DOI: 10.15625/vap.2020.00195

NHẬN DẠNG HÌNH ẢNH THỰC PHẨM BẰNG PHƯƠNG PHÁP DEEP LEARNING

Phan Anh Cang[1], Nguyễn Thanh Hoàng[1], Trần Hồ Đạt[1], Nguyễn Văn Hiếu[1], Phan Thượng Cang[2]

[1]Khoa Công nghệ thông tin, Trường Đại học Sư phạm Kỹ thuật Vĩnh Long
[2]Khoa Công nghệ thông tin và Truyền thông, Trường Đại học Cần Thơ

cangpa@vlute.edu.vn, hoangnt@vlute.edu.vn, datth@vlute.edu.vn, hieunv@vlute.edu.vn, ptcang@cit.ctu.edu.vn

TÓM TẮT: *Thói quen ăn uống không hợp lý là một trong những yếu tố nguy cơ hàng đầu dẫn đến tử vong và gánh nặng bệnh tật toàn cầu. Việc duy trì được một chế độ ăn lành mạnh trong suốt cuộc đời sẽ giúp phòng tránh được nhiều nguy cơ bệnh. Để có chế độ ăn uống lành mạnh cần nắm rõ những giá trị dinh dưỡng của các thực phẩm cũng như cách lựa chọn thực phẩm cho các bữa ăn. Hệ thống nhận dạng thực phẩm tự động và dự đoán dinh dưỡng của thực phẩm ngày càng cần thiết nhằm cung cấp được một chế độ ăn lành mạnh chính là chìa khóa để giải quyết các vấn đề dinh dưỡng bao gồm cả thừa, thiếu dinh dưỡng và thiếu vi chất dinh dưỡng. Trong nghiên cứu này, chúng tôi đề xuất hệ thống không chỉ có thể tự động nhận biết các thực phẩm mà còn có thể cho phép ước tính giá trị dinh dưỡng của chúng, làm cho chúng hữu ích trong việc lập kế hoạch ăn uống sao cho phù hợp với chế độ ăn uống của những người khác nhau. Chúng tôi thực hiện thu thập cơ sở dữ liệu ảnh thực phẩm cho hệ thống nhận dạng thực phẩm phục vụ huấn luyện và phát hiện 17 loại thực phẩm phổ biến. Bên cạnh đó, chúng tôi đề xuất mô hình mạng nơron tích chập (Faster R-CNN) sử dụng kiến trúc AlexNet và VGG16 trong nhận dạng hình ảnh thực phẩm và gợi ý giá trị dinh dưỡng của thực phẩm. Kết quả thực nghiệm cho thấy phương pháp của chúng tôi cho kết quả nhận dạng hiệu quả trên hầu hết các loại thực phẩm.*

Từ khóa: *Faster R-CNN, mạng nơron tích chập, nhận dạng hình ảnh thực phẩm.*

I. GIỚI THIỆU

Sự xuất hiện ngày càng nhiều của thực phẩm chế biến, tốc độ đô thị hóa cao cùng với sự thay đổi lối sống kéo theo sự thay đổi trong cách ăn uống, ngày nay con người tiêu thụ ngày càng nhiều các thức ăn giàu năng lượng, chất béo, đường tự do, muối. Việc lập kế hoạch ăn uống cũng phần nào cho bạn thấy được những chất dinh dưỡng mà bạn nạp vào mỗi ngày, từ đó cân bằng dinh dưỡng trong chính bữa ăn của mình tránh trường hợp bổ sung quá nhiều làm ảnh hưởng đến sức khỏe. Vì vậy, việc phát hiện, nhận dạng hình ảnh thực phẩm để đưa ra các gợi ý về các thành phần dinh dưỡng của thực phẩm trở nên cần thiết.

Trong thời gian gần đây, nhờ có sự phát triển mạnh mẽ về khả năng tính toán của các thế hệ máy tính hiện đại cũng như sự bùng nổ về dữ liệu thông qua mạng lưới Internet, ta đã chứng kiến nhiều sự đột phá trong lĩnh vực máy học, đặc biệt là trong lĩnh vực thị giác máy tính. Sự phát triển vượt bậc của các phương pháp học sâu đã giúp thị giác máy tính đạt được những thành tựu đáng kể trong lĩnh vực nhận dạng ảnh, trong đó có bài toán nhận dạng thực phẩm. Nội dung trình bày trong bài báo gồm giới thiệu các công việc liên quan; thu thập và xây dựng cơ sở dữ liệu ảnh cho hệ thống nhận dạng 17 loại thực phẩm; ứng dụng mô hình mạng Faster R-CNN phát hiện đối tượng trong ảnh; một số kết quả thực nghiệm đạt được.

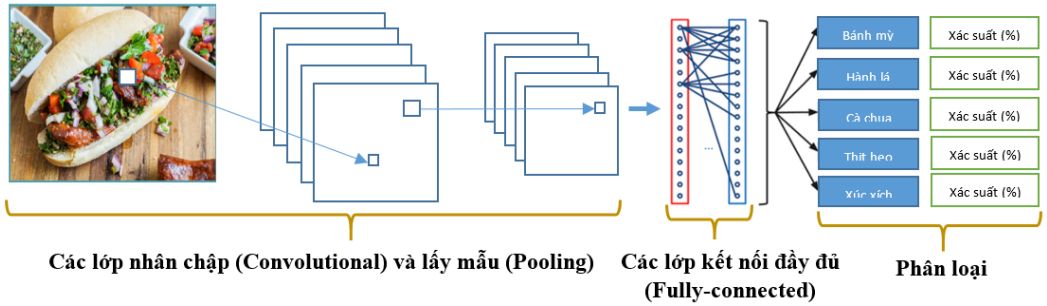
Các nhà nghiên cứu đã không ngừng cố gắng phát triển và cải thiện mô hình học sâu nhằm ngày càng nâng cao chất lượng nhận dạng ảnh hơn. Mặc dù trước đây chưa có bất kỳ hệ thống nhận dạng thành phần thực phẩm chuyên dụng nào, nhưng đã có nhiều cách tiếp cận để nhận dạng hình ảnh thực phẩm trong quá khứ sẽ được đề cập ngắn gọn về những hệ thống tiêu biểu dưới đây. Đầu tiên là nhóm nghiên cứu của Yang [1] sử dụng thuật toán STF (Semantic texton forests) ứng dụng trên 61 loại thực phẩm dựa trên tập ảnh thức ăn nhanh (Pittsburgh Fast-food Image Dataset) và kết hợp nó với mô hình SVM cho kết quả chính xác được 28,2 %. Nhóm Matsuda [2] sử dụng phương pháp mô hình biến dạng phân phối (Deformable part model) để xử lý trích chọn đặc trưng, sử dụng cửa sổ trượt trên ảnh theo định dạng kim tự tháp và áp dụng mô hình SVM để phân loại đối tượng. Họ đạt được 55,8 % cho phương pháp nhận dạng nhiều đối tượng và 68,9 % cho phương pháp nhận dạng một đối tượng, cải thiện đáng kể so với những nghiên cứu trước đây. Tuy nhiên, các nghiên cứu trước đây đều dựa trên các bộ trích xuất đặc trưng được xác định một cách thủ công, chẳng hạn như màu sắc hoặc kết cấu. Do đó, kết quả từ các công trình nghiên cứu này không đánh giá được hiệu suất trong thế giới thực do các điều kiện khác nhau trong thực tế xảy ra. Với sự ra đời của mạng nơron tích chập là sự lựa chọn tối ưu trong bải toán nhận dạng ảnh. Nhóm Yanai [3] công khai nghiên cứu của họ dựa trên mô hình mạng nơron tích chập thuần túy và đạt được kết quả là 72,26 % trên tập ảnh UEC-FOOD100 (University of Electro-Communications Food 100) được xuất bản công khai vào năm 2012, đây là độ chính xác cao nhất mà họ làm được tính tại thời điểm đó. Vào năm 2014, một phiên bản mới hơn dựa trên tập ảnh UEC-FOOD100 cũng được xuất bản, tập ảnh UEC-FOOD256 (University of Electro-Communications Food 256) chứa 256 loại thực phẩm khác nhau trong khi với tập ảnh UEC-FOOD100 thì chỉ chứa 100 loại và đạt được độ chính xác là 67,57 % trên tập này.

II. CÔNG VIỆC LIÊN QUAN

2.1. Mạng nơron tích chập (CNN - Convolutional Neural Network)

Mạng nơron tích chập (CNN - Convolutional Neural Network) là một trong những mô hình mạng phổ biến trong các hệ thống nhận dạng. Mạng CNN có khả năng xây dựng liên kết chỉ sử dụng một phần cục bộ trong ảnh kết nối đến nút trong lớp tiếp theo thay vì toàn bộ ảnh như trong mạng nơron truyền thẳng. Các lớp cơ bản trong một mạng

CNN bao gồm: lớp tích chập (Convolutional); lớp lấy mẫu (Pooling); lớp kích hoạt phi tuyến ReLU (Rectified Linear Unit) và lớp kết nối đầy đủ (Fully connected).

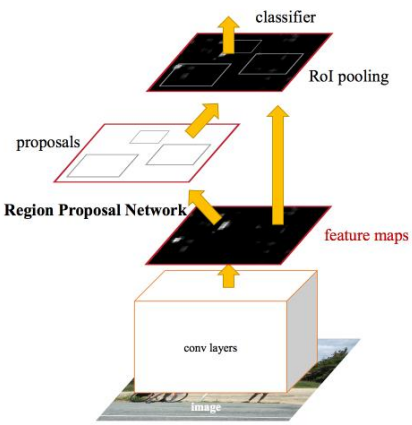


Hình 1. Cấu trúc của mạng nơron tích chập

Hình 1 mô tả cấu trúc của mạng nơron tích chập. Trong mô hình mạng nơron tích chập lan truyền thẳng thì mỗi nơron đầu vào cho mỗi nơron đầu ra trong các tầng tiếp theo, mô hình này gọi là mạng liên kết đầy đủ (Fully-connected). Các tầng liên kết được với nhau thông qua cơ chế tích chập, tầng tiếp theo là kết quả tích chập từ tầng trước đó, nhờ vậy mà ta có được các kết nối cục bộ. Như vậy mỗi nơron ở tầng kế tiếp sinh ra từ kết quả của mặt nạ chập áp lên một vùng ảnh cục bộ của nơron trước đó.

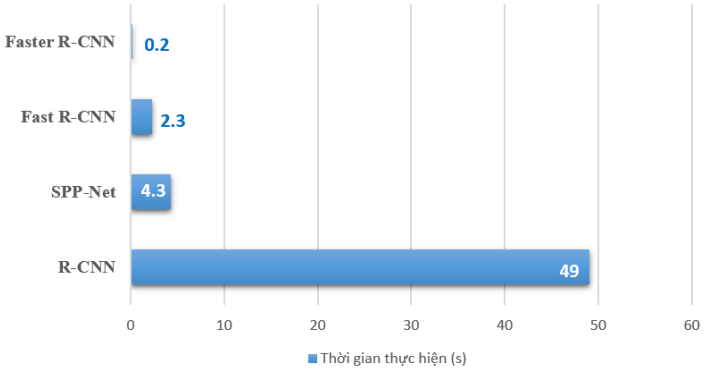
2.2. Faster R-CNN

R-CNN được giới thiệu lần đầu vào 2014 bởi Ross Girshick và các cộng sự ở UC Berkeley. Kiến trúc của R-CNN gồm 3 thành phần: vùng đề xuất hình ảnh (Region proposal); trích lọc đặc trưng (Feature Extractor) và phân loại (classifier). Một nhược điểm của phương pháp này là chậm, đòi hỏi phải vượt qua nhiều module độc lập trong đó có trích xuất đặc trưng từ một mạng CNN học sâu trên từng vùng đề xuất hình ảnh được tạo bởi thuật toán đề xuất vùng chứa ảnh. Năm 2015, mạng Fast R-CNN ra đời với sự đột phá trong phương pháp sử dụng bằng cách sử dụng một single model thay vì pipeline để phát hiện vùng và phân lớp cùng lúc. Ngay sau đó, Shaoqing Ren [4] và các cộng sự đề xuất mạng Faster R-CNN cải thiện hơn nữa về tốc độ huấn luyện và nhận dạng. Faster R-CNN là một thuật toán để tìm kiếm vị trí của vật thể trong ảnh. Thuật toán này sẽ có đầu ra là những hình hộp, cùng với vật thể bên trong hộp đó là gì. Mô hình mạng Faster R-CNN được mô tả theo Hình 2.



Hình 2. Kiến trúc mô hình mạng Faster R-CNN [5]

Theo [5] Faster R-CNN có tốc độ nhanh hơn 10 lần so với Fast R-CNN và hơn 200 lần so với khi chạy bằng thuật toán R-CNN. Kết quả thực nghiệm được tác giả trình bày trong Hình 3.

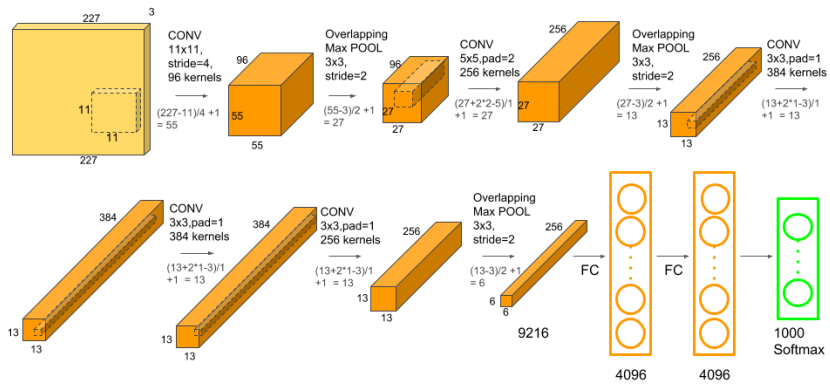


Hình 3. So sánh thời gian các thuật toán phát hiện đối tượng

Trong nội dung bài báo, chúng tôi đề xuất phương pháp phát hiện thành phần trong ảnh thực phẩm dựa trên mô hình mạng Faster R-CNN. Bước kế tiếp, chúng tôi tiến hành so sánh và đánh giá các mô hình với mục đích tìm ra mô hình tốt nhất trong nhận dạng. Một trong những lý do cần xét tới tiêu chí này vì có rất nhiều trường hợp mô hình chạy tốt trên tập dữ liệu chuẩn, nhưng bị hạn chế trên tập dữ liệu thực tế do tính phức tạp của dữ liệu. Bênh cạnh đó, tốc độ cũng như thời gian xử lý đóng vai trò quan trọng trong các ứng dụng. Hơn nữa, việc cân bằng giữa độ chính xác và tốc độ xử lý cũng là một thách thức. Để đánh giá các yếu tố này, chúng tôi sử dụng các độ đo đánh giá mô hình của bài toán nhận dạng đối tượng nhằm tìm ra mô hình có độ chính xác cao nhất ứng dụng hiệu quả trong bài toán nhận dạng ảnh thực phẩm.

2.3. Kiến trúc mạng AlexNet và VGGNet

AlexNet là kiến trúc mạng nơron tích chập đầu tiên đặt nền móng cho các kiến trúc mạng nơron sử dụng mạng nơron tích chập. Krizhevsky cha đẻ của Alexnet đã chiến thắng cuộc thi ImageNet năm 2012 với tỉ lệ lỗi khoảng 15,4 % tốt hơn hẳn so với các phương pháp được sử dụng trước đó. Kiến trúc mô hình mạng AlexNet được trình bày trong Hình 4.



Hình 4. Mô hình mạng AlexNet [6]

Kiến trúc này sử dụng 5 tầng của mạng nơron tích chập để phân loại cho 1000 lớp. Điểm đặc biệt của Alexnet không chỉ nằm ở các tầng của mạng nơron tích chập mà còn là việc sử dụng các hàm kích hoạt ReLU được chứng minh là cho tốc độ huấn luyện hiệu quả hơn so với các hàm kích hoạt khác trước đó. Kiến trúc của mô hình này được mô tả trong Hình 2.4. Đến năm 2014, VGGNet đứng hạng hai trong cuộc thi ImageNet 2014 và chỉ đứng sau mạng GoogleNet.

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
Input (224 x 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 conv1-512	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

Hình 5. Kiến trúc của mạng VGG16 [6]

Kiến trúc của mạng VGGNet được mô tả trong Hình 5 là một chuẩn thiết kế mạng học sâu của Visual Geometry Group thuộc đại học Oxford. Mô hình này đơn giản và có độ sâu hơn so với kiến trúc AlexNet. Tất cả các tầng của mạng nơron tích chập trong mô hình này gồm có bộ lọc 3x3 với bước nhảy = 1, kích thước lề = 1 và tầng tổng hợp cực đại. Chính điều này đã làm giảm số lượng các tham số của mạng. Dựa trên các nghiên cứu của tác giả [7], [5] cho thấy việc rút trích đặc trưng và nhận dạng ảnh thực phẩm dựa trên kiến trúc AlexNet và VGG16 cho kết quả tốt. Vì vậy, trong nội dung bài báo này chúng tôi đề xuất mô hình mạng nơron tích chập (Faster R-CNN) sử dụng kiến trúc AlexNet và VGG16 trong nhận dạng hình ảnh thực phẩm.

2.4. Độ đo đánh giá mô hình

Độ đo mAP (Mean Average Precision) [8] theo chuẩn đánh giá PASCAL VOC [9] được sử dụng để đánh giá cho các mô hình phát hiện đối tượng trong ảnh theo công thức (1). Khác với phương pháp đo lường theo độ chính xác thông thường, mAP cho phép chúng ta kiểm nghiệm chất lượng cho các tập ảnh không cân bằng về số lượng dữ liệu của từng loại một cách tốt hơn. Đo thời gian huấn luyện mô hình bằng giờ và thời gian dự đoán trên từng ảnh bằng giây.

Công thức tính độ đo mAP:

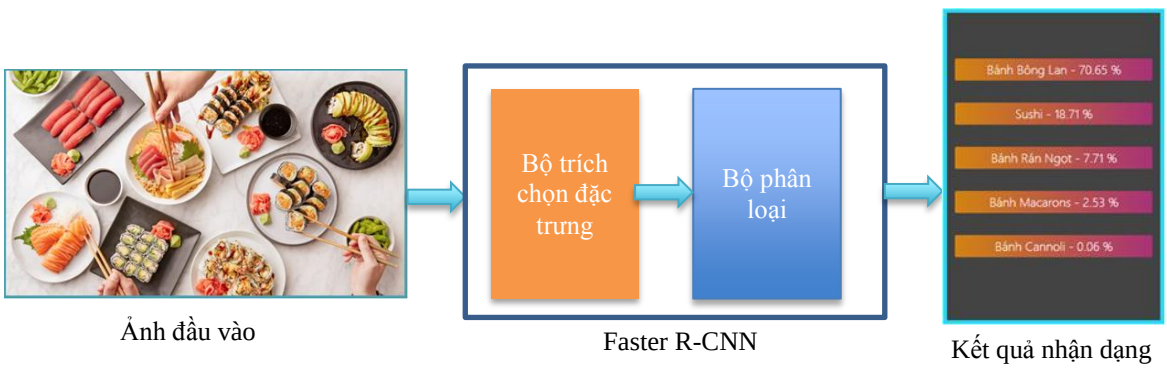
$$\text{mAP} = \sum_{q=1}^{Q} \frac{AP(q)}{Q} \quad (1)$$

Trong đó Q là số lượng lớp đối tượng (thành phần thực phẩm) có trong tập dữ liệu, AP là độ chính xác trung bình của từng lớp được tính bằng công thức như sau:

$$\text{AP} = \frac{1}{n} \sum_{r\epsilon\{0,1,\ldots,n\}} P_{inter}(r) \quad (2)$$

III. MÔ HÌNH ĐỀ XUẤT

Trong các mô hình nhận dạng thực phẩm trước đây việc nhận dạng ảnh thực phẩm được thực hiện hiện bằng phương pháp trích xuất đặc trưng trực tiếp từ ảnh đầu vào sau đó sử dụng bộ phân loại để thực hiện nhận dạng các loại thực phẩm. Tuy nhiên, nhược điểm của phương pháp này gặp khó khăn trong việc trích xuất đặc trưng từ ảnh, độ chính xác trong nhận dạng không cao và có xu hướng giảm dần khi số lượng ảnh đầu vào tăng lên. Trong các nghiên cứu gần đây [10] [7] [11] đã minh chứng việc thực hiện nhận dạng ảnh bằng phương pháp Deep Learning sẽ có kết quả tốt hơn các phương pháp trước đây và độ chính xác sẽ tăng dần khi tập dữ liệu ảnh đầu vào lớn. Do đó, chúng tôi đề xuất mô hình tổng quát nhận dạng thực phẩm bằng phương pháp Deep Learning nhằm tối ưu hóa công việc trích xuất đặc trưng trên ảnh đầu vào và tăng cường độ chính xác trong quá trình nhận dạng.



Hình 6. Mô hình tổng quát đề xuất hệ nhận dạng ảnh thực phẩm bằng phương pháp Faster R-CNN

Hình 6 mô tả quá trình nhận dạng được thực hiện bao gồm các bước: (1) Huấn luyện ảnh đầu vào bằng mô hình Faster R-CNN; (2) Thực hiện kiểm thử và tinh chỉnh các trọng số nhằm tìm ra mô hình tốt nhất; (3) Nhận dạng ảnh thực phẩm dựa trên mô hình Faster R-CNN và đánh giá độ chính xác.

Từ Hình 6 ta thấy mô hình mạng Faster R-CNN và mạng CNN đều có kiến trúc tổng quát chung, cho nên mô hình vẫn gồm ba giai đoạn chính là giai đoạn huấn luyện, giai đoạn đánh giá và giai đoạn kiểm thử, các thành phần trong hai giai đoạn huấn luyện và đánh giá gồm có bộ trích chọn đặc trưng và bộ phân loại, nhận dạng và bộ xác định vị trí bao đóng. Trong mỗi lần học, các đánh giá sẽ được chuẩn đoán và so khớp chất lượng của mỗi pha và chọn ra mô hình dự đoán tốt nhất. Riêng với pha kiểm thử, mô hình bắt đầu dự đoán, tính toán các số liệu liên quan đến chất lượng mô hình và đánh giá kết quả cho mô hình. Sau đó mô hình này sẽ được ứng dụng vào triển khai thực tế.

3.1. Phát hiện đối tượng trong ảnh bằng mô hình Faster R-CNN: Các bước thực hiện bao gồm:

1. Bộ trích chọn đặc trưng.

2. Bộ phân loại và xác định vị trí bao đóng: bộ phát hiện đối tương (Object Detector) tiếp nhận các bản đồ đặc trưng từ cả hai mạng nơron tích chập và mạng RPN và đưa qua các tầng liên kết đầy đủ (FC Layers) để thực hiện phân loại đối tượng bằng các hàm softmax và dự đoán tọa độ của tầng đối tượng đó bằng kỹ thuật hồi quy (bounding-box regression).

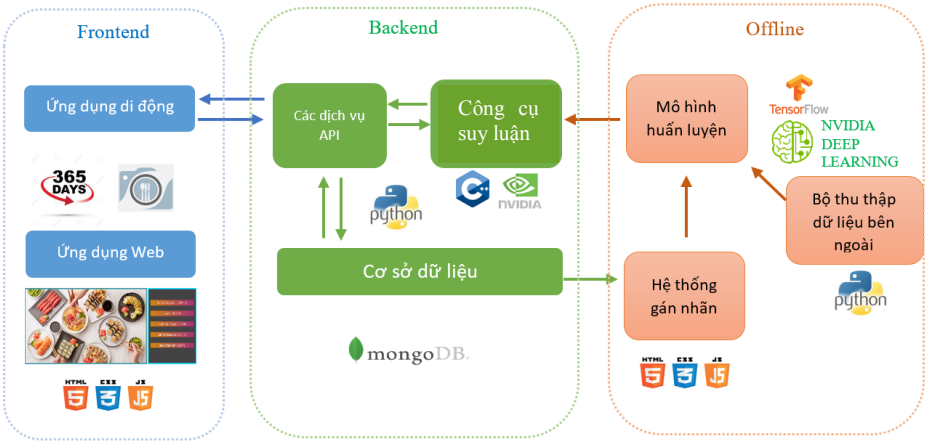
3. Lựa chọn mô hình có giá trị mAP cao nhất, tính toán độ chính xác cho mô hình.

4. Lựa chọn mô hình có độ chính xác cao đưa vào nhận dạng.

Tiếp nhận các bản đồ đặc trưng từ đầu ra của mạng nơron tích chập. Sau khi có được các đặc trưng học sâu (feature maps) từ các tầng tích chập đầu tiên (CNN), mạng RPN sử dụng cửa sổ trượt trên bản đồ đặc trưng (feature map) để rút trích đặc trưng cho mỗi vùng đề xuất. RPN được xem như là một mạng nơron tích chập đầy đủ cùng lúc thực hiện hai nhiệm vụ đó là dự đoán tọa độ cho các đối tượng (bounding box) và gán điểm số 0 (là đối tượng) hoặc 1 (không là đối tượng) cho đối tượng đó (objectness score).

3.2. Xây dựng hệ thống nhận dạng

Bên cạnh đó, chúng tôi xây dựng hệ thống nhận dạng ảnh thực phẩm bao gồm 3 module:



Hình 7. Hệ thống nhận dạng ảnh thực phẩm

Như Hình 7 hệ thống nhận dạng các thành phần thực phẩm trong món ăn được thiết kế theo mô hình client-server trong đó: Giai đoạn một là giai đoạn giao diện người dùng, cụ thể là ứng dụng client trên điện thoại di động và website, quản lý tương tác người dùng như chụp ảnh, chọn ảnh gửi lên server và hiển thị kết quả nhận dạng do server gửi về. Giai đoạn hai là giai đoạn server quản lý giao thức gửi/nhận dữ liệu với client, cụ thể giao thức được sử dụng trong hệ thống là giao thức HTTP/HTTPS. Server thực hiện xử lý các yêu cầu từ client, như quản lý và phân phối các luồng xử lý độc lập, đảm bảo hiệu năng và chất lượng tính toán nhận dạng cho nhiều client trong cùng một thời điểm. Giai đoạn ba, server còn đảm nhiệm xây dựng mô hình, tinh chỉnh và quản lý các phiên bản mô hình nhận dạng cho hệ thống và quản lý dữ liệu, bao gồm các thông tin về thành phần thực phẩm có trong món ăn, mỗi thành phần đều có các thông tin chi tiết về số lượng calo, protein, chất béo,....

IV. KẾT QUẢ THỰC NGHIỆM

4.1. Môi trường và dữ liệu kiểm thử

Cấu hình thử nghiệm: Môi trường được sử dụng để huấn luyện mô hình nhận dạng các thành phần thực phẩm món ăn là máy tính với vi xử lý intel core i5, bộ nhớ trong 16 GB, bộ xử lý đồ họa GTX 1060 với kích thước bộ nhớ là 6 GB, hệ điều hành Ubuntu 16.04, ngôn ngữ Python với framework là CNTK (Computational Network Toolkit). Sau quá trình tìm hiểu và so sánh các framework, chúng tôi đã quyết định chọn CNTK làm công cụ cài đặt triển khai ứng dụng cho bài toán nhận dạng các thành phần trong ảnh bởi Microsoft đã tích hợp các mã nguồn có sẵn cho mô hình Faster R-CNN giúp giảm thời gian xây dựng mô hình huấn luyện.

Tập dữ liệu huấn luyện: Do tập ảnh dùng để nhận dạng thành phần trong món ăn không có gán nhãn sẵn nên chúng tôi đã tự gán nhãn thủ công một tập ảnh riêng để sử dụng từ các hình ảnh được lấy trong tập ảnh Food-101 [12]. Tập ảnh Food-101 là bộ dữ liệu hình gồm 101 loại thực phẩm (Hình 8), mỗi loại thực phẩm có 1.000 hình (tổng hình trong tập ảnh là 101.000 hình). Mỗi một lớp (class) trong tập ảnh có 250 hình để kiểm tra (test) và 750 hình để huấn luyện (training). Kích thước tất cả các hình trong tập ảnh là 512 pixel.

Hình 8. Tập ảnh Food-101

Tập ảnh dùng để nhận dạng gồm 17 loại thành phần khác nhau (gọi tắt là tập ảnh) như: bánh mì, bánh ngô, bún, chanh, cà chua, cá, giá, đậu, hành lá, hành tây, khoai tây, nước sốt, phô mai, rau húng quế, thịt bò, thịt heo, tương cà, xúc xích. Trong hệ thống lưu trữ, tập ảnh được chia thành các thư mục riêng biệt: một là tập tin ảnh, hai là tập tin chứa các tầng tương ứng với các đối tượng trong ảnh và ba là tập tin chứa vị trí của các đối tượng có trong ảnh.

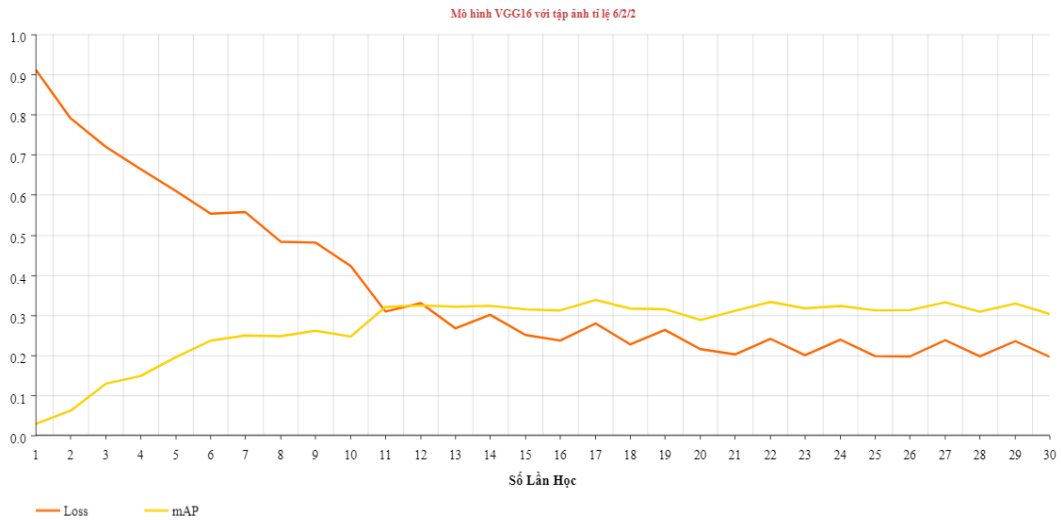
Chúng tôi sử dụng lại tập ảnh đã được xây dựng từ mô hình mạng Faster R-CNN. Tập ảnh cũng gồm 17 loại thực phẩm khác nhau. Hệ thống lưu trữ cũng được chia thành các thư mục như sau:

Thư mục gốc: Gồm các thư mục chứa các tập ảnh với tập đầu được chia theo tỉ lệ 60 % ảnh huấn luyện, 20 % ảnh đánh giá, 20 % ảnh kiểm thử. Tập còn lại với 70 % ảnh huấn luyện, 10 % ảnh đánh giá và 20 % ảnh kiểm thử. Thư mục con: Thư mục chứa các ảnh huấn luyện, ảnh đánh giá và ảnh kiểm thử nằm riêng biệt trong các thư mục. Mỗi thư mục sẽ nhận tương ứng hai tập tin, một tập tin chứa thông tin về các đường dẫn tới ảnh, tập tin thứ hai chứa các thông tin về từng đối tượng có trong ảnh như vị trí của đối tượng trong ảnh, đối tượng trong ảnh thuộc tầng nào.

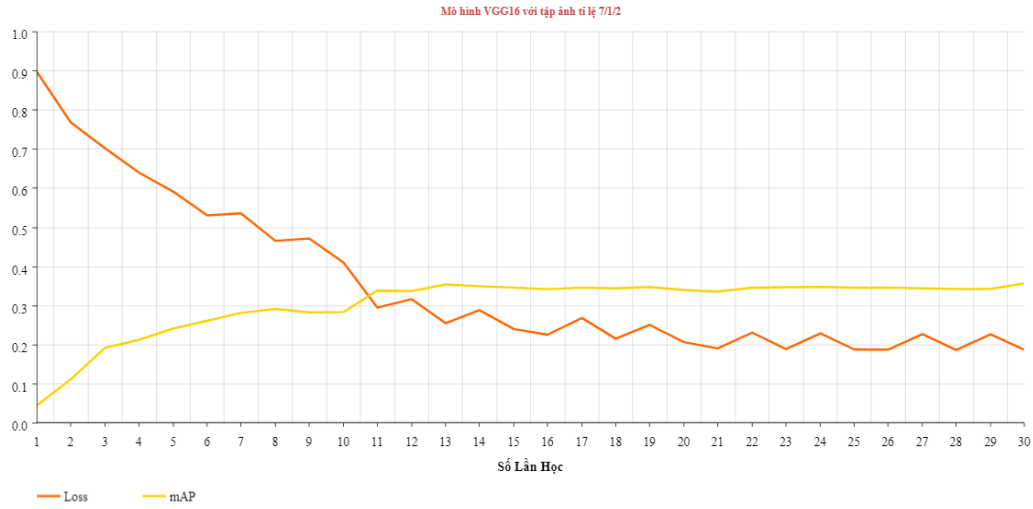
Hình 9. Hình ảnh trên tập dữ liệu và kết quả tương ứng với đầu ra của mô hình

Hình 9 mô tả ứng dụng chúng tôi đã xây dựng thực hiện nhận dạng hình ảnh với các loại thực phẩm khác nhau với kết quả đầu ra tương ứng được thể hiện bên cạnh.

Xác định tham số huấn luyện: Chúng tôi thực hiện huấn luyện trên mô hình Faster R-CNN sử dụng kiến trúc AlexNet và VGG16 với các tập ảnh tỉ lệ 6:2:2 và 7:1:2. Mô hình sẽ được huấn luyện trên ảnh đầu vào có kích thước 512x512. Tùy vào kích thước đối tượng chứa trong ảnh mà chúng ta sẽ chọn các hộp mẫu (Anchor box) phù hợp. Theo tính toán, các kích thước hộp mẫu tương đối thích hợp với tập ảnh này với tỉ lệ (4, 8, 12) kết hợp với 3 tỉ lệ (8x24, 16x16, 24x8) ta sẽ được 9 hộp mẫu = (32x96, 64x64, 96x32, 64x192, 128x128, 192x64, 96x288, 192x192, 288x96) ứng với kích thước trên ảnh gốc. Điều chỉnh tầng softmax của mô hình ứng với 17 loại thực phẩm. Về thuật toán tối ưu, mô hình sử dụng thuật toán SGD để tối ưu hàm lỗi. Với tốc độ học learning rate = 0.001 cho 10 lần học đầu tiền, 0.0001 cho 10 lần học kế tiếp và 0.00001 cho các lần học còn lại. Toàn bộ quá trình huấn luyện được thực hiện trong 30 lần học. Kết quả huấn luyện trên mô hình Faster R-CNN sử dụng kiến trúc AlexNet và VGG16 với các tập ảnh tỉ lệ 6:2:2 và 7:1:2 như sau:



Hình 10. Quá trình học của mô hình mạng Faster R-CNN sử dụng kiến trúc VGG16 trên tập dữ liệu tỉ lệ 6:2:2



Hình 11. Quá trình học của mô hình mạng Faster R-CNN sử dụng kiến trúc VGG16 trên tập dữ liệu tỉ lệ 7:1:2

Hình 10 và Hình 11 biểu diễn quá trình học của mô hình Faster R-CNN, với loss là độ lỗi trên tập dữ liệu huấn luyện, mAP là độ đo được tính trên tập dữ liệu đánh giá, ta thấy giá trị độ đo mAP ngưng tăng khi mô hình đã qua 12 lần học và sau đó bão hòa, chúng ta sẽ sử dụng mô hình được sao lưu tại lần học thứ 12 làm mô hình cho việc đánh giá cho tập dữ liệu kiểm thử. Phương pháp tương tự cũng được áp dụng cho các mô hình khác. Kết quả nhận dạng của 2 mô hình được trình bày trong phần kết quả thực nghiệm.

4.2. Kết quả thực nghiệm

Chúng tôi thực hiện so sánh trên tập dữ liệu huấn luyện sử dụng công thức (1) và (2) được trình bày trong mục 2.4. Kết quả nhận dạng trên mô hình Faster R-CNN sử dụng kiến trúc AlexNet và VGG16 với các tập ảnh tỉ lệ 6:2:2 và 7:1:2 trên 17 loại thực phẩm như sau:

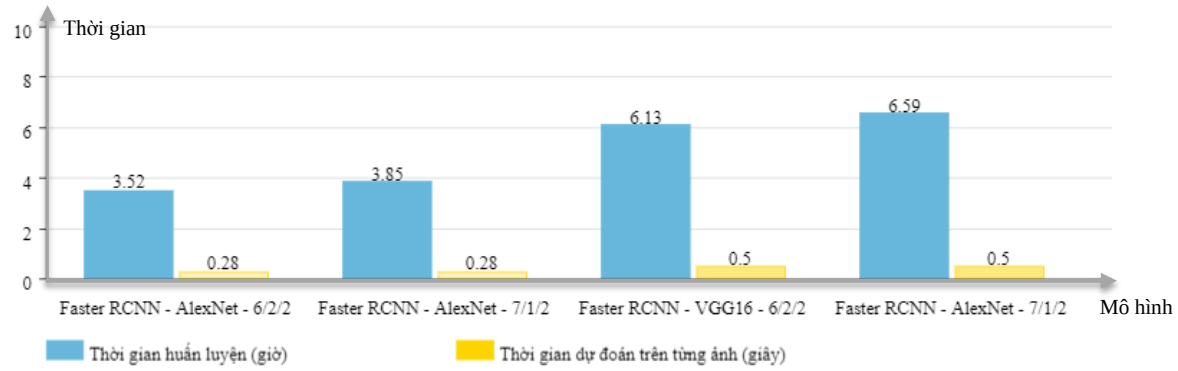
Bảng 1. So sánh kết quả từng thực phẩm giữa 2 kiến trúc AlexNet và VGG16 sử dụng độ đo AP

Thành phần thực phẩm	**AlexNet**		**VGG16**	
	6:2:2 (% AP)	**7:1:2 (% AP)**	**6:2:2 (% AP)**	**7:1:2 (% AP)**
Bánh mỳ	43,66	28,75	**55,35**	47,08
Bánh ngô	28,0	38,85	36,95	**51,17**
Bún	20,84	34,37	42,08	**49,04**
Chanh	35,63	29,03	33,6	**40,32**
Cà Chua	3,95	12,38	14,86	**17,4**
Cá	18,93	14,04	**26,43**	18,61
Giá Đậu	29,52	36,45	33,76	**43,86**
Hành Lá	**7,23**	6,12	6,17	5,4
Hành Tây	6,27	14,54	9,35	**16,56**
Khoai Tây	45,86	37,58	**57,12**	52,11
Nước Sốt	43,04	**51,83**	51,14	48,75
Phô Mai	18,49	**27,41**	23,23	25,74
Rau Húng Quế	16,45	20,03	**21,33**	16,62
Thịt Bò	2,43	4,07	3,25	**7,13**
Thịt Heo	43,09	61,91	53,89	**70,45**
Tương Cà	11,49	16,43	**28,36**	26,5
Xúc Xích	17,87	25,82	**27,52**	26,41

Bảng 2. Kết quả so sánh độ đo mAP giữa 2 kiến trúc AlexNet và VGG16

Kiến trúc	**AlexNet**		**VGG16**	
mAP(%)	23,10	27,04	30,85	**33,13**

Dựa vào kết quả Bảng 1 và Bảng 2 cho thấy kiến trúc VGG16 cho chất lượng nhận dạng tốt hơn hẳn so với kiến trúc AlexNet, trong đó tập ảnh tỉ lệ 6:2:2 với kiến trúc VGG16 tốt hơn 7,75 % so với kiến trúc AlexNet với cùng tập ảnh. Chất lượng tương tự được biểu hiện trên tập ảnh tỉ lệ 7:1:2, kiến trúc VGG16 nhận dạng tốt hơn tới 5,59 % so với AlexNet. Phần lớn chất lượng mô hình trên kiến trúc VGG16 đều cao hơn chất lượng trên kiến trúc Alex về mọi mặt, chứng minh rằng VGG16 có khả năng chọn lọc đặc trưng tốt hơn và chính xác hơn dựa vào tính chất học sâu của mô hình với 13 lớp tích chập trong khi đó AlexNet chỉ có 5 lớp tích chập. Hầu hết kết quả kiểm nghiệm trên tập dữ liệu tỉ lệ 7:1:2 đều cao hơn so với tập tỉ lệ 6:2:2 trên từng thành phần thực phẩm nhưng có một số thành phần tập 6:2:2 biểu hiện tốt hơn như bánh mỳ, cá, khoai tây, rau húng quế và tương cà.



Hình 12. Biểu đồ phân bố về thời gian thực hiện của các mô hình

Trong Hình 12, mô hình Faster R-CNN với các kiến trúc AlexNet, VGG16 là có thời gian tối đa từ 3 đến 6 giờ cho mỗi 30 lần học khác nhau. Thông qua kết quả về thời gian và độ chính xác của bảng số liệu chúng tôi thấy rằng độ chính xác của các mô hình ảnh hưởng nhiều vào mạng cơ sở khi rút trích đặc trưng và phần mạng phía sau dùng để xử lý đặc trưng cũng như hàm tính độ lỗi khi huấn luyện của các mô hình.

V. KẾT LUẬN

Trong nghiên cứu này, chúng tôi thực hiện nhận dạng và phân loại tự động thực phẩm trong các ảnh màu. Chúng tôi thực hiện thu thập cơ sở dữ liệu ảnh thực phẩm từ nhiều nguồn khác nhau cho hệ thống nhận dạng thực phẩm và hoàn thiện xây dựng bộ cơ sở dữ liệu ảnh phục vụ huấn luyện phát hiện đối tượng thực phẩm cho 17 loại phổ biến. Chúng tôi đã xây dựng hệ thống nhận dạng thực phẩm bằng phương pháp Faster R-CNN. Kết quả thực nghiệm cho thấy phương pháp đề xuất đạt kết quả chính xác cao trong nhận dạng một số loại thực phẩm phổ biến. Trong nghiên cứu sắp tới, chúng tôi tiếp tục cải tiến tập dữ liệu phong phú hơn so với 17 loại tại thời điểm hiện tại đồng thời tăng tốc độ nhận dạng của phương pháp đề xuất. Mặt khác, chúng tôi thực hiện so sánh, đánh giá với các kiến trúc mạng khác (Yolo, Mark-RCNN,…) nhằm tìm ra phương pháp tối ưu trong việc cung cấp thông tin chính xác và hiệu quả về giá trị dinh dưỡng của các loại thực phẩm để mọi người có chế độ ăn lành mạnh chính là chìa khóa giải quyết các vấn đề dinh dưỡng bao gồm cả thừa, thiếu dinh dưỡng và thiếu vi chất dinh dưỡng.

TÀI LIỆU THAM KHẢO

[1] S. C. M. P. D. &. S. R. Yang, "Food recognition using statistics of pairwise local features", Computer Society Conference on Computer Vision and Pattern Recognition, pp. 2249-2256, 2010.

[2] Y. H. H. &. Y. K. Matsuda, "Recognition of multiple-food images by detecting candidate regions", International Conference on Multimedia and Expo, pp. 25-30, 2012.

[3] Y. &. Y. K. Kawano, "Real-time mobile food recognition system", Conference on Computer Vision and Pattern Recognition Workshops, pp. 1-7, 2013.

[4] S. e. a. Ren, "Faster r-cnn: Towards real-time object detection with region proposal networks", Advances in neural information processing systems, 2015.

[5] S. H. K. G. R. &. S. J. Ren, "Faster r-cnn: Towards real-time object detection with region proposal networks", In Advances in neural information processing systems, pp. 91-99, 2015.

[6] K. &. Z. A. Simonyan, "Very deep convolutional networks for large-scale image recognition.", arXiv preprint arXiv:1409.1556., 2014.

[7] S. &. K. S. B. Mezgec, "NutriNet: a deep learning food and drink image recognition system for dietary assessment.", Nutrients, 2017.

[8] J. Hui, "mAP (mean Average Precision) for Object Detection", 2018.

[9] L. V. G. C. K. I. W. a. J. W. M. Everingham, "The PASCAL Visual Object Classes (VOC) Challenge", International Journal of Computer Vision, p. 303-338, 2010.

[10] E. E. S. A. J. B. K. &. R. S. Cust, "Machine and deep learning for sport-specific movement recognition: a systematic review of model development and performance.", Journal of sports sciences, pp. 568-600, 2019.

[11] H. &. A. K. Kagaya, Highly accurate food/non-food image classification based on a deep convolutional neural network.", International conference on image analysis and processing, pp. 350-357, 2015.

[12] K.-H. e. a. Lee, "Cleannet: Transfer learning for scalable image classifier training with label noise.", Proceedings

of the IEEE Conference on Computer Vision and Pattern Recognition, 2018.

FOOD IMAGE RECOGNITION USING DEEP LEARNING

Phan Anh Cang, Nguyen Thanh Hoang, Tran Ho Dat, Nguyen Van Hieu, Phan Thuong Cang

ABSTRACT: *In appropriate eating habits are among the leading risk factors for death and the global burden of disease. Maintaining a healthy diet throughout life will help preventing many risks of disease. To have a healthy diet needs to understand the nutritional value of foods as well as how to choose foods for meals. The automatic food identification and nutritional prediction system is necessary and is the key to solve nutritional problems including excess and lack of nutrition. and lack of micronutrients. In this study, we propose the system which is not only automatically recognize foods, but also allows for an estimate of their nutritional value, making them useful in star eating planning. to suit the diets of different people. We collected a food photo database for our food identification system for training and found 17 common foods. In addition, we propose a convolutional neural network model (Faster R-CNN) using AlexNet and VGG16 architectures in food image recognition and suggesting the nutritional value of food. Experimental results show that our method provides effective identification on almost all foods.*