Report paper: ImageNet Classification with Deep Convolutional Neural Networks

Giáp Ngọc Hiệu

**Tóm tắt nội dung**

Bài report này là phần tóm tắt lại ý tưởng, nội dung chính, những điều làm được của paper ImageNet Classification with Deep Convolutional Neural Networks

1. **Ý tưởng**

Trước đây, các phương pháp dùng để nhận dạng đối tượng chủ yếu sử dụng học máy. Nhưng cách làm chỉ hiệu quả đối với các tập dữ liệu nhỏ và đơn giản. Trong thực tế các đối tượng nhận dạng đa dạng và phức tạp hơn nhiều. Ta cần sử dụng các tập dữ liệu lớn hơn, đa dạng hơn để model có thể nắm bắt được các thông tin của đối tượng. Theo tác giả ta cần một model có khả năng học tập, tổng quát hóa trên các tập dữ liệu lớn này. Tuy nhiên với sự phức tạp cực kì lớn của nhiệm vụ nhận dạng đối tượng và tập dữ liệu lớn, mô hình cần có nhiều kiến thức trước đó bù đắp cho dữ liệu và kiến thức mà mô hình chưa được thấy và học. Để thỏa mãn các điều trên, tác giả đề xuất việc sử dụng mạng thần kinh tích chập (Convolutional neural networks hay CNNs). Khả năng của CNN có thể kiểm soát được bằng cách thay đổi độ sâu và độ rộng của mạng. Và theo tác giả, CNN cũng đưa ra các giả định mạnh mẽ và gần như đúng về bản chất của hình ảnh (stationarity of statistics và locality of pixel dependencies)

\* stationarity of statistics (tính ổn định của số liệu thống kê): trong ngữ cảnh của CNN, điều này đề cập đến việc model cho rằng các đối tượng và các đặc trưng quan trọng của hình ảnh có thể xuất hiện ở nhiều vị trí khác nhau trong không gian, bất kể vị trí, ánh sáng, nền nhưng vẫn giữ nguyên tính chất của chúng.

Ví dụ: CNN có thể học được các đặc trưng của khuôn mặt như mắt, mũi, miệng và đưa ra dự đoán chính xác dù cho khuôn mặt có thể xuất hiện ở bất kì đâu trong ảnh, hoặc bị ảnh hưởng của ánh sáng, nền.

\* locality of pixel dependencies (vị trí phụ thuộc của pixel): các pixel lân cận có xu hướng tương quan với nhau, trong khi các pixel ở xa thường không tương quan. Điều này cho phép model nắm bắt các đặc trưng cục bộ và quan trọng của hình ảnh, giúp giảm lượng thông tin cần học cũng như cải thiện độ chính xác.

Ví dụ: Thay vì quan sát toàn bộ hình ảnh, model chỉ cần tập trung vào những vùng chứa các đặc điểm cụ thể như mắt, mũi, miệng là các đặc điểm tạo nên khuôn mặt.

CNN có thể giữ được các thông tin không gian quan trọng của ảnh, điều mà standard feedforward neural networks không làm được.

Theo tác giả, so với các standard feedforward neural networks với các layer có kích thước tương tự, CNN có ít tham số hơn nhiều nên dễ huấn luyện hơn trong khi vẫn đảm bảo được độ chính xác.

1. **Phương pháp**
   1. **The Dataset**

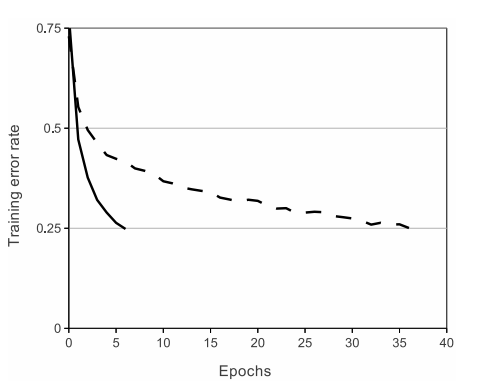
ImageNet là một bộ dữ liệu gồm hơn 15 triệu hình ảnh có độ phân giải cao được dán nhãn thuộc khoảng 22.000 danh mục. Các hình ảnh được thu thập từ nhiều trang web và được dán nhãn bởi những người từ nguồn cung ứng đám đông Mechanical Turk của Amazon. ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) là một cuộc thi thường niên sử dụng một tập hợp con của ImageNet với khoảng 1000 hình ảnh trong mỗi 1000 danh mục. Tổng cộng, có khoảng 1,2 triệu hình ảnh đào tạo, 50.000 hình ảnh xác thực và 150.000 hình ảnh thử nghiệm

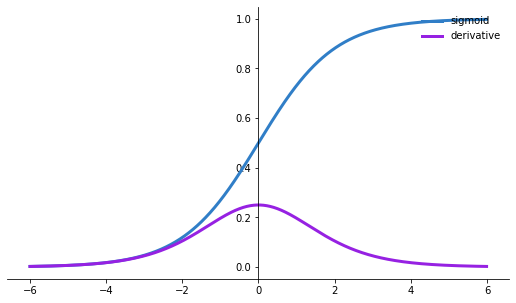
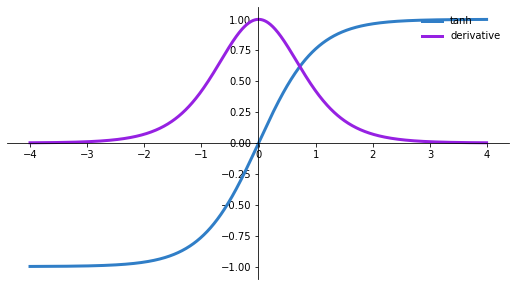
Tác giả đánh giá model chủ yếu dựa trên tập dữ liệu này. Vì tập dữ liệu chứa các ảnh có độ phân giải cao mà model của tác giả yêu cầu đầu vào với kích thước không đổi nên tác giả đã biến đổi ảnh về kích thước 256x256, thực hiện tiền xử lí bằng cách trừ đi giá trị trung bình R,G,B từ tập dữ liệu cho mỗi pixel.

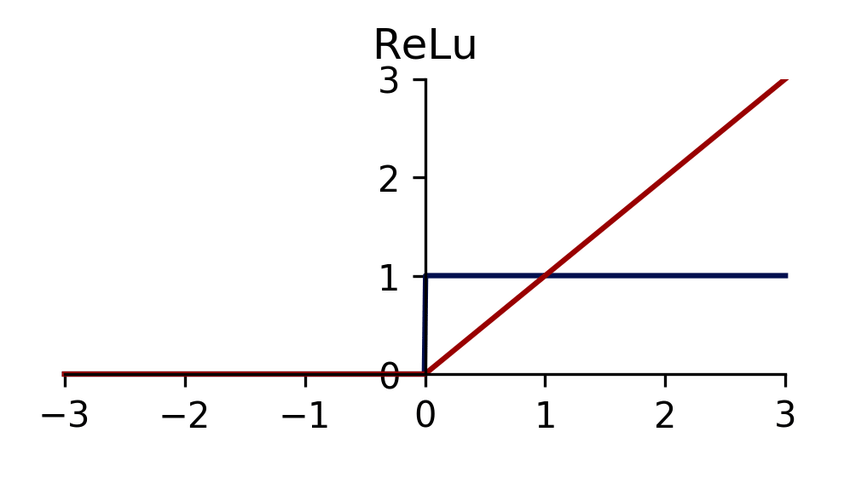
* 1. **ReLU Nonlinearity (Rectified Linear Units)**

Tác giả đã sử dụng activation ReLU (non-saturating nonlinearity hay phi tuyến không bão hòa) để thay thế cho tanh và sigmoid (saturating nonlinearity hay phi tuyến tính bão hòa) vốn được sử dụng nhiều cho các mạng thần kinh trước đó. ReLU đã trở thành một activation quan trọng và được sử dụng nhiều cho các model sau này. Theo tác giả, ReLU giúp tăng tốc độ hộ tụ của model khoảng 6 lần.

Công thức của ReLU: 





Lí giải về điều này, ta thấy hàm sigmoid và hàm tanh (a và b) bị bão hòa ở hai đầu, do đó, tốc độ hội tụ của mạng sẽ chậm khi trọng số của mạng có giá trị tuyệt đối lớn, làm cho đạo hàm trở nên vô cùng nhỏ dẫn đến hiện tượng vanishing gradient. Do đó, model hầu như không được cập nhật cũng như không học được gì.

Trong khi đó ReLU lại không bị bão hòa ở hai đầu như tanh và sigmoid, do đó có thể giải quyết được vấn đề trên.

Ưu điểm của ReLU:

+ Tốc độ hội tụ nhanh hơn sigmoid và tanh

+ Tính toán nhanh hơn: sigmoid và tanh sử dụng hàm exp và công thức phức tạp hơn ReLU nhiều

Nhược điểm:

+ Với các giá trị nhỏ hơn 0, qua ReLU sẽ thành 0, hiện tượng này gọi là “Dying ReLU”

+ Không khả vi tại 0

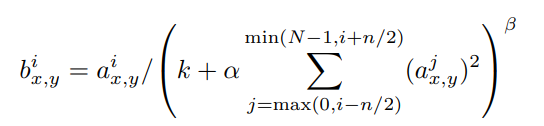
* 1. **Training on Multiple GPUs**

Vì phải xử lí tập dữ liệu lớn và độ phức tạp của mạng, tác giả thực hiện đào tạo trên 2 GPU. Các GPU đặc biệt phù hợp với quá trình song song hóa giữa các GPU, vì chúng có thể đọc và ghi trực tiếp vào bộ nhớ của nhau mà không cần thông qua bộ nhớ máy chủ. Sơ đồ song song hóa mà tác giả sử dụng về cơ bản đặt một nửa số kernel trên mỗi GPU, với điều kiện GPU chỉ giao tiếp trong một số lớp nhất định. Ví dụ, các kernel của layer 3 lấy đầu vào từ tất cả kernel map trong layer 2. Tuy nhiên, các kernel trong layer 4 chỉ nhận đầu vào từ các kernel map đó trong lớp 3 nằm trên

cùng một GPU.

* 1. **Local Response Normalization**

Mặc dù ReLU không yêu cầu chuẩn hóa đầu vào để ngăn việc bão hòa, nhưng tác giả thấy rằng việc sử dụng Local Response Normalization hỗ trợ cho quá trình tổng quát hóa của model. Kiểu chuẩn hóa này thực hiện một dạng ức chế bên (lateral inhibition) lấy cảm hứng từ các nơ-ron thực, tạo ra sự cạnh tranh cho các hoạt động lớn giữa các đầu ra neuron được tính toán bằng các kernel khác nhau. Nếu đặc trưng của một neuron nhất định rất lớn so với các đặc trưng khác trong vùng xung quanh, nó sẽ chiếm ưu thế. Điều này có nghĩa là thông tin quan trọng liên quan đến đặc trưng mạnh mẽ nhất trong hình ảnh sẽ được ưu tiên. Các đặc trưng mạnh hơn sẽ cạnh tranh và ức chế hoạt động của các đặc trưng yếu hơn, điều này giúp tăng cường tính đa dạng và tập trung vào các đặc trưng quan trọng trong ảnh.



 được tính toán bằng cách áp dụng kernel i tại vị trí (x, y) và sau đó áp dụng ReLU

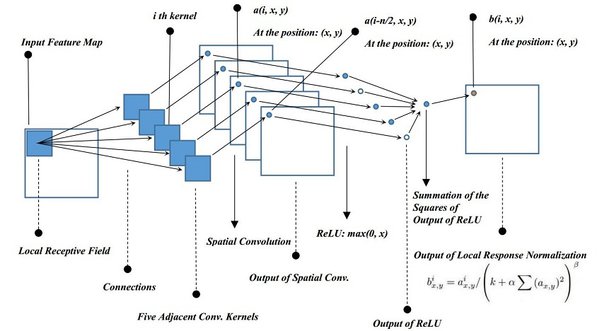
 là đầu ra sau khi chuẩn hóa

 là số kernel map liền kề ở cùng một vị trí không gian

là tổng số kernel trong layer

là các siêu tham số, trong paper tác giả thiết lập 

Tác giả thực hiện chuẩn hóa sau khi áp dụng ReLU cho một số layer nhất định.



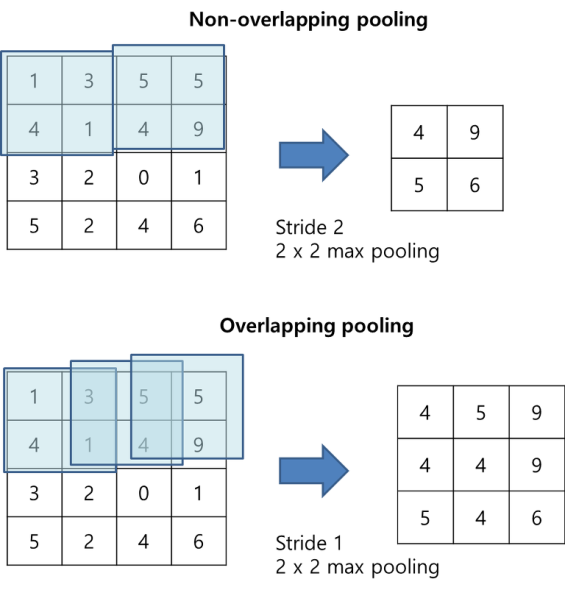
Hình. Minh họa cho việc sử dụng Local Response Normalization

* 1. **Overlapping Pooling**

Pooling layer trong CNN là layer tóm tắt kết quả đầu ra của một vùng lân cận trong cùng một kernel map. Hay đơn giản là là lấy mẫu xuống (down sampling) của một feature map hoặc một hình ảnh. Vai trò của Pooling layer là giảm độ phân giải của feature map nhưng vẫn giữ lại các đặc trưng cần thiết, trở nên ít nhạy cảm hơn với sự thay đổi vị trí của các đặc trưng trong hình ảnh và giảm đáng kể chi phí tính toán.

Có hai loại Pooling layer phổ biến là Max Pooling và Average Pooling.

Thông thường, các Pooling layer được sử dụng với các cửa sổ không chồng chéo nhau, nhưng trong paper, tác giả đã sử các Pooling layer có các cửa sổ chồng chéo nhau. Cách làm này giúp model giữ lại nhiều thông tin quan trọng hơn nhưng ngược lại làm tăng chi phí tính toán. Với việc sử dụng Overlapping Pooling, tác giả dã cải thiện được độ chính xác của model trên tập dữ liệu ImageNet.

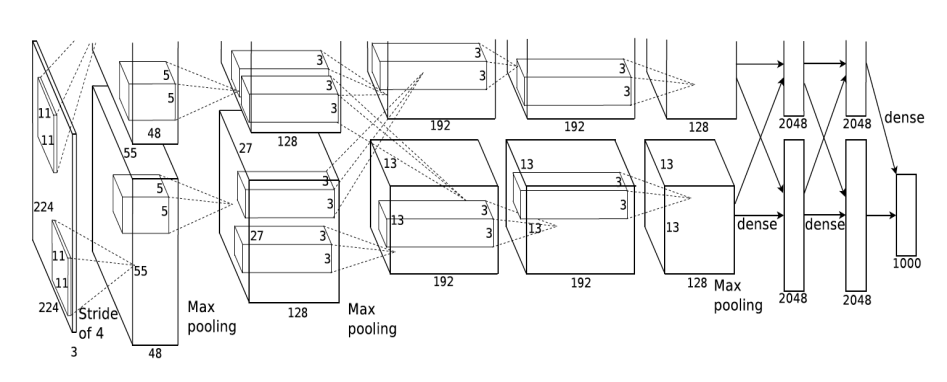


Công thức tính kích thước đầu ra của Pooling layer:



Theo công thức trên, ta có thể tùy ý điều chỉnh  để sử dụng overlapping pooling () hoặc non-overlapping pooling ()

* 1. **Overall Architecture**



Theo tác giả, mạng có tổng cộng 8 layer, 5 layer đầu là các convolution và 3 layer còn lại là các fully-connected. Đầu ra của lớp được kết nối đầy đủ cuối cùng được đưa vào softmax 1000 chiều để dự đoán trên 1000 class. Các kernel của các layer 2,4,5 chỉ được kết nối với các kernel map trong layer trước đó nằm trên cùng một GPU. Các kernel của layer 3 được kết nối với tất cả các kernel map trong layer 2. Các neurons trong các lớp fully-connected được kết nối với tất cả các neurons trong layer trước đó. Các layer Local Response Normalization theo sau các layer 1 và 2. Các layer Max Pooling thuộc dạng Overlapping Pooling, theo sau 2 layer Local Response Normalization cũng như layer 5. ReLU được áp dụng cho đầu ra của mọi convolutional và fully-connected layer.

Tổng cộng kiến trúc mạng của tác giả có 60 triệu tham số.

* 1. **Data Augmentation**

Để ngăn chặn việc overfitting, tác giả đã sử dụng phương pháp tăng cường dữ liệu. Phương pháp này tăng đáng kể sự đa dạng của dữ liệu có sẵn để đào tạo model mà không cần thu thập dữ liệu mới. Tăng cường dữ liệu gồm nhiều kĩ thuật như xoay ảnh, tăng độ sáng, làm mờ, …

Trong paper, tác giả sử dụng 2 kĩ thuật chính là:

+ Tạo ra các bản dịch hình ảnh và phản chiếu theo chiều ngang bằng cách trích xuất các bản vá (patch) ngẫu nhiên có kích thước 224 \* 224 từ ảnh có kích thước 256 \* 256 và đào tạo mạng trên các bản vá này. Đây là lí do tại sao input model của tác giả là 224x224x3.

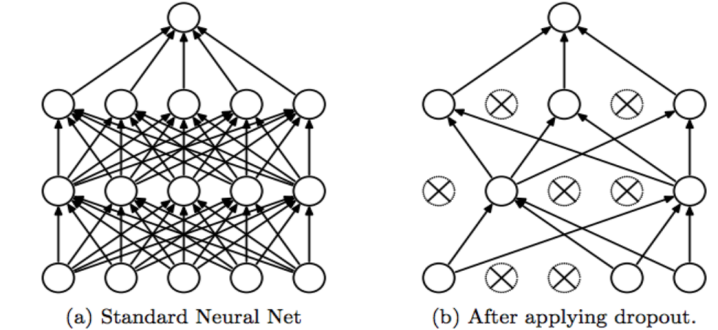
+ Thay đổi cường độ của các kênh RGB trong hình ảnh đào tạo. Tác giả thực hiện PCA trên tập hợp các giá trị pixel RGB trong toàn bộ tập dữ liệu dào tạo. Đối với mỗi hình ảnh đào tạo, thêm bội số của các thành phần chính được tìm thấy với độ lớn tỷ lệ với các giá trị riêng tương ứng nhân với một biến ngẫu nhiên được rút ra từ Gaussian với giá trị trung bình bằng 0 và độ lệch chuẩn 0,1. Do đó, với mỗi pixel ảnh RGB , thêm:



 và  lần lượt là vectơ riêng và giá trị riêng của ma trận hiệp phương sai 3 × 3 của các giá trị pixel RGB và  là biến ngẫu nhiên. Mỗi  chỉ được dùng một lần, sau đó sẽ được thay đổi. Điều này nắm bắt gần đúng một thuộc tính quan trọng của hình ảnh tự nhiên, đó là nhận dạng đối tượng là bất biến đối với những thay đổi về cường độ và màu sắc của ánh sáng.

* 1. **Dropout**

Để tăng khả năng tránh overfitting của model, tác giả sử dụng thêm một kĩ thuật chính quy hóa nữa là Dropout (bỏ học) để giảm overfitting và cải thiện việc tổng quát hóa của model.



Dropout thực hiện bằng cách loại bỏ các nút (lớp đầu vào và lớp ẩn) trong mạng. Tất cả các kết nối xuôi và ngược với một node bị loại bỏ tạm thời, do đó tạo ra một kiến ​​trúc mạng mới ngoài mạng chính. Việc làm này buộc mạng phải học các tính năng mạnh mẽ hơn, hữu ích hơn vì một số node đã bị loại bỏ. Các nút bị loại bỏ bởi xác suất bỏ học gọi là p. Trong paper, tác giả thiết lập p=0.5 và khi dự đoán, tác giả nhân trọng số với 0.5, đây là một phép tính gần đúng hợp lý để lấy giá trị trung bình hình học của các phân phối dự đoán được tạo ra bởi các mạng bỏ học theo cấp số nhân.

* 1. **Details of learning**

Tác giả đào tạo model sử dụng SGD với batch size = 128, momentum=0.9, weight decay=0.0005. Khởi tạo các trọng số trong mỗi lớp từ phân phối Gaussian có mean=0 và std=0.01. Khởi tạo bias các convolutional layer và fully-connected với hằng số 1, trong các lớp còn lại với hằng số 0.

1. **Metric**

Top-1 error rate: tỷ lệ phần trăm của số lượng ảnh trong tập kiểm tra bị phân loại sai vào class có xác suất dự đoán cao nhất.

Top-5 error rate : tỷ lệ phần trăm của số lượng ảnh trong tập kiểm tra bị phân loại sai nhưng nhãn đúng của chúng xuất hiện trong top 5 kết quả phân loại cao nhất của model.

1. **Note**

Ưu điểm của Alexnet:

+ Cải thiện độ chính xác một cách bất ngờ, vượt xa các phương pháp học máy truyền thống trước đó

+ Có thể đào tạo và xử lí trên tập dữ liệu có quy mô lớn

+ Có thể được sử dụng cho transfer learning

Nhược điểm:

+ Alexnet có độ phức tạp lớn, cần nhiều tài nguyên để tính toán

+ Có thể overfitting với các tập dữ liệu có quy mô nhỏ

+ Được thiết kế cho nhiệm vụ phân loại hình ảnh, nên có thể không phù hợp với các nhiệm vụ khác