Fundamentals of Convolutional Neural Networks

Mohammed Luqmaan Akhtar

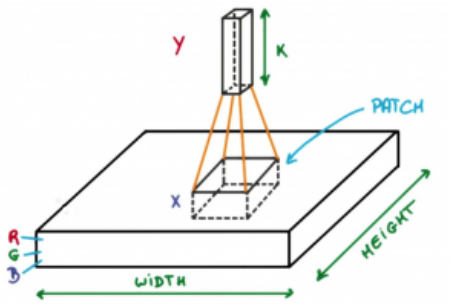
January 2026

Introduction

A Convolutional Neural Network is a type of Neural Network that is primarily used for images and visual data sets. It uses various *mechanisms* to identify smaller patterns in images, then group these patterns to find larger patterns and gradually identify the image itself. We use CNNs because using normal NNs would take a huge number of parameters. The use of CNNs decreases that number manifold. I will talk about these *mechanisms* in the following sections.

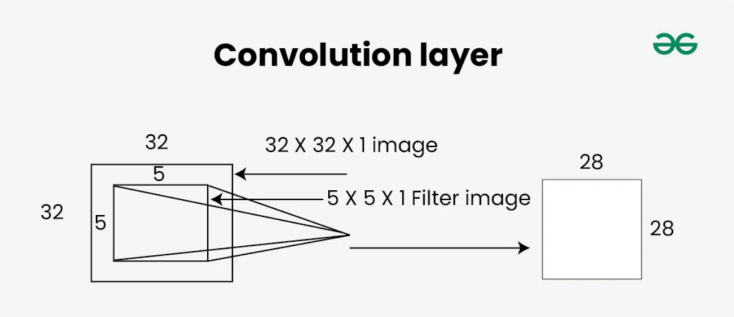
Convolution Operation

The convolutional operation applies a 2D matrix with *depth* same as that of the input image to the input. It computes a dot product of the matrix with various patches of pixels in the input image and gives us a output. This output is what identifies various patterns.



Filters and Kernels

Kernels or Filters are the 2D matrices which are the actual learnable parameters of the neural network. Their dot product with various patches of the input image is what helps recognise the patterns such as edges, shapes etc. The output matrix is called the feature map.

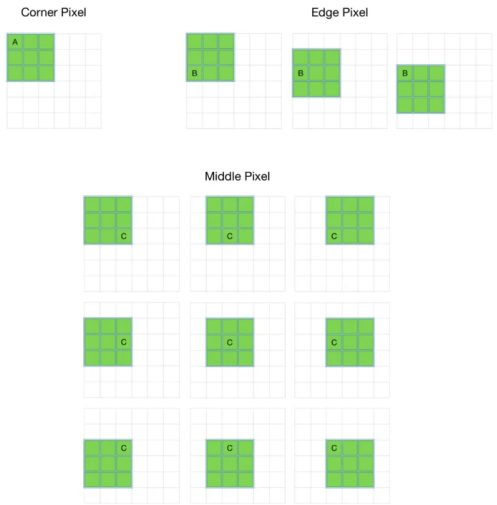


Feature Maps

The output of the dot product between Filters and the input image is called Feature Maps. Each feature map recognises different features such as either edges, shapes, corners or texture etc.

Padding

The output of te dot product is smaller in size than the input. This may lead to information loss, also some pixels may be used multiple times and some fewer times. To prevent this we do padding i.e. adding pixels to the boundaries of images so that the size of the output remains same. For example, without padding, an $8x8$ image with a $3x3$ filter will give a $6x6$ feature map; also the corner pixels will be included once, the edge pixels thrice and the middle pixel six times.

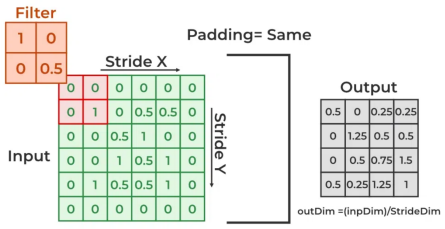


Valid Padding

No padding is done. This is useful when we want to reduce the spatial dimensions of the image.

Same Padding

Padding is done so that the feature map has the same size as the input. Pixels with value zero is added along the boundaries in appropriate number.



Strides

Strides define the steps of fiter movement, in both horizontal and vertical directon. Thus, they a play a role in the size of the feature map too.

Pooling

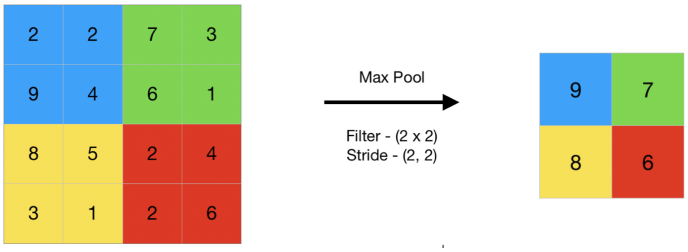
Pooling involves summarising the feature maps so as to reduce spatial dimensions which reduces the number of parameters and computational cost. It prevents overfitting and generalises features and also implements a hierarchy in which lower layers recognise finer details and higher ones capture abstract details.

Max Pooling

The pooling filter has a size and a stride and it gives the output according to it. In max pooling, it outputs the highest value in each patch.

Average Pooling

In average pooling, the pooling filter outputs the average value of each patch.



How CNNs actually work

CNNs are composed of a number of layers. The input is the first layer, it is then followed by a number of convolutional layers. Each convolutional layer is followed by an activation layer. These are then followed by pooling layers. Finally are a few layers which resemble a normal neural network connected to the final pooling layer. This in the end gives us the image classification.