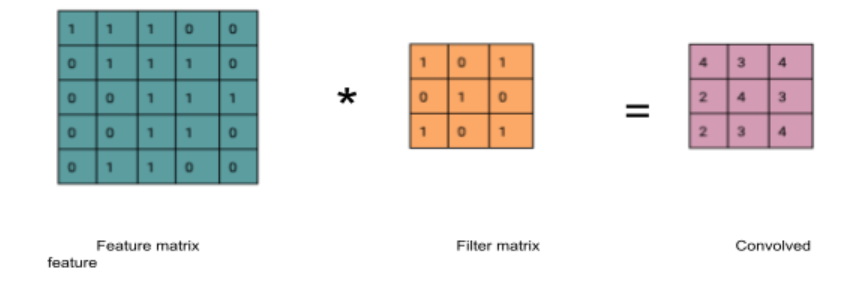
Convolution

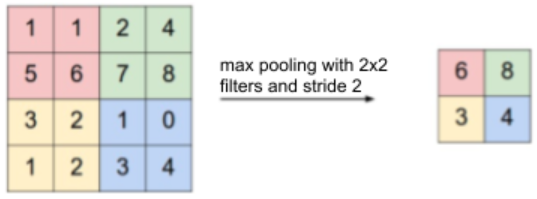
Convolutional layer, pooling layer and fully-connected layer are the three main types of layers used to build the CNN architecture.

**Convolutional layer:**This is the building block of CNNs. In mathematics, convolution refers to a mathematical operation on two functions to create a third function. Unsurprisingly, from the name, a convolution operator is used to extract features such as colour, edges orientation etc from input images in CNNs. It learns the image features using small squares of the input data to preserve the relationships between pixels.  A convolution is performed on the input data using a filter or kernel to produce a feature map. By including more layers, the network is able to learn more high-level features and have a better understanding of the images and better identifies unseen images. Considering a 5x5 image matrix with pixel values 0 and 1 and a 3x3 filter matrix, the convolution of both matrices result in a feature map as shown below



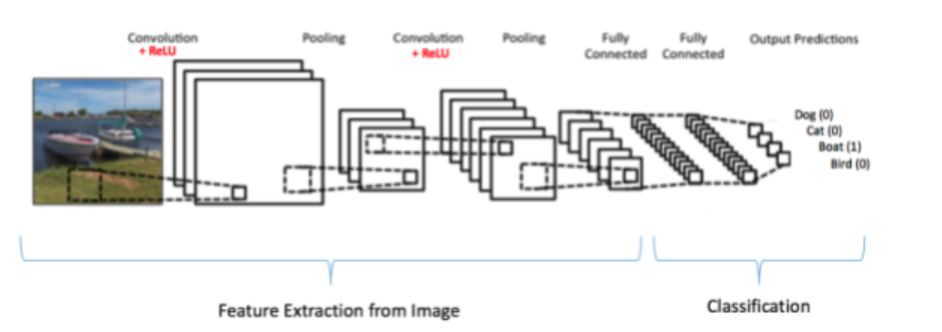
We slide the filter matrix over the feature matrix by a set number of pixels and perform an element wise multiplication for each position. It is important to note that different filters perform different operations by acting as feature detectors from the original images. The number of pixels or size of step the filter moves over is known as stride. To introduce non-linearity, a relu activation is applied after every convolution operation in the convolutional layer

**Pooling layer:** this is the layer periodically inserted in-between successive convolution layers where dimensionality reduction is performed by pooling. Pooling (also called downsampling) helps to control overfitting, train faster and reduce the number of parameters and computations in the network. Although this reduces dimensionality, the important information is still retained. Maximum, average and sum pooling are different types of pooling with max commonly used. Max pooling takes the maximum element from the feature map.



**Fully connected layer:** the features extracted from the convolution and pooling layers are fed into the fully connected layer to use for classification into different classes in the training set and also learning the non-linear combinations of the features since it is possible to obtain even better results with a combination of the features. Softmax is used as the activation function to obtain probabilities of the different classes which sum to 1.

Combining the layers described, for a CNN, an input image is provided for the convolution layer, filters are applied with strides to perform convolution, pooling is done for dimensionality reduction, more convolution and pooling layers can be added, the output of these layers is fed into a fully-connected layer and an activation function applied to obtain final probabilities for the classes.



Over time, numerous CNN architectures have been developed, most of which follow the fundamental method of applying convolution layers to the input and periodically downsampling. In this section, we briefly discuss some of these architectures.

**Image Net**

ImageNet is a large image dataset commonly used for research with the images labelled and organised by hierarchy. The dataset contains more than 14 million images labelled into over 20,000 categories. ImageNet Large Scale Visual Recognition Challenge (ILSVRC) is an annual competition (2010 - 2017) where computer vision algorithms especially for image classification and object detection  are evaluated  and compared using subsets of the ImageNet. The challenge aims to set state of the art results for various tasks and techniques. Some of the architectures we will discuss have proven to be successful in ILSVRC.

**AlexNet, VGG, ResNet and DenseNet**

AlexNet: Introduced in 2012, this network consists of 8 layers - 5 convolutional layers, 2 fully-connected layers and one fully-connected output layer won the ILSVRC challenge in 2012. Although similar to the LeNet architecture previously developed, AlexNet uses relu as the activation function as opposed to the tanh which caused the model to train about six times faster. Additionally, max pooling layers with a window size of 3x3 and a stride of 2 overlap in this architecture.

VGG - Visual Geometry Group: this was introduced by Simonyan and Zisserman in 2014. This architecture which was the second runner up for ILSVRC in 2014 is a network that is roughly twice as deep as AlexNet uses convolutions with 3x3 kernels with padding of 1, and a 2x2 max pooling matrix with a stride of 2.  It has 13 convolutional layers and 3 fully-connected layers alongside a relu activation function as in AlexNet.

ResNet - Residual Neural Network: Used in 2015 to win ILSVRC, identified that although deeper networks were performing better, they become difficult to optimise and suffer from vanishing gradients. To address these problems, skip connections were introduced in ResNet by authors from Microsoft Research with a speculation that deeper learners should be able to learn equally as shallow layers. This is a very deep network with 152 layers trained using skip connections such that the signal for a layer is included in the output of the layer located higher up.

**Transfer learning and image augmentation**

Transfer learning is a concept of exporting the knowledge obtained in a particular task to a new task. Using the features that an existing model learnt with a lot more data can improve generalization in a new task or setting. The motivation for transfer learning stemmed from the fact that although many supervised models require a large number of labeled data, many scenarios typically do not have this as it is time consuming and difficult to label data points.  When performing transfer learning, what, when and how to transfer are important questions that must be answered. The portion of the knowledge learnt by the source model that will be beneficial to the new model to improve performance should be identified. Also, although transfer learning is an option to use when training new models, it might not always improve performance in the new model instead, it might even contribute negatively. Finally, it is helpful to identify how the knowledge will be transferred between models using different techniques and algorithms. For example, in computer vision, pre-trained models for challenging tasks using ImageNet are readily available to be used in transfer learning. Pre-trained models are models that have been trained on large datasets regularly reused in similar domains and tasks. Pre-trained models can serve as feature extractors or even used as they are. As explained earlier, the convolution layers that receive the inputs in CNNs learn a lot of low-level features in images, more complex features learnt in the middle layers and the results of these computations are interpreted with the output layer. With a proper understanding of this, the relevant portions can be selected and integrated when creating new networks by either freezing the existing weights to prevent updates or update these weights when training the new model hence indirectly serving as a weight initialisation phase for the new model. Some of the models discussed above can be used for transfer learning in computer vision tasks.

Image augmentation in computer vision is not new in fact, it is extremely important for sparse datasets. As mentioned earlier, augmentation helps to create a larger dataset by adding noise, rotating, flipping, cropping, rotating etc the training dataset. This helps reduce overfitting and improve the model’s capacity to generalise to even unfamiliar scenarios in new unseen data.

**Motivations and Terms**

In object localisation tasks, the location of objects in images are identified and put in boundary boxes while classification assigns a label to each image. Object detection often referred to as object recognition  involves a combination of object classification and object localisation such that different objects in an image are found and classified. In object detection, the output is variable in length because the number of objects detected in different images may change. Object detection can be used for face detection as seen in some cameras, counting, visual search engines such as that of Pinterest, aerial image analysis. A problem with obtaining a variable number of objects is that it becomes difficult to obtain fixed-sized vectors; however, sliding windows computed convolutionally  are commonly used to resolve this.

Region-based Convolutional Neural Network (R-CNN) is a well-known architecture used for object detection such that a selective search algorithm generates about 2000 region proposals which are later passed to a CNN for feature extraction. An SVM (Support Vector Machine) is used to classify objects in the region proposal then a boundary box regression used to localise objects present. Region proposals are smaller parts of the original image identified to possibly contain the objects being searched for.  Bounding box regressor uses a scale-invariant linear regression model to create bounding boxes for the objects. They learn a target transformation between the predicted proposal and the ground truth When training the model, pairs of predicted and ground truth of four localisation dimensions are used such that the predicted bounding box *p = (px, py, pw, ph)* with *px* and *py* as the center coordinates, *pw* the width and *ph* the height while the ground truth is *g= (g, gy, gw, gh)*. The transformation between both boxes using the linear regressor can be represented as: *ĝx =pwdx(p) + px, ĝy =phdy(p) + py, ĝw =pwexp(dw(p)), ĝh =phexp(dh(p))*

A downside of R-CNN is that it is computationally expensive and slow due to the many forward computations performed by CNNs on the proposed regions for a single image.

Another object detection algorithm that is less computationally expensive is a type of single shot detection (SSD) algorithm called You Only Look Once “YOLO”. It is a cutting-edge detection algorithm that can identify distinct objects within the space of an image. It looks at the image once, divides it into grid cells, which are responsible for predicting bounding boxes, and output a score known as the Intersection Over Union (IOU). For each bounding box, the grid cells also predict a class alongside the probability distribution over all possible classes. The class-specific confidence score is a multiplication of the individual box confidence predictions and the conditional class probabilities.