TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI
VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

———————— * ————————

ĐỒ ÁN
TỐT NGHIỆP ĐẠI HỌC

NGÀNH CÔNG NGHỆ THÔNG TIN

Ứng dụng học sâu trong việc tự động xác định địa điểm du lịch nổi tiếng

Sinh viên thực hiện : **Lê Văn Mạnh**
Lớp KSVB2 – K37
Giáo viên hướng dẫn: **GV.Đinh Viết Sang**

HÀ NỘI 6-2019

Mục lục

CHƯƠNG 1 - MỞ ĐẦU 4
1.1. Tính cấp thiết của đồ án 4
1.2. Nhiệm vụ của đồ án 4
1.3. Đối tượng và phạm vi nghiên cứu 4
1.4. Phương pháp thực hiện 4
1.5. Ý nghĩa khoa học và thực tiễn 4
1.6. Kết quả dự kiến 5
1.7. Bố cục của đồ án 5
CHƯƠNG 2 - CONVOLUTIONAL NEURAL NETWORK 6
2.1. Giới thiệu về CNN 6
2.2. Cấu trúc tổng quan của mạng CNN 6
2.3. CNN hoạt động như thế nào 6
2.4. Convolution là gì 7
2.5. Stride và Padding 9
2.6. Pooling là gì 9
2.7. Fully connected neural network 10
CHƯƠNG 3 – BÀI TOÁN NHẬN DẠNG ĐỐI TƯỢNG 10
3.1. Đầu vào và đầu ra của bài toán 10
3.2. Một số kiến trúc mạng convolutional neural network 10
CHƯƠNG 4 – THIẾT KẾ HỆ THỐNG 15
4.1. Biểu đồ ca sử dụng 15
4.2. Biểu đồ hoạt động 15
4.3. Biểu đồ tuần tự 15
CHƯƠNG 5 – ĐÁNH GIÁ KẾT QUẢ 16
5.1. Giao diện trương trình 16
5.2. Minh họa chức năng phát hiện vi phạm luật giao thông 16
5.3. Độ chính xác của hệ thống 16
5.4. Hướng phát triển trong tương lai 16

CHƯƠNG 6 - TÀI LIỆU THAM KHẢO 17

CHƯƠNG 1 - MỞ ĐẦU

1.1. Tính cấp thiết của đồ án

Hiện nay để biết về một địa danh hay một điểm du lịch thông thường người dùng sẽ lên các trang tìm kiếm ví dụ google.com, bing.com …sau đó gõ từ khóa tên hoặc địa điểm du lịch muốn tới và sau đó đọc các thông tin liên quan tới địa điểm du lịch. Tuy nhiên, với sự bùng nổ của các mạng xã hội, các ứng dụng di động cùng với sự đa dạng về loại dữ liệu đặc biệt là dữ liệu ảnh nhiều trường hợp người dùng chỉ có một bức ảnh về địa điểm du lịch hoặc muốn tới một nơi có phong cảnh đẹp như trong bức ảnh mình đang có. Điều trên dẫn tới nhu cầu tìm kiếm thông tin địa danh thông qua hình ảnh ngày càn phổ biến.

1.2. Nhiệm vụ của đồ án

Đồ án được thực hiện nhằm mục đích làm cho việc tìm kiếm địa điểm du lịch và danh lam thắng cảnh của Việt Nam trở lên dễ dàng và thuận tiện. Trong phạm vi đồ án này em sẽ xây dụng một hệ thống cho phép người dùng tìm kiếm địa điểm du lịch bằng hình ảnh. Giúp đưa ra thông tin về địa điểm du lịch cũng như các địa danh khác có phong cảnh tương tự.

1.3. Đối tượng và phạm vi nghiên cứu

Đồ án thực hiện trên dữ liệu ảnh về các địa điểm du lịch nổi tiếng, mỗi địa danh sẽ chứa khoảng 1000 ảnh kèm với thông tin về địa lý liên quan tới địa danh đó. Do giới hạn về thời gian và nền tảng phần cứng nên hệ thống xây dựng để nhận diện và gợi ý 64 địa điểm du lịch khác nhau trên lãnh thổ Việt Nam.

1.4. Phương pháp thực hiện

Với bài toán nhân diện thông tin qua ảnh việc lập trình truyền thống sẽ khó có độ chính xác cao do độ phức tạp và đặc thù của thông tin dữ liệu. Qua một thời gian tìm hiểu công nghệ, em lựa chọn phương pháp xây dựng hệ thống với phần lõi nhận diện ảnh sẽ sử dụng trí tuệ nhân tạo để đạt độ chính xác cao nhất có thể.

1.5. Ý nghĩa khoa học và thực tiễn

Người dùng khi xem ảnh có thể ngay lập tức tìm kiếm thông tin địa danh thông qua hình ảnh mình đang xem. Mọi thứ sẽ trờ lên nhanh chóng và thuận tiện cho người sử dụng góp phần thúc đẩy ngành dịch vụ và du lịch của Việt Nam.

1.6. Kết quả dự kiến

Xây dựng mạng neuron nhân tạo nhận diện 64 địa danh thông qua hình ảnh với độ chính xác trên 80%, triển khai hệ thống trên nền tăng web cho người dùng truy cập và upload ảnh lên sau đó trả lại kết quả cho người dùng trên giao diện website.

1.7. Bố cục của đồ án

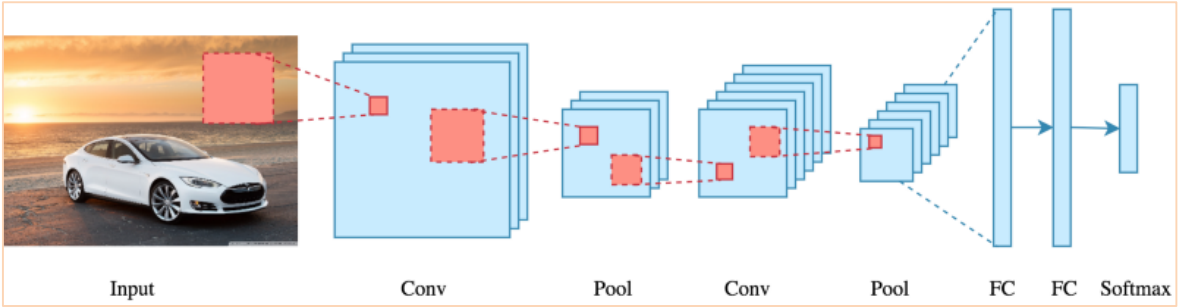
Đồ án được trình bày gồm các phần như sau:

CHƯƠNG 2 - CONVOLUTIONAL NEURAL NETWORK

2.1. Giới thiệu về CNN

Convolutional Neural Network (CNN – Mạng nơ-ron tích chập) là một trong những mô hình Deep Learning tiên tiến giúp cho chúng ta xây dựng được những hệ thống thông minh với độ chính xác cao như hiện nay như hệ thống xử lý ảnh lớn như Facebook, Google hay Amazon đã đưa vào sản phẩm của mình những chức năng thông minh như nhận diện khuôn mặt người dùng, phát triển xe hơi tự lái hay drone giao hàng tự động. CNN được sử dụng nhiều trong các bài toán nhận dạng các object trong ảnh.

2.2. Cấu trúc tổng quan của mạng CNN



Hình 2.2.1 Sở đồ tổng quát CNN

2.3. CNN hoạt động như thế nào

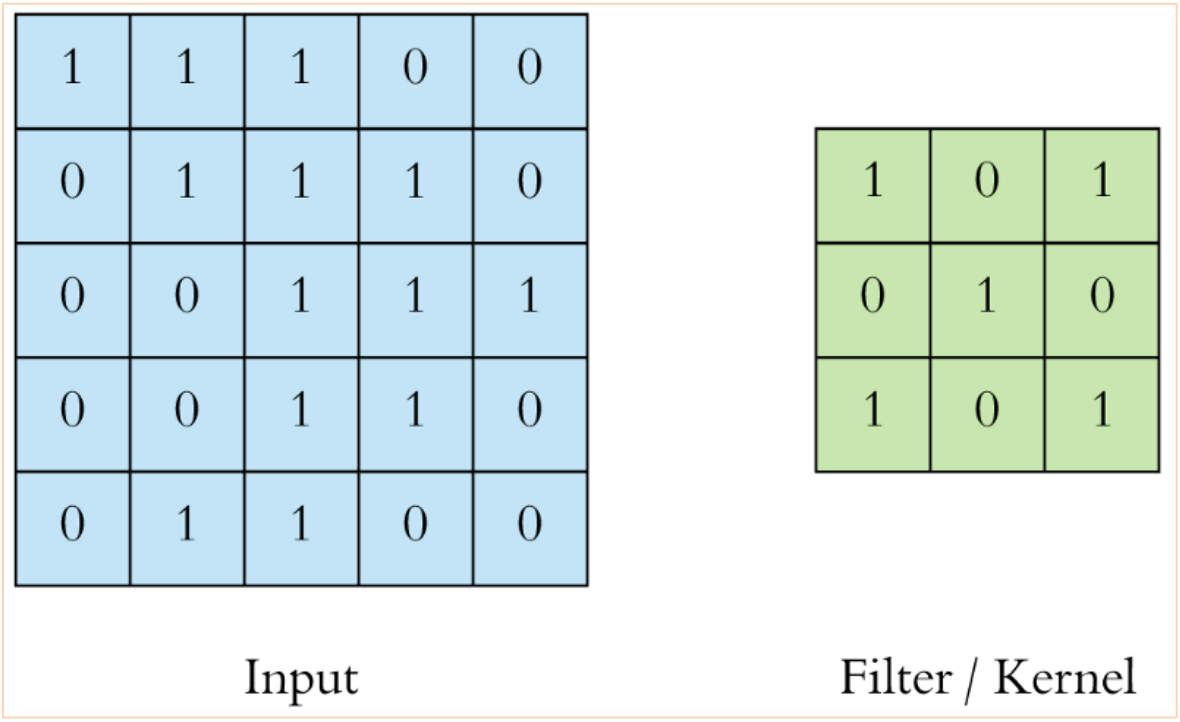
Local receptive fields: Trong mạng neural network truyền thống mỗi một neural trong input layer kết nối với một neural trọng hidden layer. Tuy nhiên trong CNN chỉ một vùng xác định trọng các neural trong input layer kết nối với một neural trong hidden layer. Những vùng xác định nêu trên gọi là Local receptive fields. Sự kết nỗi giữa input layer và hidden được chính là viêc từ Local receptive fields trên một ảnh đầu vào được biến đổi thông qua một phép toán được gọi là convolution để thu được một điểm trên hidden layer.

Shared weights và biases: Giống với mạng neural network truyền thống CNN cũng có tham số weights và biases. Các tham số này được học trong suốt quá trình trainning và liên tục cập nhật giá trị với mỗi mẫu mới (new trainning example). Tuy nhiên, các trọng số trong CNN là giống nhau đối với mọi neural trọng cùng một lớp (layer). điều này có nghĩ là tất cả các hidden neural trong cùng một lớp đang cùng tìm kiếm trung một đặc trưng (ví dụ như cạnh của ảnh) trong các vùng khác nhau của ảnh đầu vào.

Activation và pooling: Activation là một bước biến đổi giá trị đầu ra của mỗi neural thông qua việc sử dụng một số hàn ví dụ hàm ReLU. Giá trị thu được sau phép biến đổi là giá trị dương nhất có thể của output, trong trường hợp ouput mang giá trị âm thì giá trị nhận được là 0.

pooling là một bước nhằm giảm số chiều của ma trận, các thức phổ biến nhất là từ một vùng trên ma trận ta chọn ra số có giá trị lớn nhất làm kết quả thu được sau bước pooling (max pooling)

2.4. Convolution là gì

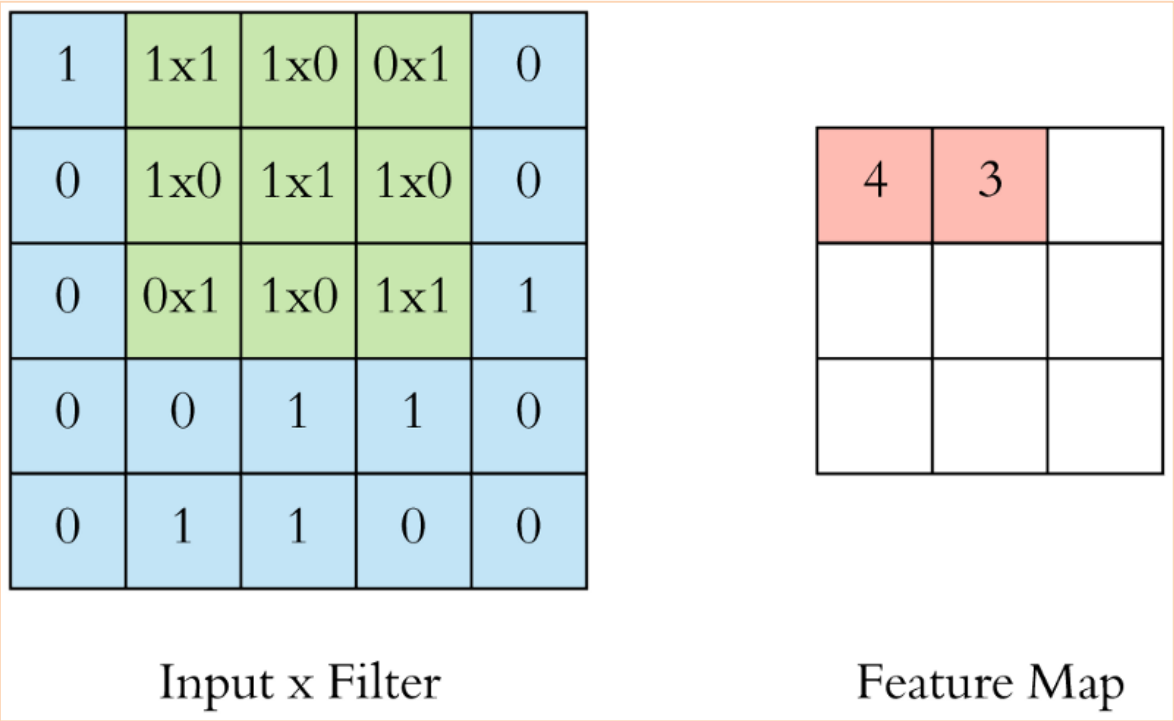


Hình 2.4.1 Convolutional là gì

Khối cơ bản tạo lên CNN là convolutional layer. Convolution là một phép toán học để kết hợp hai khối thông tin với nhau. Trong trường hợp này, convolution được áp dụng trên dữ liệu đầu vào (ma trận) và sử dụng một mặt lạ gọi là convolution filter để tạo ra một mảng mới gọi là feature map.

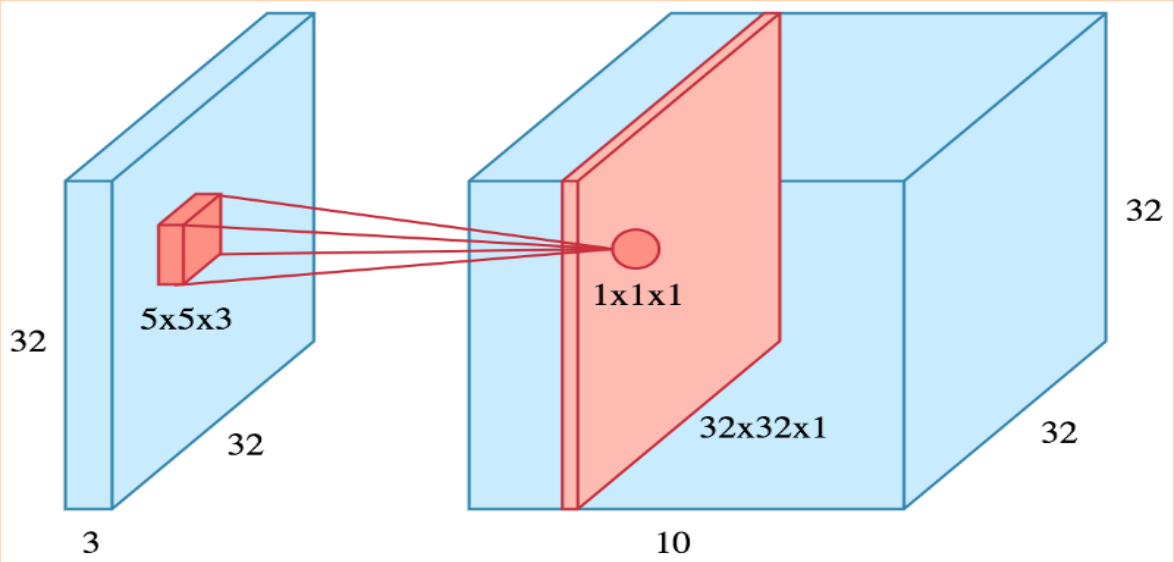
Việc thực hiện phép toán convolution được mô tả như hình dưới dây với đầu vào là một mảng hai chiều 5x5 phần tử là filter có kích thước là 3x3 phần tử. Của sổ filter sẽ được trượt từ trái qua phải, từ trên xuống dưới. Tại mỗi vị trí của cửa sổ filter ta thực hiện nhân tương ứng từng phần tử trong ma trận đầu vào với từng phần tử trong filter, sau đó cộng tổng các tích với nhau ta thu

được kết quả là một phần tử trên feature map. Quá trình thực hiện được mô tả trong hình minh họa sau đây.



Hình 2.4.2 Convolutional và mảng hai chiều

Trên đây là mô tả thực hiện phép toán convolution với ma trận hai chiều với một filter duy nhất. Trong thực tế đối với ảnh RGB ta thực hiện convolution với ma trận ba chiều ví dụ như ảnh RGB và với cùng một ảnh đầu vào ta áp dụng phép toán convolution với nhiều filter khác nhau. Mỗi một filter được áp dụng cho ta một feature layer. Nhiều feature layer xếp trồng lên nhau ta thu được một convolution layer. Ví dụ sau thể hiện ảnh có kích thước 32x32 và có ba kênh mầu, ta sử dụng 10 filter và thu được convolution layer là một ma trận 32x32x10.



Hình 2.4.3 Convolutional và mảng ba chiều

2.5. Stride và Padding

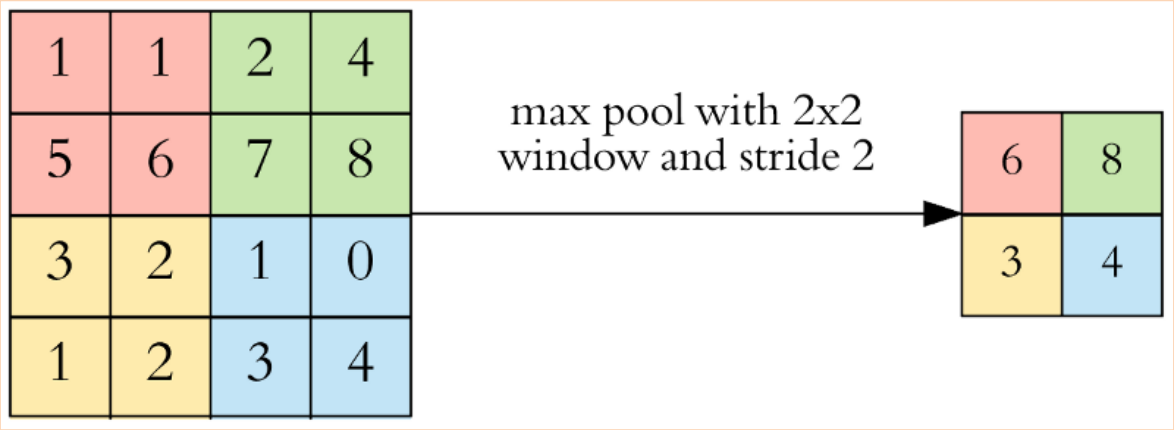
Stride là số bước nhảy của mỗi lần dịch chuyển convolution filter, trong ví dụ đầu tiên về convolution ta nhận thấy kích thước của feature map nhỏ hơn kích thước của ma trận đầu vào. Để kích thước của feature map bằng kích thước của ma trận đầu vào ta cần phải bổ xung thêm một số điểm bao quanh ma trận đầu vào thường là thêm các phần tử 0 vào xung quanh ma trận đầu vào, thao tác trên được gọi là padding.

Khi thực hiện phép toán convolution với đầu vào là ma trận vuông có kích thước là nxn, stride là s, kích thước filter là fxf, vùng padding có kích thước là p ta có kích thước của feature map thu được là:

$$Output\ size = \left(\frac{n+2p-f}{s}+1\right) x \left(\frac{n+2p-f}{s}+1\right)$$

2.6. Pooling là gì

Sau khi thực hiện phép toán convolution chúng ta thường sủ dụng pooling nhằm giảm số chiều của dữ liệu. Loại pooling thông dụng nhất là max pooling tức là trong một vùng được chọn (pooling window) được chọn của ma trận, ta lấy phần tử có kích thước lớn nhất, cũng giống với convolution thì pooling window cũng được định nghĩa kích thước (size) và bước nhảu (tride). Dưới đây là ví dụ việc áp dụng max pooling sử dụng 2x2 window và stride là 2.



Hình 2.6.1 Minh họa max pooling

2.7. Fully connected neural network

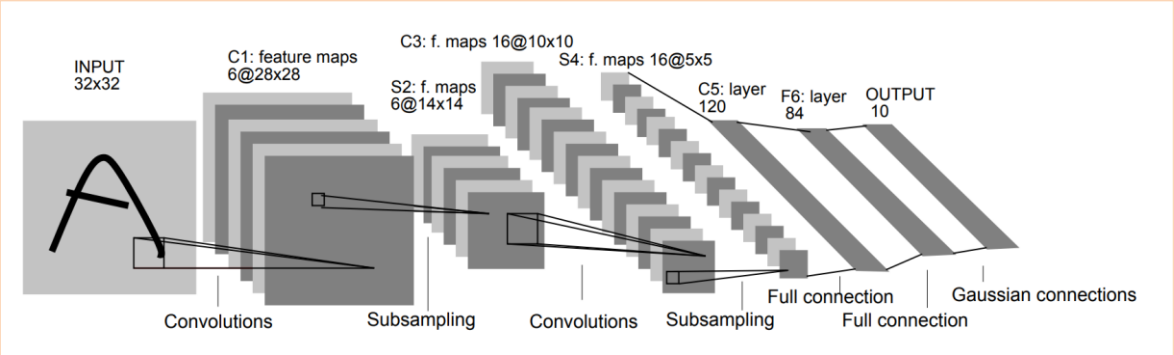
Sau các lớp convolution + pooling layers chúng ta sẽ gắn thêm một mạng Artificial Neural Network nhằm phục vụ quá trình tranning và nhận diện đối tượng. Thông qua quá trình học CNN tự động sẽ cập nhật lại giá trị cho các filter, weight matrix W và bias b.

CHƯƠNG 3 – BÀI TOÁN NHẬN DẠNG ĐỐI TƯỢNG

3.1. Đầu vào và đầu ra của bài toán

Type here

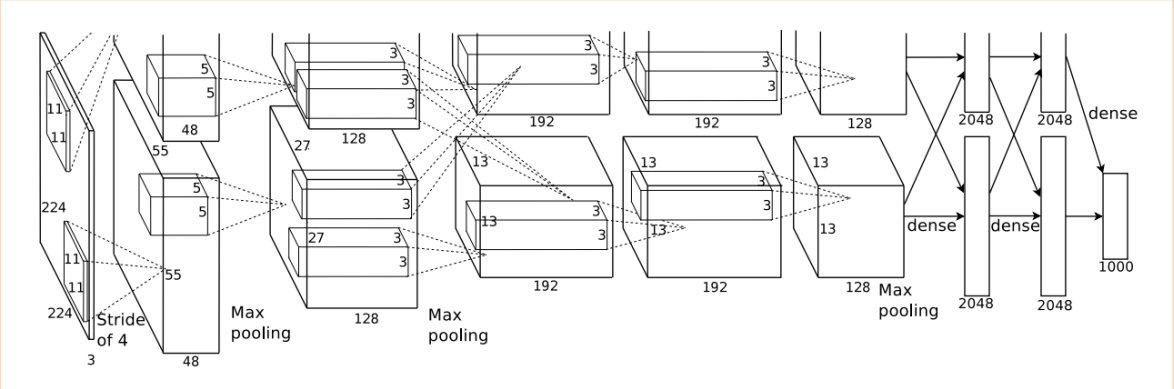
3.2. Một số kiến trúc mạng convolutional neural network



Hình 4.2.1 Sơ đồ kiến trúc mạng LeNet-5

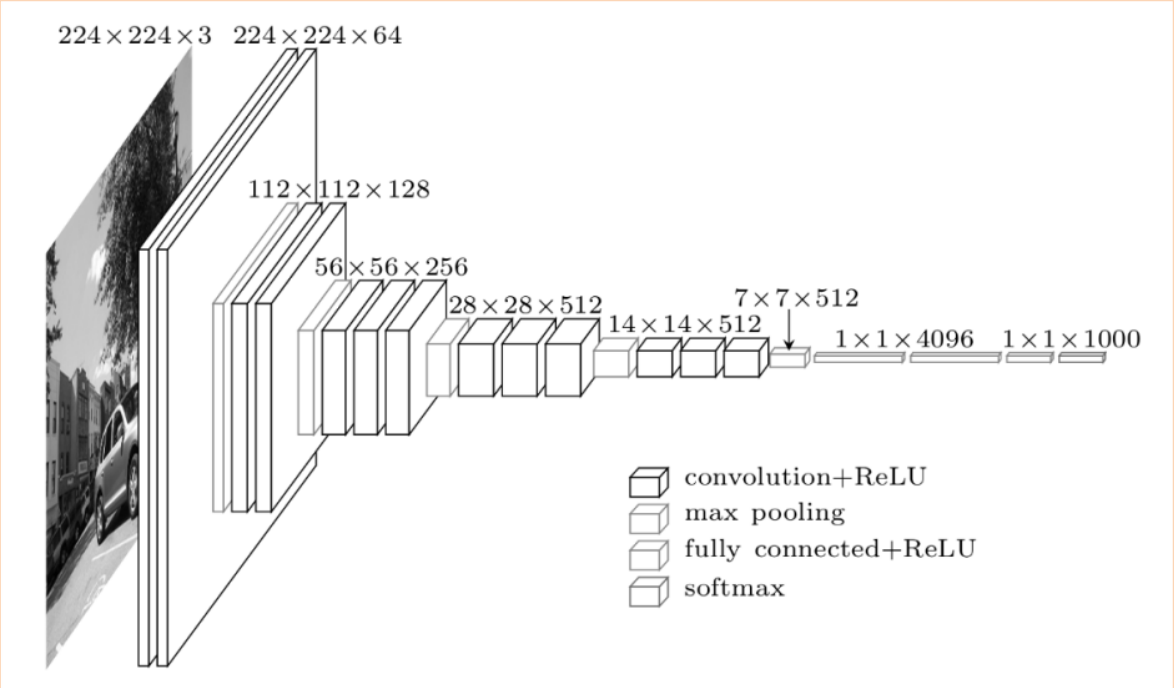
Ảnh đầu vào có kích thước 32x32x1. Layer C1 chứa 6 feature map kích thước 28x28, mỗi phần tử tại lớp này được kết nối với 5x5 phần tử tại lớp đầu vào. Số lượng parameters cần học là (5 * 5 + 1) * 6 = 156. Lớp S2 chứa 6 feature map với kích thước 14x14, mỗi phần tử tại lớp này kết nối với 2x2 phần tử tại lớp C1. Số lượng parameters cần học là 6 * 2 = 12. Lớp C3 chứa 16 feature

map kích thước là 10x10, mỗi phần tử kết nối với 5x5 phần tử tại layer S2. Số lượng parameters cần học là (5 * 5 * 6 + 1) * 16 = 1516. Lớp S4 chứa 16 featrue map kích thước 5x5, mỗi phần tử kết nối với 2x2 tương ứng ở lớp C3. Số lượng parameters cần học là 16 * 2 = 32. Lớp C5 chứa 120 feature map mỗi phần tử tại lớp này kết nối với 5x5 phần tử lại lớp S4. C5 có 120*(16*25+1) = 48120. F6 là một lớp của một full connected neural network có 84 phần tử và 84*(120+1)=10164 parameters cần học. Lớp cuối cùng là đầu ra của neural network. Các trọng số (còn được gọi là các parameters cần học) được cập nhật bởi thuật toán backpropagation.



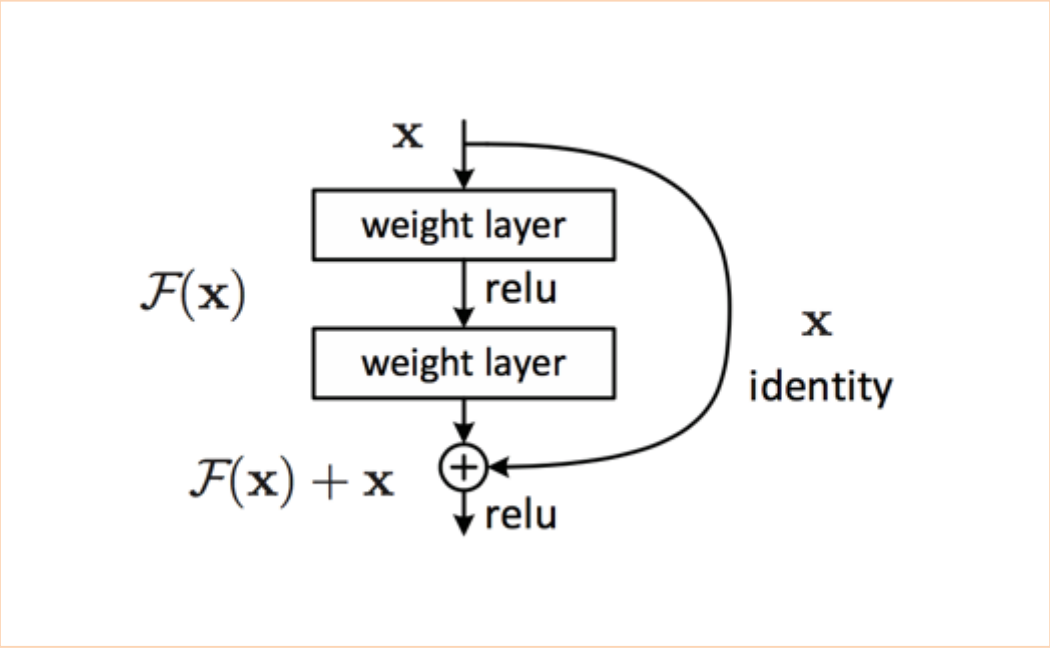
Hình 4.2.2 Sơ đồ kiến trúc mạng AlexNet

Mạng có 8 lớp với 5 lớp đầu là các convolutional layer và ba lớp tiếp theo là ba lớp thuộc mạng fully-connected neural. Đầu vào của mạng là ảnh có kích thước 224x224x3 pixel. Lớp convolutional layer đầu tiên là kết quả của việc thực hiện phép toán convolution và 96 kernel với kích thước 11x11x3 và bước nhảy stride bằng 4 pixels. Lớp thứ hai của mạng được tạo bởi việc thực hiện phép toán convolution với 256 kernel kích thước là 5x5x48. Lớp ba, bốn, năm là các lớp convoluion được kết nối từ layer trước sang layer sau không thông qua phép toán pooling. Lớp thứ ba có 384 kernels với kích thước 3x3x128. Lớp thứ tư có 384 kernels với kích thước 3x3x192. Lớp thứ năm có 256 kernels với kích thước 3x3x192. Cuối cùng mang được kết nối với một mạng fully-connected neural với 4096 neurons trong mỗi lớp.

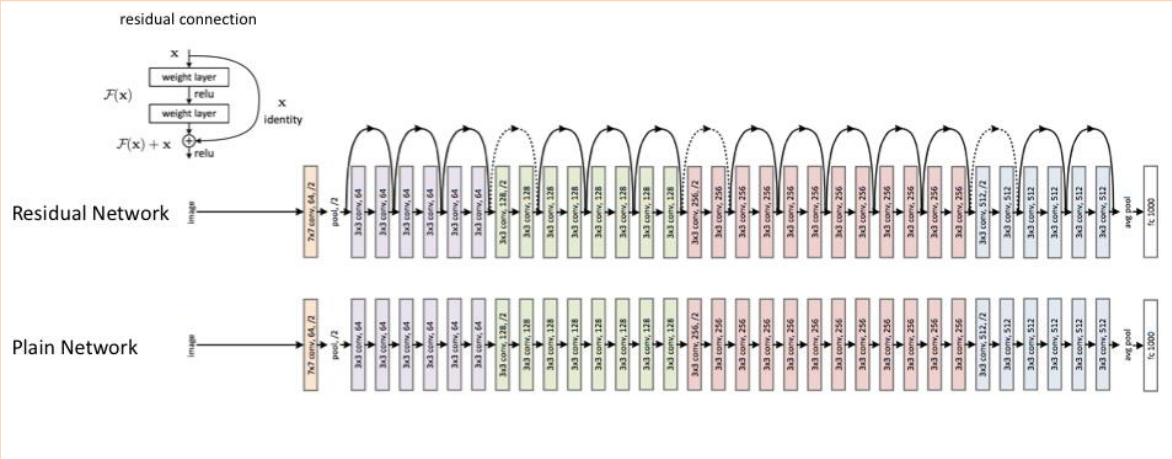


Hình 4.2.3 Sơ đồ kiến trúc mạng VGG16

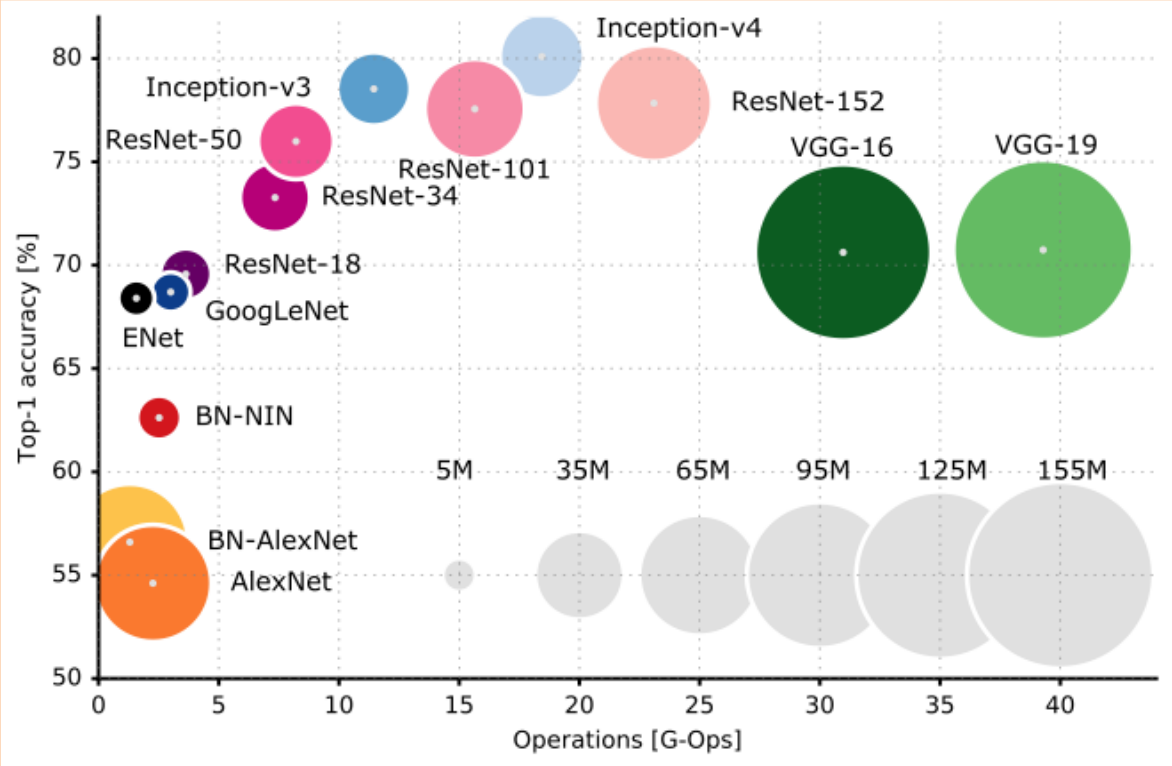
ResNet



Hình 4.2.4 Sơ đồ phần tử Residual Block



Hình 4.2.5 Sơ đồ kiến trúc mạng ResNet



Hình 4.2.6 So sánh độ chính xác và khối lương tính toán

Hình trên được lấy từ kaggle.com, chúng ta có thể thấy được sự so sánh về khối lượng tính toán và độ chính xác giữa các cách xây dựng mạng CNN phổ biến hiện nay, trong đó ResNet 50 có độ chính xác khá cao trong khi khối lượng tính toán không nhiều. AlexNet cần khối lượng tính toán thấp nhưng độ chính xác tương đối thấp so với các kiến trúc mạng khác.

3.3. Thử nghiện trên trên mạng LeNet-5 mở rộng

LeNet-5 là mạng cơ bản nhất trong việc nghiên cứu về convolutional neuron network. Mạng trên được thực hiện nhằm nhận diện các đối tượng cơ bản như số và chữ viết tay, các đối tượng đó đơn giản nên kích thước ảnh đầu vào nhỏ và dễ dàng nhận diện. Khi vào bài toán nhận diện địa danh có đặc điểm là kích thước ảnh nhận diện sẽ lớn và độ phức tạp trong chi tiết của ảnh lớn, em đã lấy ý tưởng thiết kế từ mạng này và mởi rộng số lớp convolutional cũng như kích thước của input và filter cho phù hợp với dữ liệu bài toán đang cần xử lý.

No	Layer	Input size	kernel size	stride	Bias	Activation	Learn parameter
1	convolutional	128 x 128 x 3	5 x 5 x 3 x 8	1	8	relu	608
	max pooling	128 x 128 x 8	2 x 2	2	-	-	0
2	convolutional	64 x 64 x 8	5 x 5 x 3 x 16	1	16	relu	1216
	max pooling	64 x 64 x 16	2 x 2	2	-	-	0
3	convolutional	32 x 32 x 16	5 x 5 x 3 x 32	1	32	relu	2432
	max pooling	32 x 32 x 32	2 x 2	2	-	-	0
	flatten	-	-	-	-	-	0
4	fully connect	8192	8192 x 128	-	128	-	1048704
5	fully connect	128	128x64	-	64	-	8256
							1061216

Bảng 3.3.1 thông tin thiết kế mạng lenet-5 mở rộng

Mạng được thiết kế với đầu vào là ảnh có kích thước 128x128 với 3 kênh màu RGB. Và đầu ra của mạng là một vector 64 phần tử, mỗi phần tử là giá trị sác xuất ảnh đầu vào rơi vào địa danh có mã tương ứng.

Sau đây là việc triển khai, training và thực hiện kiểm định độ chính xác của mạng bằng thư viện Tensorflow.

CHƯƠNG 4 – THIẾT KẾ HỆ THỐNG

4.1. Biểu đồ ca sử dụng

4.2. Biểu đồ hoạt động

4.3. Biểu đồ tuần tự

CHƯƠNG 5 – ĐÁNH GIÁ KẾT QUẢ

5.1. Giao diện trương trình

5.2. Minh họa chức năng phát hiện vi phạm luật giao thông

5.3. Độ chính xác của hệ thống

5.4. Hướng phát triển trong tương lai

CHƯƠNG 6 - TÀI LIỆU THAM KHẢO

[1]https://towardsdatascience.com/applied-deep-learning-part-4-convolutional-neural-networks-584bc134c1e2

[2]https://www.mathworks.com/videos/introduction-to-deep-learning-what-are-convolutional-neural-networks--1489512765771.html

[3] https://medium.com/machine-learning-bites/deeplearning-series-convolutional-neural-networks-a9c2f2ee1524

[4] https://www.jefkine.com/general/2016/09/05/backpropagation-in-convolutional-neural-networks/

[5] http://www.ee.surrey.ac.uk/CVSSP/demos/chars74k/

[6] https://www.kaggle.com/shivamb/cnn-architectures-vgg-resnet-inception-tl