Hệ Thống Khuyến Nghị Sản Phẩm Dựa Trên Hình Ảnh Với Mạng Nơ-Ron Tích Chập

1st Lê Thanh Phong
Khoa Công nghệ thông tin
Bộ môn Khoa học dữ liệu
thanhphong27092001@gmail.com
MSSV: 19475611

***Tóm tắt nội dung*—Hầu hết các công cụ tìm kiếm mua sắm trực tuyến vẫn chủ yếu phụ thuộc vào nền tảng kiến thức và sử dụng đối sánh từ khóa làm chiến lược tìm kiếm của họ để tìm sản phẩm có khả năng mà người tiêu dùng muốn mua nhất. Điều này không hiệu quả theo cách mô tả sản phẩm có thể khác nhau rất nhiều từ phía người bán sang phía người mua.**

Trong bài báo này, chúng tôi trình bày một công cụ tìm kiếm thông minh để mua sắm trực tuyến.Về cơ bản, chúng tôi sử dụng hình ảnh làm đầu vào và cố gắng để hiểu thông tin về sản phẩm từ những hình ảnh này. Đầu tiên, chúng tôi sử dụng mạng nơ-ron để phân loại hình ảnh đầu vào là một trong các danh mục sản phẩm. Sau đó, sử dụng một mạng nơ-ron khác để lập mô hình điểm giống nhau giữa các cặp hình ảnh, mà chúng được sử dụng để chọn sản phẩm gần nhất trong bộ dữ liệu sản phẩm của chúng tô. Chúng tôi sử dụng Jaccard similarity để tính toán điểm giống nhau cho quá trình đào tạo dữ liệu. Chúng tôi thu thập dữ liệu thông tin sản phẩm (bao gồm hình ảnh, nhãn lớp, giá, tên sản phẩm, ...) từ Amazon để tìm hiểu các mô hình này. Cụ thể, tập dữ liệu của chúng tôi chứa thông tin về 9,2 triệu sản phẩm có hình ảnh và có tổng cộng có hơn 20 danh mục. Phương pháp của chúng tôi đạt được độ chính xác phân loại là 0.48. Cuối cùng, chúng tôi có thể giới thiệu các sản phẩm có mức độ tương tự cao hơn 0,48 và cung cấp hỗ trợ mua sắm trực tuyến nhanh chóng và chính xác.

I. Giới thiệu

Cuộc khủng hoảng dịch COVID-19 mang đến những thách thức nhưng đồng thời cũng là cơ hội để các nhà bán lẻ thay đổi nhằm theo kịp sự phát triển của thương mại điện tử. Mua sắm trực tuyến đang thay thế cho việc mua sắm tại cửa hàng khi hàng tỷ người phải ở nhà vì dịch viêm đường hô hấp cấp COVID-19. Qua đó, thương mại điện tử hay mua hàng trực tuyến trong bối cảnh đó đang bùng nổ ở nhiều khu vực trên toàn cầu.

Tuy nhiên, sự phát triển bùng nổ về việc mua hàng trực tuyến đã tạo ra thách thức về tình trạng quá tải thông tin đối với người mua, điều này hạn chế khả năng tiếp cận kịp thời với các mặt hàng quan tâm trên Internet. Điều này đã làm tăng nhu cầu về hệ thống khuyến nghị. Mặc dù hầu hết mọi công ty thương mại điện tử ngày nay đều có hệ thống đề xuất riên cho mình để có thể được sử dụng để cung cấp tất cả các loại đề xuất, chúng chủ yếu dựa trên văn bản sử dụng hệ thống đối sánh từ chính. Điều này yêu cầu người mua sắm trực tuyến cung cấp chi tiết mô tả về sản phẩm, có thể khác nhau rất nhiều từ phía người bán sang phía người mua.

Trong những năm gần đây, với sự phát triển nhanh chóng của mạng nơ-ro. Giờ đây, chúng ta có thể thay đổi cách tìm kiếm thủ công bằng cách mô tả sản phẩm hay tìm kiếm bằng tên sản phẩm. Thì bây giờ chúng ta có thể chụp nhanh một hình ảnh sản phẩm để tìm kiếm chúng một cách dễ dàng hơn. Phương pháp tìm kiếm bằng hình ảnh thì đã được áp dụng rất nhiều nhưng riêng với lĩnh vực thương mại điện tử nói chung và mua sắm trực tuyến nói riêng thì vẫn chưa được ứng dụng rộng rãi. Dựa trên ý tưởng này, ở đây chúng tôi xây dựng một hệ thống khuyến nghị thông minh, từ việc lấy hình ảnh của các đối tượng làm dầu thay vì mô tả văn bản như một cách truyền thống.

Đầu vào cho thuật toán của chúng tôi là hình ảnh của bất kỳ đối tượng nào khách hàng muốn mua. Sau đó, chúng tôi sử dụng Convolutional Neural Network (CNN) để phân loại danh mục đối tượng này có thể thuộc về và sử dụng vectơ đầu vào của lớp được kết nối đầy đủ cuối cùng dưới dạng vectơ đặc trưng để cấp dữ liệu trong quá trình tính toán độ tương tự mô hình CNN để tìm các sản phẩm gần nhất trong bộ dữ liệu của chúng tôi. Cụ thể hơn là ,hai chức năng mà chúng tôi muốn đạt được trong hệ thống khuyến nghị là:

1. Phân loại: Hình ảnh của sản phẩm được khách hàng cung cấp và tìm xem sản phẩm đó thuộc danh mục nào trong tổng số 20 danh mục mà chúng tôi hiện có trong bộ dữ liệu của mình. Ví dụ, một hình ảnh của chiếc điện thoại SamSung sẽ được phân loại là "Điện thoại di động & Phụ kiện".

2. Khuyến nghị: Cung cấp các đặc trưng của hình ảnh và danh mục mà sản phẩm này thuộc về. Tính toán điểm tương tự và tìm các sản phẩm tương tự nhất trong bộ dữ liệu. Tốt nhất là khác hàng tìm kiếm điện thoại SamSung thì nên đề xuất điện thoại SamSung.

II. Nghiên cứu liên quan

Bài báo [6] đã trình bày một ý tưởng về việc kết hợp gợi ý hình ảnh và đề xuất hình ảnh từ nhiều thập kỷ trước. Trong này dự án, chúng tôi sử dụng tập dữ liệu sản phẩm của Amazon, được sử dụng để xây dựng hệ thống tư vấn điển hình sử dụng phương pháp lọc cộng tác trong [4] và [8]. Trong lĩnh vực đề xuất hình ảnh, [5] có xu hướng đề xuất hình ảnh bằng cách sử dụng Tuned perceptual truy xuất (PR), bổ sung đồng thuận láng giềng gần nhất (CNNC), mô hình hỗn hợp Gaussian (GMM), chuỗi Markov (MCL) và truy xuất bất khả tri về kết

cấu (TAR), v.v. CNNC, GMM, TAR và PR rất dễ đào tạo, nhưng CNNC và GMM rất khó để kiểm tra trong khi PR, GMM và TAR rất khó để phổ biến hóa. Ngoài ra, vì dữ liệu bao gồm các hình ảnh, công việc mạng thần kinh nên là một phương pháp đáng thử.

Bài báo [7] đã trình bày mô hình AlexNet có thể phân loại hình ảnh thành 1000 loại khác nhau. Ngoài ra, Bài báo [9] đã trình bày mạng nơ-ron VGG phân loại hình ảnh trong ImageNet Challenge 2014. Trong phần đầu tiên của dự án, chúng tôi sử dụng cả hai mô hình để phân loại các danh mục của sản phẩm. Đã bao giờ, cả hai bài báo đều không đưa ra phương pháp sửa lỗi gợi ý hình ảnh.

Mặc dù có những bài báo nghiên cứu sự giống nhau về hình ảnh chẳng hạn như [12] và [11], hầu hết chúng đều dựa trên danh mục tương tự, tức là các sản phẩm được coi là tương tự nếu chúng ở cùng thể loại. Tuy nhiên, các sản phẩm đến từcùng một danh mục vẫn có thể khác nhau rất nhiều. Do đó, một chiến lược đáng tin cậy là trước tiên phải phân loại hình ảnh mục tiêu vào một danh mục nhất định và sau đó giới thiệu hình ảnh từ danh mục đã phân loại này.

Trong bài báo [13], họ đã xem xét việc sử dụng mạng nơ-ron để tính toán các điểm tương đồng trong danh mục. Tuy nhiên, mỗi người chỉ xem xét ConvNet, DeepRanking, v.v. Vì chúng tôi có tập dữ liệu lớn hơn, mạng nơ ron phức hợp sâu hơn chẳng hạn như AlexNet và VGG sẽ tốt hơn so với ConvNets ngây thơ. Ý tưởng cũng có thể được tìm thấy trong [3] và [10].

Bài báo [1] cũng đang tập trung vào việc học tính tương tự bằng cách sử dụng CNN. Tuy nhiên, nó xem xét nhiều hơn về trường hợp nhiều sản phẩm chứa trong một hình ảnh duy nhất. Trong dự án của chúng tôi, chúng tôi giả định rằng người dùng đang tìm kiếm một sản phẩm và hình ảnh đó sẽ chỉ chứa một sản phẩm.

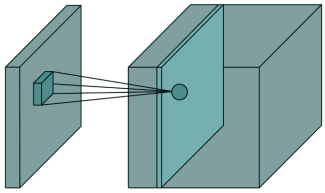
Trước khi chúng tôi đề xuất, chúng tôi cần trả lời phép đo độ giống nhau. Câu trả lời tốt nhất là cosine similarity hoặc L2 norm similarity. Cách khác để đo mức độ tương tự là bằng cách giới thiệu sự kết hợp thông tin ngữ nghĩa. Bài báo [2] chỉ ra rằng sự tương đồng về hình ảnh và sự giống nhau giữa các hình ảnh có mối tương quan với nhau. Vì vậy, chúng tôi giới thiệu một mô hình để tính toán sự tương đồng giữa các hình ảnh dựa trên thông tin se mantic. Bài báo [15] và [14] có cùng ý tưởng như chúng tôi làm ở đây.

III. Phương pháp tiếp cận

Có hai vấn đề lớn mà chúng tôi muốn giải quyết trong dự án của mình. Đầu tiên, xác định danh mục mà một hình ảnh nhất định thuộc về. Thứ hai, tìm và giới thiệu các sản phẩm tương tự nhất theo hình ảnh đã cho trước. Dự án của chúng tôi chủ yếu dựa trên mạng nơ-ron phức hợp, chúng tôi sẽ lần đầu tiên giới thiệu mạng nơ-ron phức hợp được sử dụng phổ biến các lớp.

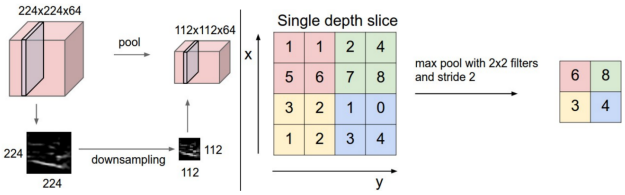
A. *CNN Layers*

Bước quan trọng nhất của CNN là lớp tích chập (Conv). Như chúng ta có thể thấy từ Hình 1, lớp tích chập sẽ dịch hình chữ nhật nhỏ của lớp đầu vào thành một số lớp đầu ra bằng cách sử dụng phép nhân ma trận. Tích chập là lớp đầu tiên để trích xuất các tính năng từ hình ảnh đầu vào. Tích chập duy trì mối quan hệ giữa các pixel bằng cách tìm hiểu các tính năng hình ảnh bằng cách sử dụng các ô vương nhỏ của dữ liệu đầu vào. Nó là 1 phép toán có 2 đầu vào như ma trận hình ảnh và 1 bộ lọc hoặc hạt nhân.



Hình 1. Lớp tích chập

Pooling layers tương tự như các lớp chập, ngoại trừ rằng nó sẽ sử dụng phương thức không tham số để biến đổi nhỏ hình chữ nhật thành một số. Max pooling thường được sử dụng trong CNN, sẽ xuất ra số lượng tối đa trong hình chữ nhật của lớp đầu vào.



Hình 2. Pooling layers

B. *Phân loại*

Trong bước này, chúng tôi muốn phân loại một hình ảnh đầu vào là một trong 20 loại danh mục trong bộ dữ liệu của mình. Chúng tôi xây dựng mô hình AlexNet và mô hình VGG-16 cho nhiệm vụ phân loại và so sánh chúng với mô hình SVM làm mô hình cơ sở.

- **Support Vector Machine:** Mô hình phân loại tuyến tính được sử dụng làm mô hình cơ sở ở đây. Mô hình này về cơ bản là một lớp được kết nối đầy đủ. Chúng tôi sử dụng độ lỗi Multi-class Support Vector Machine (SVM) cộng với L2 Norm để làm hàm mất mát. Đối với một hình ảnh i, chúng tôi sử dụng các pixel RGB làm đặc trưng đầu vào $x_i \in \mathbb{R}^d$ với $d = 244 \times 244$. Chúng tôi tính điểm lớp cho $n = 20$ danh mục thông qua một phép biến đổi tuyến tính

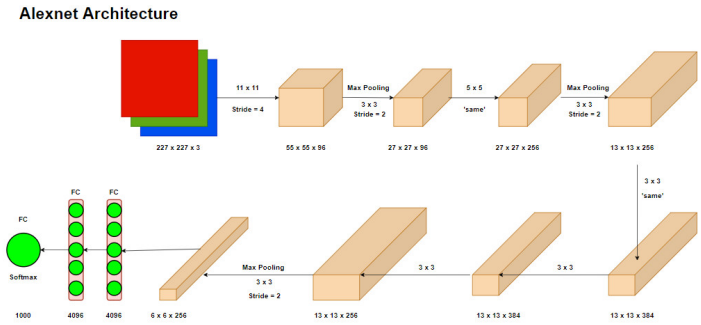
$$s = Wx_i + b \tag{1}$$

Trong đó $W \in \mathbb{R}^{n \times d}$ là ma trận trọng số và $b \in \mathbb{R}^n$ là trọng số. Tổn thất của SVM được tính bởi công thức

$$L_{SVM}(W, b; x_i) = \sum_{j \neq y_i} \max(0, s_j - s_{y_i} + 1) \tag{2}$$

Trong đó y_i là nhãn đúng của lớp thực sự.

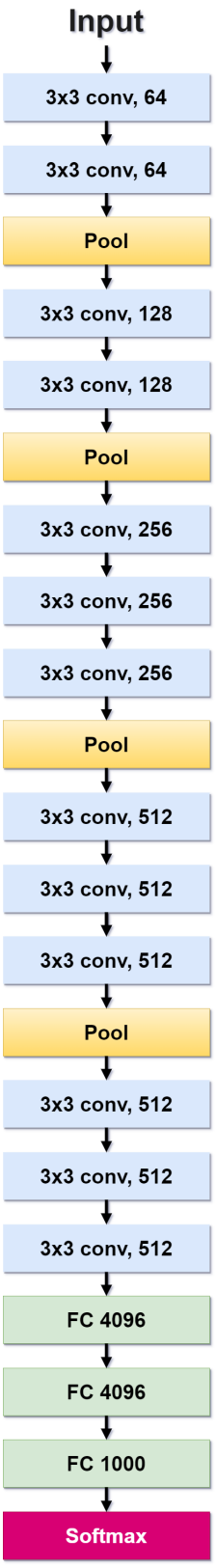
- **AlexNet:** Một mô hình phân loại mạng nơ-ron tích chập sâu được đề xuất bởi [7]. Như chúng ta thấy, (Hình 3) Mô hình AlexNet đầu tiên chứa 2 lớp tích chập với tính năng tổng hợp tối đa và chuẩn hóa hàng loạt; sau đó là 3 lớp phức hợp với tính năng tách biệt và một gộp tối đa trước 3 lớp fully connected.



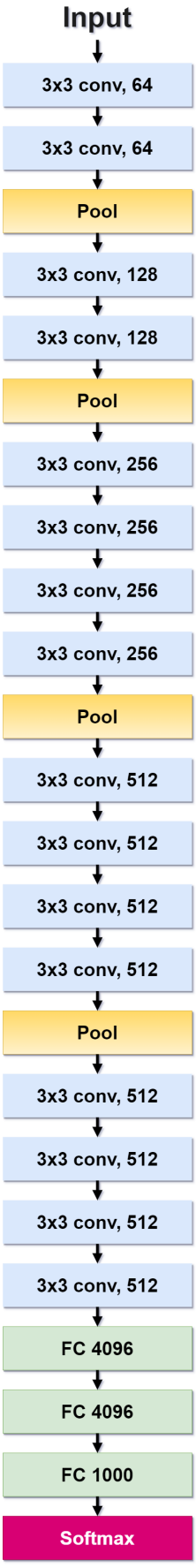
Hình 3. Mô hình AlexNet

Mô hình ban đầu được đào tạo để phân loại hình ảnh trong cuộc thi ImageNet LSVRC-2010, có 1000 danh mục. Vì vấn đề của chúng tôi chỉ chứa 20 danh mục, chúng tôi thay đổi lớp được kết nối đầy đủ cuối cùng thành 4096 × 20. Để tiết kiệm thời gian, chúng tôi sử dụng lại các trọng số được đào tạo trước đó trong số 5 tế bào thần kinh đầu tiên và đào tạo ba lớp fully connected cuối cùng.

- **VGG-16:** Một mô hình phân loại mạng nơron tích chập sâu được đề xuất bởi [9]. Như hình bên dưới (Hình 4), VGG 16 chứa 13 lớp tích chập với max pooling mỗi 2 hoặc 3 lớp tích chập; sau đó là 3 lớp fully connected và softmax là lớp cuối cùng. Mô hình ban đầu được huấn luyện để phân loại hình ảnh trong cuộc thi ImageNet ILSVRC-2014, có 1000 danh mục. Chúng tôi thay đổi lớp được kết nối đầy đủ cuối cùng thành 4096 × 20. Chúng tôi cũng sử dụng các trọng số được huấn luyện trước khi khởi tạo các tham số và đào tạo ba lớp fully connected cuối cùng. Chúng tôi cũng thêm các lớp chuẩn hóa hàng loạt sau các hàm kích hoạt trong hai lớp fully connected đầu tiên.
- **VGG-19:** chứa 16 lớp tích chập với max pooling mỗi 2 hoặc 4 lớp tích chập (5 lớp MaxPool); sau đó là 3 lớp fully connected và softmax là lớp cuối cùng. Như hình bên dưới (Hình 5). Mô hình ban đầu được huấn luyện để phân loại hình ảnh trong cuộc thi ImageNet ILSVRC-2014, có 1000 danh mục. Chúng tôi thay đổi lớp được kết nối đầy đủ cuối cùng thành 4096 × 20. Chúng tôi cũng sử dụng các trọng số được huấn luyện trước khi khởi tạo các tham số và đào tạo ba lớp fully connected cuối cùng. Hình ảnh đầu vào có kích thước cố định là 224 x 224 pixel với ba kênh (R, G và B), có nghĩa là ma trận có hình dạng (224,224,3).
  Ngoài ra chúng tôi còn thêm cũng thêm các lớp chuẩn hóa hàng loạt (batch normalization) sau các hàm kích hoạt để tăng tốc độ huấn luyện cho mô hình và trách trường hợp over-fitting. Và công thức được định nghĩa như sau



Hình 4. Mô hình VGG 16



Hình 5. Mô hình VGG 19

$$\mu_i = \frac{1}{m}\sum_{j=1}^{m} x_{ij}$$
$$\sigma_i^2 = \frac{1}{m}\sum_{j=1}^{m}(x_{ij} - \mu_i)^2 \quad (3)$$
$$\hat{x}_{ij} = \frac{x_{ij} - \mu_i}{\sqrt{\sigma_i^2 + \epsilon}}$$

- **ResNet-50:** ResNet (Residual Network) được giới thiệu đến công chúng vào năm 2015 và thậm chí đã giành được vị trí thứ 1 trong cuộc thi ILSVRC 2015 với tỉ lệ lỗi top 5 chỉ 3.57%. ResNet cũng là kiến trúc sớm nhất áp dụng batch normalization. ResNet50 là một biến thể của mô hình ResNet có 48 lớp tích chập cùng với 1 lớp MaxPool và 1 lớp Average Pool. ResNet sử dụng các kết nối tắt (kết nối trực tiếp đầu vào của lớp (n) với (n+x) được hiển thị dạng mũi tên cong thông qua Hình 6. Qua mô hình nó chứng minh được có thể cải thiện hiệu suất trong quá trình training model khi mô hình có hơn 20 lớp.

C. *Khuyến nghị*

Đối với bước khuyến nghị, chúng tôi sử dụng lớp fully connected cuối cùng trong mô hình phân loại của chúng tôi dưới dạng vectơ đặc trưng hình ảnh. Đối với bất kỳ hình ảnh nào trong tập dữ liệu, sẽ có một vector đặc trưng tương ứng. Và vectơ đặc điểm này sẽ là đầu vào cho mô hình đề xuất của chúng tôi. Luồng công việc của bước này được hiển thị trong các gạch đầu dòng sau.

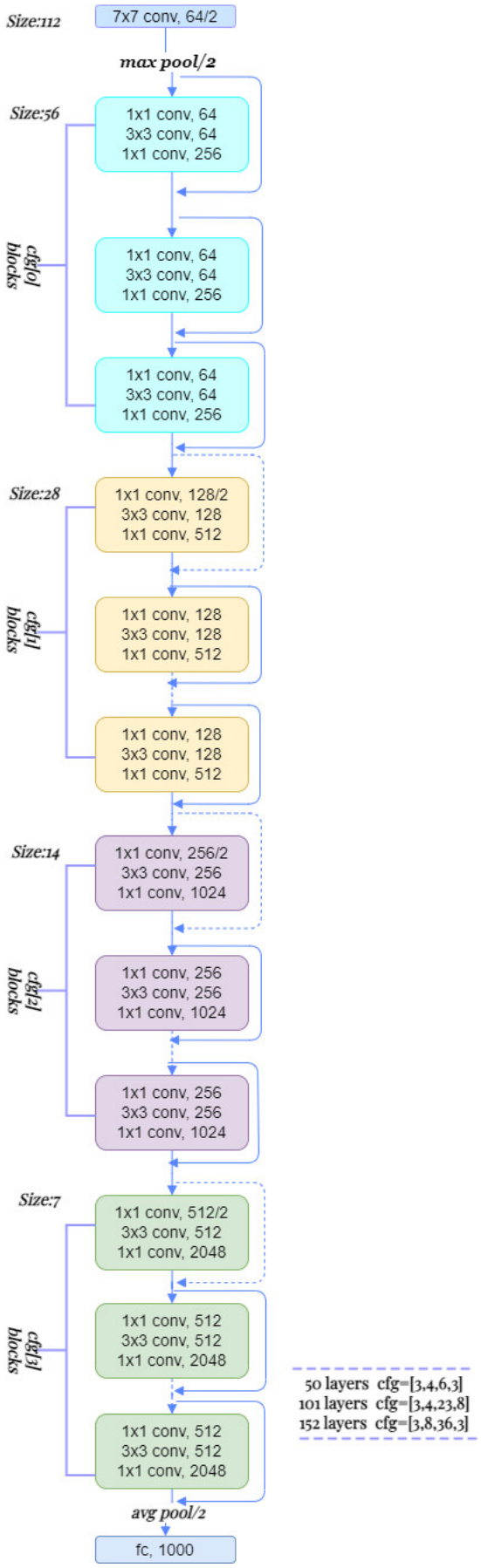
- Trích xuất đặc trưng: Mô hình phân loại được sử dụng để xác định các hình ảnh sẽ thuộc danh mục nào. Sau đó, chúng tôi trích xuất đầu vào từ lớp fully connected cuối cùng của mô hình phân loại.
- Đầu vào của mô hình: Vector đặc trưng của hình ảnh được trích xuất ở trên.
- Similarity calculation (tính độ tương tự):sử dụng các thước đo khác nhau để tính toán điểm tương đồng giữa vectơ đặc trưng của hình ảnh và vectơ đặc trưng của tất cả các ảnh trong toàn bộ danh mục để đo mức độ giống nhau giữa các cặp hình ảnh. Chúng tôi đã thử L2 distance, cosine distance và các mô hình mạng nơ-ron để tính độ tương đồng. Vì 2 hình ảnh i và j khác nhau thì điểm L2 distance sẽ được định nghĩa như sau

$$S_{L_2} = \left\| v_i - v_j \right\|_2 \quad (4)$$

Trong đó $v_i, v_j \in \mathbb{R}^l$ là hai tính năng tương ứng Vectơ, và $l = 4096$ là độ dài của vectơ đặc trưng.
Điểm S_{L_2} càng nhỏ thì hai hình ảnh càng giống nhau
Điểm cosine distance được định nghĩa như sau

$$S_{cosine} = \frac{v_i^T v_j}{\left\| v_i \right\| \left\| v_j \right\|} \quad (5)$$

Điểm S_{cosine} càng lớn thì hai hình ảnh càng giống nhau



Hình 6. Mô hình kiến trúc ResNet-50

Cách tiếp cận theo hướng dữ liệu để tính toán điểm tương tự là đào tạo mạng nơ-ron 3 lớp sau:

$$h_1 = f(v.W_1 + b_1)$$
$$h_2 = f(v.W_2 + b_2) \quad (6)$$
$$S_{model} = sigmoid(h_2.W_3 + b_3)$$

Trong đó $v = [v_1, v_2] \in \mathbb{R}^{l\times 2}$ thu được bằng cách nối hai vectơ đặc trưng. $f(x) = max(0.01x, x)$ là hàm leaky ReLU. Lớp đầu tiên có thể được coi là một lớp tích chập 1-d với Leaky ReLU làm hàm kích hoạt và $W_1 \in \mathbb{R}^2$ và $b_1 \in \mathbb{R}$ làm tham số. Lớp thứ hai là một lớp fully connected với Leaky ReLU làm hàm kích hoạt và $W_2 \in \mathbb{R}^l$ và $b_2 \in \mathbb{R}$ làm tham số. Lớp đầu ra là một phép chuyển đổi tuyến tính với hàm sigmoid là hàm kích hoạt. Không có cách nào dễ dàng để xác định điểm tương đồng hoàn toàn dựa trên các pixel hình ảnh. May mắn thay, các hình ảnh đầu vào có tiêu đề tương ứng mô tả sản phẩm. Để mô tả mức độ tương tự của hai hình ảnh, chúng tôi sử dụng mức độ tương tự Jaccard của hai tiêu đề của hình ảnh như là sự tương đồng của hai ảnh. Mức độ tương tự Jaccard của hai tập A và B được định nghĩa như sau

$$S_{Jaccard} = \frac{|A \cap B|}{|A \cup B|} \quad (7)$$

Nó là một số từ 0 đến 1. Đây cũng là lý do mà chúng ta sử dụng hàm sigmoid làm hàm kích hoạt cho layer cuối cùng.Chúng tôi đào tạo mô hình này bằng cách giảm thiểu độ lỗi L_2 $\left\|S_{model} - S_{Jaccard}\right\|_2^2$.

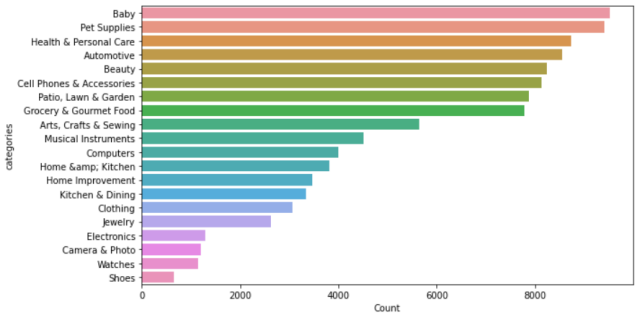
- Đầu ra: hình ảnh k (sản phẩm) giống với hình ảnh mục tiêu nhất.

IV. Dữ liệu và đặc trưng

Để xây dựng hệ thống khuyến nghị, chúng tôi sử dụng dữ liệu hình ảnh sản phẩm của Amazon, kéo dài từ tháng 5 năm 1996 đến tháng 7 năm 2014, trong đó bao gồm 9,4 triệu sản phẩm. Không bao gồm những hình ảnh bị thiếu, chúng tôi đã thu thập một tập dữ liệu gồm 4 triệu sản phẩm, với tổng số 20 danh mục. Hình 7 cho thấy biểu đồ phân phối của tất cả các nhãn của tập dữ liệu nhưng chúng tôi chỉ lấy một phần dữ liệu để trực quan. Chi tiết thông tin của mỗi hình ảnh bao gồm:

- **asin** - ID của sản phẩm, ví dụ: 0000027091
- **title** - tên của sản phẩm
- **price** - giá bằng đô la Mỹ (tại thời điểm thu thập dữ liệu)
- **imUrl** - link của hình ảnh sản phẩm
- **related** - sản phẩm liên quan (also bought, also viewed, bought together, buy after viewing)
- **salesRank** - thông tin xếp hạng bán hàng
- **brand** - tên thương hiệu
- **categories** - danh sách các danh mục sản phẩm thuộc về

Vì bị giới hạn về tài nguyên nên chúng tôi chỉ có thể lấy ngẫu nhiên 250 ảnh trong mỗi danh mục và thu thập 5000 ảnh cho quá trình phân loại. Sau đó, chúng tôi chia tập dữ liệu thành 7:2:1 cho training, validation and testing tương ứng. Mỗi hình ảnh trong tập dữ liệu có 300 × 300 pixel.Chúng tôi sử dụng



Hình 7. Biểu đồ phân phối các danh mục trong tập dữ liệu

các pixel thô của hình ảnh làm đầu vào cho mô hình mạng nơ-ron phân loại của chúng tôi. Ví dụ về dữ liệu được hiển thị trong Hình 8. Để thuận tiện cho việc điều chỉnh các siêu tham số, chúng tôi thay đổi kích thước hình ảnh thành 224 × 224 × 3 bằng cách sử dụng "scipy.misc" cho VGG, ResNet và thay đổi kích thước thành 227 × 227 cho AlexNet.

Hình 8. Ví dụ về dữ liệu. Đây là ba sản phẩm từ danh mục "Cell Phones & Accessories"

V. Thử nghiệm

A. *Xử lí dữ liệu*

Các hình ảnh thô cần được xử lý trước trước khi được sử dụng làm đầu vào của các mô hình phân loại. Đầu tiên, một hình ảnh gốc được thay đổi kích thước thành kích thước đầu vào tiêu chuẩn của mô hình VGG, ResNet (224 × 224) hoặc mô hình AlexNet (227 × 227) (Hình 9)

Đối với mô hình khuyến nghị, sự tương tự thật sự được xác định bằng cách sử dụng sự tương tự Jaccard (6) của các bộ mã thông báo trong tiêu đề của hai hình ảnh. (thông tin tiêu đề được đính kèm với mỗi hình ảnh trong tập dữ liệu). Tuy nhiên, chúng tôi quan tâm nhiều hơn đến cặp hình ảnh tương tự. Nếu chúng ta sử dụng tất cả dữ liệu, phần lớn các cặp sẽ có điểm giống nhau gần bằng 0. Do đó, thay vì sử dụng tất cả các cặp, chúng tôi chỉ xem xét các cặp trong cùng một danh mục. Hơn nữa, chúng tôi đặc biệt muốn tìm các cặp hình ảnh có độ tương đồng Jaccard tương đối lớn.Chúng tôi tìm thấy những cặp có điểm giống nhau trên 0,5 và có khoảng 800 cặp như vậy. Chúng tôi cũng lấy mẫu 1000 cặp có điểm giống nhau bằng 0. Và sử dụng các cặp này làm ví dụ cho quá trình huấn luyện.

Hình 9. Xử lý hình ảnh đầu vào. Bên trái là hình ảnh gốc ban đầu. Bên phải là hình ảnh đã thay đổi kích thước (224 × 224 pixel).

B. *Đánh giá*

Chúng thôi chia tập dữ liệu thành 7:2:1 cho training, validation và test tương ứng. Để đánh giá các mô hình, chúng tôi chạy các mô hình của mình trên tập dữ liệu thử nghiệm và so sánh kết quả dầu ra với kết quả thực.

Đối với vấn đề phân loại, chúng tôi đánh giá mô hình bằng cách tính toán độ chính xác của phân loại:

$$Accuracy = \frac{\#\text{correctly classified images}}{\#\text{images in validationdataset}} \quad (8)$$

C. *Phân loại*

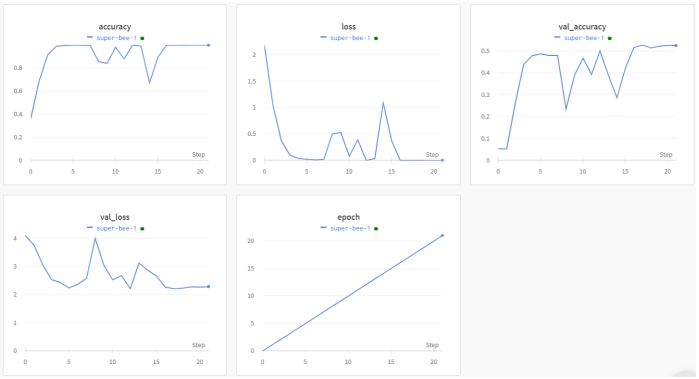
Đối với nhiệm vụ phân loại, chúng tôi đã đào tạo hai Mạng Nơ-Ron tích chập (VGG16, VGG19, ResNet50 và AlexNet) để phân loại các loại hình ảnh sản phẩm so với mô hình phân loại tuyến tính theo mô hình cơ sở của chúng tôi (mô hình SVM).

Bảng 1 cho thấy độ chính xác tốt nhất của chúng tôi về training data, validation data và test data cho ba mô hình này tương ứng. Đối với mô hình SVM, chúng tôi sử dụng learning rate 0.0005, regularization coefficients 0.001. Đối với AlexNet, chúng tôi sử dụng mini-batch size 128, regularization coefficient 0.01, learning rate 0.00005 và dropout 0.5. Đối với mô hình VGG16, chúng tôi sử dụng mini-batch size 64, regularization coefficient 0.01, learning rate 0.0007 và dropout 0.2. Đối với mô hình VGG19, chúng tôi sử dụng mini-batch size 64, regularization coefficient 0.01, learning rate 0.0001. Đối với mô hình ResNet50, chúng tôi sử dụng learning rate 0.001.Từ kết quả, chúng tôi có thể thấy rằng các mô hình của chúng tôi gặp phải vấn đề over-fitting. Độ chính xác đào tạo của cả mô hình AlexNet và mô hình VGG gần như gấp 1,5 lần độ chính xác đối với validation set và test set. Đó là lý do tại sao chúng ta cần regularization coefficients (0.01) cao hơn. Nhưng khi chúng tôi tăng regularization coefficients thì độ chính xác của quá trình huấn luyện không tăng lên nữa. Vì lý do tương tự, dropout mà chúng tôi chọn cho hai mô hình này cũng tương đối (tương ứng là 0,5 và 0,2). Tuy nhiên, ở một mức độ nào đó, chúng ta vẫn phải chịu đựng over-fitting.

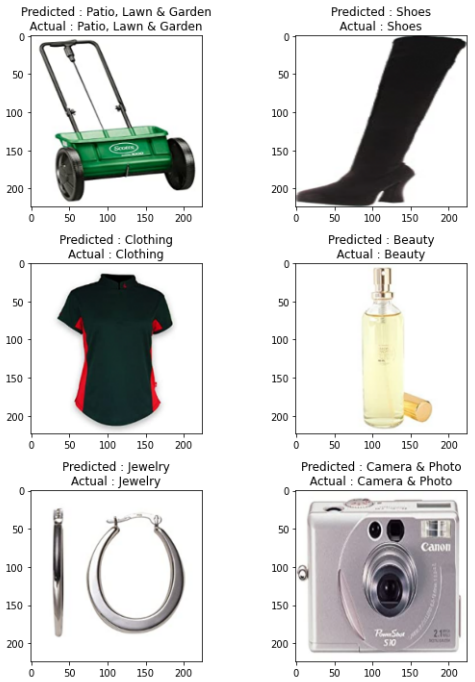
Hình 10, cho ta thấy được quá trình phân loại danh mục hình ảnh từ mô hình ResNet50

Bảng I
KẾT QUẢ ĐỘ CHÍNH XÁC CỦA CÁC MÔ HÌNH

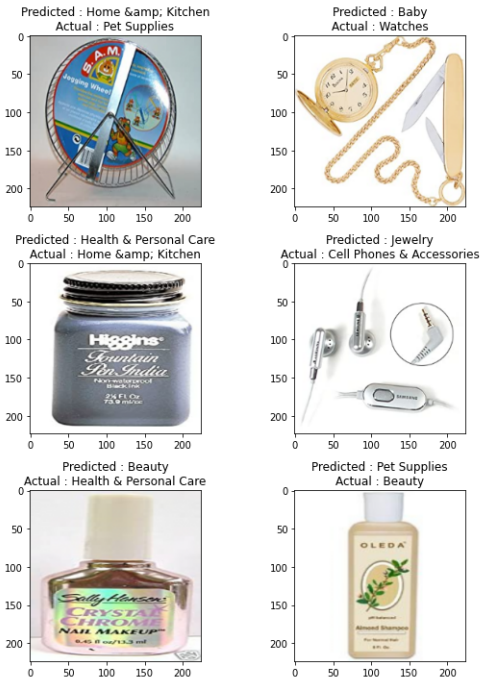
Model	Train acc.	Valid acc.	Test acc.
SVM (baseline)	0.29	0.25	0.22
AlexNet	0.7927	0.2948	0.3020
VGG16	0.9768	0.3861	0.3717
VGG19	0.9986	0.4472	0.4300
VGG19 have BatchNormalization	0.9994	0.4549	0.4400
ResNet50	0.9989	0.4872	0.4820



Hình 10. Phân tích sau khi huấn luyện mô hình phân loại



Hình 11. Hình ảnh các sản phẩm được phân loại đúng



Hình 12. Hình ảnh các sản phẩm được phân loại không đúng

Hình 11 và Hình 12 là những hình ảnh mà chúng tôi phân loại hình ảnh mà chúng tôi đã huấn luyện được trên các mô hình. Hình 11 là những sản phẩm phân loại đúng danh mục và Hình 12 là những sản phẩm được phân loại sai danh mục.

D. Khuyến nghị

Đối vối nhiệm vụ khuyến nghị, để tránh trường hợp over-fitting chúng tôi có thêm thuật ngữ chính quy và chọn một hệ số tương đối lớn. Ở đây, tỷ lệ bỏ cuộc là tương đối nhỏ (0,1) vì số lượng hình ảnh trong một số danh mục bị hạn chế, không giống như tình huống trong nhiệm vụ phân loại, mà chúng tôi có dữ liệu hình ảnh trong tất cả 20 danh mục

Chúng tôi sử dụng mô hình VGG19 để dự đoán danh mục và trích xuất các đặc trưng. Đề xuất của chúng tôi dựa trên điểm tương tự cosine, vì chúng tôi phát hiện ra nó tốt hơn điểm tương tự L2

Qua hình 13 và hình 14, cho ta thấy được đầu vào của hệ thống là một hình ảnh và đầu ra là năm hình ảnh tương tự. Đối với Hình 13, đầu vào ở đây là một chiếc gương trong danh mục "Beauty" và đầu ra là năm hình ảnh nhưng có ba ảnh đúng là chiếc gương còn hai hình ảnh còn lại thì cho đề xuất sai một hình là chiếc quạt máy và một hình là sản phẩm thuộc danh mục "Home & Kitchen". Dựa vào Hình 13, đã cho thấy được hệ thống khuyến nghị chúng tôi đối với sản phẩm này chỉ có được 60%. Đối với Hình 14, đầu vào là hình ảnh của chiếc giày thuộc danh mục "Shoes" và đầu ra là hình ảnh 5 chiếc giày với màu sắc khác nhau và cho chúng ta thấy được hệ thống đề xuất đối với sản phẩm này là 100%.

Hình 13. Ví dụ về kết quả hệ thống đề xuất của chúng tôi

Hình 14. Ví dụ về kết quả hệ thống đề xuất của chúng tôi

VI. Kết luận

Trong dự án này, chúng tôi xây dựng một đề xuất mua sắm thông minh để tìm kiếm hình ảnh. Chúng tôi đã thử các mô hình mạng nơ-ron khác nhau để phân loại hình ảnh và các cách khác nhau để định lượng mức độ giống nhau giữa hai hình ảnh. Chúng tôi có thể đạt được độ chính xác phân loại là 0,5 và đề xuất các sản phẩm có điểm tương tự cao hơn 0,5. Có một vấn đề quá phù hợp trong mô hình của chúng tôi, đây có thể là một trong những điều cần làm trong công việc trong tương lai.

Như đã trình bày trong phần **Dữ liệu và đặc trưng**, mặc dù chúng ta có một bộ dữ liệu khổng lồ nhưng do giới hạn về thời gian và bộ nhớ máy nên chúng ta chỉ sử dụng được 5.000 trong tổng số 4 triệu hình ảnh. Trong bước tiếp theo, chúng tôi có thể cố gắng đào tạo mô hình của mình trên một lượng dữ liệu lớn hơn bằng cách sử dụng các lô. Điều này có thể làm tăng độ chính xác của mô hình

Hiện tại chúng tôi chỉ sử dụng 20 danh mục khi thực hiện phân loại. Tuy nhiên, các sản phẩm trong danh mục khác nhau rất nhiều, điều này giải thích cho việc phân loại của chúng tôi có độ chính xác thấp. Chúng tôi sẽ cố gắng tìm thông tin danh mục cụ thể hơn và đào tạo mô hình của chúng tôi về nó.

Bên cạnh đó, chúng tôi cũng muốn thử các mạng thần kinh sâu hơn như DenseNet.

Tài liệu

[1] S. Bell and K. Bala. Learning visual similarity for product design with convolutional neural networks. ACM Transactions on Graphics (TOG), 34(4):98, 2015.

[2] T. Deselaers and V. Ferrari. Visual and semantic similarity in imagenet. In Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, pages 1777–1784.IEEE, 2011.

[3] A. Dosovitskiy and T. Brox. Generating images with perceptual similarity metrics based on deep networks. In Advances in Neural Information Processing Systems, pages 658–666, 2016.

[4] R. He and J. McAuley. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In Proceedings of the 25th International Conference on World Wide Web, pages 507–517. International World Wide Web Conferences Steering Committee, 2016.

[5] V. Jagadeesh, R. Piramuthu, A. Bhardwaj, W. Di, and N. Sundaresan. Large scale visual recommendations from street fashion images. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 1925–1934. ACM, 2014.

[6] I. Kanellopoulos and G. Wilkinson. Strategies and best practice for neural network image classification. International Journal of Remote Sensing, 18(4):711–725, 1997.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, pages 1097–1105, 2012.

[8] J. McAuley, C. Targett, Q. Shi, and A. Van Den Hengel. Image-based recommendations on styles and substitutes. In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 43–52. ACM, 2015.

[9] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[10] M. Tan, S.-P. Yuan, and Y.-X. Su. A learning-based approach to text image retrieval: using cnn features and improved similarity metrics. arXiv preprint arXiv:1703.08013, 201

[11] G. W. Taylor, I. Spiro, C. Bregler, and R. Fergus. Learning invariance through imitation. In Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, pages 2729–2736. IEEE, 2011.

[12] G. Wang, D. Hoiem, and D. Forsyth. Learning image similarity from flickr groups using stochastic intersection kernel machines. In Computer Vision, 2009 IEEE 12th International Conference on, pages 428–435. IEEE, 2009.

[13] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu. Learning fine-grained image similarity with deep ranking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1386–1393, 2014.

[14] J. Yang, J. Fan, D. Hubball, Y. Gao, H. Luo, W. Ribarsky, and M. Ward. Semantic image browser: Bridging information visualization with automated intelligent image analysis. In Visual Analytics Science And Technology, 2006 IEEE Symposium On, pages 191–198. IEEE, 2006.

[15] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics, 2:67– 78, 2014.