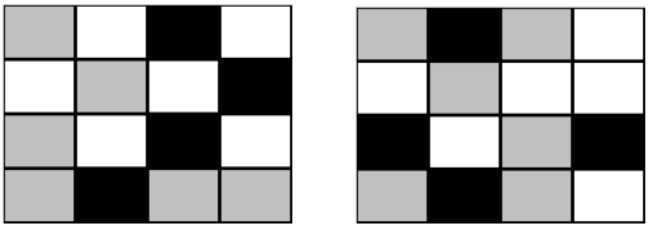
Trong khi đó sử dụng đặc trưng được trích chọn bởi CNN có thể khắc phục được những hạn chế nêu trên do sử dụng các bộ lọc với kích thước khác nhau trượt trên ảnh, do đó tạo ra được nhiều dữ liệu hơn.

a. Kiến trúc mạng CNN

Chúng tôi sử dụng mô hình CNN do Krizhevsky & cs. (2012) đề xuất, mô hình này được cung cấp sẵn trong thư viện Caffe CNN (Jia, 2014).



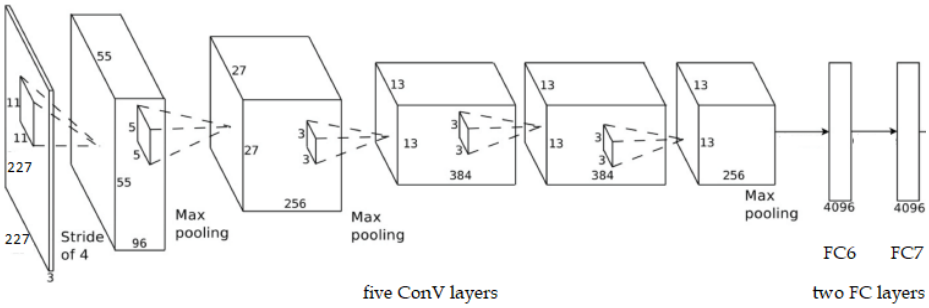
Hình 6. Ví dụ về sử dụng toán tử pooling



Nguồn: Văn Thế Thành, 2017.

Hình 7. Hai ảnh khác nhau nhưng có cùng lược đồ màu

Hình 8. Số 8 ở các vị trí khác nhau trong ảnh



Hình 9. Mô hình mạng CNN do Krizhevsky và các cộng sự (2012) đề xuất

Trong kiến trúc trên, lớp nhân chập thứ nhất thực hiện lọc ảnh đầu vào có kích thước [227 × 227 × 3] bởi 96 bộ lọc có kích thước [11 × 11 × 3] với bước dịch chuyển S = 4 pixel, P = 0 kết quả thu được đầu ra có kích thước [55 × 55 × 96]. Lớp nhân chập thứ hai thực hiện lọc đầu vào có kích thước [55 × 55 × 96] bởi 256 bộ lọc có kích thước [5 × 5 × 96] với bước dịch chuyển S = 1 pixel, P = 2 và thu được đầu ra có kích thước [27 × 27 × 256]. Tương tự với các lớp nhân chập khác. Các lớp pooling sử dụng toán tử maxpooling với cửa sổ có kích thước [3 × 3] và bước dịch chuyển S = 2.

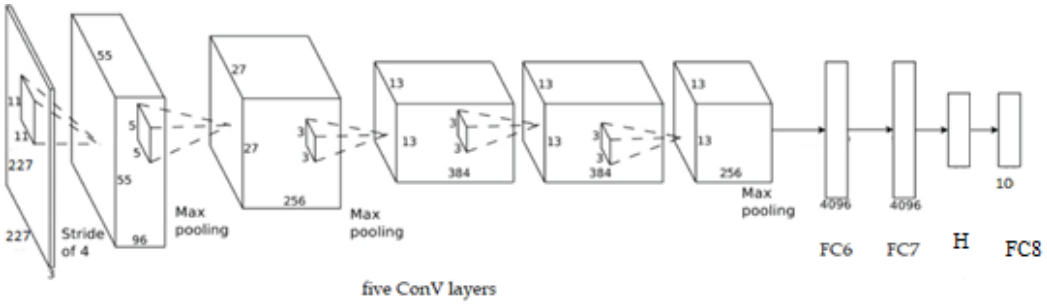
Sau khi mạng được huấn luyện, đầu ra của lớp FC7 được sử dụng như véctơ đặc trưng của ảnh và được sử dụng cho bài toán tìm kiếm ảnh. Với ảnh truy vấn I_q và tập cơ sở dữ liệu ảnh P, gọi V_q và V_i^P tương ứng là các véctơ đặc trưng của ảnh truy vấn I_q và của ảnh Ii trong tập P. Chúng tôi xác định mức độ tương tự giữa I_q và I_i là khoảng cách Euclidean giữa hai véctơ đặc trưng tương ứng của chúng.

$$S_i = \left\| V_q - V_i^P \right\| \quad (1)$$

Khoảng cách Euclidean càng nhỏ thì mức độ giống nhau của hai ảnh càng cao. Dựa vào khoảng cách này có thể xác định được top k hình ảnh giống với ảnh truy vấn nhất.

b. Nâng cao hiệu quả tìm kiếm ảnh

Việc sử dụng đầu ra của lớp FC7 làm véctơ đặc trưng cho thấy kết quả tìm kiếm ảnh tốt. Tuy nhiên, nó không hiệu quả trong hệ thống tìm kiếm ảnh với cơ sở dữ liệu lớn do việc tính khoảng cách Euclidean trên các véctơ chiều cao mất nhiều thời gian. Để cải thiện thời gian truy xuất hình ảnh, một cách đơn giản là chuyển đổi các vectơ đặc trưng của ảnh thành dạng nhị phân. Khi đó việc so sánh độ giống nhau của hai hình ảnh có thể được thực hiện nhanh chóng bằng cách sử dụng khoảng cách Hamming.



Hình 10. Mô hình mạng CNN sau khi chèn lớp ẩn H

Để thực hiện ý tưởng này, chúng tôi chèn thêm một lớp ẩn H sau lớp FC7 và lớp ra FC8 là phân loại softmax ứng với 10 lớp dữ liệu để phù hợp với các bộ dữ liệu mà chúng tôi sử dụng để thực nghiệm. Các nơron trong lớp H được kết nối đầy đủ với các nơron ở lớp trước và lớp sau và được kích hoạt bởi hàm truyền sigmoid. Các trọng số liên kết từ lớp thứ nhất đến lớp FC7 được lấy từ mạng CNN đã được huấn luyện trước đó, trọng số kết nối từ lớp FC7 tới lớp H và từ lớp H đến lớp FC8 ban đầu được khởi tạo ngẫu nhiên và được cập nhật trong quá trình huấn luyện mạng.

Do các nơron trong lớp H sử dụng hàm truyền sigmoid nên đầu ra của lớp H ký hiệu là O(H) là các giá trị trong khoảng {0,1}, để đưa về dạng mã nhị phân chúng tôi thực hiện phân ngưỡng cho mỗi bit j = 1,…, h (với h là số nơron trong lớp H) như sau:

$$H^j = \begin{cases} 1 \text{ nếu } O^j(H) \geq 0,5 \\ 0 \text{ nếu ngược lại} \end{cases} \quad (2)$$

Gọi P = {I_1, I_2,… , I_n} là tập cơ sở dữ liệu ảnh bao gồm n ảnh, mã nhị phân tương ứng của tập ảnh được ký hiệu là $\mathbf{P}_H$ = {H_1, H_2,…, H_n} với $H_i \in \{0, 1\}^h$. Cho một hình ảnh truy vấn I_q với mã nhị phân tương ứng là H_q. Chúng tôi xác định mức độ tương tự giữa I_q và I_j là khoảng cách Haming giữa H_q và H_j.

2.3. Kết quả thực nghiệm

2.3.1. Phương pháp đánh giá

Với mỗi bộ dữ liệu, chúng tôi sử dụng 10.000 ảnh trong tập test làm ảnh truy vấn. Với một ảnh truy vấn q và một phép đo độ tương tự, chúng tôi tính độ đo tương tự giữa ảnh truy vấn và từng ảnh trong tập train, sau đó lấy ra 1.000 ảnh có độ đo tương tự cao nhất để đánh giá độ chính xác của việc tìm kiếm bằng một độ đo Precision như sau:

$$\text{Precision} = \frac{\sum_{i=1}^{1.000} \text{Rel}(i)}{1.000} \quad (3)$$

Trong đó Rel(i) biểu thị sự liên quan giữa ảnh truy vấn q và ảnh được xếp thứ hạng i trong số 1.000 ảnh được lấy ra. Ở đây, chúng tôi sử dụng nhãn lớp của ảnh để đo mức độ liên quan. Rel(i) = 1 nếu ảnh truy vấn q và ảnh thứ i thuộc cùng một lớp và bằng 0 nếu ngược lại.

Thời gian truy vấn là thời gian trung bình để tính độ đo tương tự giữa ảnh truy vấn và từng ảnh trong tập train.

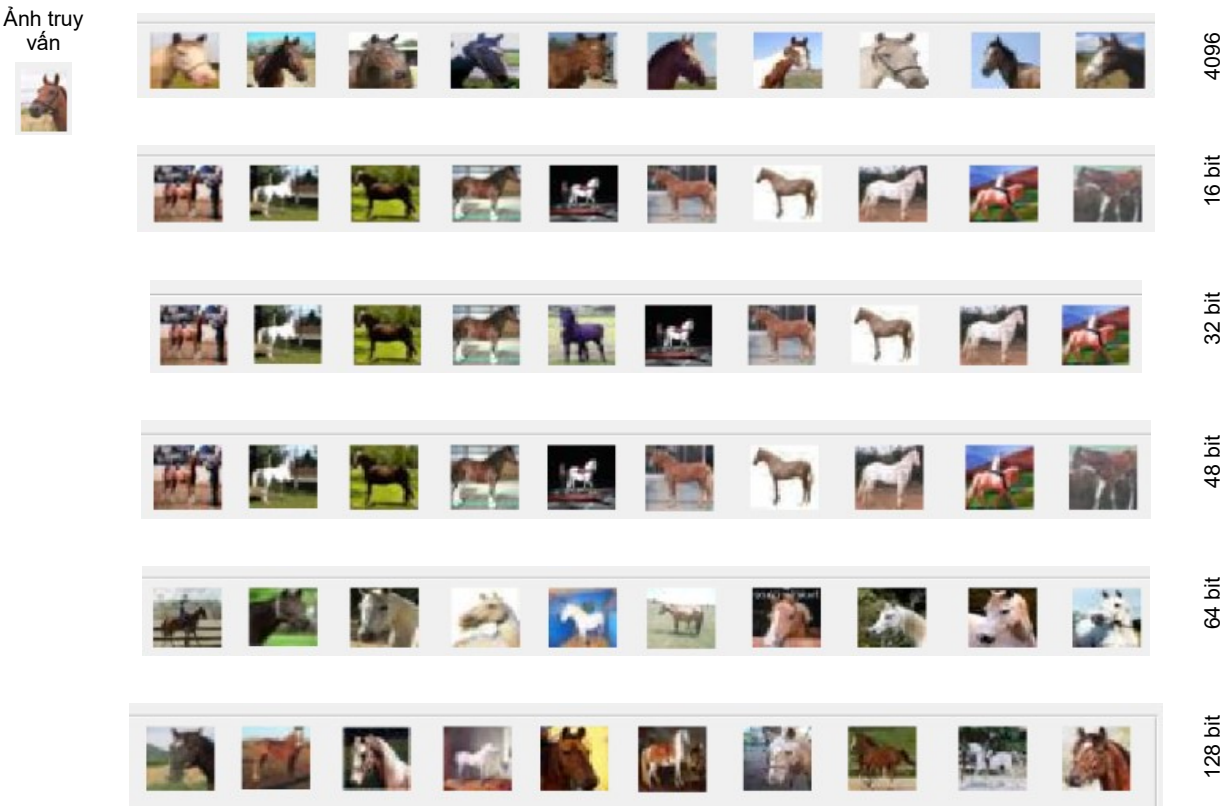
2.3.2. Kết quả thực nghiệm

Chúng tôi thực nghiệm các trường hợp: (1) Sử dụng đầu ra của lớp FC7 làm đặc trưng ảnh, khi đó, mỗi ảnh sẽ được biểu diễn bằng véctơ có độ dài 4096 và sử dụng khoảng cách Euclidean để đo độ tương tự giữa hai ảnh; (2) sử dụng phương pháp sinh mã nhị phân với số nơron trong lớp ẩn H lần lượt là 16, 32, 48, 64 và 128 chính là độ dài của vec-tơ đặc trưng (hash code), trong trường hợp này, chúng tôi sử dụng khoảng cách Hamming để đo độ tương tự giữa hai ảnh. Kết quả thực nghiệm được cho trong bảng 1.

Kết quả thực nghiệm cho thấy việc sử dụng mạng CNN vào bài toán tìm kiếm ảnh theo nội dung cho kết quả tìm kiếm với độ chính xác cao, tuy nhiên thời gian truy vấn khá lâu. Việc áp dụng mạng CNN kết hợp với phương pháp sinh mã nhị phân không những làm tăng hiệu suất tìm kiếm mà còn cải thiện rất nhiều về thời gian truy vấn ảnh.

Bảng 1. Kết quả thực nghiệm

Bộ dữ liệu	Độ dài véctơ đặc trưng	Độ đo	Độ chính xác (%)	Thời gian truy vấn (giây)
CIFAR-10	4096	Euclidean	87,48	17,43
	16 (bit)	Hamming	89,77	0,02
	32 (bit)	Hamming	89,72	0,04
	48 (bit)	Hamming	89,74	0,06
	64 (bit)	Hamming	89,79	0,07
	128 (bit)	Hamming	89,79	0,13
MNIST	4096	Euclidean	98,12	15,32
	48 (bit)	Hamming	98,31	0,05
	128 (bit)	Hamming	98,43	0,11



Hình 11. Top 10 ảnh kết quả tìm kiếm ảnh trên bộ dữ liệu CIFAR-10

Hình 12. Top 10 ảnh kết quả tìm kiếm ảnh trên bộ dữ liệu MNIST

3. KẾT LUẬN

Khi sử dụng mạng CNN trên tập hai tập dữ liệu CIFAR-10 và MNIST cho kết quả tìm kiếm với độ chính xác cao (~ 87% trên bộ dữ liệu CIFAR-10 và (~ 98% trên bộ dữ liệu MNIST) nhưng thời gian truy vấn lâu (trên 10 giây). Thực hiện sinh mã nhị phân bằng cách chèn thêm một lớp ẩn vào mạng CNN cho thấy thời gian truy vấn đã giảm đáng kể còn chưa đến 1 giây. Nghiên cứu này đã cho thấy việc áp dụng mạng CNN và phương pháp sinh mã nhị phân đáng được quan tâm để góp phần nâng cao hiệu quả cho các hệ thống tìm kiếm ảnh.

TÀI LIỆU THAM KHẢO

Babenko A., Slesarev A., Chigorin A. & Lempitsky V. (2014). Neural codes for image retrieval. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 8689 LNCS(PART 1). pp. 584-599.

Cheng Z., Yang Q. & Sheng B. (2015). Deep Colorization. Proceedings of the IEEE International Conference on Computer Vision. pp. 415-423.

Ciressan D., Meier U. & Schmidhuber J. (2012). Multi-column Deep Neural Networks for Image Classification. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Jia Y., Shelhamer E., Donahue J., Karayev S., Long J., Girshick R., Guadarame S. & Darrell T. (2014). Caffe: Convolutional Architecture for Fast Feature Embedding. UC Berkeley EECS, Berkeley, CA 94702

Krizhevsky A., Nair V., & Hinton G. (2009). CIFAR-10 and CIFAR-100 dataset. Retrieved from https://www.cs.toronto.edu/~kriz/cifar.html on May 12, 2018.

Krizhevsky A., Sutskever I. & E. Hinton G. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In Proc. NIPS.

Lecun Y., Cortes C., Burges C. (1998). MNIST handwritten digit database. Retrieved from http://yann.lecun.com/exdb/mnist/ on May 12, 2018.

Lecun Y., Bottou L., Bengio Y. & Haffner P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE.

Li H., Lin Z., Shen X., Brandt J. & Hua G. (2015). A Convolutional Neural Network Cascade for Face Detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5325-5334.

Luo P., Tian, Y., Wang X. & Tang X. (2014). Switchable Deep Network for Pedestrian Detection Ping. Computer Vision Fundation.

Lê Minh Phúc & Trần Công Án (2017). Tìm kiếm ảnh theo nội dung và ngữ nghĩa. Tạp chí Khoa học, Trường Đại học Cần Thơ. Số chuyên đề: Công nghệ thông tin. tr. 58-64.

Photoindustrie-Verband e.V. (2016). Photo and imaging market: Trend report. p. 63.

Văn Thế Thành (2017). Tìm kiếm ảnh dựa trên đồ thị chữ ký nhị phân. Luận án tiến sĩ khoa học máy tính. Đại học Huế.

Văn Thế Thành & Lê Mạnh Thạnh (2016). Một số cải tiến cho hệ truy vấn ảnh dựa trên cây S-Tree. Kỷ yếu hội nghị khoa học quốc gia lần thứ IX - Nghiên cứu cơ bản và ứng dụng công nghệ thông tin (FAIR'9).