Hệ Thống Đề Xuất Sản Phẩm Dựa Trên Hình Ảnh Với Mạng Thần Kinh Tích Chập

Đỗ Văn Thắng
Khoa công nghệ thông tin
Bộ môn khoa học dữ liệu
dovanthang08072k@gmail.com
MSSV: 19469481

Trần Xuân Thủy
Khoa công nghệ thông tin
Bộ môn khoa học dữ liệu
19446021.thuy@student.iuh.edu.vn
MSSV: 19446021

Trương Nguyễn Duy Tân
Khoa công nghệ thông tin
Bộ môn khoa học dữ liệu
truongnguyenduytan.2016@gmail.com
MSSV: 19485441

***Tóm tắt nội dung*—- Hầu hết các công cụ tìm kiếm mua sắm trực tuyến hiện nay vẫn phụ thuộc vào cơ sở kiến thức và kết hợp sử dụng từ khóa như cách tìm kiếm sản phẩm có nhiều khả năng muốn mua. - Điều này là Không hiệu quả bởi cách mô tả sản phẩm hay lựa chọn từ khóa có thể có sự khác nhau giữa bên bán và bên mua.**

Trong bài báo này , Chúng tôi sẽ trình bày một công cụ tìm kiếm thông minh hơn cho mua sắm trực tuyến. Nó sử dụng hình ảnh làm đầu vào và cố gắng hiểu thông tin về các sản phẩm từ hình ảnh. Sau đây là các bước chúng tôi thực hiện công cụ tìm kiếm thông minh này:

1) Đầu tiên, chúng tôi sử dụng bộ dữ liệu chứa thông tin 3.5 triệu sản phẩm thuộc 20 danh mục chính cùng với hình ảnh của Amazone.

2) Trên tập dữ liệu trên chúng tôi đào tạo một mô hình mạng nơ-ron để phân loại hình ảnh đầu vào thuộc một trong các danh mục sản phẩm có trong cơ sở dữ liệu.

3)Tiếp theo, Với mỗi hình ảnh sản phẩm, chúng tôi sẽ trích xuất đầu ra của lớp cuối cùng trong mô hình phân loại trên (Ứng với mỗi hình sẽ là một vector có kích thước 20x1 và tổng các phần tử trong vector này bằng 1) như là một vector đặc trưng.

4) Với hình ảnh đầu vào là một hình ảnh được người dùng cung cấp chúng tôi sẽ làm tương tự như ở bước 2, sau đó sử dụng độ đo khoảng cách cosin similarity để tính toán điểm tương đồng giữa vector đặc trưng của hình ảnh này với các vector đặc trưng của các hình ảnh trong cơ sở dữ liệu của chúng tôi. Hình ảnh được khuyến nghị sau đó sẽ là top n hình ảnh có khoảng cách cosin similarity so với hình ảnh đầu vào lớn nhất.

I. LỜI MỞ ĐẦU

Cuộc Cách mạng công nghiệp 4.0 diễn ra đã trở thành động lực cho thương mại điện tử (TMĐT) của thế giới cũng như Việt Nam ngày càng phát triển, đặc biệt là hoạt động thương mại điện tử xuyên biên giới thu hút được sự tham gia của nhiều thành phần trong xã hội. Trong bối cảnh dịch bệnh COVID-19, thị trường TMĐT càng trở nên sôi động hơn. khảo sát của Bộ Công Thương cho thấy, tính đến năm 2020, Việt Nam có 49,3 triệu người tham gia mua sắm trực tuyến (số liệu này năm 2016 mới chỉ ghi nhận 32,7 triệu người). Điều này cũng kéo theo sự bùng nổ thông tin gây quá tải thông tin cho người mua sắm trực tuyến bởi có rất nhiều chủ cửa hàng dần chuyển qua hình thức kinh doanh qua mạng. Để giải quyết vấn đề này, những hệ thống khuyến nghị cho người dùng đã được ra đời, tuy nhiên, chúng chủ yếu dựa trên sự mô tả kết hợp với từ khóa. Điều này chỉ giải quyết được phần nhỏ của vấn đề trên khi mô tả sản phẩm và lựa chọn từ khóa có thể có sự khác nhau giữa bên mua và bên bán. Nhờ sự phát triển của Machine learning và Deep learning cho ta cách tiếp cận mới, xử lý bài toán dựa trên dữ liệu. Giờ đây, chúng tôi có thể thay đổi cách tìm kiếm sản phẩm một cách trực quan và dễ dàng hơn. Xây dựng một hệ khuyến nghị sản phẩm dựa trên hình ảnh. Đầu vào cho thuật toán của chúng tôi là hình ảnh của bất kỳ sản phẩm nào mà khách hàng muốn mua $\Rightarrow$ Sử dụng mạng neu-ron CNN để phân loại danh mục $\Rightarrow$ đối tượng có thể thuộc về và sử dụng vector đầu vào của lớp được kết nói đầy đủ cuối cùng dưới dạng vector đặc trưng để cấp vào tính toán độ tương tự để tìm ra các sản phẩm gần nhất , Cụ thể 2 chức năng muốn đạt được

1) Phân loại: Đưa ra một bức ảnh của sản phẩmđuuọc chụp bởi khách hàng và từ đầu vào ấy sẽ tìm ra danh mục mà sản phẩm này có khả năng thuộc về nhất

2) Đề xuất: các đặc điểm của bức ảnh và danh mục sản phẩm này thuộc về, tính toán độ chẵn lẻ của sim và tìm các sản phẩm tương tự nhất trong cơ sỡ dữ liệu

II. CÔNG VIỆC 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 bằng cách sử dụng các nghiên cứu hợp 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.

Identify applicable funding agency here. If none, delete this.

Bài báo [7] đã trình bày mô hình AlexNet có thể phân loại độ tuổi của tôi thành 20 loại khác nhau. Ngoài ra, giấy [9] đã trình bày mạng nơ-ron VGG phân loại hình ảnh trong Im ageNet 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 chữa 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 từ cùng 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 đề xuất 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 naive ConvNets. Ý tưởng cũng có thể được tìm thấy trong [3] và [10].

Bài báo [1] cũng 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 bản chất nhất là độ tương tự cosin ei hoặc độ tương tự định mức L_2. Cách khác để đo mức độ tương đồng là bằng cách giới thiệu 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. Paper [15] và [14] có cùng ý tưởng như chúng tôi làm ở đây.

III. PHƯƠNG PHÁP TIẾP CẬN

Có hai vấn đề chính mà chúng tôi muốn giải quyết trong dự án của chúng tôi. Đầu tiên, xác định danh mục mà một hình ảnh nhất định thuộc về. Thứ hai, tìm và đề xuất những thứ tương tự nhất sản phẩm theo hình ảnh cho trước. Kể từ khi 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 lớp

A. Chập các lớp mạng nơ-ron

Bước quan trọng nhất của mạng nơ-ron phức hợp là Chuyển đổi lớp(Conv) . Như chúng ta có thể thấy từ Hình 1, lớp chuyển đổi sẽ dịch hình chữ nhật nhỏ của lớp đầu vào thành một số lớp đưa ra bằng cách sử dụng phép nhân ma trận

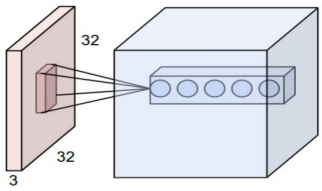


Figure1. Conv Layer

Các lớp gộp 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. Tổng hợp tối đa thường được sử dụng trong mạng nơ-ron phức hợp, sẽ xuất ra số lượng tối đa trong hình chữ nhật của lớp đầu vào.

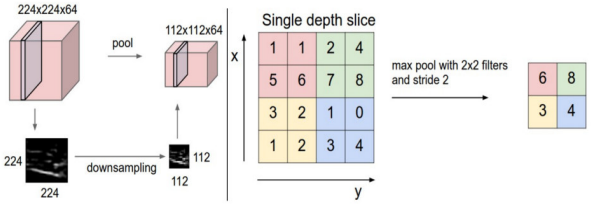


Figure2. Pooling Layer

Đánh giá trên mô hình VGG16:

B. Phân Loại

Trong bước này, chúng tôi muốn phân loại hình ảnh đầu vào thành một trong 20 loại. Chúng tôi xây dựng mô hình AlexNet và VGG cho nhiệm vụ phân loại và so sánh chúng với mô hình SVM làm mô hình cơ sở.

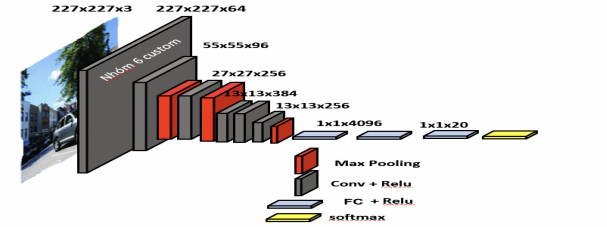
- **Support Vector Machine:** một mô hình phân loại tuyến tính, được sử dụng làm mô hình cơ sở ở đây. Mô hình này về cơ bản là một lớp kết nối đầy đủ. Chúng tôi sử dụng hàm loss SVM cho phân loại nhiều lớp cộng với điều kiện định mức L_2làm hàm tổn thất. Đối với hình ảnh i, chúng tôi sử dụng các pixel RGB làm đặc điểm đầu vào $x_i \in \mathbb{R}^d$, trong đó d = 224 × 224. Chúng tôi tính toán vô hướng cho n = 20 lớp thông qua một phép biến đổi tuyến tính

$$s = Wx_i + b$$

  trong đó $W \in \mathbb{R}^{n \times d}$ là ma trận trọng số và $b \in \mathbb{R}^n$ là độ lệch. Tổn thất SVM được đưa ra bởi

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

  trong đó y_i là nhãn cho lớp đúng.
- **AlexNet:** một mô hình phân loại mạng nơ-ron tích chập sâu được đề xuất bởi [7]. Như chúng ta có thể 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 có 3 lớp tích chập với tính năng tách biệt; 1 lớp gộp tối đa trước 3 lớp kết nối đầy đủ.



Mô hình ban đầu được đào tạo để phân loại hình ảnh trong nội dung ImageNet LSVRC-2010, nơi có 1000 danh mục. Vì vấn đề thiết bị khởi chạy mô hình nên bộ dữ liệu của chúng tôi chỉ chứa 20 danh mục, chúng tôi thay đổi lớp kết nối đầy đủ cuối cùng thành 4096 × 20. Để tiết kiệm thời gian, chúng tôi sử dụng trọng số huấn luyện trước đó của 5 nơ-ron đầu tiên và huấn luyện 3 lớp kết nối đầy đủ cuối cùng.

- **VGG 16:** Một mô hình phân loại mạng nơ ron tích tụ sâu được đề xuất bởi [9]. Như hình bên dưới (Hình4), VGG chứa 13 lớp phức hợp với tối đa gộp mỗi 2 hoặc 3 lớp chập; sau đó 3 đầy đủ các lớp được kết nối và softmax là lớp cuối cùng.

Bản mẫu ban đầu được huấn luyện để phân loại hình ảnh trong nội dung ImageNet ILSVRC-2014, nơi có là 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 đào tạo trước khi khởi tạo các tham số và để đào tạo ba lớp cuối cùng được kết nối đầy đủ. Chúng tôi cũng thêm các lớp chuẩn hóa hàng loạt sau khi kích hoạt chức năng trong hai lớp đầu tiên được kết nối đầy đủ.

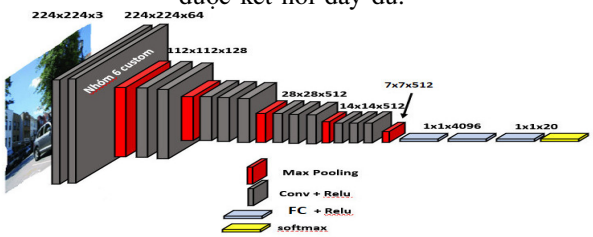


Figure4.VGG16 Model

Mô hình này bao gồm 16 lớp,bao gồm 13 lớp tích chập đều có kernel(3x3) và 3 lớp kết nối đầy đủ với lớp softmax là lớp cuối cùng.Sau mỗi lớp tích chập thì có giảm kích thước tối đa xuống 0.5 và hàm kích hoạt Relu để đơn giãn lọc các giá trị nhỏ hơn 0.

Mô hình ban đầu được huấn luyện để phân loại hình ảnh trong nội dung ImageNet ILSVRC-2014, nơi có là 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 đào tạo trước khi khởi tạo các tham số và để đào tạo ba lớp cuối cùng được kết nối đầy đủ. Chúng tôi cũng thêm các lớp chuẩn hóa hàng loạt sau khi kích hoạt chức năng trong hai lớp đầu tiên được kết nối đầy đủ.

Kết hợp 2 conv 3×3 có hiểu quả hơn 1 cov 5×5 về receptive field giúp mạng deeper hơn lại giảm tham số tính toán cho model. 3 Conv 3×3 có receptive field same 1 conv 7×7. Input size giảm dần qua các conv nhưng tăng số chiều sâu. Làm việc rất tốt cho task classifier và localizer (rất hay được sử dụng trong object detection). Với VGG-16, quan điểm về một mạng nơ ron sâu hơn sẽ giúp ích cho cải thiện độ chính xác của mô hình tốt hơn. Về kiến trúc thì VGG-16 vẫn dữ các đặc điểm của AlexNet nhưng có những cải tiến:

- Kiến trúc VGG-16 sâu hơn, bao gồm 13 layers tích chập 2 chiều (thay vì 5 so với AlexNet) và 3 layers fully connected VGG-16 cũng kế thừa lại hàm activation ReLU ở AlexNet.
- VGG16 chỉ sử dụng các bộ lọc kích thước nhỏ 3x3 thay vì nhiều kích thước bộ lọc như AlexNet. Kích thước bộ lọc nhỏ sẽ giúp giảm số lượng tham số cho mô hình và mạng lại hiểu quả tính toán hơn

C. Hệ Khuyến Nghị

Đối với bước đề xuất, chúng tôi sử dụng lớp được kết nối hoàn chỉnh cuối cùng trong mô hình phân loại của chúng tôi làm vectơ đặc trưng của 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 sử dụng để xác định hình ảnh mục tiêu mà đối tượng thuộc về. Sau đó chúng tôi trích xuất đầu vào từ lớp kết nối đầy đủ cuối cùng lớp của mô hình phân loại các tính năng.
- Đầu vào của mô hình: Vector mục tiêu đặc trưng của hình ảnh được trích xuất ở trên.
- Tính toán tương tự:Sử dụng các biện pháp khác nhau để tính toán điểm tương đồng giữa vectơ đặc trưng của mục tiêu lấy hình ảnh và vector đặc trưng của tất cả các hình ảnh mục tiêu thể loại để đo mức độ giống nhau giữa các cặp hình ảnh chúng tôi đã thử khoảng cách L_2 khoảng cách cosin và mô hình mạng nơ-ron để tính toán điểm tương đồng. Vì hai hình i và j khác nhau thì điểm khoảng cách L_2 là:

$$S_{L_2} = ||v_i - v_j||_2 \quad (3).$$

Tại $v_i, v_j \in \mathbb{R}_<$ là hai tính năng tương ứng vector, và l= 4096 là độ dài của vector đặc trưng, điểm S_{L_2} càng nhỏ là cả hai ảnh càng giống nhau.Điểm khoảng cách cosine được định nghĩa là:

$$s_{\text{cosine}} = \frac{v_i^{\mathsf{T}} vj}{||v_i||\ ||v_j||}$$

Điểm S_{cosine} càng lớn thì hai hình ảnh càng giống nhau. Cách tiếp cận theo hướng dữ liệu để tính toán điểm tương tự là đào tạo mạng nơ-ron 3 lớp sau:

$$h1 = f(vW1 + b1)$$
$$h2 = f(h1W2 + b2) \quad (5)$$
$$s_{model} = \text{sigmoid}\,(h_2 \cdot W_3 + b_3)$$

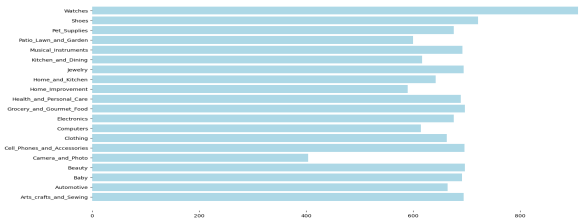
trong đó $v = [v_1, v_2] \in \mathbb{R}^{l\times 2}$ nhận được bằng cách ghép hai vectơ đặc trưng. $f\,(x) = \max\,(0.01x, x)$ là hàm ReLU bị rò rỉ. Lớp đầu tiên có thể được coi là lớp tích chập 1-d với

Leaky ReLU là hàm kích hoạt và $W_1 \in \mathbb{R}^2$ và $b_1 \in \mathbb{R}$ là tham số. Lớp thứ hai là lớp được kết nối đầy đủ với Leaky ReLU là chức năng kích hoạt và $W_2 \in \mathbb{R}_l$ và $b_2 \in \mathbb{R}$ là tham số. Lớp đầu ra là một phép biến đổi tuyến tính với hàm sigmoid là hàm kích hoạt. Điểm luyện càng lớn thì hai hình ảnh càng giống nhau. Không có cách nào dễ dàng để xác định điểm tương đồng hoàn toàn dựa trên các pixel hình ảnh. Ngoài ra, các hình ảnh đầu vào có tiêu đề tương ứng mô tả sản phẩm. là một số từ 0 đến 1. Đây cũng là lý do mà chúng tôi sử dụng hàm sigmoid làm hàm hoạt động cho lớp cuối cùng. Chúng tôi đào tạo mô hình này bằng cách giảm thiểu tổn thất $L_2 = ||s_{model} - s_{cosin}||_2^2$.
• Đầu ra: k ảnh hàng đầu (sản phẩm) giống nhất với ảnh mục tiêu.

IV. DỮ LIỆU VÀ THUỘC TÍNH

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, bao gồm 9,4 triệu sản phẩm. Do sự giới hạn về bộ nhớ chúng tôi chỉ lấy 13K sản phẩm với tổng số 20 danh mục. "Hình 5 cho thấy biểu đồ phân bổ của tất cả các nhãn của tập dữ liệu." Thông tin chi tiết của mỗi hình ảnh bao gồm:

- asin - ID của sản phẩm, ví dụ: 0000031852
- title - tên của sản phẩm
- price - giá bằng đô la Mỹ (tại thời điểm thu thập thông tin)
- imUrl - url của hình ảnh sản phẩm
- related - sản phẩm liên quan (đã mua, đã xem, mua chung, mua sau khi xem)
- salesRank - thông tin xếp hạng bán hàng
- brand - tên thương hiệu
- categories - danh sách các danh mục sản phẩm thuộc về



Xem xét sự mất cân bằng giữa các lớp khác nhau và do giới hạn của bộ nhớ máy, chúng tôi lấy mẫu ngẫu nhiên 500 hình ảnh trong mỗi lớp và thu thập 10000 hình ảnh cho nhiệm vụ phân loại. Sau đó, chúng tôi chia tập dữ liệu thành 7: 2: 1 tương ứng để đào tạo, xác nhận và kiểm tra. Chúng tôi sử dụng 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 6. 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 cho VGG và thay đổi kích thước thành 227 × 227 cho AlexNet.

V. THỰC NGHIỆM

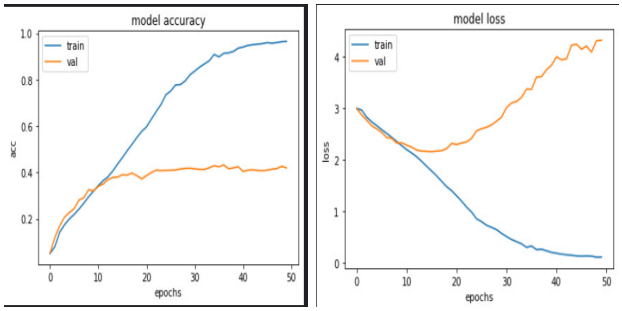
A. Tiền Xử Lí Dữ Liệu

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, hình ảnh gốc được thay đổi kích thước thành kích thước đầu vào tiêu chuẩn của mô hình VGG (224 × 224) hoặc mô hình AlexNet (227 × 227). Đối với mô hình khuyến nghị, sản phẩm tương tự được xác định bằng độ tương tự cosin similarity của các vector đặc trưng được trích xuất từ mô hình phân lớp cho hai hình ảnh. Cặp hình ảnh có độ tương tự cosin lớn hơn sẽ giống nhau hơn.

B. Đánh Giá

Chúng tôi chia tập dữ liệu thành 7: 2: 1 để đào tạo, thẩm định và kiểm tra 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 đầu ra với thực tế. Đối với vấn đề phân loại, chúng tôi đánh giá mô hình bằng cách tính độ chính xác phân loại:

$$Accuracy = \frac{\#correctly\ classified\ images}{\#images\ in\ validation\ dataset} \quad (7)$$



Đối với nhiệm vụ khuyến nghị, chúng tôi đánh giá mô hình sử dụng sai số bình phương trung bình căn (RSME).

$$RSME = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(s_{\text{model}} - s_{\text{cosin}})^2} \quad (8)$$

C. Phân Loại

Đối với nhiệm vụ phân loại, chúng tôi đã đào tạo hai Mạng thần kinh chuyển đổi (VGG16 và AlexNet) để phân loại các loại hình ảnh sản phẩm so với mô hình cơ sở của chúng tôi - mô hình phân loại tuyến tính (mô hình SVM). Bảng 1 cho thấy độ chính xác tốt nhất của chúng tôi về dữ liệu đào tạo, dữ liệu xác nhận và dữ liệu thử nghiệm cho ba mô hình này tương ứng. Đối với mô hình SVM, chúng tôi sử dụng tỷ lệ học 0,0005, hệ số điều chỉnh 0,001. Đối với AlexNet, chúng tôi sử dụng kích thước lô nhỏ 128, hệ số chính quy 0,01, tỷ lệ học 0,00001 và bỏ học 0,5. Đối với mô hình VGG, chúng tôi sử dụng kích thước lô nhỏ 100, hệ số chính quy 0,01, tỷ lệ học 0,0007 và bỏ học 0,5. Từ kết quả, chúng ta có thể thấy rằng các mô hình của chúng ta gặp phải vấn đề phù hợp quá mức. Khả năng đào tạo của cả mô hình AlexNet và mô hình VGG gần như gấp 1,5 lần độ chính xác đối với bộ xác nhận và bộ kiểm tra của chúng. Đó là lý do tại sao chúng ta cần hệ số chính quy tương đối cao hơn (0,01).

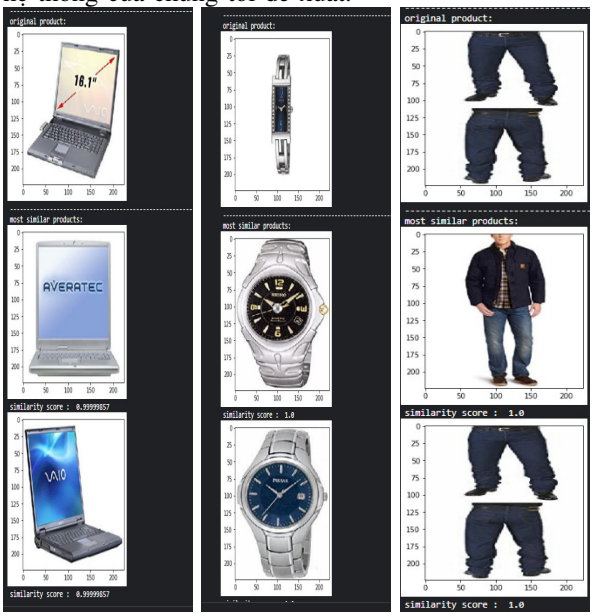
Nhưng khi chúng tôi tăng hệ số chính quy hóa, độ chính xác của thử nghiệm không tăng lên nữa. Vì lý do tương tự, tỷ lệ bỏ học mà chúng tôi chọn cho hai mô hình này cũng tương đối lớn (tương ứng là 0,5 và 0,5). Tuy nhiên, chúng tôi vẫn gặp phải tình trạng quá phù hợp với một số ngoại cảnh.

Mô Hình	training acc	validation acc	test acc
SVM	0.3583	0.0578	0.0613
AlexNet	0.6413	0.3567	0.3774
VGG	0.8932	0.4223	0.4431

Bảng 1. Kết quả độ chính xác của mô hình của AlexNet và VGG so với mô hình cơ sở

D. Hệ Khuyến Nghị

Chúng tôi sử dụng mô hình VGG để dự đoán danh mục và trích xuất các tính năng. Đề xuất của chúng tôi dựa trên điểm tương tự cosine, vì chúng tôi phát hiện ra nó tạo ra điểm tương đồng giữa các sản phẩm. Hình ảnh đầu tiên hiển thị các hình ảnh đầu vào mà người dùng đã chụp. Các hình ảnh tiếp theo hiển thị bốn sản phẩm tương tự hàng đầu mà hệ thống của chúng tôi đề xuất.



E. 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 đánh giá sự 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 Dataset and Features, 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 10.000 trong tổng số 3.5 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 ResNet.

VI. TÀI LIỆU THAM KHẢO

TÀI LIỆU

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

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

[3] A.DosovitskiyandT.Brox.Generatingimageswithpercep- tual 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 fil- tering. 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.KanellopoulosandG.Wilkinson.Strategiesandbestprac- tice 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 Confer- ence 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,andY.-X.Su.Alearning-basedapproach to text image retrieval: using cnn features and improved sim- ilarity metrics. arXiv preprint arXiv:1703.08013, 2017.

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

[12] G. Wang, D. Hoiem, and D. Forsyth. Learning image sim- ilarity from flickr groups using stochastic intersection ker- nel machines. In Computer Vision, 2009 IEEE 12th Interna- tional 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 Sym- posium On, pages 191–198. IEEE, 2006.

[15] P.Young,A.Lai,M.Hodosh,andJ.Hockenmaier.Fromim- age descriptions to visual denotations: New similarity met- rics for semantic inference over event descriptions. Transac- tions of the Association for Computational Linguistics, 2:67– 78, 2014.