

TÌM KIẾM SẢN PHẨM THEO ẢNH BẰNG HỌC SÂU

Dương Trần Đức

Học viện Công nghệ Bưu chính Viễn thông

Tóm tắt: Tìm kiếm thông tin là một trong những nhu cầu cần thiết của con người. Trước kia, hình thức tìm kiếm chủ đạo là tìm kiếm bằng từ khóa, với ưu điểm là đơn giản và dễ dùng, tuy nhiên đôi khi lại không đủ khả năng để diễn tả toàn bộ ý muốn của con người. Trong những năm gần đây, nhận biết được nhược điểm trên, cùng với sự phát triển của khoa học công nghệ, việc tìm kiếm bằng ảnh đã bắt đầu được nghiên cứu và áp dụng. Bài báo này trình bày một phương pháp tìm kiếm sản phẩm thông qua ảnh sử dụng thuật toán học sâu và kỹ thuật so sánh ảnh tương tự trên tập dữ liệu tự thu thập từ Internet. Các kết quả cho thấy phương pháp có tiềm năng ứng dụng trong thực tế.

Từ khóa: tìm kiếm ảnh, nhận dạng ảnh, học sâu, mạng nơ ron tích chập.

I. MỞ ĐẦU

Trong các hệ thống thương mại điện tử (TMĐT), chức năng tìm kiếm là một trong các chức năng quan trọng, giúp cho hệ thống có thể giới thiệu các sản phẩm, hàng hoá đến người dùng một cách chính xác nhất. Các hệ thống TMĐT truyền thống thường chỉ cung cấp chức năng tìm kiếm theo từ khóa, dựa trên các mô tả bằng văn bản về sản phẩm cần tìm. Phương pháp tìm kiếm này trong nhiều trường hợp gây ra sự khó khăn và đạt hiệu quả không cao, do nhiều loại sản phẩm, hàng hoá khó có thể mô tả hết các đặc điểm bằng cách sử dụng văn bản. Khi đó, tìm kiếm bằng hình ảnh là một chức năng hữu hiệu, giúp cho người dùng có thể dễ dàng hơn trong việc mô tả đặc điểm của sản phẩm muốn tìm. Một hình ảnh có thể chứa đựng nhiều thông tin mô tả ẩn, khó có thể nhanh chóng mô tả bằng văn bản. Ngoài ra, nó còn giúp ích trong một số trường hợp, người dùng chỉ có thể biết về hình ảnh sản phẩm mà không biết cách mô tả chính xác nó bằng văn bản như thế nào.

Kỹ thuật tìm kiếm bằng hình ảnh không phổ biến trước đây do sự phức tạp trong xử lý cùng với khối lượng tính toán, cũng như việc yêu cầu người dùng phải có thiết bị chụp và lưu trữ ảnh dễ dàng thì mới có thể thực hiện được thao tác tìm kiếm. Trong những năm gần đây, do sự phát triển nhanh chóng và mạnh mẽ của phần cứng máy tính, cùng với sự ra đời của các thuật toán xử lý ảnh cũng như các công nghệ mới như trí tuệ nhân tạo, mạng học sâu v.v, việc xử lý bài toán tìm kiếm ảnh đã trở nên khả thi, và chúng tỏ được hiệu quả của nó so với tìm kiếm bằng từ khóa thông thường. Các máy tìm kiếm phổ biến cũng như một số hệ thống lớn như Google, Bing (Microsoft), Amazon,

Pinterest v.v. đã triển khai áp dụng hệ thống tìm kiếm bằng hình ảnh, và đạt được nhiều thành công. Ngoài ra, khi các thiết bị di động tích hợp camera như điện thoại, máy tính bảng trở nên cực kỳ phổ biến, việc thực hiện tìm kiếm ảnh đối với người dùng cũng trở nên dễ dàng, điều này khiến cho tìm kiếm ảnh ngày càng được sử dụng rộng rãi hơn.

Bài báo này đề xuất phương pháp tìm kiếm các sản phẩm bằng ảnh sử dụng thuật toán học sâu, trong đó người dùng có thể cung cấp một hình ảnh bất kỳ và hệ thống sẽ trả về danh sách các ảnh với các mức độ tương tự được sắp xếp từ cao xuống thấp. Công việc chính là việc nhận diện ảnh người dùng cung cấp thuộc thể loại nào và độ tương tự của nó với các ảnh thuộc thể loại đó. Việc phân loại ảnh người dùng được thực hiện dựa trên một mạng nơ ron tích chập (Convolution Neural Networks - CNN) và việc tính toán độ tương tự của ảnh được thực hiện dựa trên kỹ thuật tìm “láng giềng gần nhất” (Nearest-Neighbors) để trả về danh sách các ảnh sản phẩm có độ tương tự cao nhất với ảnh đầu vào.

Bài báo có cấu trúc như sau. Phần II trình bày về các nghiên cứu liên quan trong lĩnh vực phân loại và tìm kiếm ảnh bằng học sâu. Phần III mô tả phương pháp. Phần IV trình bày về các kết quả và thảo luận. Cuối cùng, các kết luận sẽ được trình bày trong phần V của bài báo.

II. TỔNG QUAN

A. Mô hình so sánh độ tương đồng hình ảnh

Ban đầu, các mô hình tìm kiếm hình ảnh thường tìm kiếm sự tương đồng theo nhãn (category). Hai bức ảnh được đánh giá là tương đồng nhau nếu như chúng có cùng nhãn, ví dụ như bức ảnh về hoa, bức ảnh về con người, v.v. Tuy nhiên, mức độ tương đồng này vẫn chưa đủ đối với các ứng dụng cấp cao hơn. Các bài toán sau này cần tìm kiếm sự tương đồng ở mức độ lớn, chi tiết hơn, kể cả khi hai bức ảnh có cùng nhãn, nhằm nhận diện sự khác nhau dù nhỏ nhất.

Một phương pháp xây dựng sự giống nhau về hình ảnh đó là sử dụng các thuật toán trích xuất đặc trưng (feature) của ảnh như bộ lọc SIFT (Scale-Invariant Feature Transform) [14], HOG (Histogram of Oriented Gradient) [6], rồi sử dụng các đặc trưng đó để tính toán sự tương đồng giữa hai bức ảnh. Phương pháp này đã được sử dụng trong các nghiên cứu [3, 4]. Tuy nhiên, những mô hình này bị giới hạn bởi khối lượng tính toán quá lớn. Trong những năm gần

đây, mô hình mạng nơ ron tích chập (CNN) được sử dụng phổ biến cho việc nhận dạng và phân loại hình ảnh đã đem lại một cách tiếp cận mới cho việc tính toán độ tương đồng hình ảnh [12, 15]. Các mô hình học sâu, đặc biệt là CNN, có khả năng tìm các đặc trưng từ bậc thấp bậc cao với độ chính xác ổn định, điều này giúp rất nhiều trong việc trích xuất các đặc trưng chính của bức ảnh để phục vụ quá trình so sánh. Các lớp CNN kế tiếp nhau sẽ biểu thị hình ảnh theo các mức độ trừu tượng khác nhau. Lớp cuối cùng làm một véc tơ đại diện cho ảnh, có thể dùng để làm đặc trưng tính toán độ tương đồng hình ảnh.

Để so sánh độ tương đồng hình ảnh từ các véc tơ đặc trưng thu được, có thể sử dụng các phép đo khoảng cách. Một số phép đo thông dụng bao gồm L1-norm (khoảng cách Manhattan), L2-norm (khoảng cách Euclid), hay độ tương đồng Cosine (cosine similarity). Cần chú ý rằng, khoảng cách và độ tương đồng tỉ lệ nghịch với nhau, khoảng cách càng lớn thì độ tương đồng càng nhỏ và ngược lại.

Với 2 vector x, y độ dài m , khoảng cách Manhattan được tính như sau:

$$l_1 = \sum_{i=1}^m |x_i - y_i| \quad (1)$$

Công thức cho khoảng cách:

$$l_2 = \sqrt{\sum_{i=1}^m (x_i - y_i)^2} \quad (2)$$

Công thức cho khoảng cách Cosine:

$$similarity = \frac{\sum_{i=1}^m x_i y_i}{\sum_{i=1}^m (x_i)^2 \sum_{i=1}^m (y_i)^2} \quad (3)$$

Để làm tăng độ chính xác của việc tìm kiếm ảnh thông qua tính toán độ tương đồng như trên, có thể áp dụng bước phân loại ảnh trước, sau đó tiến hành tính toán độ tương đồng trên các ảnh có cùng nhãn (loại) với ảnh được tìm kiếm. Các ảnh không cùng nhãn cũng có thể được tính toán độ tương đồng, nhưng nằm ở phần sau của kết quả tìm kiếm.

Một phương pháp khác cũng được sử dụng để tính toán độ tương đồng hình ảnh đó là sử dụng mạng bộ ba (triplet networks) [13, 16]. Phương pháp này sử dụng các đầu vào gồm 3 thành phần là ảnh đầu vào (anchor), ảnh dương (positive) giống ảnh đầu vào, và ảnh âm (negative) khác với ảnh đầu vào. Các mạng loại này có chức năng khá đặc biệt là dùng để tính toán độ tương đồng hình ảnh chứ không phải gắn nhãn phân loại ảnh như các mạng CNN khác. Bộ ba ảnh đầu vào được đưa vào ba mạng riêng biệt (có trọng số chia sẻ), và sẽ được tạo ra các đặc trưng của từng ảnh ở lớp cuối của mạng. Sau đó, các chuỗi này sẽ được so sánh độ tương đồng dựa trên các thuật toán đã được nêu ở trên. Mạng có nhiệm vụ sinh ra các đặc trưng sao cho khoảng cách giữa ảnh đầu vào tới ảnh dương phải lớn hơn khoảng cách tới ảnh âm. Ưu điểm của phương pháp này là tạo ra được các đặc trưng có thể thể hiện nhiều đặc tính của ảnh hơn, nhưng quá trình chuẩn bị dữ liệu tốn nhiều công sức hơn và thời gian huấn luyện lâu hơn.

Trong nghiên cứu này, chúng tôi sử dụng phương pháp thứ nhất do nó có ưu điểm về thời gian và phù hợp với hạn chế về mặt chuẩn bị dữ liệu huấn luyện.

B. Tìm kiếm ảnh

Vấn đề tìm kiếm sản phẩm theo ảnh đã được quan tâm và thực hiện trong một số nghiên cứu trước đây [1, 2, 11].

Kiapour et al. [11] thực hiện nghiên cứu việc tìm các sản phẩm tương tự trên các trang TMĐT. Các tác giả đã thực hiện và so sánh một số phương pháp, trong đó nổi bật là phương pháp sử dụng mạng CNN hai lớp ẩn và thực nghiệm trên tập dữ liệu Exact Street2Shop. Borras et al. [1] đề xuất cách kết hợp 5 đặc tính của sản phẩm quần áo thời trang trong một cấu trúc đồ họa nhằm xác định xem một người mặc đồ như thế nào từ các hình ảnh thu được, tuy nhiên độ chính xác chỉ đạt được 64%. Bossard et al. [2] cũng giải quyết vấn đề liên quan đến nhận dạng và tìm kiếm ảnh sản phẩm thời trang, tuy nhiên kết quả đạt được cũng còn hạn chế về độ chính xác.

Vấn đề tìm kiếm ảnh cũng được quan tâm nghiên cứu và áp dụng trong các hệ thống như máy tìm kiếm, mạng xã hội v.v. Jing et al. [8] phát triển một hệ thống tìm kiếm theo ảnh có tính hiệu quả và ổn định cao và đã áp dụng cho mạng xã hội Pinterest. Phương pháp này có hiệu quả về chi phí nhưng có năng lực biểu cảm hình ảnh không cao. Các máy tìm kiếm như Google hay Bing [9] cũng đã nghiên cứu và áp dụng tính năng tìm kiếm ảnh bằng mạng nơ ron học sâu, nhưng phải cân đối giữa độ chính xác và tốc độ phản hồi.

Trong nghiên cứu này, chúng tôi thực hiện tìm kiếm ảnh sản phẩm, có tính đặc thù hơn so với các hệ thống như mạng xã hội hay máy tìm kiếm, nhưng có tính tổng quát hơn các nghiên cứu thực hiện trên các tập dữ liệu ảnh sản phẩm thời trang.

III. PHƯƠNG PHÁP

Phương pháp tìm kiếm theo ảnh áp dụng trong bài báo này được bao gồm hai giai đoạn: phân loại ảnh và so sánh độ tương đồng với các ảnh trong cùng loại để tìm ra các ảnh có độ tương đồng cao nhất. Với một ảnh đầu vào của một sản phẩm được cung cấp, nó sẽ được phân loại thành loại sản phẩm gì. Sau đó, các hình ảnh sản phẩm khác cùng loại giống nó nhất sẽ được tính toán và trả về kết quả tìm kiếm. Phần này sẽ trình bày về phương pháp được áp dụng để phân loại và tìm kiếm ảnh tương đồng như đã nói ở trên.

A. Phân loại ảnh bằng mạng nơ ron tích chập

Mạng nơ ron tích chập (CNN) cho phân loại ảnh nhận đầu vào là một ảnh với 3 chiều biểu diễn là dài, rộng, sâu (chiều dài, rộng của ảnh và chiều sâu thể hiện các màu sắc ảnh). Mỗi lớp của mạng CNN sẽ chuyển đổi 1 khối 3D (ma trận 3 chiều) thành 1 khối 3D khác. Có 3 loại lớp chính để xây dựng nên mạng CNN, đó là lớp tích chập (Convolution), lớp hợp nhất (Pooling), và lớp kết nối đầy đủ (Fully-Connected).

Lớp tích chập (CONV) là khối quan trọng nhất trong mạng neuron tích chập, nó thực hiện hầu hết khối lượng tính toán trong mạng. Nó dựa trên phép tích chập trên ma trận, phép toán này giúp giảm số lượng tính toán đi đáng kể so với các lớp kết nối đầy đủ. Với ma trận A có kích thước $h \times w \times d$, phép tính tích chập của A với một bộ lọc (filter) kích cỡ $f_h \times f_w \times d$ sẽ tạo ra một đầu ra có kích thước $(h - f_h + 1) \times (w - f_w + 1) \times 1$. Để thực hiện được một phép tính tích chập hoàn chỉnh trên một lớp CONV, ngoài tham số là số bộ lọc K , kích thước bộ lọc F , thì còn các tham số khác là kích thước bước nhảy mỗi lần dịch bộ lọc S , và kích thước lẻ P .

Các lớp hợp nhất (POOL) thường được sắp xếp xen kẽ với các lớp CONV một cách đều đặn. Lớp này có chức năng làm giảm nhanh chóng kích thước khối dữ liệu nhằm giảm số lượng hệ số những như khối lượng tính toán của toàn mạng, qua đó tránh được vấn đề quá khớp. Phép hợp nhất

đơn giản nhất thường được sử dụng đó là lấy giá trị lớn nhất của một vùng để đại diện cho vùng đó. Một hàm MAX trên bộ lọc kích thước 2x2 và kích thước bước nhảy 2 sẽ làm giảm đi 75% kích thước của khối dữ liệu đầu vào. Ngoài phép lấy giá trị lớn nhất thì các phép hợp nhất khác cũng được sử dụng như lấy giá trị trung bình hoặc hàm chuẩn hoá L2. Tuy nhiên, phép hợp nhất lấy giá trị lớn nhất được sử dụng phổ biến nhất hiện nay do tính hiệu quả của nó trong thực tế.

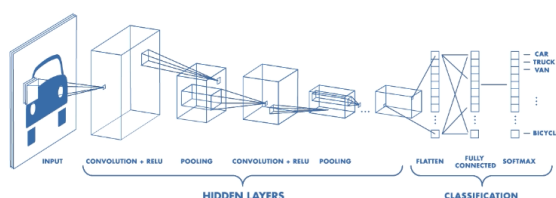
Lớp kết nối đầy đủ (FC) là lớp cuối cùng trong mạng nơ ron tích chập, có đầy đủ các kết nối tới các nơ ron liên trước như trong mạng nơ ron thông thường. Hàm kích hoạt của chúng có thể được tính bằng phép nhân ma trận cùng với một tham số là độ lệch (bias).

Dạng thông dụng nhất của một mạng CNN bao gồm một vài lớp CONV, tiếp sau đó là lớp POOL, và tiếp tục lặp lại chuỗi này cho tới khi ảnh được giảm tới kích thước đủ nhỏ. Khi đó lớp cuối cùng sẽ được duỗi thẳng thành một véc tơ dọc và thêm vào các lớp FC như mạng nơ ron truyền thống.

Nhìn chung, mạng CNN tổng quát có dạng như sau:

INPUT -> [CONV*N->POOL]*M -> FC*K -> FC

Trong đó N, M, K lần lượt là số lượng các lớp CONV, POOL, và FC.



Hình 1. Kiến trúc mạng CNN [mathworks.com]

Để thực hiện huấn luyện cho mạng CNN, có thể sử dụng tập dữ liệu riêng và thực hiện huấn luyện mạng từ đầu, tối ưu các tham số để mạng đạt kết quả phân loại tốt nhất. Phương pháp này cần một tập dữ liệu khá lớn và tài nguyên tính toán lớn, tỷ lệ với độ sâu của mạng. Đây là phương án cơ bản của các bài toán phân loại nói chung và sử dụng mạng nơ ron nói riêng: tự huấn luyện một bộ phân loại và tối ưu tham số. Tuy nhiên, đối với mạng CNN cho phân loại ảnh, phương pháp này không thật sự hiệu quả do dữ liệu đầu vào thường không được chuẩn bị tốt. Phương pháp tiếp cận khác là sử dụng một mạng CNN đã huấn luyện từ trước, và tối ưu lại tham số trên tập dữ liệu riêng theo phương pháp học chuyển giao (transfer learning) [10]. Phương pháp này vẫn cần thực hiện khối lượng xử lý khá lớn, nhưng có thể chấp nhận một tập dữ liệu huấn luyện nhỏ hơn, do phần lớn khối lượng xử lý đã được thực hiện trong quá trình huấn luyện mạng trước đó. Khối lượng xử lý còn lại được thực hiện trong quá trình học chuyển giao trên tập dữ liệu riêng.

Việc đánh giá độ chính xác của một mạng CNN cũng khá đơn giản. Sử dụng một tập dữ liệu kiểm tra, có thể đánh giá mạng tạo ra các kết quả có độ chính xác như thế nào nhờ các chỉ số đo thông thường như độ đo chính xác (accuracy). Để đảm bảo tính khách quan khi đánh giá mạng, tập dữ liệu kiểm tra sẽ được trích ngẫu nhiên từ tập dữ liệu ban đầu và không được dùng để huấn luyện mạng.

B. Thu thập ảnh tương tự

Quá trình thu thập ảnh tương tự cũng cần sử dụng mạng CNN. Mục tiêu chính của việc cho các hình ảnh qua mạng CNN không phải để phân loại mà để chuyển đổi hình ảnh về một đối tượng đại diện có kích thước và số chiều nhỏ hơn. Tất cả các ảnh trong cùng phân loại với ảnh đầu vào sẽ được cho qua mạng CNN để tạo ra các véc tơ đại diện X_i . Sau đó, véc tơ đại diện X' của ảnh đầu vào sẽ được so sánh với từng véc tơ X_i thu được ở trên bằng một phép đo độ tương đồng nào đó và các ảnh giống ảnh đầu vào nhất sẽ được trả về làm kết quả tìm kiếm theo phương pháp “láng giềng gần nhất” (k-nearest neighbors). Điểm mấu chốt của phương pháp này là cần tạo được véc tơ đại diện phản ánh chính xác và đầy đủ đặc trưng của ảnh và độ đo đánh giá sự tương đồng tốt. Độ đo sự tương đồng của 2 véc tơ được sử dụng trong bài báo là độ đo L2, do nó có tính phổ biến và đơn giản khi tính toán. Các véc tơ đại diện được tạo thông qua mạng CNN đã trình bày ở phần trước, nhưng không phải để phân loại mà được sử dụng như một bộ tạo đặc trưng. Theo đó, véc tơ đặc trưng ở lớp FC cuối cùng sẽ được sử dụng như véc tơ đại diện cho ảnh. Tất cả các ảnh trong tập dữ liệu sẽ được cho qua mạng CNN để tạo các véc tơ đại diện theo phương pháp trên. Khi một ảnh đầu vào được tìm kiếm, véc tơ đại diện của nó cũng được tạo theo phương pháp tương tự và được so sánh với tất cả các véc tơ đại diện của các ảnh trong tập dữ liệu. Các ảnh có độ tương đồng cao nhất (độ đo L2 thấp nhất) sẽ được chọn làm kết quả tìm kiếm.

Việc đánh giá độ chính xác của tác vụ thu thập ảnh tương tự khó khăn hơn so với đánh giá độ chính xác của tác vụ phân loại, do bản thân khái niệm “tương tự” trên thực tế đã có tính tương đối. Trong bài toán phân loại, một hình ảnh rõ ràng là thuộc lớp này hay lớp kia, làm cho việc đánh giá kết quả phân loại được thực hiện dễ dàng hơn. Tuy nhiên, việc đánh giá một hình ảnh nhìn có “giống” một hình ảnh khác không lại mang nhiều tính chủ quan, trong khi kết quả tìm kiếm hình ảnh liên quan đến việc đánh giá bề ngoài của hình ảnh. Do đó, việc đánh giá độ chính xác trong tác vụ này được thực hiện qua các thao tác lấy mẫu và đánh giá mang tính chủ quan.

IV. THỰC NGHIỆM VÀ KẾT QUẢ

A. Dữ liệu và môi trường thực nghiệm

Trong nghiên cứu này, chúng tôi sử dụng tập dữ liệu tự thu thập từ các trang ảnh và thương mại điện tử phổ biến như Pinterest, Mediamart, Hoà Phát, Canifa v.v. Các ảnh được thu thập đa dạng nguồn nhằm tăng tính khách quan cho vấn đề phân loại và tính toán độ tương tự hình ảnh. Tổng số 11.539 ảnh với 11 nhãn được phân bố như sau:

Bảng 1. Thống kê về tập dữ liệu thực nghiệm

Nhãn	Số lượng	Nguồn ảnh
Điều hoà	1.080	Pinterest, Mediamart
Quần	679	Pinterest, Canifa.com
Ghế	721	Pinterest, Hoà Phát, Xuân Hòa
Bàn	1.129	Pinterest, Nội thất Hòa Phát
Váy	1.219	Pinterest, Loza.vn
Tủ lạnh	1.465	Pinterest, Mediamart
Lò vi sóng	415	Pinterest, Mediamart

Bút	946	Pinterest, Thiên Long Flex Office
Nồi cơm điện	514	Pinterest, Mediamart
Áo	2.206	Pinterest, Canifa.com
Máy giặt	1.165	Pinterest, Mediamart

Các ảnh được phân bố với tỉ lệ 80% để huấn luyện, 20% để kiểm chứng mô hình. Ngoài ra, mỗi nhãn còn có thêm 200 ảnh với nguồn từ Google Images để làm bộ dữ liệu test.

Các thực nghiệm được thực hiện trên 2 hệ thống, dành cho 2 loại tác vụ khác nhau.

1) Môi trường thực hiện quá trình học máy: Sử dụng Google Colab:

- CPU: 1x Single core hyper threaded Xeon Processor @2.3Ghz
- GPU: 1x Tesla K80, 12GB GDDR5 VRAM
- RAM: 13GB
- Disk: 30GB

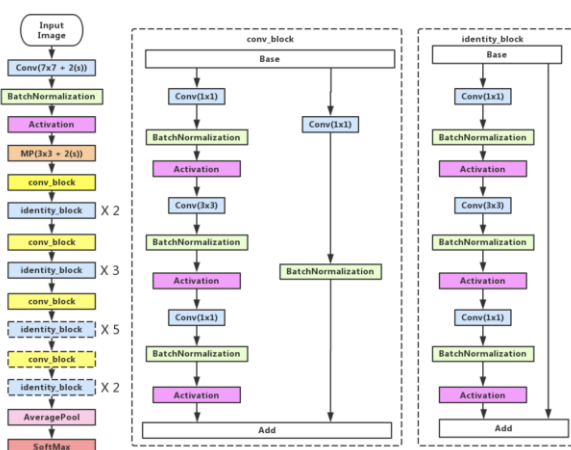
2) Môi trường thực hiện quá trình đưa dữ liệu ảnh qua mô hình học máy:

- CPU: Intel Core i5-4200H (2 cores, 4 threads) @2.8Ghz
- GPU: Nvidia GTX 950M, 4GB GDDR3 VRAM
- RAM: 12GB DDR3L
- Ổ cứng: SSD 128GB

B. Kiến trúc mạng

Mô hình học sâu của hệ thống sẽ sử dụng kiến trúc của mạng ResNet50. ResNet có tên đầy đủ là Residual Network, được phát triển bởi Kaiming He và các cộng sự. Nó nổi bật bởi nó có khả năng skip connection, tức là một phần dữ liệu đầu vào có thể tiếp tục đi qua các lớp sau mà không qua xử lí. Ngoài ra nó còn sử dụng một lượng lớn các lớp chuẩn hóa theo lô (Batch Normalization). ResNet cũng không sử dụng các lớp kết nối đầy đủ ở cuối mạng. ResNet là một trong những mạng CNN hiện đại nhất cho tới ngày nay, và là sự lựa chọn được tin dùng khi sử dụng CNN trong thực tế.

Mạng ResNet có kiến trúc như sau:



Hình 2: Kiến trúc mạng ResNet [7]

Các đặc điểm chính của mạng ResNet bao gồm [7]:

- Mỗi khối conv_block bao gồm hàng loạt các lớp CONV nhỏ, cùng với hàm kích hoạt ReLU và lớp chuẩn hoá theo lô.
- Ngoài ra, khối này còn có sử dụng skip connection (nhằm tránh giảm hiệu năng khi mạng quá sâu).
- Identity block có trật tự thực hiện gần giống như conv_block, tuy nhiên skip connection của nó không bao gồm tích chập 1x1 và chuẩn hoá theo lô, và quan trọng nhất, là các hàm tích chập của nó đều được chỉnh kích thước đệm sao cho kích thước của khối 3D trước và sau khi qua các lớp này vẫn không đổi.

ResNet có nhiều biến thể như ResNet50, ResNet101, ResNet152, ... Trong bài báo này, hệ thống sử dụng mạng ResNet50 để có thể có thời gian huấn luyện cũng như tìm kiếm ở mức vừa phải, hơn nữa tránh vấn đề quá khớp do lượng dữ liệu không lớn.

Tiếp sau mạng CNN, các lớp FC được sử dụng để huấn luyện bao gồm:

- Lớp ẩn: 128 nơ ron, hàm kích hoạt ReLU
- Lớp drop-out: giá trị dropout [0.1, 0.25, 0.4]
- Lớp ẩn: 64 nơ ron, hàm kích hoạt ReLU
- Lớp đầu ra: 11 nơ ron (tương ứng với số nhãn), hàm kích hoạt softmax.

Cuối cùng, mô hình sẽ sử dụng hàm chi phí là hàm cross-entropy.

$$L(y, \hat{y}) = - \sum_{j=0}^M \sum_{i=0}^N (y_{ij} * \log(\hat{y}_{ij})) \quad (4)$$

Hàm cross-entropy sẽ so sánh phân bố của đầu ra, với xác suất của nhãn đúng là 1 và xác suất của các nhãn khác là 0. Nói một cách khác, nhãn kết quả sẽ được dịch thành 1 véc tơ với 1 là nhãn đúng, 0 là nhãn sai, sau đó véc tơ này sẽ được so sánh với đầu ra của nơ ron. Độ sai lệch giữa 2 véc tơ càng nhỏ, giá trị hàm chi phí cũng nhỏ theo.

Một số các tham số được sử dụng:

- Tốc độ học: [0.01, 0.001, 0.0001]
- Giá trị drop-out: [0.1, 0.25, 0.4]
- Hàm tối ưu: nadam
- Hàm chi phí: cross-entropy
- Số vòng: [10, 20, 50]

C. Kết quả và đánh giá

Các thực nghiệm được thực hiện sử dụng thư viện Keras [5]. Với mạng ResNet50 được Keras cung cấp, độ chính xác được tính toán và so sánh trong 20 vòng (epochs) và mô hình có độ chính xác tốt nhất được lưu lại để làm kết quả thực nghiệm. Do mạng ResNet không cung cấp các lớp FC nên cần thêm một vài lớp FC ở cuối mạng để thực hiện phân loại.

Đối với hoạt động tính toán véc tơ đại diện cho các ảnh nhằm thực hiện so sánh và thu thập ảnh tương đồng, tất cả các ảnh trong tập dữ liệu được đưa qua mô hình mạng đã huấn luyện. Các véc tơ đại diện thu thập được tại lớp FC cuối cùng sẽ được lưu lại sử dụng trong quá trình tìm kiếm sau này.

Các tham số được thử nghiệm để tối ưu trong các khoảng bao gồm kích thước bộ lọc, số bộ lọc, kích thước đệm, kích thước bước nhảy, tốc độ học, giá trị drop out.

Các tham số trên được thực hiện tối ưu qua các thực nghiệm và bảng 2 cho thấy kết hợp tốt nhất của các tham số.

Bảng 2. Giá trị tham số tối ưu

Tham số	Giá trị tối ưu
Kích thước bộ lọc	3x3
Số bộ lọc	1000
Kích thước đệm	1
Kích thước bước nhảy	2
Số drop-out	0,4
Tốc độ học	0,001

Với môi trường thực nghiệm như trình bày ở trên, kết quả phân loại sản phẩm đạt được độ chính xác tổng thể là 85.09%, trong đó loại sản phẩm Váy có độ chính xác tốt nhất (94.57%) và loại sản phẩm Bàn có độ chính xác thấp nhất (65.83%).

Bảng 3 cho thấy kết quả phân loại chi tiết cho từng loại sản phẩm:

Bảng 3. Kết quả phân loại

Nhãn	Độ chính xác (%)	Nhãn	Độ chính xác (%)
Điều hoà	83.82	Lò vi sóng	80.90
Quần	88.37	Bút	78.79
Ghế	92.50	Nồi cơm điện	90.05
Bàn	65.82	Áo	78.24
Váy	94.57	Máy giặt	87.18
Tủ lạnh	90.82		

Như đã trình bày ở phần III, việc đánh giá kết quả thu thập ảnh tương tự có sự khó khăn hơn, do dựa nhiều vào đánh giá chủ quan. Trong nghiên cứu này, chúng tôi thực hiện đánh giá tương tự như trong [11], theo đó lấy ngẫu nhiên 100 sản phẩm và dùng làm ảnh đầu vào cho quá trình tìm kiếm. Thu thập 5 kết quả đầu tiên của mỗi ảnh đầu vào, tiến hành đánh giá chủ quan theo thang điểm 1-5 về độ tương tự của nó với ảnh đầu vào. Các kết quả 4, 5 được xem là tương đồng và nhỏ hơn 4 được xem là không tương đồng. Độ chính xác được tính là tổng số ảnh được đánh giá tương đồng trên tổng số ảnh thu được.

$$\text{Độ chính xác tìm kiếm} = \frac{\text{Số ảnh tương đồng}}{\text{Tổng số ảnh thu được}}$$

Hình 2 cho thấy ví dụ về kết quả tìm kiếm với hình ảnh nồi cơm điện. Với kết quả trong hình 2, có thể đánh giá chủ quan là có 4 ảnh tương đồng và 1 ảnh không tương đồng trong 5 ảnh đầu tiên của kết quả tìm kiếm.

Kết quả cuối cùng cho thấy độ chính xác tìm kiếm theo đánh giá chủ quan trên toàn bộ 100 mẫu là 75%.

Các kết quả nhận được cho thấy sự khả quan của phương pháp phân loại cũng như thu thập ảnh tương đồng được áp dụng trên tập dữ liệu tương ứng. Mặc dù có sự chênh lệch giữa kết quả phân loại của các loại sản phẩm khác nhau, nhưng kết quả trung bình cũng như kết quả riêng của hầu hết các loại sản phẩm là có nhiều tiềm năng trong khoảng chấp nhận được của các ứng dụng phân loại ảnh. Kết quả thu thập ảnh tương tự cũng có độ chính xác khá tốt, dù vậy mới trải qua đánh giá mang tính chủ quan.

Về thời gian chạy, việc huấn luyện mô hình phân loại mất khoảng 80 phút, trong khi thời gian để đưa toàn bộ hơn 11.539 ảnh qua mô hình để thu thập véc tơ đại diện là 70 phút. Thời gian để thực hiện tìm kiếm từ khi cung cấp ảnh đầu vào đến khi trả về kết quả là 10 giây (sau khi đã có mô hình và có các véc tơ đại diện của các ảnh trong tập ảnh để so sánh).



Hình 2. Kết quả tìm kiếm với hình ảnh nồi cơm điện

V. KẾT LUẬN

Trong nghiên cứu này, chúng tôi đã trình bày phương pháp sử dụng mạng nơ ron tích chập để phân loại và so sánh độ tương đồng để tìm kiếm các ảnh tương tự trên tập các ảnh sản phẩm gồm 11 nhãn. Các kết quả thực nghiệm cho thấy tính khả thi của phương pháp khi áp dụng trên các ứng dụng thương mại điện tử hoặc máy tìm kiếm, cả về độ chính xác tìm kiếm và thời gian thực hiện.

Phương pháp có ưu điểm là tận dụng được khả năng xử lý của mạng nơ ron tích chập cho cả thao tác phân loại và tính toán véc tơ đại diện cho các ảnh trong tập ảnh tìm kiếm. Tuy nhiên, nhược điểm của phương pháp là việc sử dụng mạng nơ ron tích chập đơn thuần có thể không tạo ra các véc tơ đại diện có thể thể hiện tốt được sự tương đồng về mặt thị giác giữa các hình ảnh.

Hướng phát triển tiếp theo có thể là tiến hành nghiên cứu việc thực hiện tìm kiếm ảnh sản phẩm sử dụng một mạng bộ ba (triplet network) nhằm tạo ra một cơ chế có thể nắm bắt được tốt hơn sự tương đồng về các chi tiết thị giác của sản phẩm so với việc sử dụng mạng nơ ron tích chập

thông thường. Ngoài ra, các tham số của mạng nơ ron tích chập cũng cần được bổ sung và mở rộng khoảng khi thực hiện tối ưu nhằm tìm ra bộ tham số tốt nhất. Do việc huấn luyện trên mạng nơ ron tích chập 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.

TÀI LIỆU THAM KHẢO

- [1] Agnes Borras, Francesc Tous, Josep Lladós, Maria Vanrell, High-Level Clothes Description Based on Color-Texture and Structural Features, In: Lecture Notes in Computer Science, Iberian Conference, Pattern Recognition and Image Analysis (2003)
- [2] Lukas Bossard, Matthias Dantone, Christian Leistner, Christian Wengert, Till Quack, Luc Van Gool, Apparel Classification with Style”, In: Computer Vision-ACCV 2012, Springer (2013)
- [3] Y-Lan Boureau, Francis Bach, Yann LeCun, and Jean Ponce, Learning Mid-Level Features for Recognition, In Proc. CVPR (2010)
- [4] Gal Chechik, Varun Sharma, Uri Shalit, and Samy Bengio, Large Scale Online Learning of Image Similarity Through Ranking, Journal of Machine Learning Research 11, p. 1109–1135 (2010)
- [5] F. Chollet, & others, Keras, Available at: <https://github.com/fchollet/keras> (2015)
- [6] Navneet Dalal and Bill Triggs, Histograms of Oriented Gradients for Human Detection, In Proc. CVPR. p.886–893 (2005)
- [7] Q. Ji, J. Huang, W. He, Y. Sun, Optimized Deep Convolutional Neural Networks for Identification of Macular Diseases from Optical Coherence Tomography Images, Algorithms 12(3), 51 (2019).
- [8] Yushi Jing, David Liu, Dmitry Kislyuk, Andrew Zhai, Jiajing Xu, Je Donahue, and Sarah Tavel, Visual Search at Pinterest, In Proc. KDD, p.1889–1898 (2015)
- [9] H. Hu, Y. Wang, L. Yang, P. Komlev, L. Huang, X. S. Chen, Web-scale Responsive Visual Search at Bing, Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, pp. 359–367 (2018)
- [10] Andrej Karpathy, Transfer Learning, (2015), URL: <http://cs231n.github.io/transfer-learning/>
- [11] N. Khosla, and V. Venkataraman, Building Image-Based Shoe Search Using Convolutional Neural Networks, CS231N Course Project Reports, (2015)
- [12] M. Hadi Kiapour, Xufeng Han, Svetlana Lazebnik, Alexander C. Berg, and Tamara L. Berg, Where to Buy It: Matching Street Clothing Photos in Online Shops, In Proc. ICCV, (2015)
- [13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, ImageNet Classification with Deep Convolutional Neural Networks, In Proc. NIPS, p.1106–1114 (2012)
- [14] Hanjiang Lai, Yan Pan, Ye Liu, and Shuicheng Yan, Simultaneous Feature Learning and Hash Coding with Deep Neural Networks, In Proc. CVPR, p.3270–3278 (2015)
- [15] David G. Lowe, Object Recognition from Local Scale-Invariant Features, In Proc. ICCV, p.1150–1157 (1999)
- [16] K. Simonyan and A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, In Proc. ICLR (2015)
- [17] Jiang Wang, Yang Song, Omas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu, Learning Fine-Grained Image Similarity with Deep Ranking, In Proc. CVPR, p.1386–1393 (2015)

PRODUCT IMAGE SEARCH USING DEEP LEARNING

Abstract: Author profiling is the task of identify the characteristics of authors such as gender, age, etc. only based on analysis of their texts. This paper show research

on author profiling of Vietnamese texts using deep learning based on Convolutional Neural Networks (CNN). The experiments were conducted on the datasets which was experimented in the previous research on author profiling of Vietnamese texts [8]. The experiments show that the new method has better results than the traditional machine learning methods such as SVM (Support Vector Machine) and Random Forest on author profiling task.

Keywords: deep learning, convolutional neural network, author profiling.



Dương Trần Đức Tốt nghiệp Đại học KHTN, Đại học Quốc gia Hà Nội ngành Công nghệ thông tin năm 1999, Thạc sĩ chuyên ngành Hệ thống thông tin tại Đại học Tổng hợp Leeds, Vương Quốc Anh năm 2004, và Tiến sĩ chuyên ngành Kỹ thuật máy tính tại Học viện Công nghệ Bưu chính Viễn thông năm 2018. Hiện đang công tác tại Khoa Công nghệ Thông tin, Học viện Công nghệ Bưu chính Viễn thông.