Very Deep Convolutional Networks for Large-Scale Image Recognition

Evandro Dessani

Dez-2017

Intro

- Convolutional networks (ConvNets) have recently enjoyed a great success in large-scale image and video recognition.
  - Krizhevsky - 2012;
  - Zeiler & Fergus - 2013;
  - Sermanet - 2014; Simonyan & Zisserman - 2014)
- utilise smaller receptive window size and smaller stride of the first convolutional layer.
- Another line of improvements dealt with training and testing the networks densely over the whole image and over multiple scales

ConvNet General Premisses

- Fix other parameters other than depth;
- Steadly increase the depth of the network by adding more convolutional layers;
- Use of the very small convolution filters in all layers;

Results

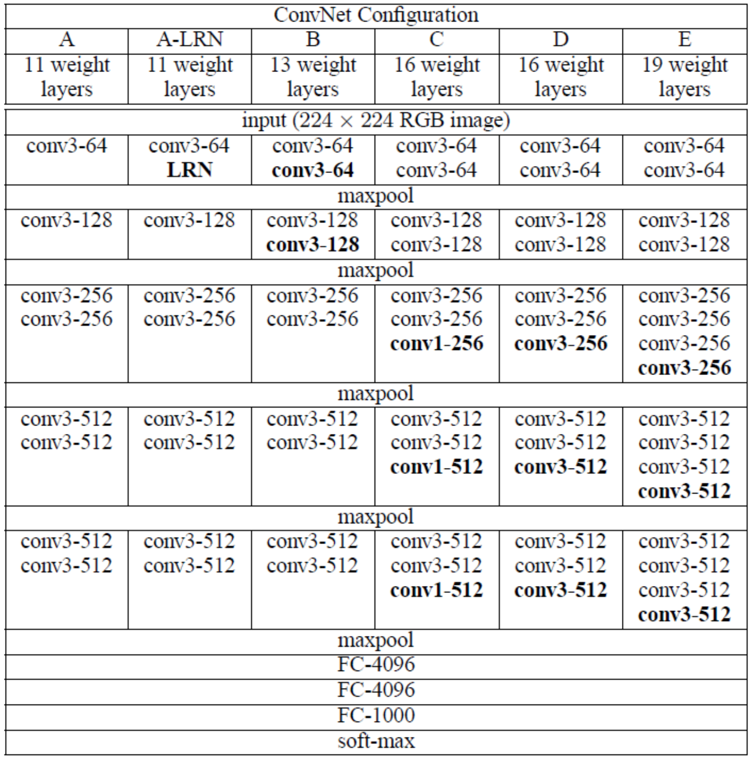
- significantly more accurate ConvNet architectures;
- achieve the state-of-the-art accuracy on ILSVRC classification and localisation tasks;
- applicable to other image recognition datasets;
- achieve excellent performance even when used as a part of a relatively simple pipelines;

Architecture

- The input is as fixed-size 224 x 224 RGB image.
- 3 x 3 filters on convolutional layers wich is smallest size for notion of left/right, up/down/center;
- The convolution stride is fixed to 1 pixel;
- The padding is 1 pixel for 3 × 3 conv. Layers;
- Spatial pooling is carried out by five max-pooling layers follow by some convolutional layers;
- All hidden layers are equipped with the rectification (ReLU)
- Max-pooling is performed over a 2 × 2 pixel window, with stride 2.
- Three Fully-Connected (FC) layers:
  - the first two have 4096 channels each, followed by a dropout with 0.5 ratio
  - the third performs 1000-way ILSVRC classification and thus contains 1000 channels (one for each class).
  - The final layer is the soft-max layer.

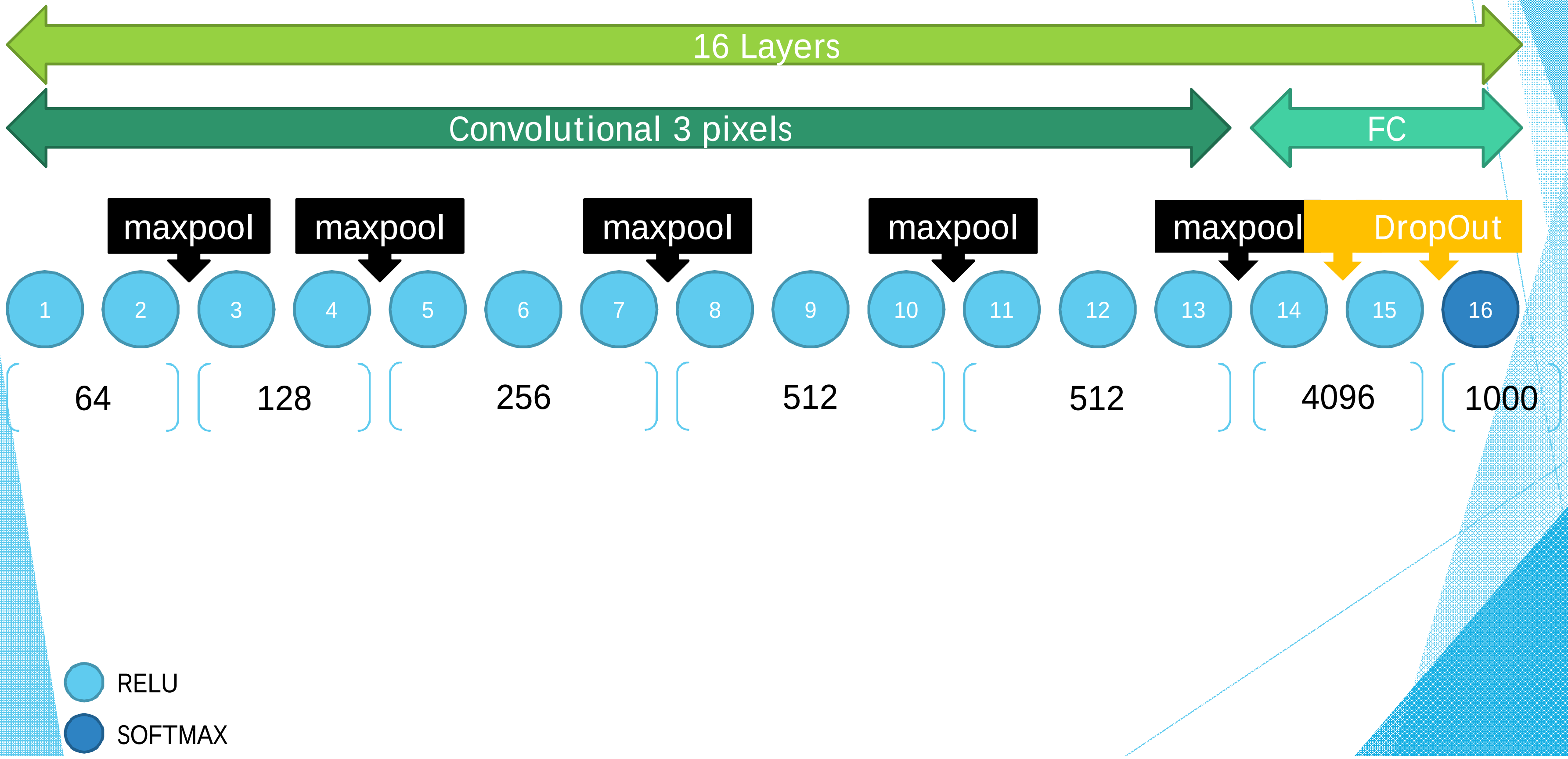
Configurations

- All configurations follow the generic design and differ only in depth:
  - From 11 to 19 weight layers where 3FC layers are fixed;
- The width of conv. layers (the number of channels), start from 64 in the first layer and then increasing by a factor of 2 after each max-pooling layer, until it reaches 512.



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 × 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					

ConvNet Configuration D



16 Layers

Convolutional 3 pixels

FC

maxpool maxpool maxpool maxpool maxpool DropOut

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16

64 128 256 512 512 4096 1000

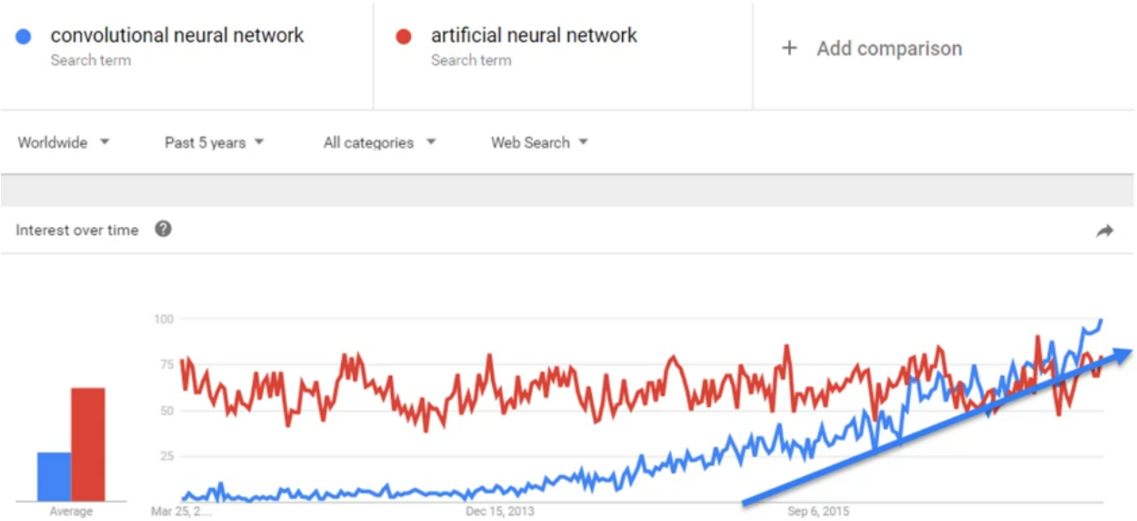
RELU

SOFTMAX

Training

- Image input sizes = 224 x 224
- Batch Size = 256
- Momentum 0.9
- Dropout Ratio = 0.5
- Penalty Multiplier = 0.0005
- Initial Learning Rate = 0.01
- Final Learning Rate = 0.00001
  - The learning rate was set to decrease by a fator of 10 when the validation set accuracy stopped improving
  - In total, the learning rate was decreased 3 times, and the learning was stopped after 370K iterations (74 epochs).
- Initialization Weights
  - First four Convolutional Layers and the last three conected layers was initialized with the layers of netconf A
  - The intermediate layers were initialized randomly

Coding



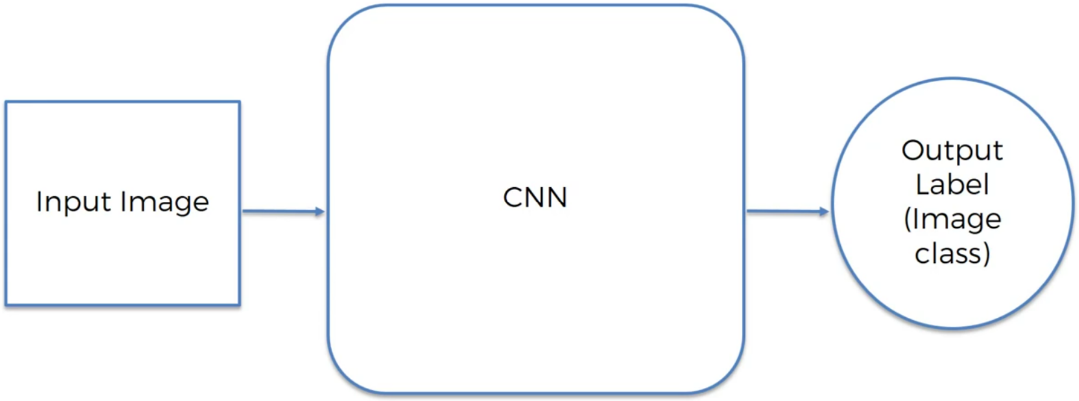
● convolutional neural network
Search term

● artificial neural network
Search term

+ Add comparison

Worldwide ▾ Past 5 years ▾ All categories ▾ Web Search ▾

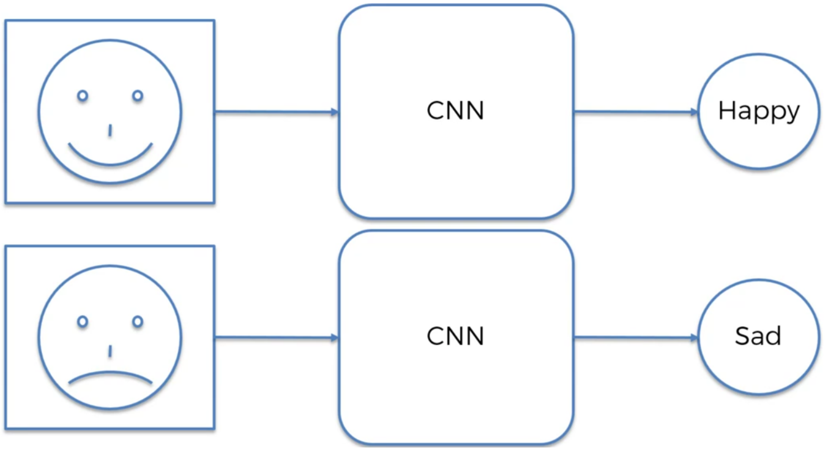
Interest over time



Input Image

CNN

Output Label (Image class)

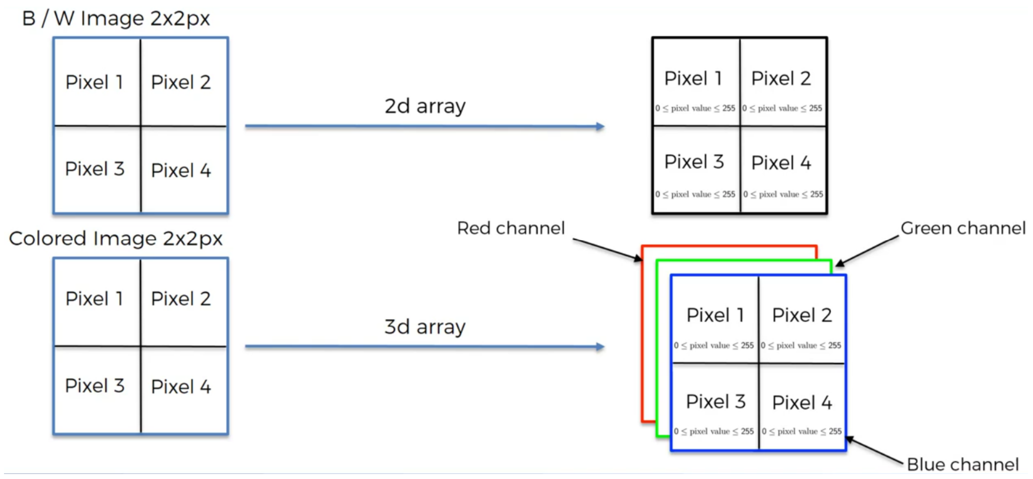


CNN

Happy

CNN

Sad



B / W Image 2x2px

Pixel 1 Pixel 2

Pixel 3 Pixel 4

2d array

Pixel 1 Pixel 2

0 ≤ pixel value ≤ 255 0 ≤ pixel value ≤ 255

Pixel 3 Pixel 4

0 ≤ pixel value ≤ 255 0 ≤ pixel value ≤ 255

Colored Image 2x2px

Pixel 1 Pixel 2

Pixel 3 Pixel 4

3d array

Red channel

Green channel

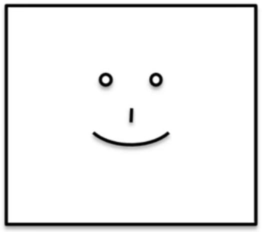
Pixel 1 Pixel 2

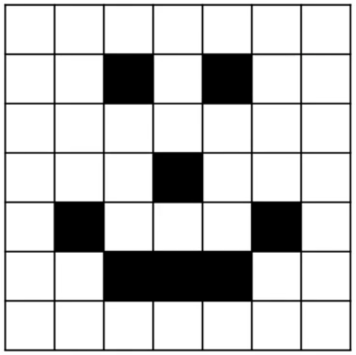
0 ≤ pixel value ≤ 255 0 ≤ pixel value ≤ 255

Pixel 3 Pixel 4

0 ≤ pixel value ≤ 255 0 ≤ pixel value ≤ 255

Blue channel





0	0	0	0	0	0	0
0	1	0	0	0	1	0
0	0	0	0	0	0	0
0	0	0	1	0	0	0
0	1	0	0	0	1	0
0	0	1	1	1	0	0
0	0	0	0	0	0	0

STEP 1: Convolution

↓

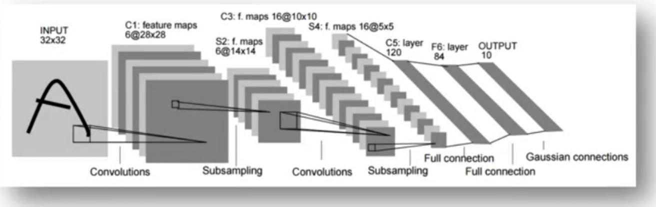
STEP 2: Max Pooling

↓

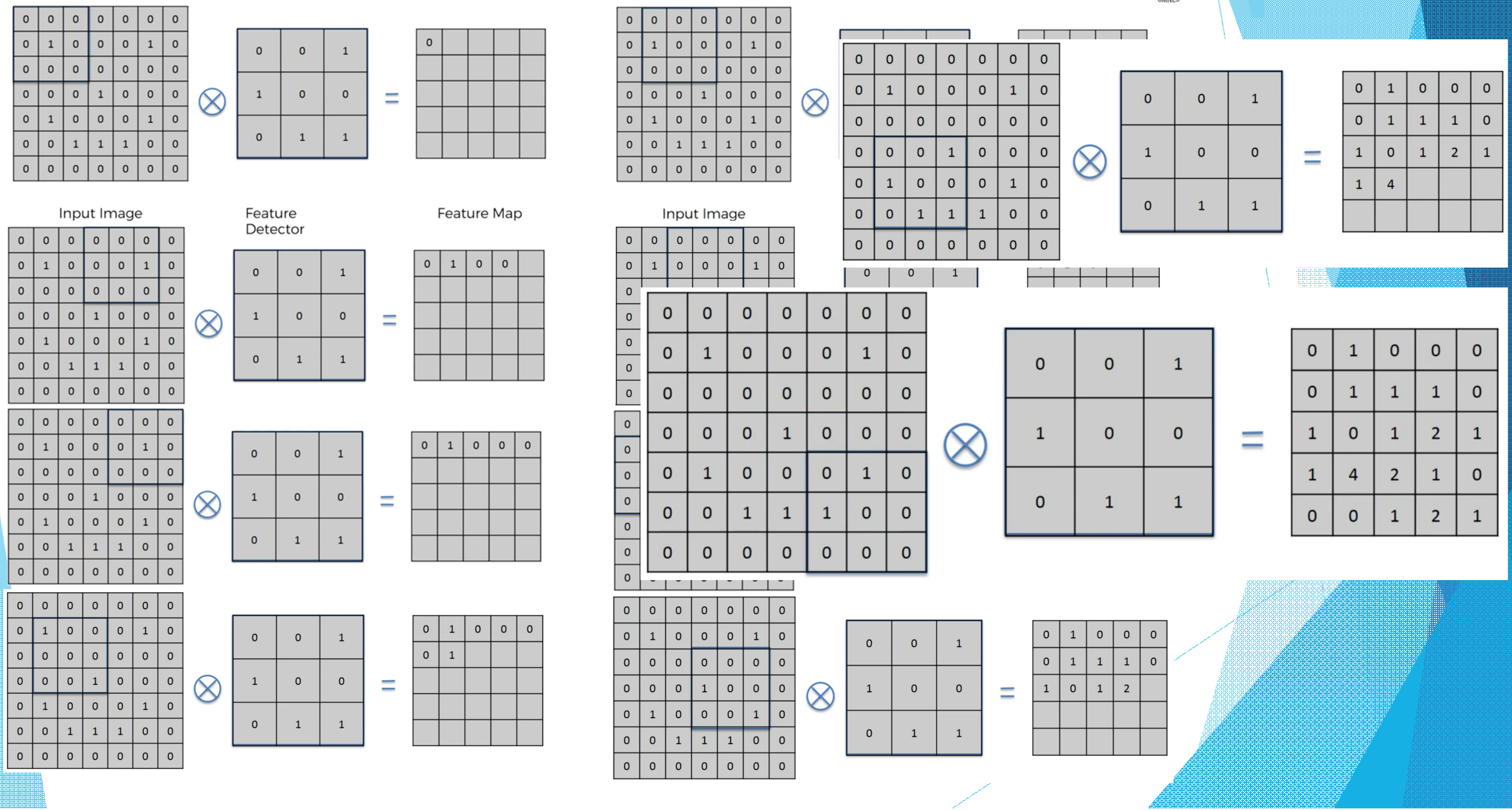
STEP 3: Flattening

↓

STEP 4: Full Connection



Step 1 - Convolution



Input Image

0	0	0	0	0	0	0
0	1	0	0	0	1	0
0	0	0	0	0	0	0
0	0	0	1	0	0	0
0	1	0	0	0	1	0
0	0	1	1	1	0	0
0	0	0	0	0	0	0

⊗

Feature Detector

0	0	1
1	0	0
0	1	1

=

Feature Map

0				

0	1	0	0	

0	1	0	0	0

0	1	0	0	0
0	1			

0	1	0	0	0
0	1	1	1	0
1	0	1	2	

0	1	0	0	0
0	1	1	1	0
1	0	1	2	1
1	4			

0	1	0	0	0
0	1	1	1	0
1	0	1	2	1
1	4	2	1	0
0	0	1	2	1