CONVOLUTIONAL NEURAL NETWORK (CNN) TRONG TRÍCH XUẤT ĐẶC TRƯNG ẢNH

NGUYỄN HOÀI NAM

Ngày 11 tháng 12 năm 2023

Tóm tắt nội dung

Học sâu là một lĩnh vực học máy mới nổi đã phát triển nhanh chóng và áp dụng cho nhiều lĩnh vực với tần suất thành công cao bao gồm xử lý hình ảnh, nhận dạng giọng nói và xử lý văn bản. Các thử nghiệm cho thấy khả năng ứng dụng cao và hiệu suất rõ rệt so với các phương pháp học máy truyền thống. Các thuật toán học sâu chủ yếu là sự kế thừa của kiến trúc mạng nơ-ron nhân tạo với số lượng lớp ẩn (hidden layers) cao hơn, do đó được gọi là mạng nơ-ron sâu. Mạng thần kinh chuyển đổi (Convolutional Neural Network - CNN), là một trong những mô hình mạng thần kinh sâu phổ biến nhất. Bài báo cáo này trình bày về CNN với các layer cơ bản, đồng thời, một thử nghiệm cũng được trình bày để so sánh CNN với biểu diễn đặc trưng ảnh qua Scale-Invariant Feature Transform (SIFT) và Bag of Visual Words (BoVW).

1 Giới thiệu

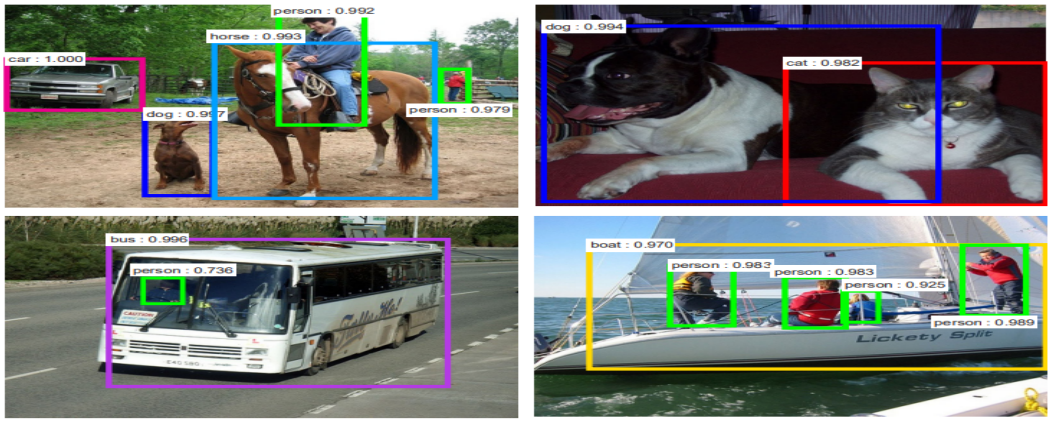
Convolutional Neural Network (CNN hoặc ConvNets) là một loại kiến trúc mạng nơ-ron được thiết kế đặc biệt để xử lý và phân loại dữ liệu có cấu trúc lưới, như hình ảnh và video.

Năm 1998, Yann LeCun đã đưa ra ví dụ đầu tiên về việc áp dụng phương pháp lan truyền ngược (backpropagation) và học tập dựa trên độ dốc (gradient-based learning) để huấn luyện các mạng thần kinh tích chập hoạt động rất tốt trong việc nhận dạng tài liệu [8]. Đặc biệt là họ đã làm tốt công việc nhận dạng các chữ số của mã zip. Do đó, CNN được sử dụng khá rộng rãi để nhận dạng mã zip trong dịch vụ bưu chính. Nhưng, nó vẫn chưa thể mở rộng quy mô sang những dữ liệu phức tạp và khó khăn hơn.

Những kết quả thành công đầu tiên bằng cách sử dụng mạng lưới thần kinh, điều thực sự khơi dậy cơn sốt sử dụng những loại mạng CNN này một cách thực sự rộng rãi là vào khoảng năm 2012, là bài báo mang tính bước ngoặc của Alex Krizhevsky trong phòng thí nghiệm của Geoff Hinton, giới thiệu kiến trúc mạng thần kinh tích chập đầu tiên có thể thực hiện được và thu được kết quả thực sự mạnh mẽ về phân loại ImageNet [1], mô hình này được gọi là AlexNet. Ngoài mô hình LeNet [8] và mô hình AlexNet [1], còn có các mô hình CNN khác như ZFNet[9], GoogleNet[7], VGGNet[6] và ResNet[2]. Những mô hình này phổ biến vì chúng cung cấp hiệu suất cao và đáng kể trên các nền tảng khác nhau tiêu chuẩn phân loại hình ảnh.

Ngày nay, mạng CNN được sử dụng ở khắp nơi, ví dụ như phân loại và truy xuất ảnh như hình 1, trích xuất và nhận dạng đối tượng như hình 2, ...

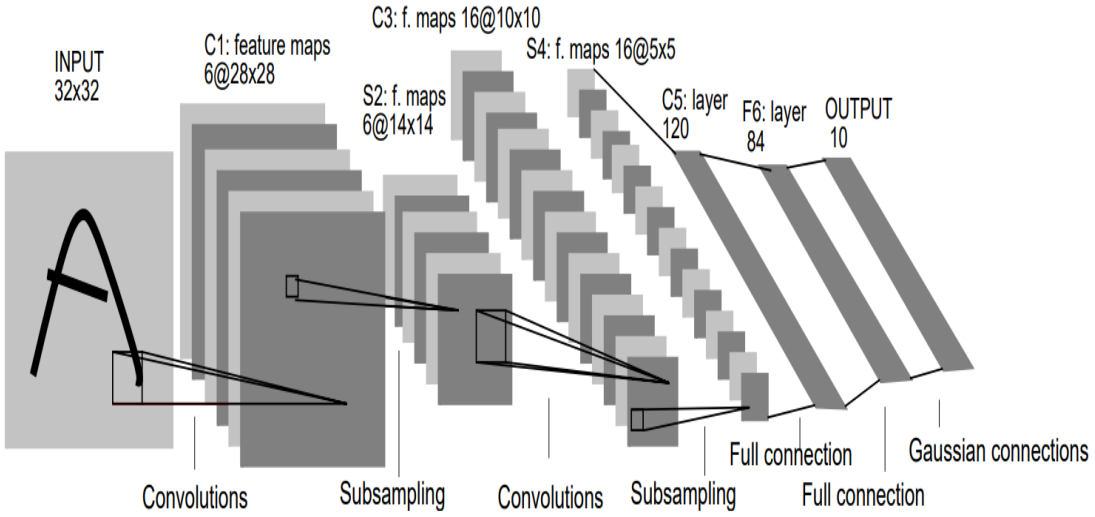
Hình 1: Trái: Tám hình ảnh thử nghiệm ILSVRC-2010 và năm nhãn được mô hình AlexNet coi là có khả năng xảy ra nhất. Nhãn chính xác được ghi dưới mỗi hình ảnh và xác suất được gán cho nhãn chính xác cũng được hiển thị bằng thanh màu đỏ (nếu nó nằm trong top 5); Phải: Năm hình ảnh thử nghiệm ILSVRC-2010 ở cột đầu tiên. Các cột còn lại hiển thị sáu ảnh huấn luyện tạo ra các vectơ đặc trưng ở lớp ẩn cuối cùng có khoảng cách euclide nhỏ nhất tới vectơ đặc trưng cho ảnh thử nghiệm (Hình ảnh tham khảo ở [1]).



Hình 2: Ví dụ về phát hiện đối tượng bằng cách sử dụng đề xuất RPN trong bài kiểm tra PASCAL VOC 2007 (Hình ảnh tham khảo ở [5]).

2 Kiến trúc

CNN sở hữu khả năng học các tính năng tự động từ dữ liệu đầu vào và nó loại bỏ việc trích xuất tính năng thủ công. CNN yêu cầu bộ dữ liệu được dán nhãn khổng lồ để đào tạo vì nó là phương pháp học có giám sát và về cơ bản nó được lấy cảm hứng từ vỏ não thị giác của động vật.



Hình 3: Kiến trúc của LeNet-5, để nhận dạng chữ số. Mỗi mặt phẳng (plane) là một bản đồ đặc trưng (feature map), tức là một tập hợp các đơn vị có trọng số bị ràng buộc giống hệt nhau (Hình ảnh tham khảo [8]).

Dựa vào hình 3, có thể thấy được cơ bản kiến trúc của CNN, lớp chập (convolutional layer) bao gồm tập hợp các bộ lọc (filter) đã được áp dụng trên hình ảnh đầu vào và sau khi dữ liệu được tạo được truyền qua đến lớp gộp (polling layer). Bản đồ đặc trưng (feature map) được tạo ở mỗi lớp chập, nó có được thông qua tính toán tích chập giữa các mảng cục bộ và vectơ trọng số thường được gọi là bộ lọc. Bản đồ đặc trưng là nhóm các tổng trọng số cục bộ. Để nâng cao hiệu quả đào tạo, các bộ lọc được được áp dụng nhiều lần. Quá trình này giúp giảm số lượng tham số trong quá trình học.

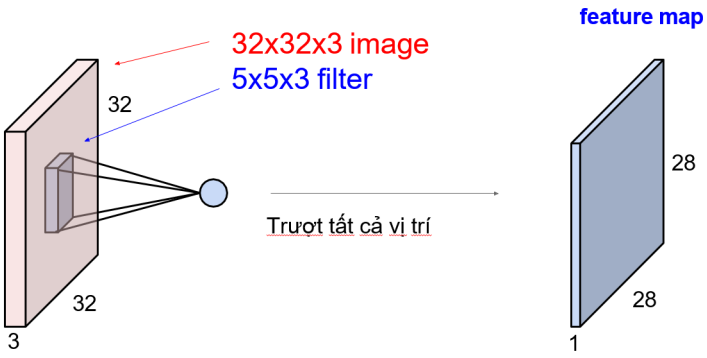
Việc lấy mẫu con (subsampling) tối đa hoặc trung bình của các vùng không chồng chéo trong bản đồ đặc trưng được thực hiện tại mỗi lớp gộp (polling layer). Cuối cùng, các lớp kết nối đầy đủ (fully connected layer) đang hoạt động như một mạng lưới thần kinh thông thường. Giai đoạn cuối cùng, CNN vẫn có hàm mất mát (ví dụ: SVM/Softmax) trên lớp kết nối đầy đủ.

Như vậy, kiển trúc CNN thông thường được xây dựng bởi các lớp bao gồm: lớp chập (convolutional layer), lớp gộp (pooling layer), lớp kích hoạt (activation layer), lớp kết nối đầy đủ (fully connected layer).

2.1 Lớp Chập (Convolutional Layer)

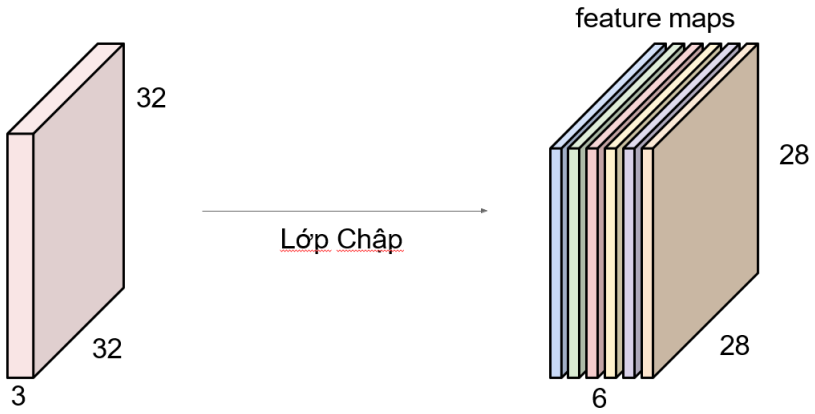
Lớp chập là cốt lõi của Mạng tích chập, thực hiện hầu hết các công việc tính toán nặng nhọc.

Các tham số của lớp chập bao gồm một tập hợp các bộ lọc có thể học được. Mọi bộ lọc đều có kích thước nhỏ về mặt không gian (dọc theo chiều rộng và chiều cao), nhưng kéo dài đến toàn bộ chiều sâu của đầu vào. Ví dụ: bộ lọc thông thường trên lớp đầu tiên của CNN có thể có kích thước [5x5x3] (tức là chiều rộng và chiều cao 5 pixel và 3 vì hình ảnh có độ sâu 3, tức là các kênh màu).



Hình 4: Ảnh minh hoạ ảnh đầu vào [32x32x3] và bộ lọc [5x5x3], thực hiện tích chập bộ lọc trên tất cả các vị trí của ảnh đầu vào, ta được bản đồ đặc trưng (feature map).

Trong quá trình chuyển tiếp, CNN thực hiện tích chập từng bộ lọc theo chiều rộng và chiều cao của đầu vào và tính toán tích số chấm (dot product) giữa các mục của bộ lọc và đầu vào ở bất kỳ vị trí nào (Như hình 4). Khi trượt bộ lọc theo chiều rộng và chiều cao của đầu vào, CNN sẽ tạo ra bản đồ đặc trưng (feature map) 2 chiều cung cấp phản hồi của bộ lọc đó ở mọi vị trí. Theo trực quan, CNN sẽ tìm hiểu các bộ lọc kích hoạt khi chúng nhìn thấy một số loại tính năng trực quan, chẳng hạn như cạnh của một số hướng hoặc một đốm màu nào đó trên lớp đầu tiên. Bây giờ, chúng ta sẽ có toàn bộ bộ lọc trong mỗi lớp CONV (ví dụ: 6 bộ lọc) và mỗi bộ lọc sẽ tạo ra một bản đồ đặc trưng 2 chiều riêng biệt. Chúng tôi sẽ xếp chồng các bản đồ đặc trưng này dọc theo chiều sâu và tạo ra khối lượng đầu ra (Như hình 5).

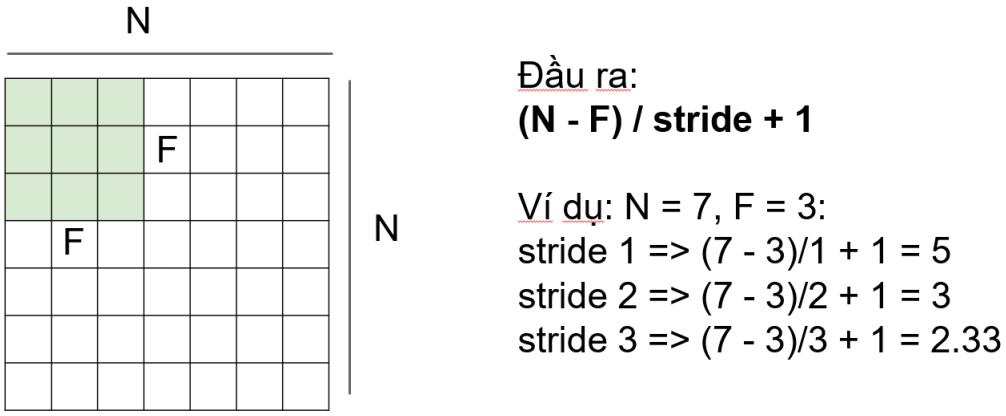


Hình 5: Ảnh minh hoạ ảnh đầu vào [32x32x3] với 6 bộ lọc, mỗi bộ lọc sẽ tạo ra bản đồ đặc trưng 2 chiều riêng biệt, sau đó xếp chồng các bản đồ đặc trưng, ta được đầu ra [28x28x6].

Bây giờ, chúng ta vẫn chưa rõ về việc có bao nhiêu nơ-ron trong khối lượng đầu ra hoặc cách chúng được sắp xếp, ta có ba siêu tham số kiểm soát kích thước của âm lượng đầu ra: độ sâu (depth), bước nhảy (stride) và khoảng đệm 0 (zero-padding).

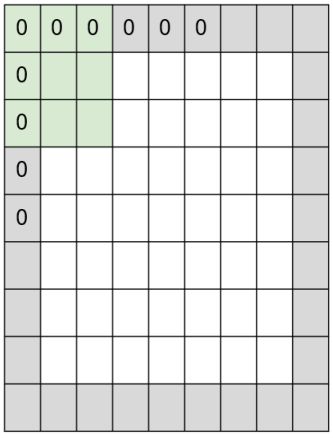
- Đầu tiên, độ sâu của đầu ra là một siêu tham số: nó tương ứng với số lượng bộ lọc mà chúng ta muốn sử dụng, mỗi bộ lọc sẽ tìm kiếm thứ gì đó khác nhau trong đầu vào. Ví dụ: nếu Lớp chập đầu tiên lấy hình ảnh thô làm đầu vào, thì các nơ-ron khác nhau dọc theo chiều sâu có thể kích hoạt khi có các cạnh định hướng khác nhau hoặc các đốm màu. Điều này cũng có thể giải thích theo hình 5, ta có thể thay đổi số lượng bộ lọc để thay đổi giá trị đầu ra.
- Thứ hai, chúng ta phải chỉ định bước nhảy (stride) mà chúng ta trượt bộ lọc. Khi bước nhảy là 1 thì bộ lọc di chuyển qua từng pixel một. Khi bước nhảy là 2 (hoặc hiếm gặp là 3 hoặc nhiều hơn) thì các bộ lọc sẽ nhảy 2 pixel mỗi trượt chúng. Điều này sẽ tạo ra khối lượng đầu ra nhỏ hơn về mặt không gian. Hình 6 minh hoạ giá trị đầu ra khi thay đổi bước nhảy (stride), trong đó N = 7,

F = 3 và với bước nhảy là 1, ta nhận được đầu ra là 5x5; bước nhảy là 2, ta nhận được đầu ra là 3x3; bước nhảy là 3, có lỗi không khớp xảy ra.



Hình 6: Ảnh minh hoạ giá trị đầu ra thay đổi khi thay đổi bước nhảy (stride).

- Đôi khi sẽ rất thuận tiện khi đệm đầu vào bằng các số 0 xung quanh đường viền. Kích thước của phần đệm số 0 (zero-padding) này là một siêu tham số. Tính năng hay của phần đệm bằng 0 là nó sẽ cho phép chúng ta kiểm soát kích thước không gian của âm lượng đầu ra. Theo ví dụ hình 7, ta có một đầu vào [7x7], với bước nhảy là 1; bộ lọc [3x3]; đệm 1 lớp 0 xung quanh;, ta được đầu ra sẽ là [7x7]. Thông thường, giá trị kích thước của phần đệm số 0 sẽ được tính dựa vào F (kích thước bộ lọc), dựa vào công thức $(F-1)/2$. Ví dụ, với F = 3 thì sẽ đệm $(F-1) = (3-1)/2 = 1$ lớp 0, với F = 3 thì sẽ đệm $(F-1) = (5-1)/2 = 2$ lớp 0.



Hình 7: Ảnh minh hoạ giá trị đầu ra thay đổi khi thay đổi bước nhảy (stride).

Như chúng ta có thể thấy trong CNN, việc xác định tham số cho mô hình CNN một cách thích hợp sao cho tất cả các kích thước đều "hoàn thiện" có thể thực sự là một vấn đề đau đầu.

Tổng kết lại, đầu ra sau khi thông qua lớp chập có thể tính theo công thức hình 8:

Bước 1: Nhận đầu vào W_1 x H_1 x D_1
Bước 2: Nhập 4 tham số:
- Số bộ lọc K,
- Phạm vi bộ lọc F,
- Bước nhảy S,
- Kích thước của bộ đệm 0.

Bước 3: Tính đầu ra W_2 x H_2 x D_2, với:
- $W_2 = (W_1 - F + 2P)/S + 1$,
- $H_2 = (H_1 - F + 2P)/S + 1$,
- $D_2 = K$

Hình 8: Ảnh tổng kết lại công thức tính đầu ra của lớp chập cùng với các tham số.

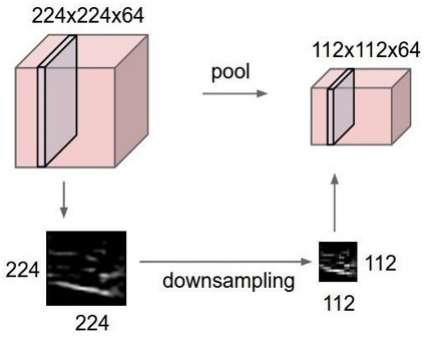
Ví dụ, có một ảnh đầu vào [32x32x3], 10 bộ lọc [5x5], bước tiến bằng 1, kích cớ đệm 0 bằng 2, như vậy, ta có: $W1 = 32; H1 = 32; D1 = 3; K = 10; F = 5; S = 1; P = 2$. Áp dụng công thức hình 8:

$$W2 = (W1 - F + 2P)/S + 1 = (32 - 5 + 2 * 2)/1 + 1 = 32$$
$$H2 = (H1 - F + 2P)/S + 1 = (32 - 5 + 2 * 2)/1 + 1 = 32$$
$$D2 = K = 10$$

Như vậy, đầu ra sẽ là [32x32x10].

Bên cạnh đó, một số bài báo sử dụng tích chập 1x1, như Network in Network [4] đã nghiên cứu lần đầu tiên. Một số người lúc đầu bối rối khi thấy các tích chập [1x1], thông thường các tín hiệu là 2 chiều nên các tích chập [1x1] không có ý nghĩa (nó chỉ là tỷ lệ theo điểm). Tuy nhiên, trong CNN, điều này không xảy ra vì cần phải nhớ rằng CNN hoạt động trên các khối 3 chiều và các bộ lọc luôn mở rộng đến toàn bộ độ sâu của khối đầu vào. Ví dụ: nếu đầu vào là [32x32x3] thì việc thực hiện tích chập [1x1] sẽ thực hiện các tích điểm 3 chiều một cách hiệu quả (vì độ sâu đầu vào là 3 kênh).

2.2 Lớp Gộp (Pooling Layer)



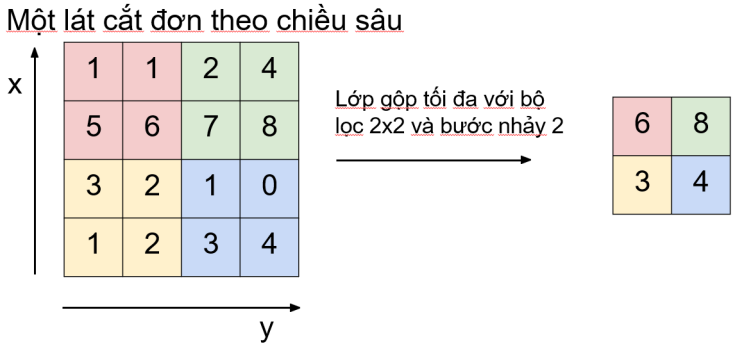
Hình 9: Ví dụ về lớp gộp (pooling layer).

Người ta thường luôn chèn một lớp gộp vào giữa các lớp chập liên tiếp trong CNN. Chức năng của nó là giảm dần kích thước không gian của biểu diễn để giảm số lượng tham số và tính toán trong mạng và do đó cũng kiểm soát việc overfitting.

Dựa vào hình 9, lớp gộp nhận thông tin đầu vào của ảnh [224x224x64] và tiến hành giảm kích thước xuống, cuối cùng nhận được [112x112x64] (độ sau không thay đổi vì chỉ gộp lại theo chiều rộng (width) và chiều cao (height)).

Phổ biến nhất là gộp tối đa (Max Pooling), theo ví dụ như hình 10. Các thí nghiệm thường làm cho lớp gộp không có bất kỳ sự trùng lặp nào vì về cơ bản là chỉ muốn lấy mẫu xuống, từ đó sẽ hợp lý hơn khi nhìn vào khu vực (region) này và chỉ lấy một cái giá trị để đại diện cho vùng này và sau đó chỉ cần nhìn vào vùng tiếp theo.

Ngoài sử dụng gộp tối đa (max pooling), một số khác cũng nảy lên ý tưởng gộp trung bình (average pooling), tuy nhiên, cách này không phổ biến và không hiệu quả. Trong bài toán trích xuất đặc trưng ảnh hay nhận diện đối tượng, dù là ánh sáng hay khía cạnh nào đó của hình ảnh mà bạn đang tìm kiếm, các thí nghiệm luôn muốn nhắm đến với giá trị cao nhằm tìm thấy giá trị đại diện cho vùng đó, từ đó, giữ lại đặc trưng quan trọng cho vùng đó.



Hình 10: Ví dụ về lớp gộp (pooling layer).

2.3 Lớp Kích Hoạt (Activation Layer)

Lớp kích hoạt trong CNN thường được sử dụng để thêm tính phi tuyến tính vào mô hình. Mục tiêu của lớp này là giúp mô hình học được các biểu diễn phức tạp hơn từ dữ liệu đầu vào.

Lớp kích hoạt được áp dụng cho mỗi giá trị đầu ra của một neuron trong lớp trước đó và thường được thiết kế để giữ lại các giá trị dương và loại bỏ các giá trị âm (hoặc ngược lại). Hàm kích thích phổ biến nhất là Rectified Linear Unit (ReLU).

Công thức của hàm ReLU được mô tả như sau:

$$f(x) = max(0, x)$$

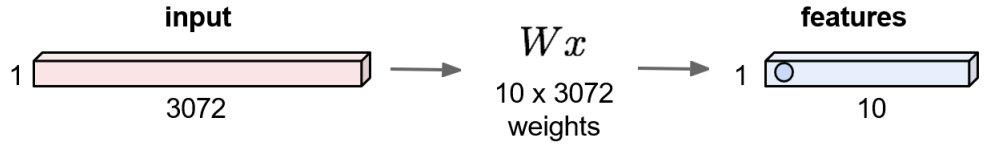
Trong đó, x là giá trị đầu vào; hàm max là hàm lấy giá trị lớn nhất giữa 0 và x.

ReLU giúp mô hình học được các đặc trưng phi tuyến tính, làm cho nó linh hoạt hơn trong việc biểu diễn các mối quan hệ phức tạp trong dữ liệu. Ngoài ra, ReLU còn giúp giảm vấn đề mất mát đạo hàm bằng cách tránh giá trị đạo hàm tiến gần 0 ở các giá trị dương. Vì ReLU làm cho một số đơn vị trong mạng không hoạt động (output là 0), nó có thể làm tăng tốc quá trình huấn luyện bằng cách giảm số lượng tham số cần được cập nhật.

2.4 Lớp Kết Nối Đầy Đủ (Fully Connected Layer)

Lớp kết nối đầy đủ (Fully-Connected Layer hay còn gọi là Dense Layer) đóng vai trò quan trọng trong việc kết hợp thông tin từ các đặc trưng đã được trích xuất từ các lớp trước đó để tạo ra dự đoán hoặc phân loại cuối cùng.

Trong lớp kết nối đầy đủ, mỗi nơ-ron kết nối với mọi nơ-ron trong lớp trước đó. Đầu vào của lớp này là một vector một chiều của các giá trị từ các đơn vị trước đó. Nếu lớp trước có n đơn vị, và lớp kết nối đầy đủ có m đơn vị, thì sẽ có n×m trọng số (weights) cần được học (như hình 11).



Hình 11: Hình mình hoạ đầu vào của lớp kết nối đầy đủ có 3072 đơn vị, lớp kết nối đầy đủ có 10 đơn vị thì sẽ có 10x3072 trọng số cần được hoc.

3 So sánh với SIFT và BoVW

3.1 Dữ liệu

Để thực hiện so sánh với phương pháp trích xuất đặc trưng ảnh SIFT và BoVW, bài báo cáo sử dụng tập dữ liệu CIFAR-10 [3]. CIFAR-10 là tập hợp ảnh được gắn nhãn của bộ dữ liệu 80 triệu hình ảnh nhỏ. Chúng được sưu tầm bởi Alex Krizhevsky, Vinod Nair và Geoffrey Hinton. Bộ dữ liệu CIFAR-10 bao gồm 60000 hình ảnh màu [32x32] trong 10 lớp, với 6000 hình ảnh mỗi lớp. Có 50000 hình ảnh đào tạo và 10000 hình ảnh thử nghiệm.

Bộ dữ liệu được chia thành năm tập huấn luyện và một tập thử nghiệm, mỗi tập có 10000 hình ảnh. Tập dữ liệu thử nghiệm (test datasets) chứa chính xác 1000 hình ảnh được chọn ngẫu nhiên từ mỗi lớp. Các tập huấn luyện chứa các hình ảnh còn lại theo thứ tự ngẫu nhiên, nhưng một số tập huấn luyện có thể chứa nhiều hình ảnh từ lớp này hơn lớp khác. Giữa chúng, các tập huấn luyện chứa chính xác 5000 hình ảnh từ mỗi lớp.

Tập dữ liệu gồm 10 lớp gồm có: máy bay (airplane); automobile(ô tô); chim (bird); mèo (cat); hươu (deer); chó (dog); ếch (frog); ngựa (horse); tàu (ship); xe tải (truck). Các lớp học hoàn toàn loại trừ lẫn nhau. Không có sự chồng chéo giữa ô tô và xe tải. "Ô tô" bao gồm xe sedan, SUV, những thứ tương tự. “Xe tải” chỉ bao gồm xe tải lớn, không bao gồm xe bán tải.

3.2 So sánh CNN với SIFT và BoVW

Bảng 1: Bảng số liệu độ chính xác của hai mô hình trích xuất đặc trưng ảnh CNN và SIFT kết hợp BoVW.

Nhãn	CNN (phần trăm)	SIFT + BoVW (k=300) (phần trăm)
plane	56.1	35.2
car	54.1	29.91
bird	44.5	25.5
cat	31.9	20.4
deer	30.5	26.5
dog	40.3	29.98
frog	57.8	45.9
horse	63.4	21.1
ship	57.3	33.1
truck	59.5	30.0
Tất cả mẫu	49.54	32.5

Giống nhau:

- Cả hai phương pháp CNN và SIFT kết hợp BoVW đều chú trọng vào việc trích xuất đặc trưng từ ảnh. CNN tự động hóa việc học các đặc trưng trong quá trình huấn luyện, trong khi SIFT kết hợp BoVW sử dụng kỹ thuật cục bộ để tìm các điểm keypoint và mô tả chúng.
- Cả hai phương pháp đều được sử dụng trong các ứng dụng nhận diện vật thể và phân loại hình ảnh.
- Cả hai phương pháp đều yêu cầu một lượng dữ liệu lớn để huấn luyện mô hình hiệu quả.
- Cả hai phương pháp đều có thể chịu ảnh hưởng bởi biến động và góc nhìn khác nhau trong ảnh.

Khác nhau:

- Về mặt mục tiêu, CNN thường được sử dụng chủ yếu trong các ứng dụng nhận dạng và phân loại hình ảnh, SIFT kết hợp BoVW thường được sử dụng chủ yếu trong các ứng dụng nhận dạng và phân loại hình ảnh.
- Về phương pháp, CNN tự động hóa việc học các đặc trưng từ dữ liệu thông qua quá trình huấn luyện end-to-end, SIFT kết hợp BoVW sử dụng SIFT để trích xuất các điểm keypoint và mô tả chúng, sau đó sử dụng BoVW để tạo biểu diễn histogram của các từ visual words.
- Về tài nguyên, CNN đòi hỏi tài nguyên tính toán lớn và cần có phần cứng mạnh mẽ, SIFT kết hợp BoVW phù hợp cho việc triển khai trên các thiết bị có tài nguyên hạn chế và yêu cầu tính toán thấp.
- CNN có khả năng học các đặc trưng toàn cục và phi tuyến tính, SIFT kết hợp BoVW thường tập trung vào các đặc trưng cục bộ và không tự nhiên.

Để so sánh ưu và nhược điểm của hai phương pháp trích xuất đặc trưng ảnh CNN và SIFT kết hợp với BoVW, bài báo cáo sử dụng bảng 1 có từ kết quả chạy mô hình trên tập dữ luệ CIFAR-10.

Ưu điểm:

- Đối với tập dữ liệu lớn như CIFAR-10, phương pháp CNN có lợi thế hơn, thể hiện sự phù hợp (hiệu suất cao) đối với dữ liệu lớn với khả năng học đặc trưng từ hàng triệu hoặc thậm chí hàng tỷ ảnh.
- CNN tự động hóa quá trình học đặc trưng từ dữ liệu trong quá trình huấn luyện. Mô hình tự động học được các đặc trưng cấp cao và phức tạp từ dữ liệu, giảm sự phụ thuộc vào sự can thiệp thủ công.
- Có khả năng học đặc trưng toàn cục và tìm hiểu các mối quan hệ phức tạp trong dữ liệu, điều mà SIFT kết hợp BoVW không thể tự động hóa một cách hiệu quả.
- CNN có khả năng tích hợp tốt trong các ứng dụng thực tế và hỗ trợ chức năng Transfer Learning, cho phép sử dụng các mô hình đã được đào tạo trước để giảm yêu cầu về dữ liệu huấn luyện.

- CNN có khả năng học đặc trưng phi tuyến tính, giúp mô hình thích ứng với các quan hệ phức tạp trong dữ liệu.

Nhược điểm:

- CNN yêu cầu một lượng lớn dữ liệu huấn luyện được gán nhãn để học các đặc trưng cụ thể cho mỗi lớp, trong khi SIFT kết hợp BoVW thích hợp cho các tập dữ liệu nhỏ và không đòi hỏi lượng lớn dữ liệu gán nhãn.
- CNN đòi hỏi tài nguyên tính toán lớn và cần có phần cứng mạnh mẽ. Điều này làm tăng khả năng tích hợp và triển khai chúng trên các thiết bị có tài nguyên hạn chế.
- Thường được xem xét là "hộp đen" khó giải thích cách một quyết định cụ thể được đưa ra, trong khi SIFT kết hợp BoVW có khả năng giải thích và hiểu được quyết định do sự can thiệp thủ công trong quá trình chọn keypoint và cụm từ visual words.
- Mặc dù hiệu suất cao, nhưng có thể không hiệu quả trong một số tình huống, đặc biệt là khi dữ liệu không đồng nhất.

4 Thảo luận và nhận xét

4.1 Xu hướng trong tương lai

Ngày nay, từ nền tảng CNN đem lại, thế giới ngày càng có xu hướng hướng tới các bộ lọc nhỏ hơn và kiến trúc sâu hơn. Các mô hình như EffcientNet, ResNet, ViT(Vision Transformer) đã chứng minh được sự hiệu quả và có thể cải thiện thêm.

Một số ý tưởng táo bạo nhằm giảm độ phức tạp của mô hình cũng đang được nghiên cứu như loại bỏ lớp gộp (Pooling layer) và lớp kết nối đầy đủ (Fully Connected Layer), chỉ sử dụng lớp chập để tối ưu hoá mô hình.

Cấu trúc đặc trưng của một mô hình CNN được miêu tả như

$$INPUT -> [(CONV -> RELU) * N -> POOL] * M -> (FC -> RELU) * K -> SOFTMAX$$

với N thông thường có giá trị tới bằng 5, M là lớn (large), 0 <= K <= 2. Nhưng, một số mô hình gần đây như ResNet/GoogleNet đang thách thức cấu trúc này. Ví dụ:

- $INPUT -> FC$, bổ sung một bộ phân loại tuyến tính. Với $N = M = K = 0$.
- $INPUT -> CONV -> RELU -> FC$
- $INPUT -> [CONV -> RELU -> POOL] * 2 -> FC -> RELU -> FC$. Ví dụ này có một lớp CONV duy nhất giữa mỗi lớp POOL.
- $INPUT -> [CONV -> RELU -> CONV -> RELU -> POOL] * 3 -> [FC -> RELU] * 2 -> FC$. Ví dụ này có hai lớp CONV được xếp chồng lên nhau trước mỗi lớp POOL. Nhìn chung, đây là một ý tưởng hay cho các mạng CNN lớn hơn và sâu hơn, vì nhiều lớp CONV xếp chồng lên nhau có thể phát triển các tính năng phức tạp hơn của đầu vào trước lớp gộp làm nhỏ đi.

4.2 Hạn chế

Convolutional Neural Networks (CNN) đã đạt được nhiều thành công trong nhiều ứng dụng thị giác máy tính, nhưng cũng tồn tại một số hạn chế như:

- CNN yêu cầu một lượng lớn dữ liệu huấn luyện để học các đặc trưng phức tạp. Điều này có thể là một thách thức trong các tình huống nơi lượng dữ liệu hạn chế.
- Với lượng dữ liệu lớn, có nguy cơ mô hình bị overfitting, đặc biệt là khi kiến trúc mô hình quá phức tạp so với độ phức tạp của tập dữ liệu.

- Các CNN phức tạp thường yêu cầu tính toán cao, đặc biệt là trong quá trình huấn luyện. Điều này có thể làm tăng yêu cầu về phần cứng, đặc biệt là khi triển khai trên thiết bị có tài nguyên hạn chế.
- CNN có thể gặp khó khăn khi xử lý các dữ liệu không đồng nhất hoặc có biến động lớn trong chất lượng ảnh.

5 Kết luận

Trong bài báo cáo này, chúng ta đã tim hiểu CNN với các lớp (layer) cơ bảm, đồng thời, xem xét và so sánh giữa Convolutional Neural Network (CNN) và phương pháp truyền thống SIFT kết hợp với Bag of Visual Words (BoVW) trong trích xuất đặc trưng ảnh, đặc biệt là trên tập dữ liệu CIFAR-10. CNN thường là sự lựa chọn ưu tiên đối với các nhiệm vụ phức tạp và có lượng dữ liệu lớn. Nó tự động hóa quá trình học đặc trưng và có khả năng tìm hiểu các mối quan hệ phức tạp trong dữ liệu. Ngược lại, SIFT kết hợp BoVW thường phù hợp với các tập dữ liệu nhỏ và đòi hỏi tính toán thấp hơn.

Với sự phát triển của công nghệ, CNN ngày càng trở thành lựa chọn phổ biến trong nhiều ứng dụng thị giác máy tính, từ nhận dạng vật thể cho đến phân loại hình ảnh. Tuy nhiên, việc lựa chọn giữa các phương pháp vẫn phụ thuộc vào bối cảnh cụ thể và yêu cầu của dự án.

Tài liệu

[1] Geoffrey E. Hinton Alex Krizhevsky, Ilya Sutskever. Imagenet classification with deep convolutional neural networks. 2012.

[2] Kaiming He. Deep residual learning for image recognition. 2016.

[3] Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009.

[4] Shuicheng Yan Min Lin, Qiang Chen. Network in network. 2014.

[5] Ross Girshick Shaoqing Ren, Kaiming He and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. 2016.

[6] Karen Simonyan and Andrew Zisserman. Deep convolutional networks for large-scale image recognition. 2014.

[7] Christian Szegedy. Going deeper with convolutions. 2015.

[8] Yoshua Bengio Yann LeCun, Leon Bottou and Patrick Haffner. Gradientbased learning applied to document recognition. 1998.

[9] M. D. Zeiler and Fergus. Visualizing and understanding convolutional networks. 2013.