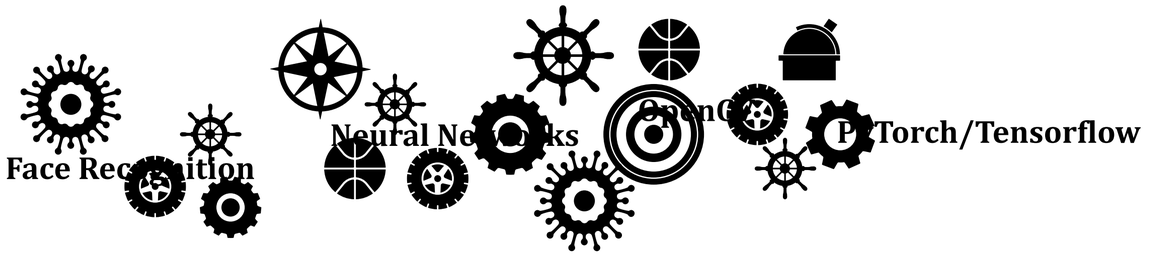
MPSSDEGB
कौशल से रोजगार, समृद्धि अपार

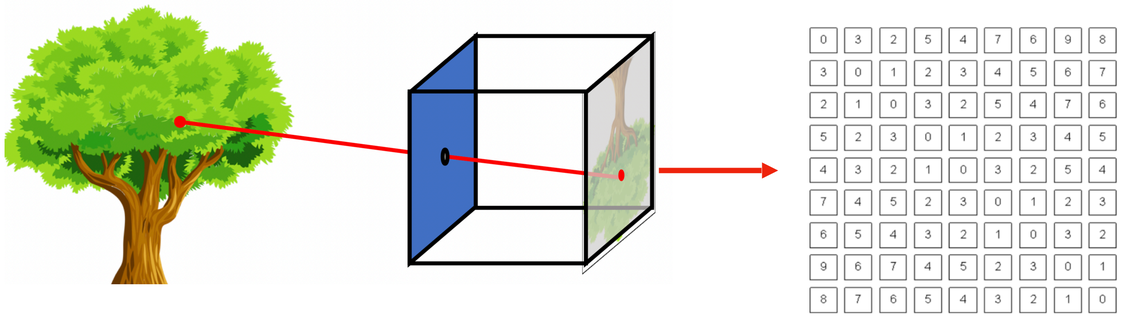
FOUNDATION FOR INNOVATION AND TECHNOLOGY TRANSFER
भारतीय प्रौद्योगिकी संस्थान दिल्ली
Indian Institute of Technology Delhi

Computer Vision

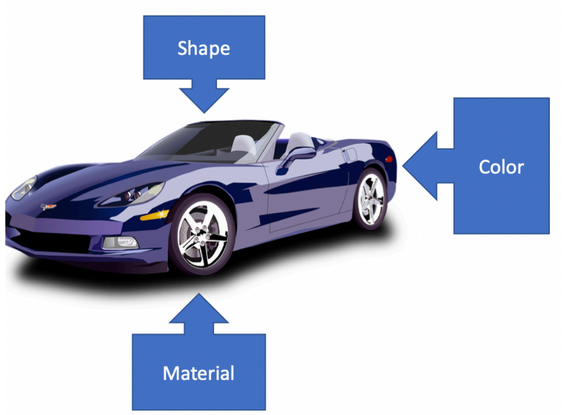


Deep Learning
Object Detection
Segmentation
Image Classification

Computer Vision

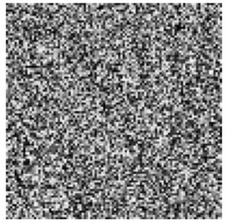


0	3	2	5	4	7	6	9	8
3	0	1	2	3	4	5	6	7
2	1	0	3	2	5	4	7	6
5	2	3	0	1	2	3	4	5
4	3	2	1	0	3	2	5	4
7	4	5	2	3	0	1	2	3
6	5	4	3	2	1	0	3	2
9	6	7	4	5	2	3	0	1
8	7	6	5	4	3	2	1	0



- Computer vision process 2D matrix of picture element (pixel)
- In contrast, real world objects does not deal with pixels but with physical attributes such as shape, colour and material.

Not all 2D arrays are images



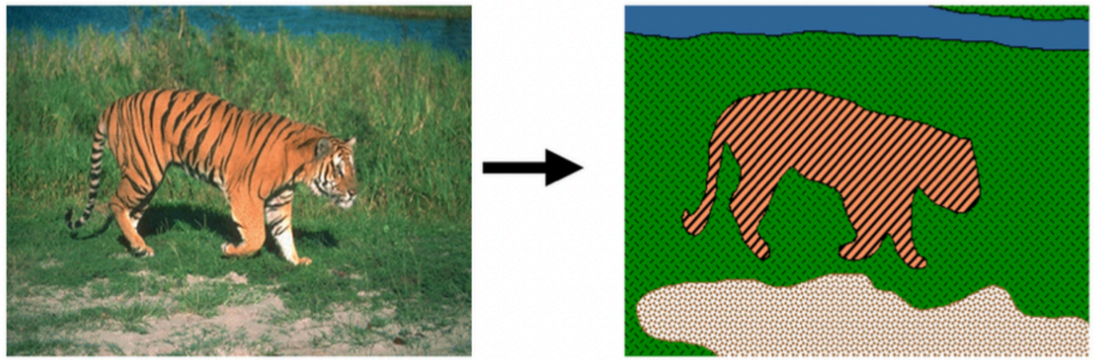
https://www.cs.cornell.edu/courses/cs4670/2018sp/lec02-imageformn.pdf

the content are taken from here and there

Computer Vision

- Natural images are *not* arbitrary 2D arrays
- They have properties resulting from physics / math of image formation

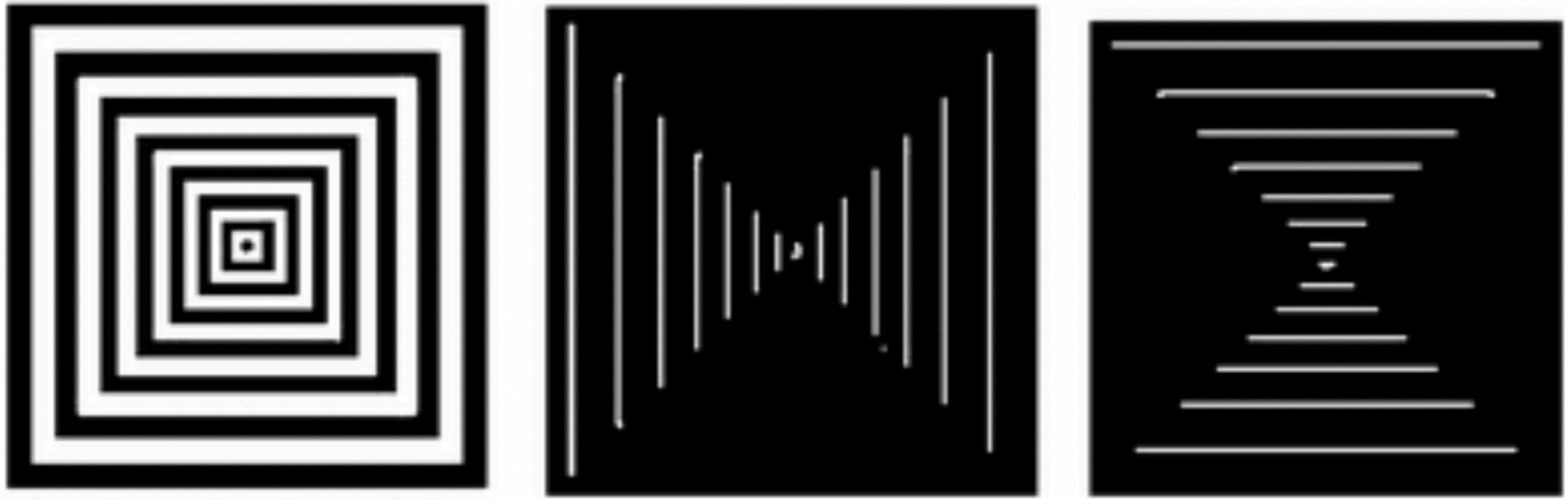
- Convert from “pixels” to “objects”: which groups of pixels correspond to objects?



https://www.cs.cornell.edu/courses/cs4670/2018sp/lec02-imageformn.pdf

the content are taken from here and there

Computer Vision



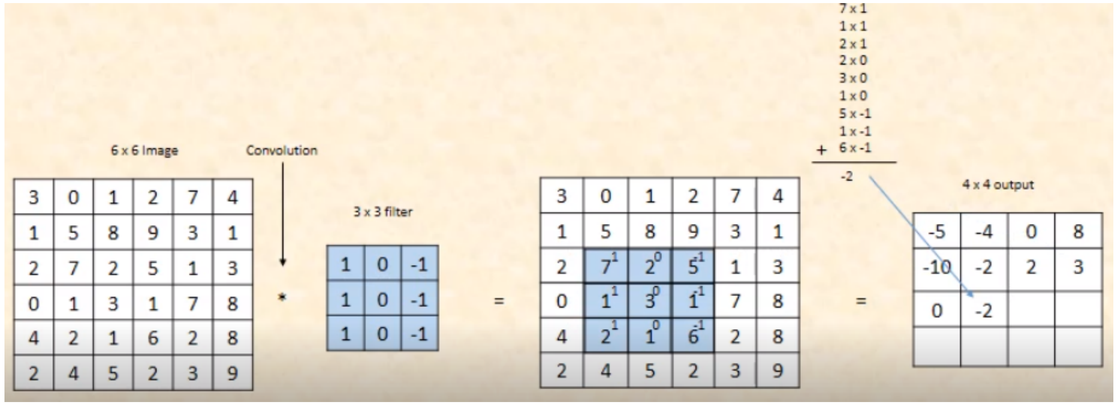
https://www.researchgate.net/figure/a-Input-binary-image-b-Measured-results-of-horizontal-edge-detection-using-Prewitt_fig10_313372308 the content are taken from here and there

Mathematics

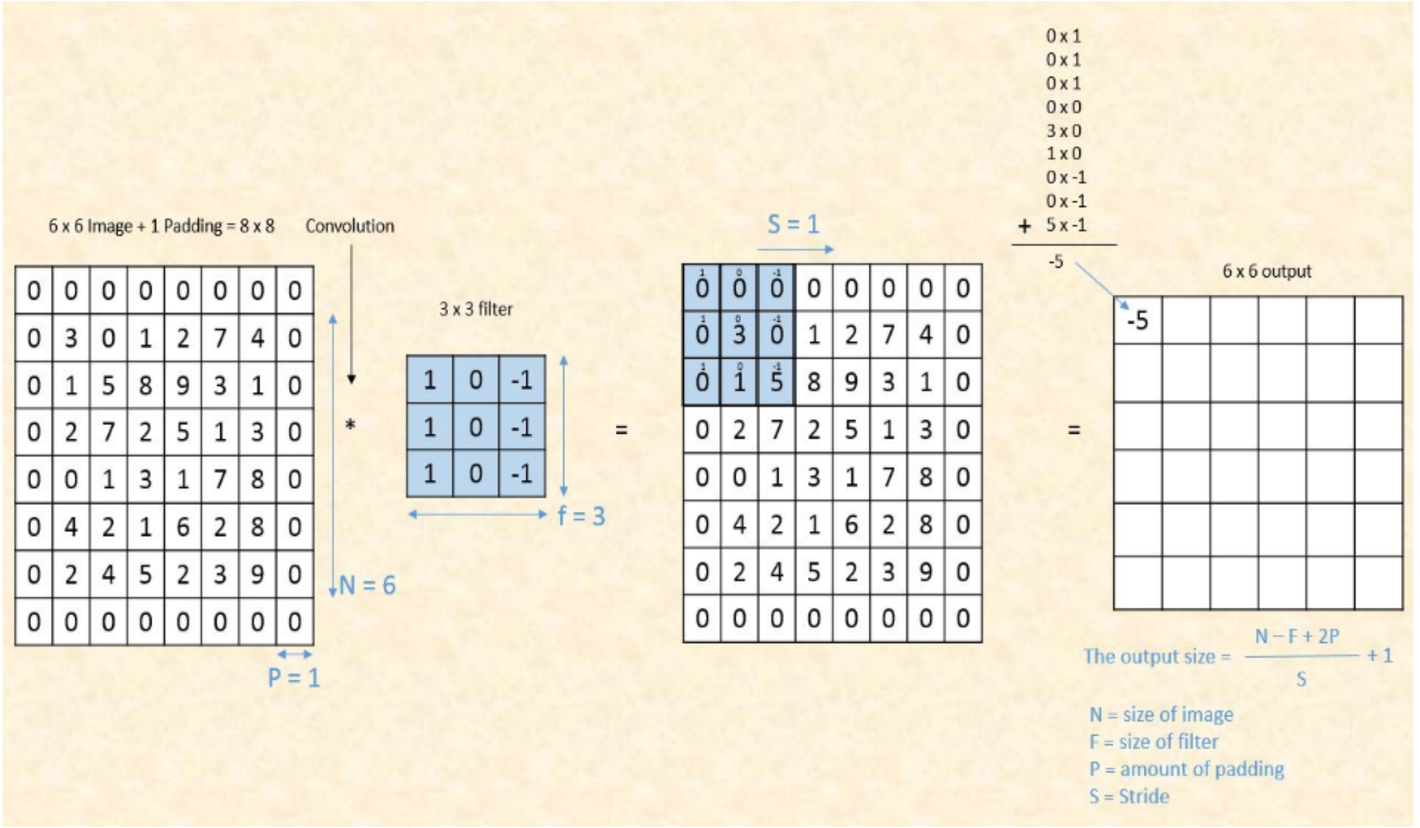
Convolution operation -

- convolving a 6 x 6 grayscale image with a 3 x 3 matrix called filter or kernel to produce a 4 x 4 matrix.

- dot product between the filter and the first 9 elements of the image matrix and fill the output matrix.



the content are taken from here and there



the content are taken from here and there

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

6 x 6 image

1	-1	-1
-1	1	-1
-1	-1	1

Filter 1

-1	**1**	-1
-1	1	-1
-1	1	-1

Filter 2

⋮ ⋮

Each filter detects a small pattern (3 x 3).

the content are taken from here and there

Convolution

1	-1	-1
-1	1	-1
-1	-1	1

Filter 1

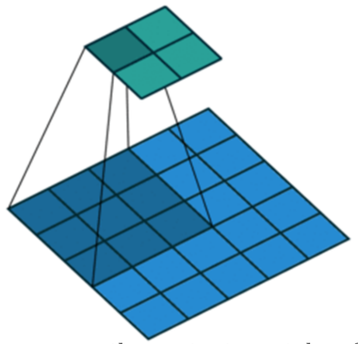
stride=1

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

Dot product → 3 -1

1*1 + 0*(-1) + 0*(-1) +
0*(-1) + 1*1 + 0*(-1) +
0*(-1) + 0*(-1) + 1*1

6 x 6 image



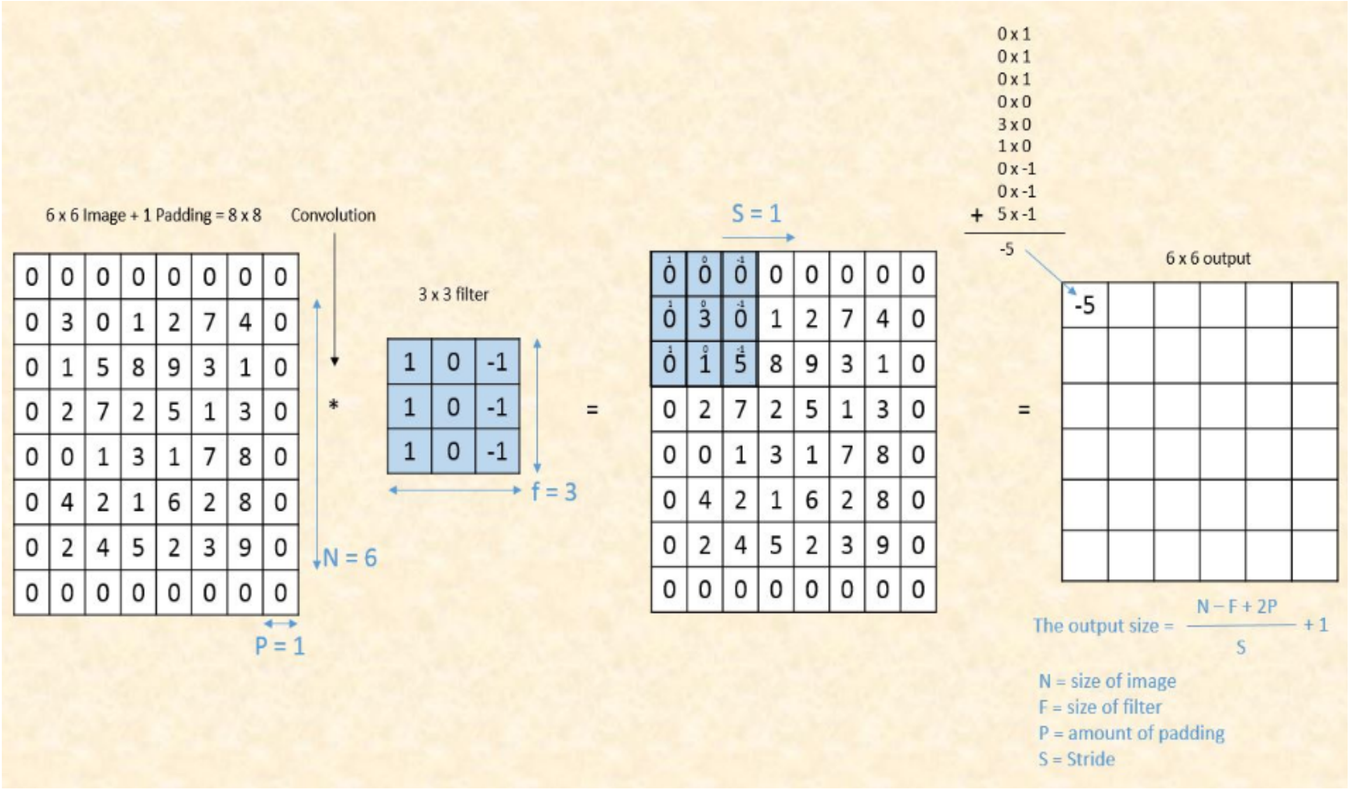
the content are taken from here and there

Challenges with convolution operation:

- Output may shrink.
- Data loss from image corners.

Solution - Padding (size of the output image is equal to input image size.)

the content are taken from here and there

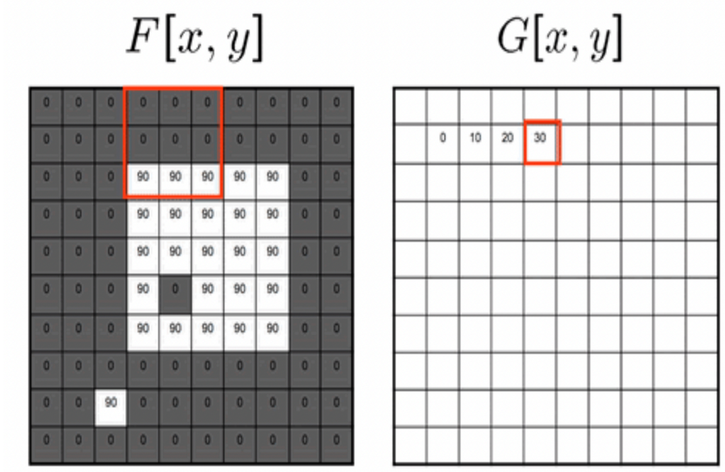


the content are taken from here and there

Computer Vision

1. Filtering and Convolution

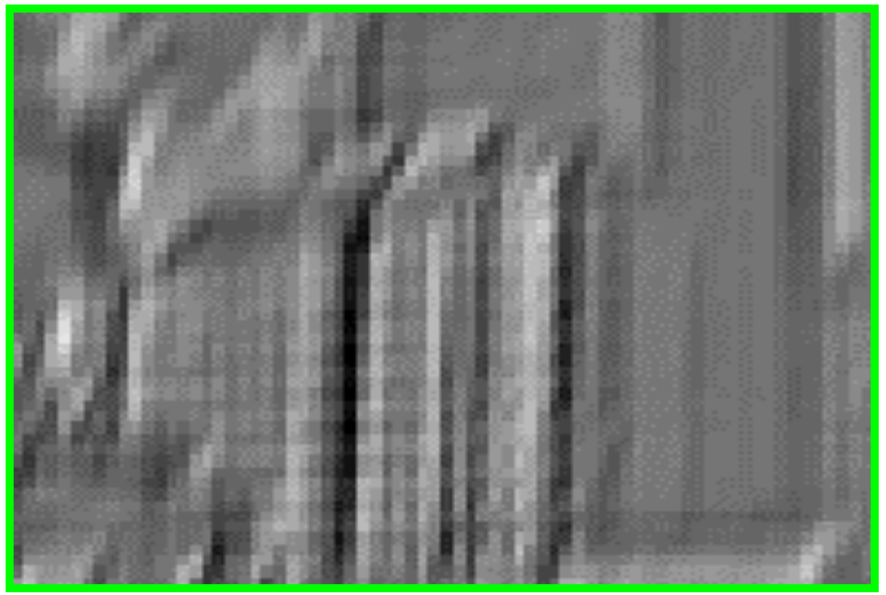
Filtering operations such as blurring, sharpening and noise reduction are applied to images using convolution. Convolution involves sliding a filter or kernel over an image and performing mathematical operations on each pixel. This process enables various improvements such as anti-aliasing, edge detection and texture removal.

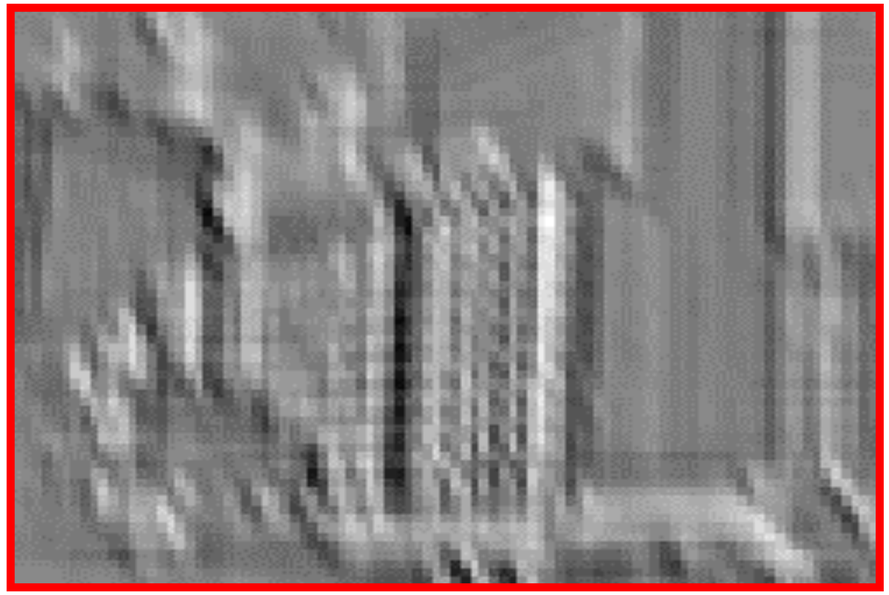


https://medium.com/@naveenpandey2706/lesson-2-understanding-image-processing-basics-and-techniques-in-computer-vision-c1d7c9763966

https://ai.stanford.edu/~syyeung/cvweb/tutorial1.html

the content are taken from here and there





- Weighted moving sum generating features

slide credit: S. Lazebnik

the content are taken from here and there

Biological Vision

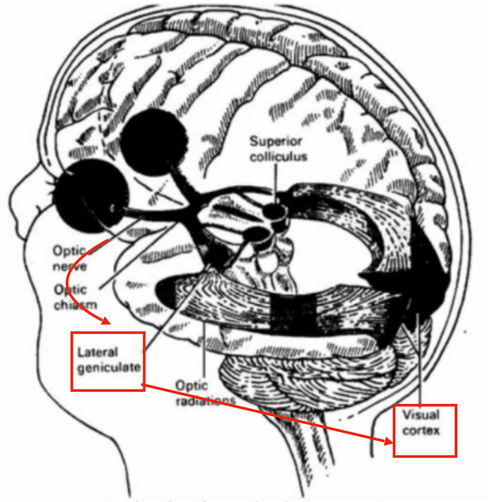


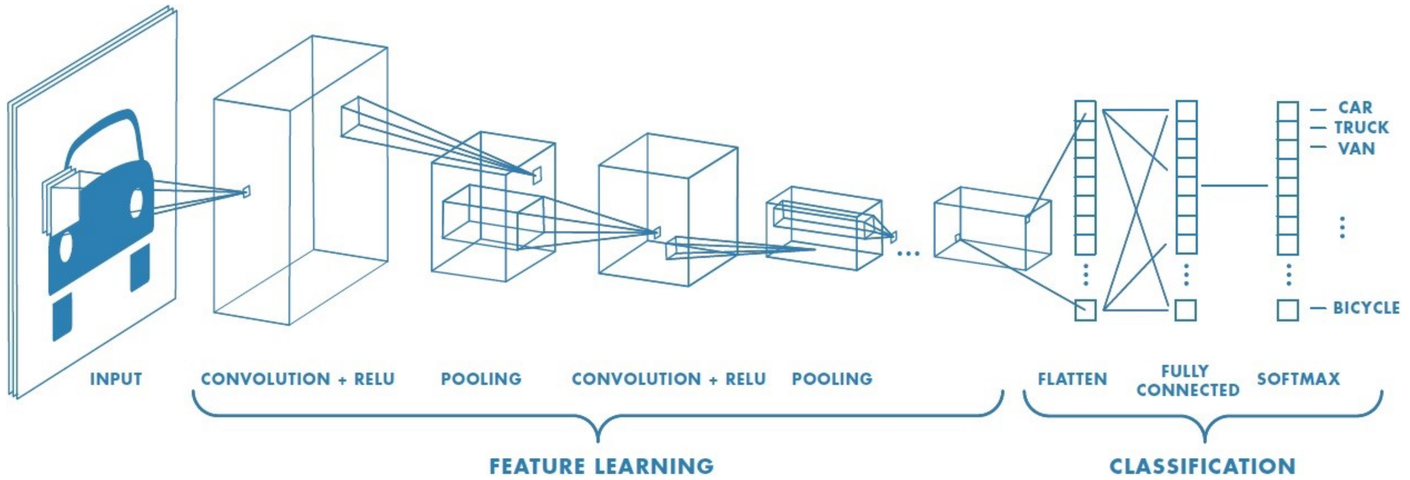
Image Credit:The Three R's of Computer Vision, Jitendra Malik
UC Berkeley

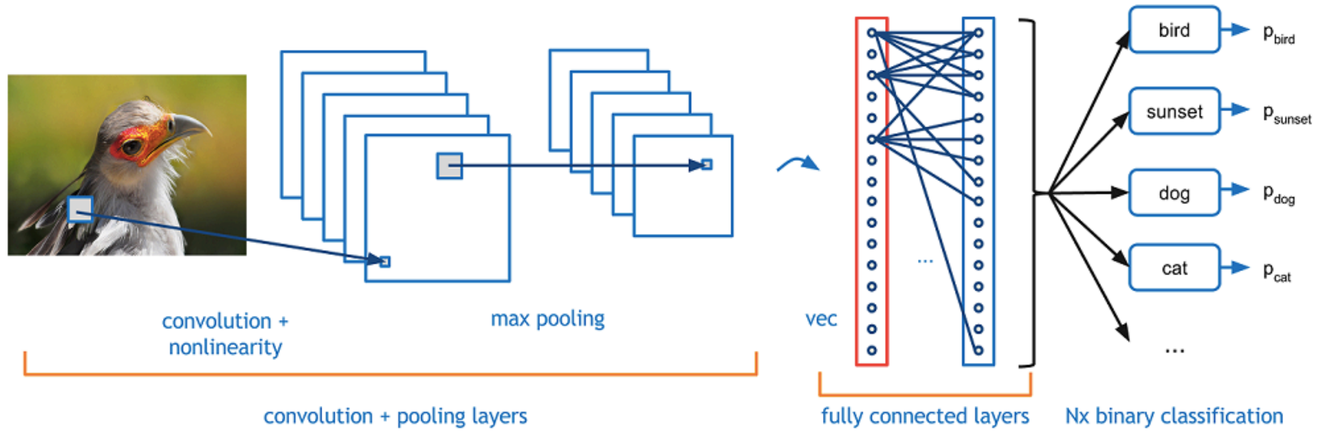
simple visual forms, edges, corners

Pathway of Information Processing

high-level object descriptions, faces, objects

the content are taken from here and there

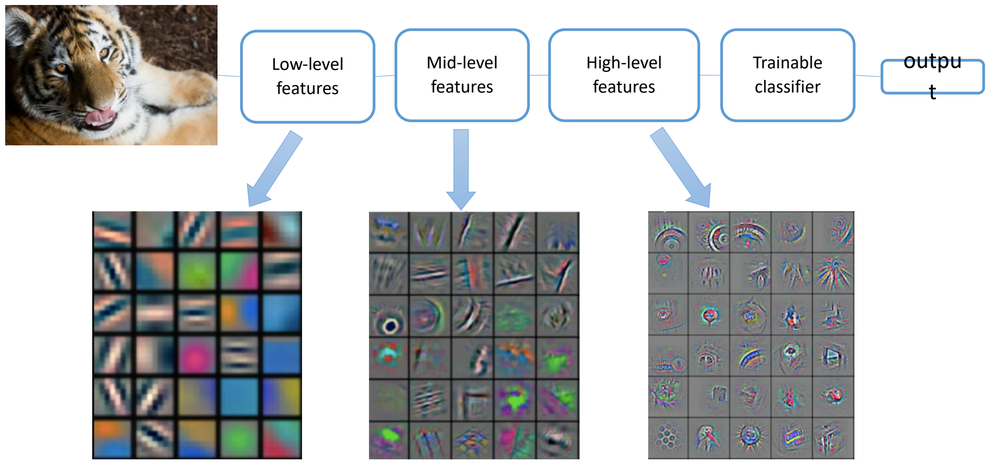




the content are taken from here and there

Deep Learning

- Deep learning (a.k.a. representation learning) seeks to learn rich hierarchical representations (i.e. features) automatically through multiple stage of feature learning process.



the content are taken from here and there

Convolutional Neural Networks

- Each layer in a CNN applies a different set of filters, typically hundreds or thousands of them, and combines the results, feeding the output into the next layer in the network.

- During training, a CNN automatically learns the values for these filters.

- In the context of image classification, our CNN may learn to:

✓ Detect edges from raw pixel data in the first layer.

✓ Use these edges to detect shapes (i.e., “blobs”) in the second layer.

✓ Use these shapes to detect higher-level features such as facial structures, parts of a car, etc. in the highest layers of the network.

- The last layer in a CNN uses these higher-level features to make predictions regarding the contents of the image.

the content are taken from here and there

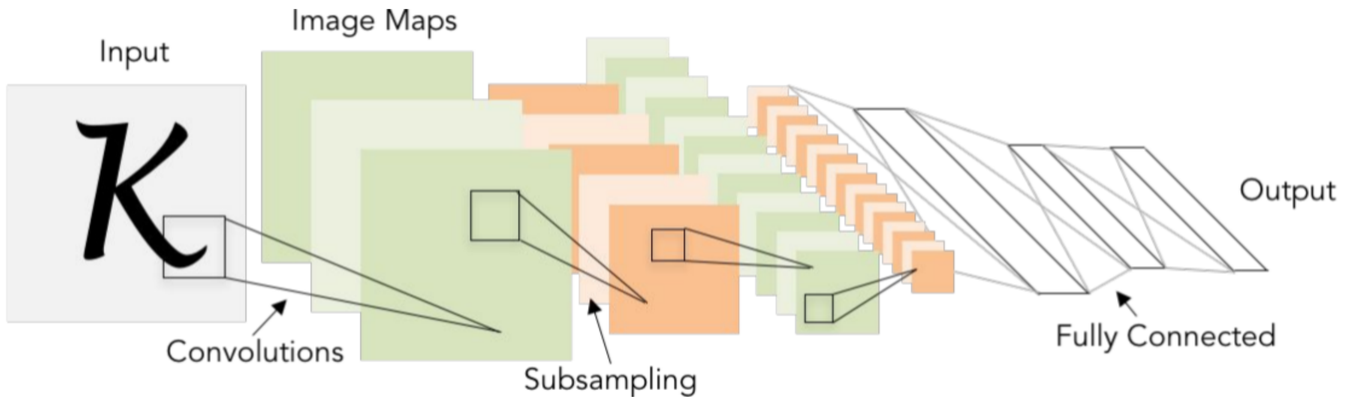
Convolutional Neural Networks

CNN consist of three layers -

1. Convolution Layer,
2. Pooling Layer, and
3. Fully Connected Layer.

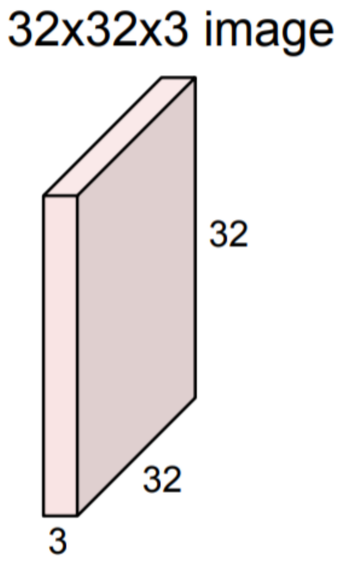
ConvNet is architecture is formed stacking these layers.

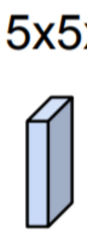
- The output can be a softmax layer indicating whether there is a cat or something else. You can also have a sigmoid layer to give you a probability of the image being a cat.



the content are taken from here and there

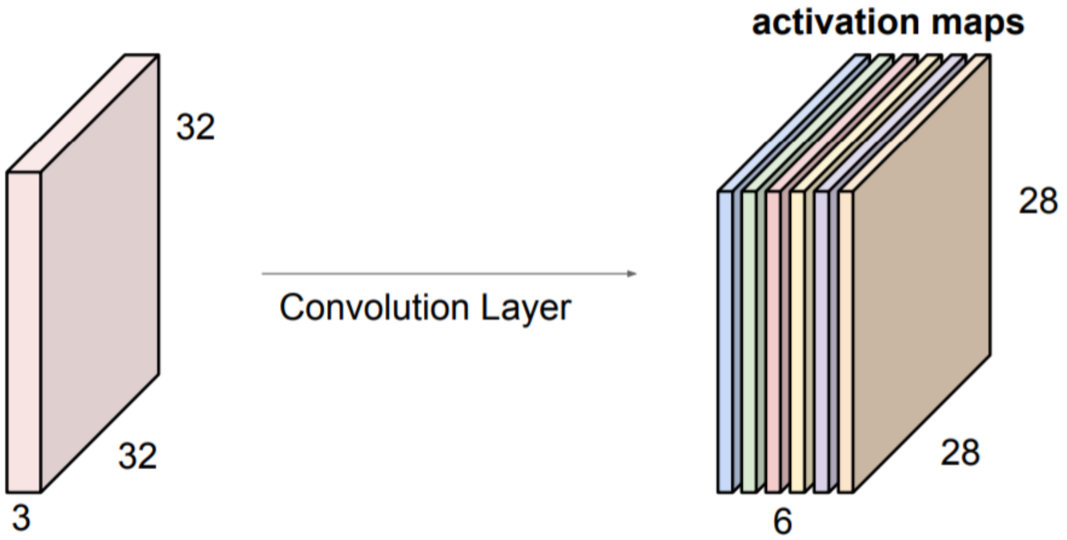
Convolve the filter with the image
i.e. “slide over the image spatially,
computing dot products”





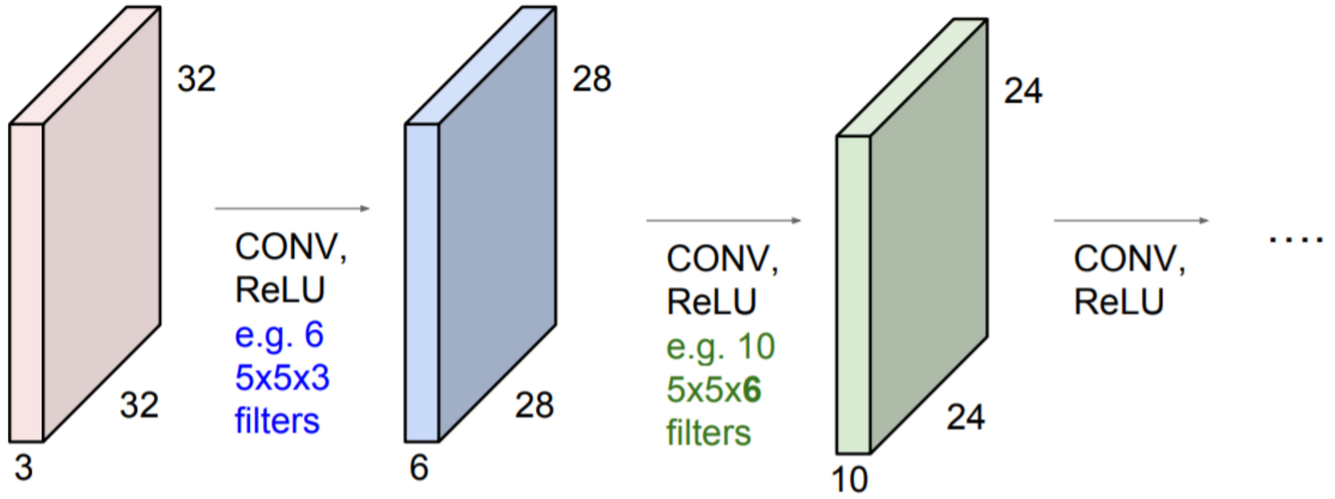
Convolve the filter with the image
i.e. “slide over the image spatially,
computing dot products”

the content are taken from here and there



Ex: we have 6 - > 5x5 filters, we'll get 6 separate activation maps. "new image" of size 28x28x6!

the content are taken from here and there



the content are taken from here and there

Convolution over volume

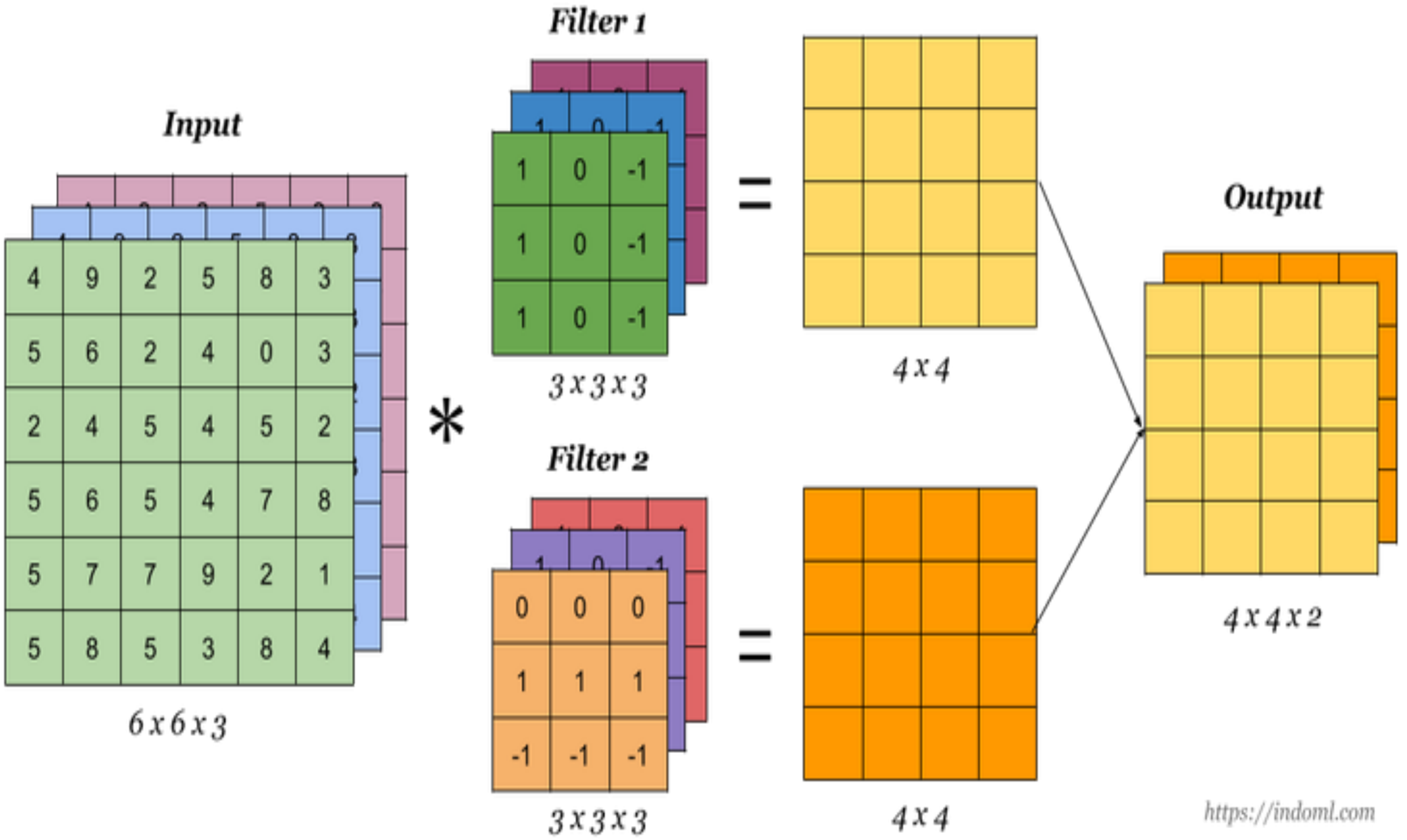
depth of the filter = depth of image (necessary condition)

- For every filter one feature map is created
- Depth of the output image represents the number of filters

With every increasing layer size of image decreases and number of filter increases

the content are taken from here and there

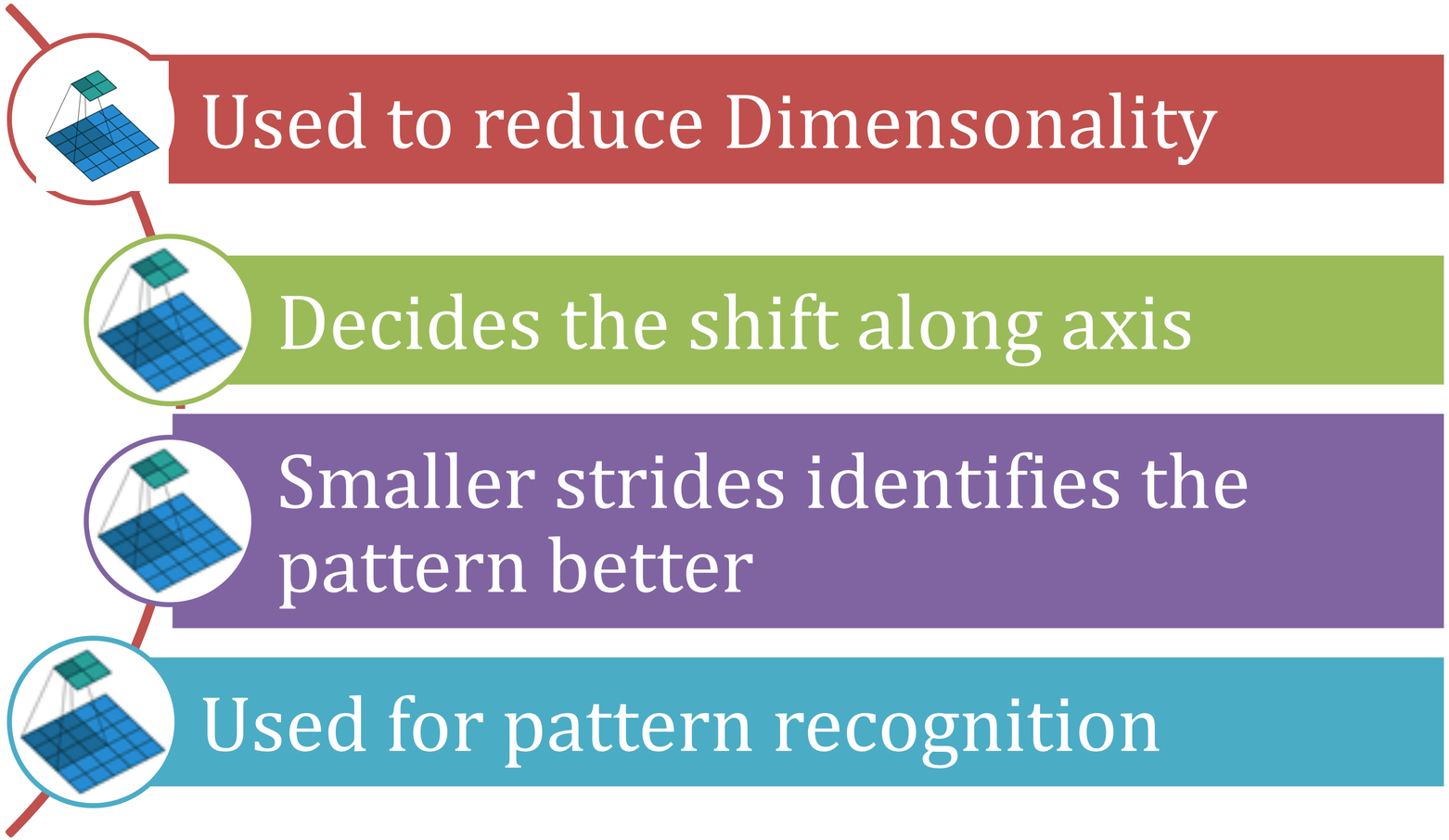
Convolution over volume



https://images.app.goo.gl/kGgsx37M3rcgtgsS9

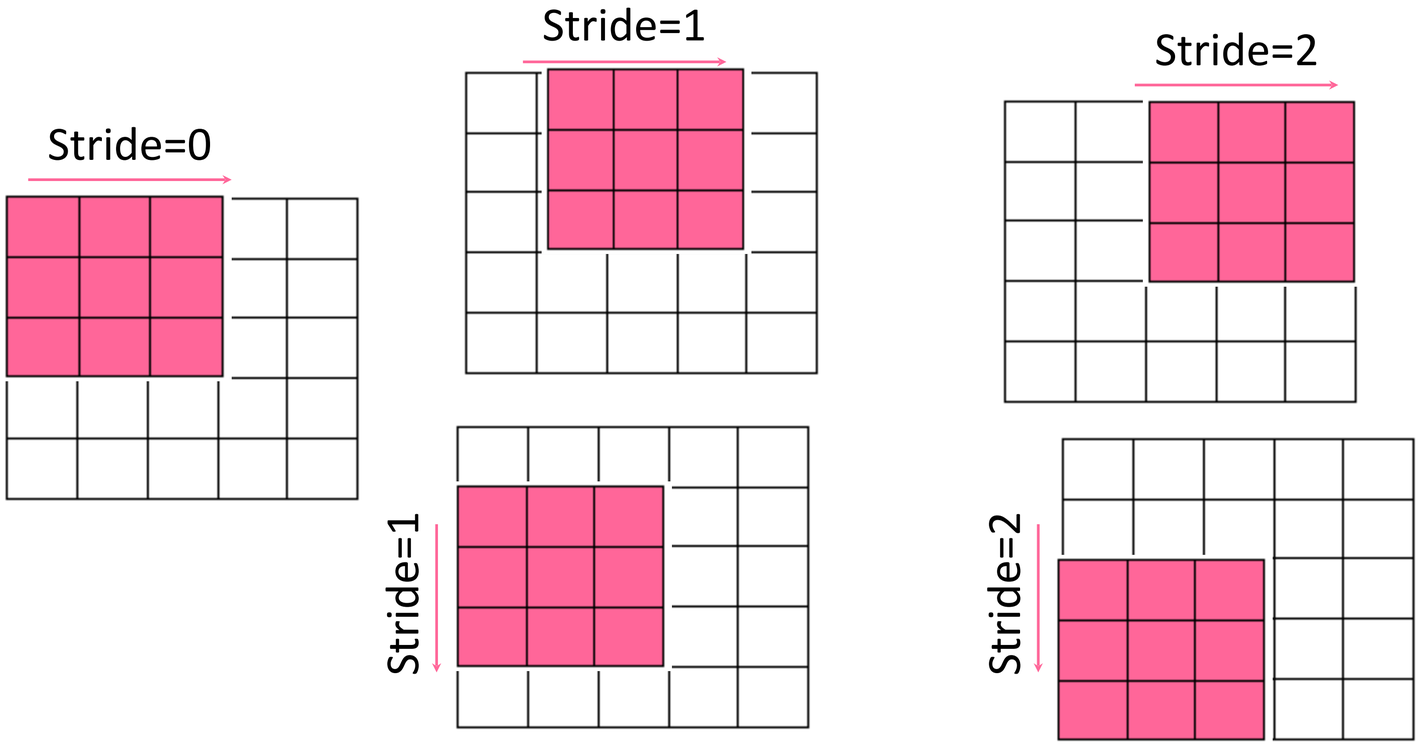
the content are taken from here and there

Strides

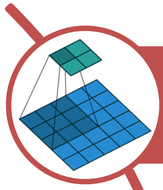


the content are taken from here and there

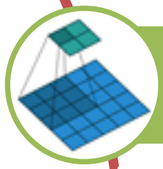
Stride- Visualisation



the content are taken from here and there



Used to avoid reduction in size



Retaining the original size

Protect the information along the border

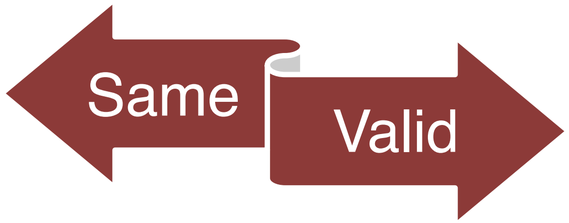
Input : image (h*w*d), filter($f_h * f_{w*d}$)

the content are taken from here and there

Types of padding

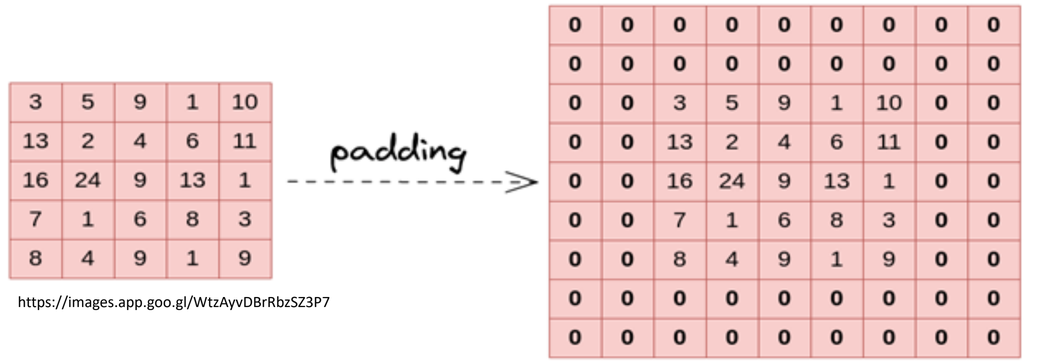
Same

after padding the size of image is uneffected.

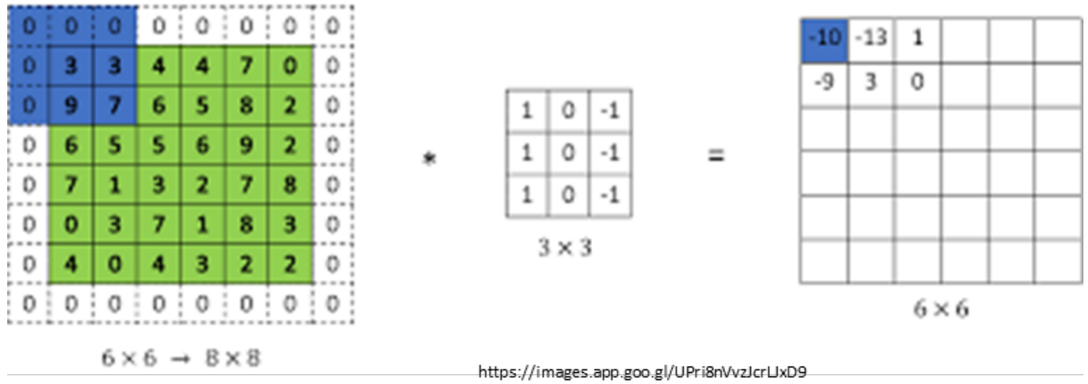


Valid

no padding done.



https://images.app.goo.gl/WtzAyvDBrRbzSZ3P7



https://images.app.goo.gl/UPri8nVvzJcrLJxD9

the content are taken from here and there

Padding Numerical

- If the image matrix is $6 * 6$
- If the size of the filter is $3 * 3 * 2$,
- If stride is 1 and padding is 1
- The output is calculated as

$$\left(\frac{h + (2 * p) - fh}{s} + 1\right) * \left(\frac{w + (2 * p) - fw}{s} + 1\right) * d$$

- So, output image is

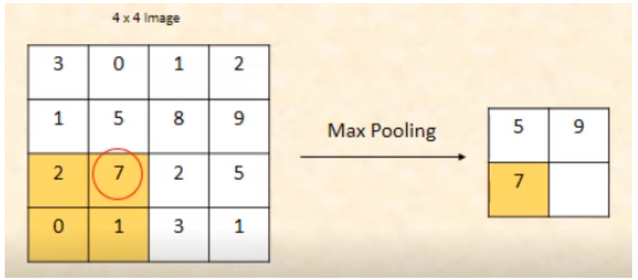
$$\left(\frac{6 + (2 * 1) - 3}{1} + 1\right) * \left(\frac{6 + (2 * 1) - 3}{1} + 1\right) * 2$$

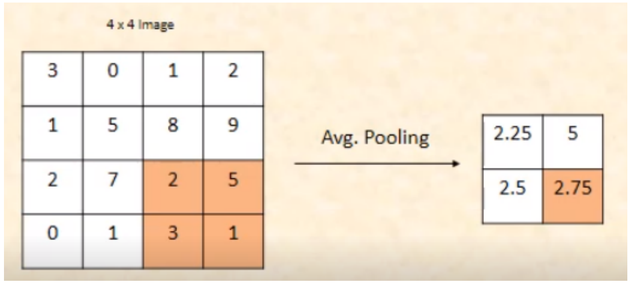
=6*6

the content are taken from here and there

Pooling Layer:

- Reduce number of parameters.
- In pooling layer, we have two hyperparameters filter size and stride which are fixed only once.



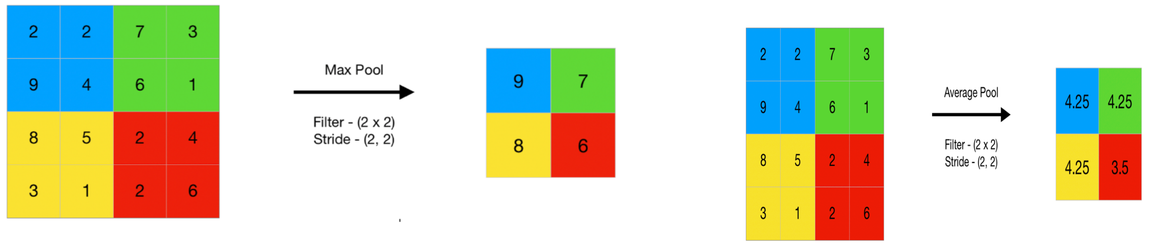


the content are taken from here and there

- Pooling layers are used to reduce the dimensions of the feature maps. Reduces the number of parameters to learn and the amount of computation performed in the network.

- The pooling layer summarises the features present in a region of the feature map generated by a convolution layer. So, further operations are performed on summarised features instead of precisely positioned features generated by the convolution layer. This makes the model more robust to variations in the position of the features in the input image.

- Average pooling computes the average of the elements present in the region of feature map covered by the filter. Thus, while max pooling gives the most prominent feature in a particular patch of the feature map, average pooling gives the average of features present in a patch.



the content are taken from here and there

Why Pooling

- Subsampling pixels will not change the object

bird

Subsampling

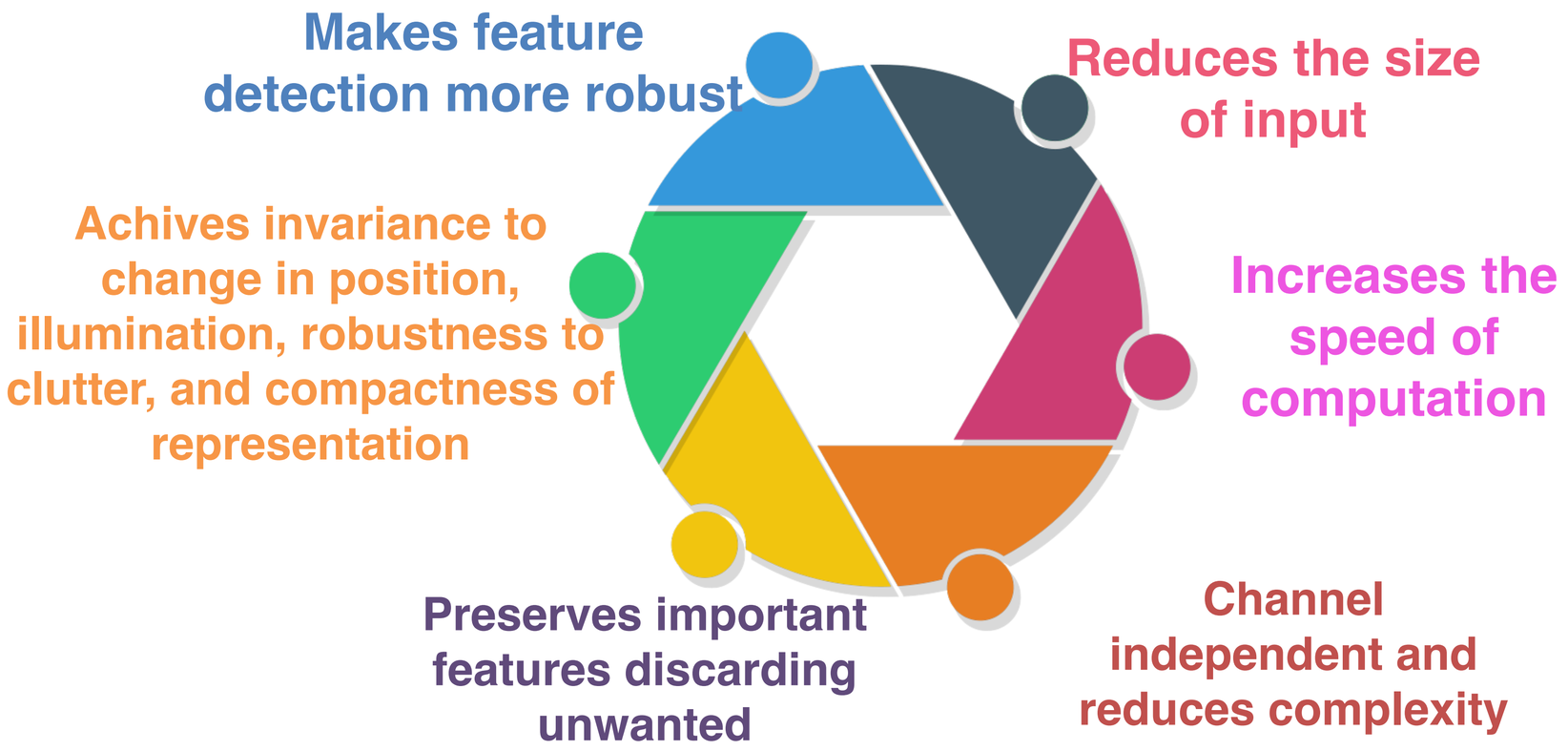
bird

We can subsample the pixels to make image smaller

→ fewer parameters to characterize the image

the content are taken from here and there

Importance of pooling



the content are taken from here and there

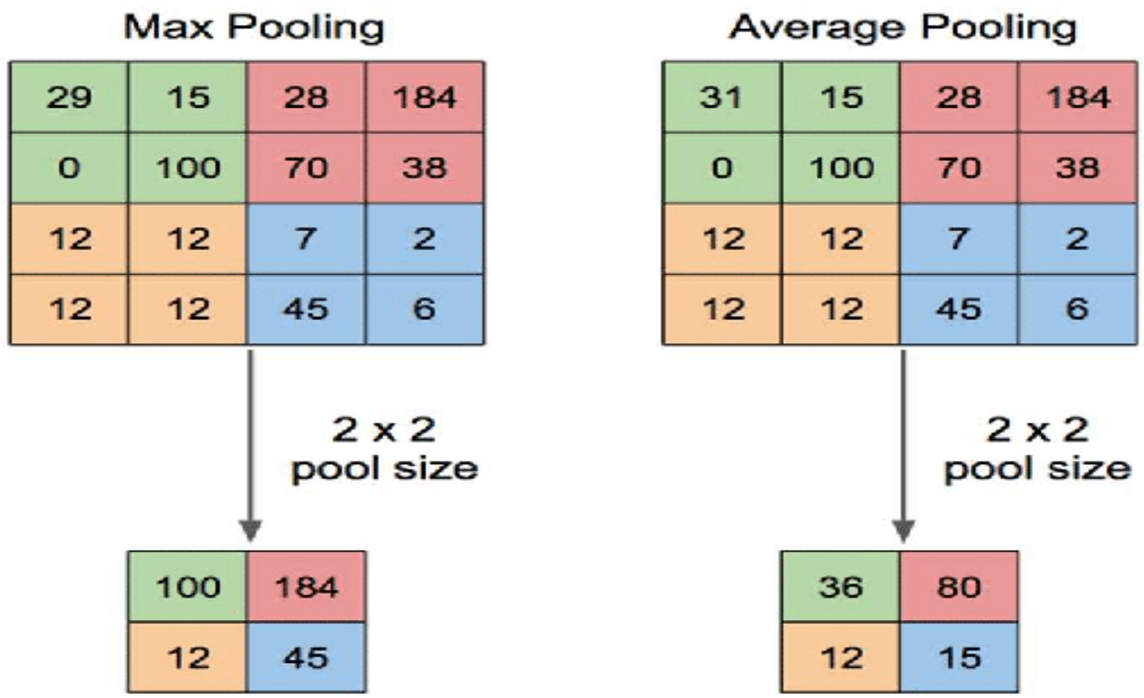
Types of Pooling

Max

Take the max value in each block

Average

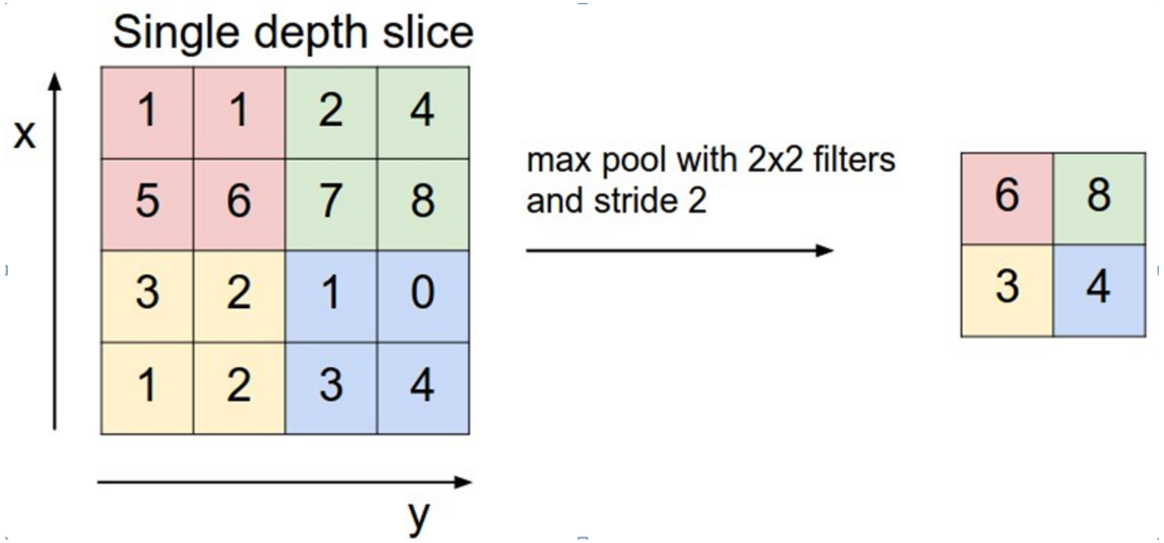
average all values in each block

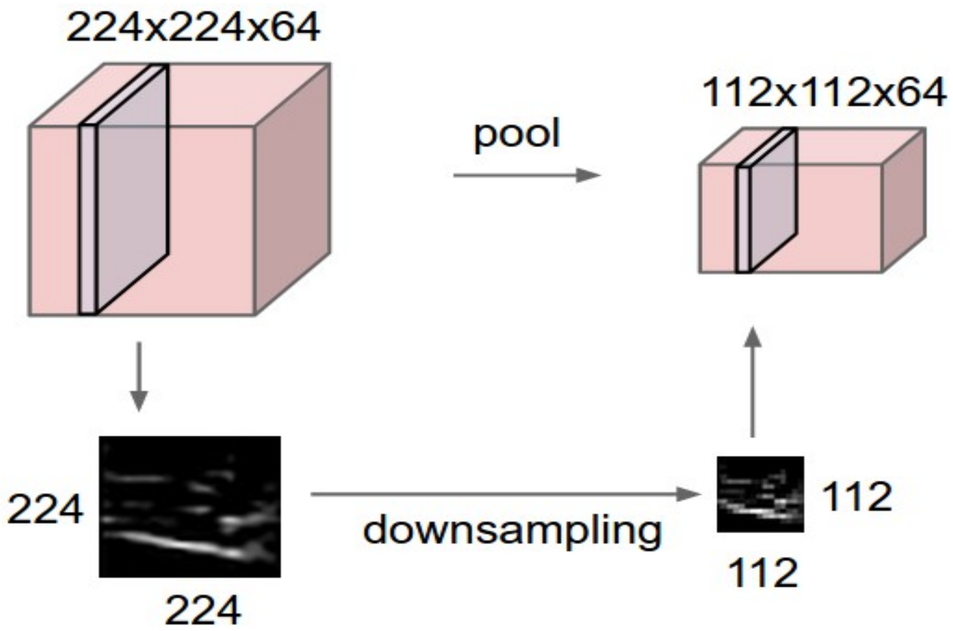


https://images.app.goo.gl/MWnbBi743TrW6EtYA

the content are taken from here and there

Visualisation of Pooling



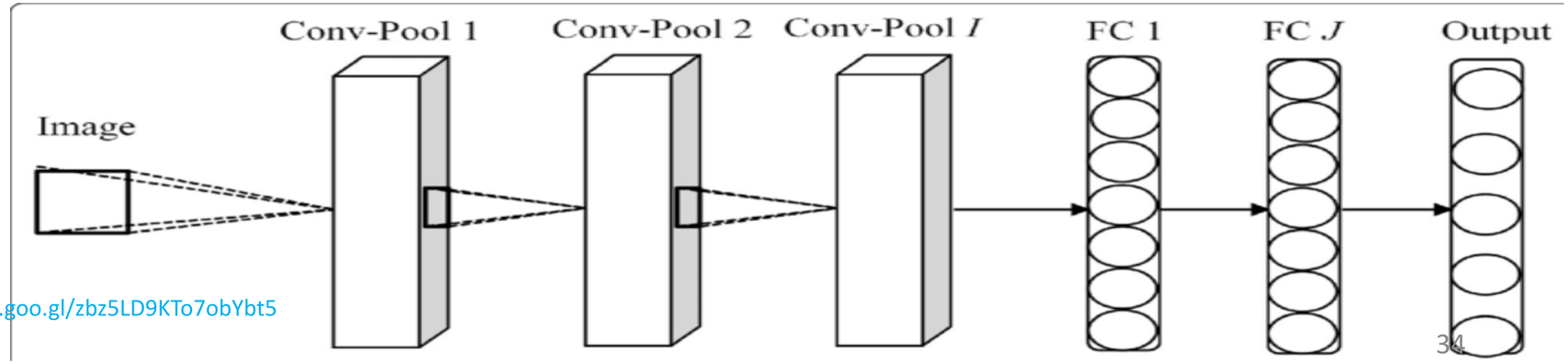


the content are taken from here and there

Flattening or FC layer

Hidden layers are called classification layer

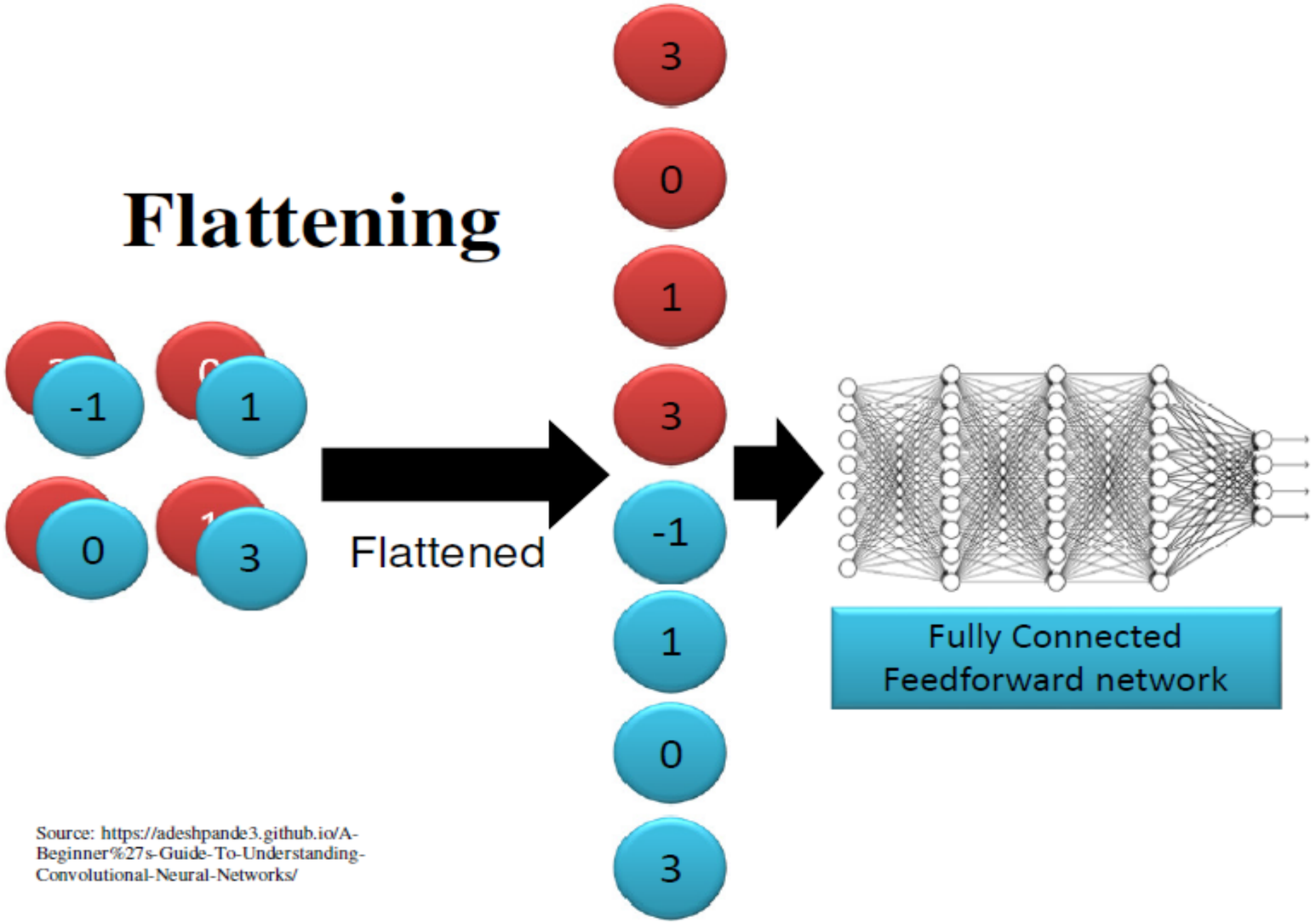
Number of neurons in the output layer is equal to the number of classes.



https://images.app.goo.gl/zbz5LD9KTo7obYbt5

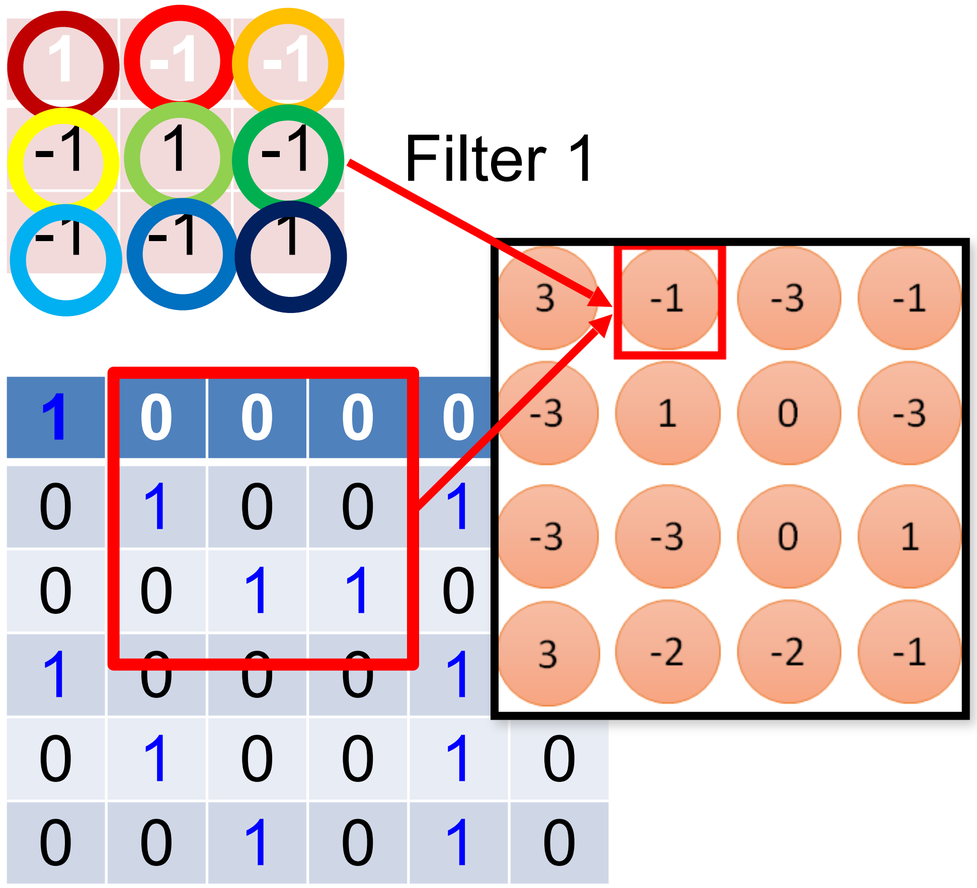
the content are taken from here and there

Flattening or FC layer



Source: https://adeshpande3.github.io/A-Beginner%27s-Guide-To-Understanding-Convolutional-Neural-Networks/

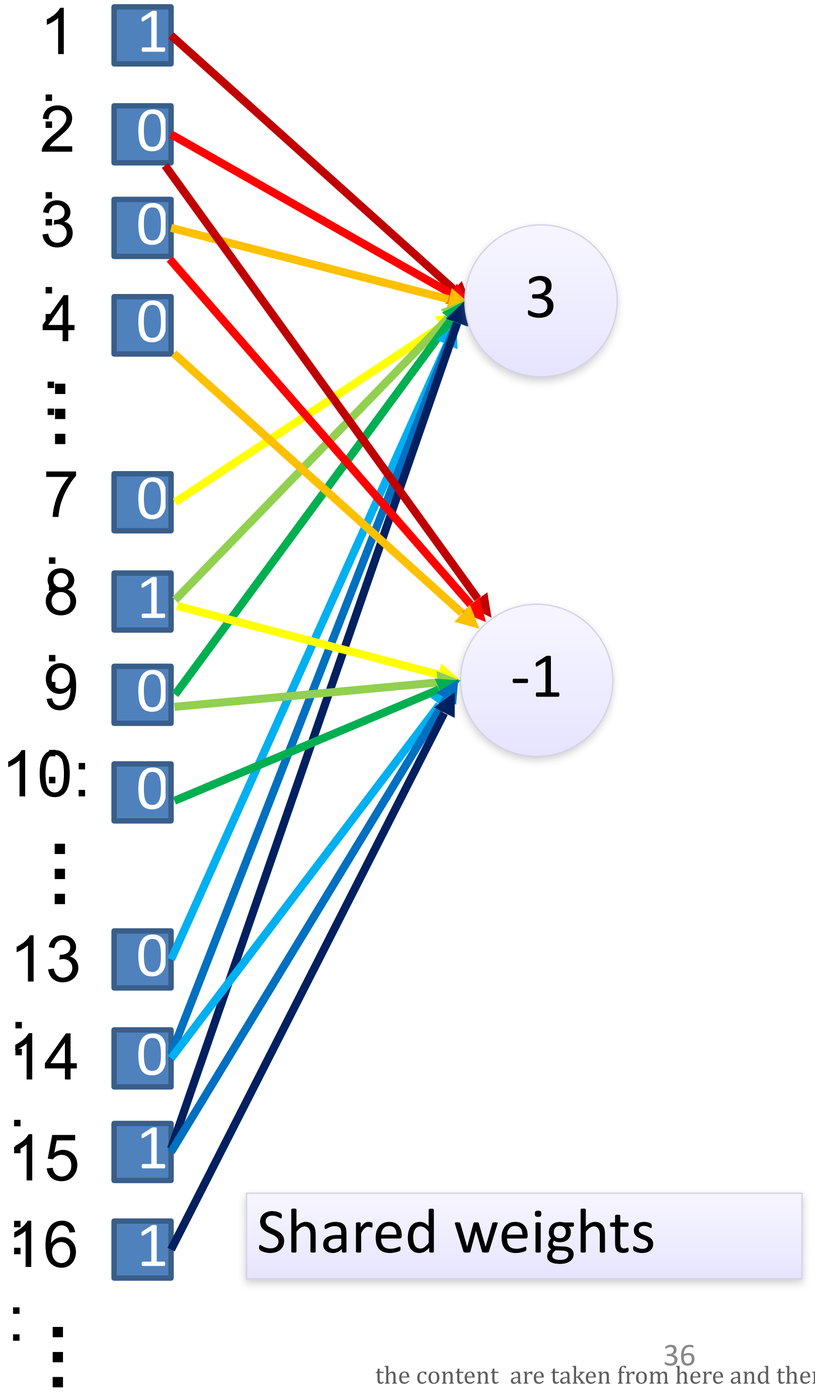
the content are taken from here and there



6 x 6 image

Fewer parameters

Even fewer parameters



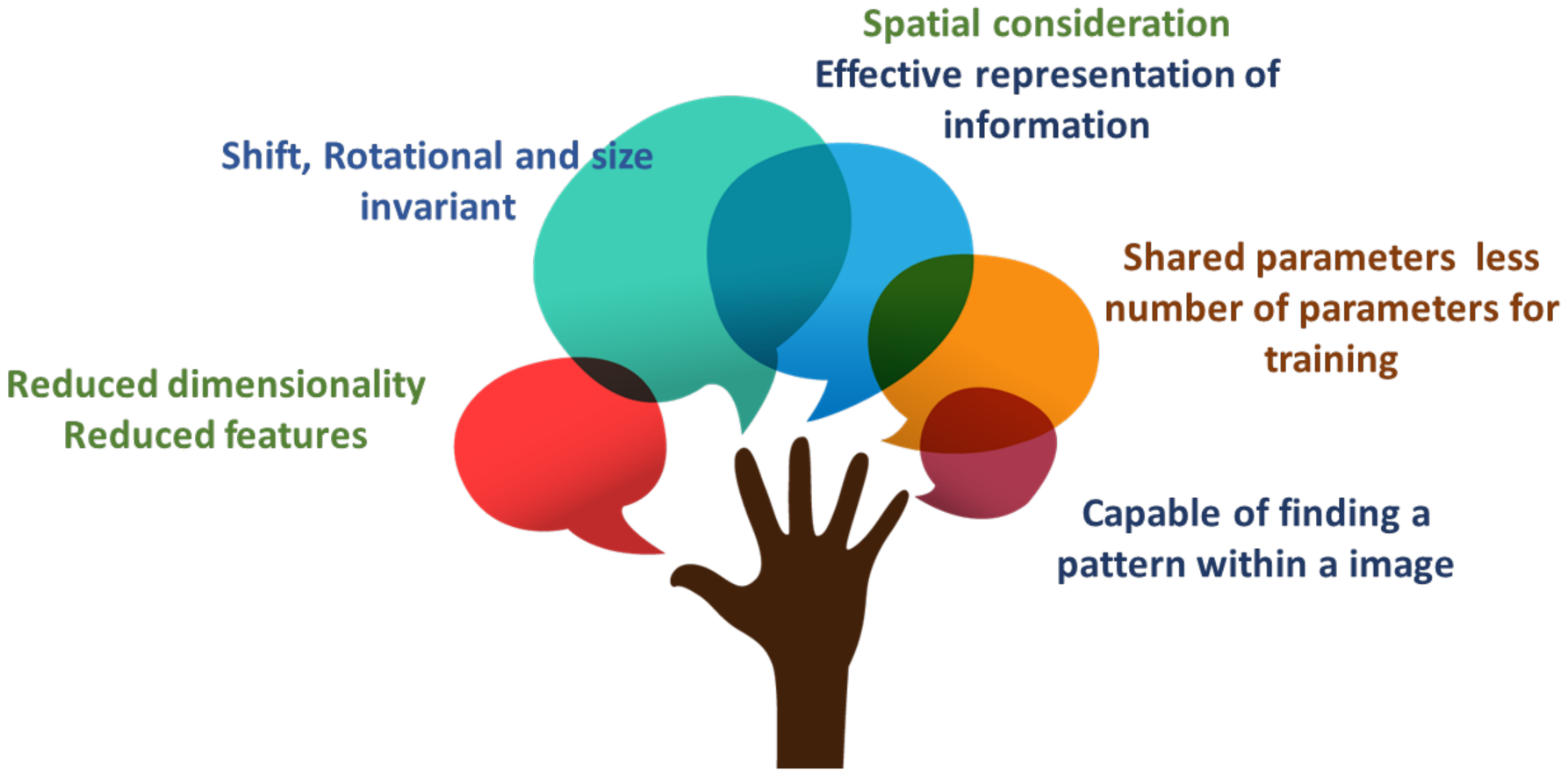
the content are taken from here and there

CNN Overview

- Better than fully connected network as it considers the spatial aspect .
- Use shared weights to reduce the number of parameters and hence number of trainable parameters are reasonable
- Represent a small region with fewer parameters hence reducing dimensionality
- The kernels (shared weights) are employed for learning the patterns, giving importance to only those features that are important
- The patterns in the input data will be further refined
- Little or no invariance to shifting, scaling, and other forms of distortion
- Network is capable of finding a object/ feature in any part of the image.

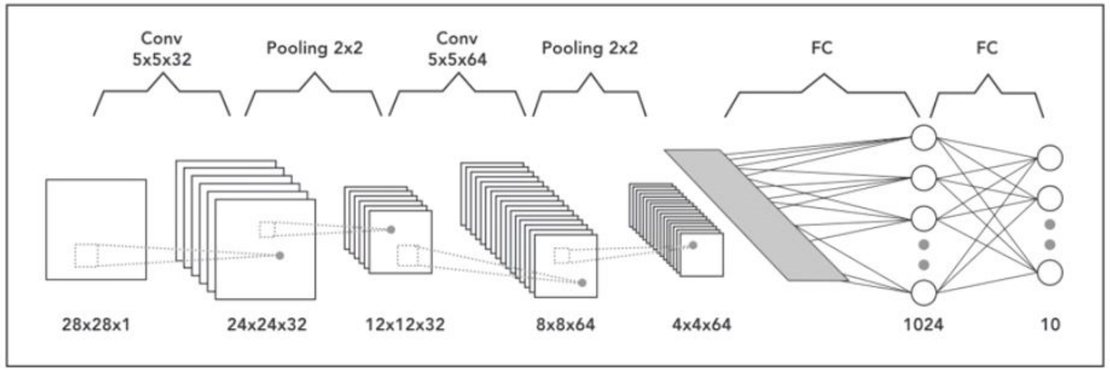
the content are taken from here and there

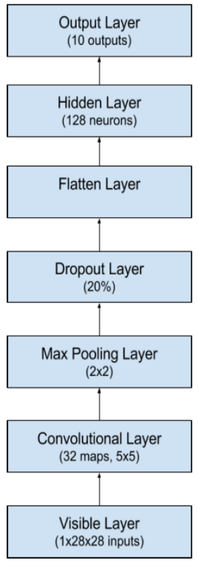
Convolution Layer



the content are taken from here and there

Architecture:





CNN Architecture for MNIST

the content are taken from here and there

Summary of CNN

- CNN helps in analyzing the imaginary.
- Q -> What human see vs what system see?
- Size of image reducing with increase in stride value.
- Padding the input image with zero - retains depth.
- Kernal/filter that extracts useful info (edges, color, ...)
- Pooling helps in reducing the spatial size of the image.
- Output volume is controled by 3 parameters (no. of filters, stride, zero padding) - ([W-F+2P]/S) + 1.
- Output layer in CNN is a FC layer.

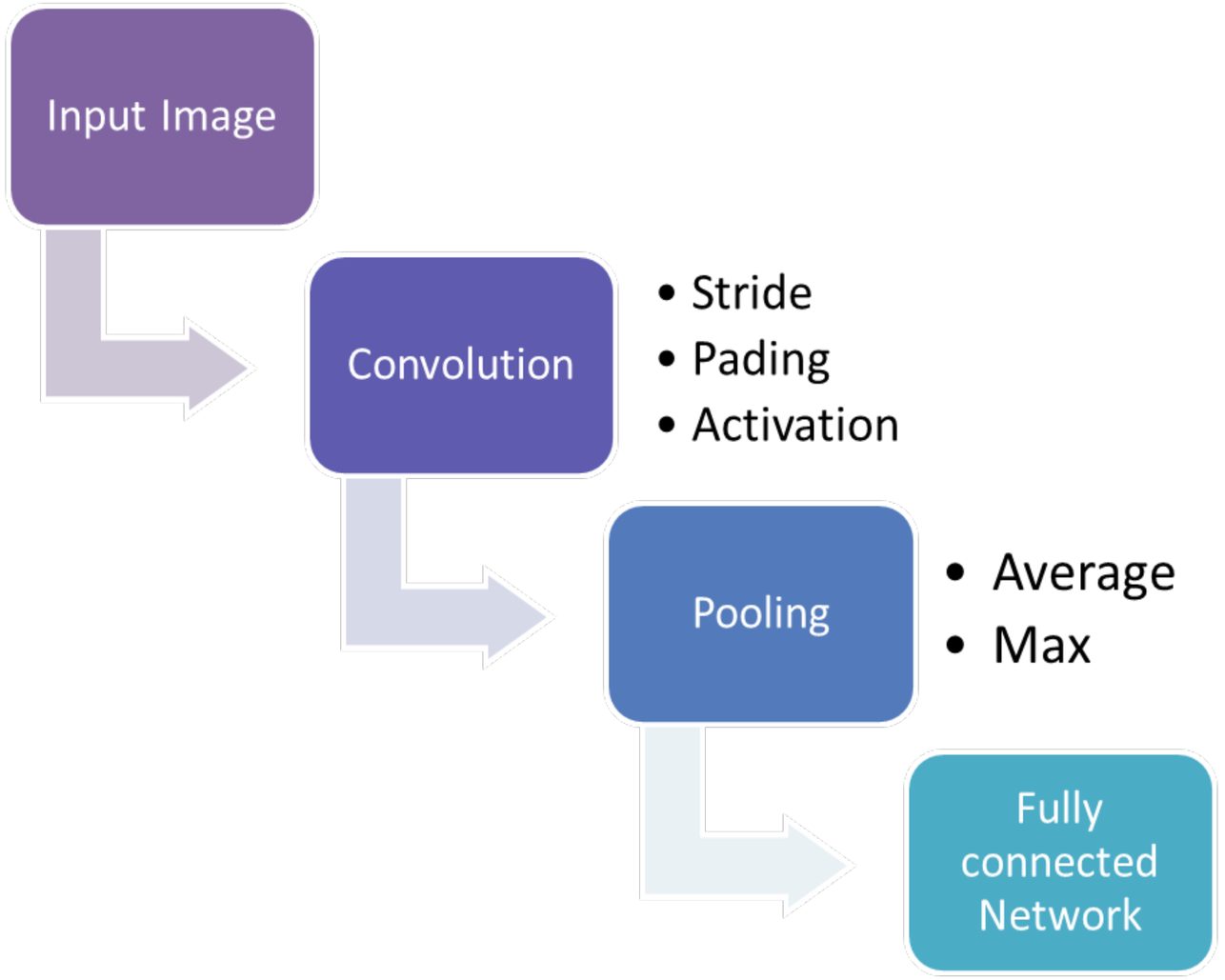
CNN: INPUT => CONV => RELU => FC => SOFTMAX

CNN that accepts an input, applies a convolution layer, then an activation layer, then a fully-connected layer, and, finally, a softmax classifier to obtain the output classification probabilities.

the content are taken from here and there

Architecture of CNN

- A typical CNN has 4 layers



the content are taken from here and there

Quiz

Ex:

Q1. Consider input image of 32X32X3 and we have 10 -> 5x5 filters with stride 1, pad 2.

What is output volume?

Q2 . Number of parameters in this layer?

the content are taken from here and there

Key

- (32+2*2-5)/1+1 = 32 spatially, so 32x32x10.

- each filter has 5*5*3 + 1 = 76 params (+1 for bias) => 76*10 = 760

the content are taken from here and there