ĐÁNH GIÁ ĐỘ TƯƠNG ĐỒNG HÌNH ẢNH BẰNG HỌC SÂU SỬ DỤNG MẠNG BỘ BA

Dương Trần Đức
Học viện Công nghệ Bưu chính Viễn thông

Tóm tắt: Đánh giá độ tương đồng hình ảnh là một trong các vấn đề quan trọng của thị giác máy tính, đặc biệt là khi ứng dụng vào vấn đề tìm kiếm theo ảnh. Trong các phương pháp được nghiên cứu, phương pháp đánh giá độ tương đồng sử dụng mạng bộ ba (triplet networks) là một phương pháp đem lại nhiều ưu điểm. Một mạng bộ ba thường bao gồm 3 mạng nơ ron tích chập CNN (Convolutional Neural Network) thành phần được chia sẻ trọng số và biểu thị các đặc trưng bậc cao của ảnh sao cho các ảnh giống nhau thì có khoảng cách gần và các ảnh khác nhau thì có khoảng cách xa. Bài báo này trình bày phương pháp sử dụng mạng bộ ba để đánh giá độ tương đồng của các ảnh sản phẩm và ứng dụng vào trong bài toàn tìm kiếm sản phẩm theo ảnh. Các kết quả thực nghiệm trên tập dữ liệu ảnh sản phẩm tự thu thập được cho thấy phương pháp có nhiều tiềm năng.

Từ khóa: học sâu, mạng nơ ron tích chập, mạng bộ ba, tìm kiếm theo ảnh.

I. MỞ ĐẦU

Đánh giá độ tương đồng hình ảnh (image similarity) là việc so sánh các đặc trưng (về màu sắc, bố cục, hình dáng, .v.v.) của ảnh để kết luận hai ảnh có giống nhau hay không [22]. Việc đánh giá độ tương đồng hình ảnh phụ thuộc rất lớn vào phương pháp trích chọn đặc trưng từ ảnh và phương pháp đánh giá sự tương đồng của các đặc trưng này.

Các kỹ thuật trích chọn đặc trưng ảnh để so sánh độ tương đồng đã được nghiên cứu từ hàng thập kỷ trước đây. Đây là một vấn đề có nhiều thách thức, và trong thời kỳ đầu, các kỹ thuật trích chọn chưa thực sự biểu diễn được các đặc trưng mức cao của ảnh mà thường chỉ được cảm nhận tốt bởi con người. Trong những năm gần đây, các mô hình học sâu đã được nghiên cứu và sử dụng để giải quyết các bài toán học máy nói chung và xử lý ảnh nói riêng. Điển hình là mạng nơ ron tích chập CNN đã có các ứng dụng đột phá xong các vấn đề như phân loại ảnh, nhận dạng ảnh, phát hiện vật thể, tìm kiếm ảnh, v.v. Sử dụng các cấu trúc của mạng nơ ron sâu, các đặc trưng bậc cao gần với cảm nhận của con người có thể được trích chọn từ các ảnh, nhằm phục vụ cho việc so sánh độ tương đồng tốt hơn. Nhiều mô hình mạng CNN phức tạp và hiệu quả đã được các nghiên cứu phát triển và công bố như mạng Alexnet, VGGNet, GoogLetNet, ResNet, v.v.

Mặc dù việc sử dụng các mạng CNN để trích chọn đặc trưng ảnh đã cho các biểu diễn tốt hơn so với các phương pháp truyền thống, việc sử dụng một mạng CNN duy nhất chưa thực sự đem lại hiệu quả cao nhất trong việc phân biệt các ảnh giống và khác nhau. Một số mô hình mạng cải tiến được sử dụng trong việc đánh giá độ tương đồng hình ảnh là mạng Siamese [5, 17] và mạng bộ ba (triplet networks) [8, 22]. Thay vì sử dụng chỉ một mạng CNN, các mô hình này thường sử dụng hai hoặc ba mạng CNN để thu được các đặc trưng biểu thị một ảnh sao cho các ảnh giống nhau thì có khoảng cách gần và các ảnh khác nhau thì có khoảng cách xa. Nhờ đó, phương pháp này mang lại hiệu quả tốt hơn trong việc phân biệt các ảnh giống hay khác nhau.

Bài báo này trình bày phương pháp đánh giá độ tương đồng hình ảnh qua mạng bộ ba (triplet networks) và ứng dụng trong bài toán tìm kiếm ảnh sản phẩm. Các thực nghiệm được thực hiện trên tập ảnh được thu thập từ mạng Internet với 1.200 ảnh các sản phẩm có gán nhãn thủ công và phân chia vào các bộ ảnh tương tự và khác nhau làm đầu vào cho mạng bộ ba. Sau khi sử dụng mạng bộ ba để trích chọn đặc trưng cho các ảnh, độ tương tự của các ảnh được tính 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. Các phương pháp trích chọn đặc trưng ảnh

Trong thời kỳ đầu, các phương pháp được sử dụng phổ biến là các thuật toán trích xuất đặc trưng (feature) của ảnh như bộ lọc SIFT (Scale-Invariant Feature Transform) [15], 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 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 [13, 16]. 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.

Tác giả liên hệ: Dương Trần Đức,
Email: duongtranduc@gmail.com
Đến tòa soạn: 28/7/2021, chỉnh sửa: 17/11/2021, chấp nhận đăng: 27/11/2021.

Một cải tiến của phương pháp sử dụng mạng CNN là phương pháp sử dụng các mạng gồm nhiều mạng CNN nhánh, như mạng Siamese[5, 17] hoặc mạng bộ ba [8, 22]. Phương pháp này sử dụng các đầu vào gồm 2 hoặc 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 gán 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. Phần tiếp theo sẽ trình bày chi tiết hơn về loại mạng này.

Nghiên cứu này là một mở rộng của nghiên cứu trước [7]. Trong [7], chúng tôi sử dụng mạng CNN thông thường và đã đem lại những kết qủa khả quan. Nghiên cứu này khai thác mở rộng phương pháp sử dụng mạng bộ ba. Mặc dù thời gian huấn luyện và vấn đề chuẩn bị dữ liệu phức tạp hơn, nhưng kết quả được cải thiện cho thấy ưu điểm của phương pháp mạng bộ ba so với mạng CNN thông thường.

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, 13]. Kiapour et al. [13] 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 hoạ 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. [10] 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 [11] 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 *x* w *x* d, phép tính tích chập của A với một bộ lọc (filter) kích cỡ f_h *x* f_w *x* d sẽ tạo ra một đầu ra có kích thước $(h - f_h + 1)$ *x* $(w - f_w + 1)$ *x* 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 lề 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 liền 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.

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). 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 qủa 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.

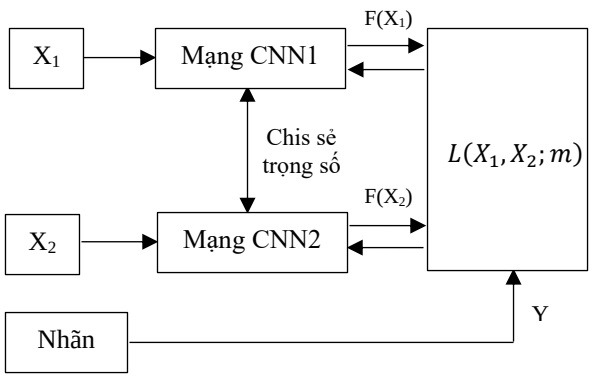
B. Trích chọn đặc trưng ảnh bằng mạng Siamese và mạng bộ ba

Mạng Siamese là một loại mạng học sâu cho phép học các đặc trưng tương đồng của một ảnh bằng cách tối ưu khoảng cách đặc trưng giữa các cặp ảnh. Mạng Siamese bao gồm hai mạng CNN nhánh có chia sẻ trọng số và các tham số. Mỗi mạng CNN này được loại bỏ đi lớp cuối cùng (lớp phân loại). Mô hình của mạng được biểu thị trong hình xx, trong đó hàm F biểu thị đặc trưng từ mỗi ảnh được trích xuất bởi mạng CNN. Mô hình mạng Siamese sử dụng một cặp ảnh X_1 và X_2 làm đầu vào và xây dựng một hàm mất mát L có công thức như trong công thức (1). Hàm mất mát này cố gắng tối thiểu hoá khoảng cách giữa các đặc trưng của cặp ảnh giống nhau và tối đa hoá khoảng cách đặc trưng giữa các cặp ảnh khác nhau. Mục tiêu cuối cùng là làm tối thiểu hoá giá trị làm mất mát và chọn lọc được các tham số tốt nhất cho mạng từ tập dư liệu huấn luyện.

$$L(X_1, X_2; m) =$$

$$\frac{1}{2} Y * D(X_1, X_2) + \frac{1}{2}(1 - Y) * \max(0, m - D(X_1, X_2)) \quad (1)$$ [22]

Trong đó Y là nhãn nhị phân của cặp ảnh đầu vào X_1, X_2. Y = 0 nếu cặp ảnh khác nhau và Y=1 nếu cặp ảnh giống nhau. Tham số m là ngưỡng lề giữa cặp ảnh giống và cặp ảnh khác.



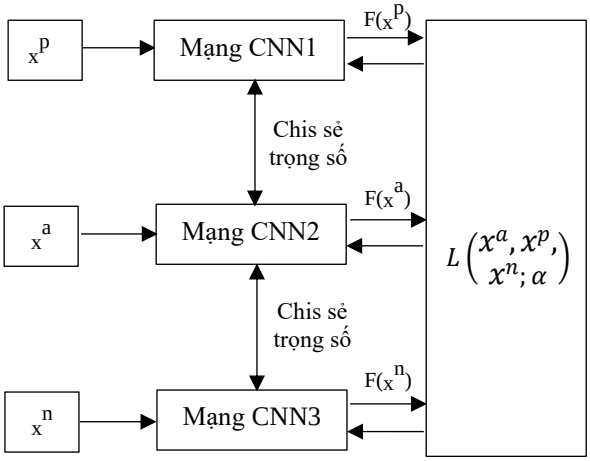
Hình 1. Mạng Siamese [22]

Mạng bộ ba là một cải tiến của mạng Siamese, được Wang đề xuất năm 2014. Khác với mạng Siamese, mạng này bao gồm ba mạng CNN nhánh có chia sẻ trọng số. Như vậy, mạng bộ ba sẽ nhận đồng thời ba ảnh đầu vào, gọi là ảnh chuẩn (x^a), ảnh giống (x^p), và ảnh khác (x^n). Cặp ảnh x^a và x^p là cặp ảnh giống nhau (hoặc cùng loại). Cặp ảnh x^a và x^n là cặp ảnh khác nhau (hoặc khác loại). Tương tự như mạng Siamese, hàm mất mát của mô hình này được xây dựng sao cho tối thiểu hoá khoảng cách giữa ảnh giống nhau và tối đa hoá khoảng cách giữa 2 ảnh khác nhau, nhưng khác ở chỗ nó dựa trên đồng thời khoảng cách giữa 2 cặp ảnh. Công thức cho hàm mất mát của mạng bộ ba được cho như trong công thức sau [22]:

$$L(x^a, x^p, x^n; \alpha) =$$

$$\frac{1}{N}\sum_i^N \max\{D(x^a, x^p) - D(x^a, x^p) + \alpha, 0\} \quad (2)$$

Trong đó, $D(x^a, x^p)$ là khoảng cách giữa cặp ảnh giống và $D(x^a, x^n)$ là khoảng cách giữa cặp ảnh khác, α là ngưỡng lề của 2 khoảng cách và N là số lượng bộ ba mẫu.



Hình 3. Mạng bộ ba [22]

Từ các khoảng cách đặc trưng giữa các ảnh đầu vào, mục tiêu của mạng bộ ba là xây dựng khoảng cách giữa ảnh chuẩn và ảnh giống nhỏ hơn khoảng cách giữa ảnh chuẩn và ảnh khác. Với mọi bộ 3 ảnh x^a, x^p, x^n như nói ở trên, mối quan hệ giữa khoảng cách đặc trưng giữa các cặp ảnh trong mọi bộ ảnh cần thoả mãn công thức sau [22]:

$$D(x^a, x^p) < D(x^a, x^n) + \alpha \quad (3)$$

C. Đánh giá độ tương đồng giữa các ảnh

Sau khi mạng bộ ba được huấn luyện, một trong ba mạng CNN nhánh của nó có thể được chọn làm mạng dùng để trích xuất đặc trưng ảnh. Các đặc trưng ảnh có thể được tạo ra bằng cách cho ảnh qua mạng CNN này sau khi đã loại bỏ đi lớp cuối cùng. 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.

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

$$l_1 = \sum_{i=1}^{m} |x_i - y_i| \quad (4)$$

Công thức cho khoảng cách Euclid:

$$l_2 = \sqrt{\sum_{i=1}^{m}\left(x_i - y_i\right)^2} \quad (5)$$

Công thức cho khoảng cách Cosine:

$$similarity = \frac{\sum_{i=1}^{m} x_i y_i}{\sum_{i=1}^{m}(x_i)^2 \sum_{i=1}^{m}(y_i)^2} \quad (6)$$

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, bao gồm các loại sản phẩm gia dụng như lò vi sóng, nồi cơm điện, các sản phẩm thời gian như quần, áo, váy .v.v. Mỗi nhãn có số lượng từ 500 đến 2.000 sản phẩm.

Để tạo ra các bộ ảnh làm đầu vào cho mạng bộ ba, các ảnh từ tập ban đầu được chọn ngẫu nhiên để tạo bộ ba ảnh. Mỗi bộ ba ảnh bao gồm 1 ảnh chuẩn x^a, 1 ảnh giống x^p, và 1 ảnh khác x^n. Trong đó, ảnh x^a và x^p được chọn từ cùng 1 loại, ảnh x^n được chọn từ 1 loại khác. Tổng số 15.000 bộ ảnh được tạo ra từ tập ảnh ban đầu.

Các bộ 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 để trích xuất đặc trưng:

- 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.

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.

Đầu tiên, mạng ResNet trên sẽ được sử dụng để phân loại các ảnh sản phẩm. Để thực hiện phân loại sản phẩm, mạng này cần được bổ sung một lớp FC ở cuối để tiến hành phân loại. Việc phân loại sản phẩm được thực hiện tương tự như trong [ ]. 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%).

Tiếp theo, mạng ResNet này sẽ được sử dụng làm các mạng CNN thành phần để xây dựng mạng Siamese và mạng bộ ba phục vụ cho việc trích chọn đặc trưng ảnh để tiến hành so sánh độ tương đồng. Các bộ ảnh mẫu được tạo ra ở bước trước sẽ được đưa vào để huấn luyện các mạng này. Sau khi mạng được huấn luyện xong, một trong các mạng CNN nhánh sẽ được dùng làm mạng trích xuất đặc trưng của một ảnh mới. Lưu ý rằng bất kỳ mạng CNN nhánh nào được chọn đều cho kết quả giống nhau do các mạng này đã được chia sẻ cấu trúc chung và các trọng số/tham số. Khác với mạng ResNet dùng để phân loại ở trên, mạng này không cần bổ sung thêm lớp FC để phân loại, vì mục đích của mạng này không phải để phân loại mà để trích chọn đặc trưng ảnh. Đặc trưng của ảnh mới thu được khi cho qua mạng này sẽ được so sánh với các đặc trưng ảnh trong cùng phân loại để chọn ra các ảnh có độ tương đồng cao nhất với ảnh mới (thường là các ảnh do người dùng cung cấp trong một hệ thống tìm kiếm theo ảnh).

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 [12], 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.

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

Bảng 1 cho thấy kết quả độ chính xác tìm kiếm theo đánh giá chủ quan trên toàn bộ 100 mẫu với 3 mô hình thử nghiệm là mạng CNN thông thường, mạng Siamese, và mạng bộ ba.

STT	Loại mạng	Độ chính xác (%)
1	Mạng CNN thường	75%
2	Mạng Siamese	75.8%
3	Mạng bộ ba	76.3%

Các kết quả nhận được cho thấy việc sử dụng các mạng nhiều nhánh như Siamese hoặc mạng bộ ba đem lại kết quả tốt hơn so với mạng CNN thông thường. Mặc dù độ chênh lệch chưa lớn, nhưng kết quả chứng tỏ việc sử dụng các mạng có tính chất phân biệt ảnh giống và ảnh khác nhau đã đem lại những kết quả tốt hơn.

Về thời gian chạy, việc huấn luyện mô hình phân loại CNN thông thường mất khoảng 80 phút, còn thời gian để huấn luyện các mạng nhiều nhánh như Siamese hay mạng bộ ba mất thời gian lâu hơn, khoảng 150 phút (với số ảnh mẫu là 11.539 ảnh). Trong khi đó, 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).

V. KẾT LUẬN

Trong nghiên cứu này, chúng tôi đã trình bày phương pháp sử dụng mạng bộ ba để huấn luyện và trích chọn đặc trưng cho các ảnh, nhằm đánh giá độ tương đồng giữa chúng. Các kết quả thực nghiệm cho thấy các mạng có nhiều nhánh và được thiết kế để tạo ra các đặc trưng ảnh sao cho các ảnh mẫu giống nhau thì có khoảng cách gần và các ảnh mẫu khác nhau thì có khoảng cách xa như mạng Siamese và mạng bộ ba đã cho kết quả tích cực hơn mạng CNN thông thường.

Mặc dù các mạng học sâu đã cố gắng trích xuất các đặc trưng ảnh và tính toán độ tương đồng theo cách mô phỏng lại cách đánh giá của con người, nhưng vẫn còn nhiều khoảng cách về ngữ nghĩa trong cách đánh giá độ tương đồng của ảnh giữa máy và người. Các hướng phát triển tiếp theo của nghiên cứu có thể là kết hợp khai thác nhiều hơn các đặc trưng ngữ nghĩa của ảnh để có sự đánh giá tương đồng tốt hơn về khía cạnh này.

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] S. Chopra, R. Hadsell, Y. Lecun, Learning a Similarity Metric Discriminatively, with Application to Face Verification. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, 20–06 June 2005; Volume 1, pp. 539–546.

[6] Navneet Dalal and Bill Triggs, Histograms of Oriented Gradients for Human Detection,.In Proc. CVPR. p.886–893 (2005)

[7] Dương Trần Đức, Tìm kiếm sản phẩm theo ảnh bằng học sâu, Tạp chí Khoa học Công nghệ Thông tin và Truyền thông, Học viện Công nghệ Bưu chính Viễn thông, Tập 1, Số 2 (2020).

[8] E. Hoffer, N. Ailon, Deep Metric Learning Using Triplet Network. In Proceedings of the International Workshop on Similarity-based Pattern Recognition, Copenhagen, Denmark, 12–14 October 2015.

[9] 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).

[10] 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)

[11] 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)

[12] N. Khosla, and V. Venkataraman, Building Image-Based Shoe Search Using Convolutional Neural Networks, CS231N Course Project Reports, (2015)

[13] 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)

[14] Alex Krizhevsky, Ilya Sutskever, and Geo rey E. Hinton, ImageNet Classification with Deep Convolutional Neural Networks, In Proc. NIPS, p.1106–1114 (2012)

[15] 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)

[16] David G. Lowe, Object Recognition from Local Scale-Invariant Features, In Proc. ICCV, p.1150–1157 (1999)

[17] I. Melekhov, J. Kannala, E. Rahtu, Siamese network features for image matching. In Proceedings of the International Conference on Pattern Recognition, Cancun, Mexico, 4–8 December 2016.

[18] K. Simonyan and A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, In Proc. ICLR (2015)

[19] 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)

[20] Wang, J.; Song, Y.; Leung, T.; Rosenberg, C. Learning Fine-Grained Image Similarity with Deep Ranking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014, pp. 1386–1393.

[21] Yuan X, Liu Q, Long J, Hu L, Wang Y, Deep Image Similarity Measurement Based on the Improved Triplet Network with Spatial Pyramid Pooling, Information (2019)

[22] X. Yuan, Q. Liu, J. Long, L. Hu, Y. Wang, Deep Image Similarity Measurement Based on the Improved Triplet Network with Spatial Pyramid Pooling, Information (2019)

IMAGE SIMILARITY MEASUREMENT BASED ON DEEP LEARNING USING TRIPLE NETWORK

Abstracrt: Image similarity measurement is one of the most important in computer vision, specially in image search field. Among the proposed methodologies, the method using triple network has some remarkable advantages. A triple nework often contains 3 CNN (Convolutional Neural Network) branches, which have been shared weights and parameters. It presents the high level feature of image in which the similar images are close to each other and the different images are far from each other. This paper reports the method of using triple network in image similarity measurement and apply in the product image search problem. The experiments showed the promissing results.

Keywords: deep learning, convolutional neural network, triple network, image search.

Dương Trần Đức Tốt nghiệp Đại học KHTN, Đại học Quốc gia Hà Nội ngành Công nghệ thông tin năm 1999, Thạc sỹ chuyên ngành Hệ thống thông tin tại Đại học Tổng hợp Leeds, Vương Quốc Anh năm 2004, và Tiến sỹ 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.