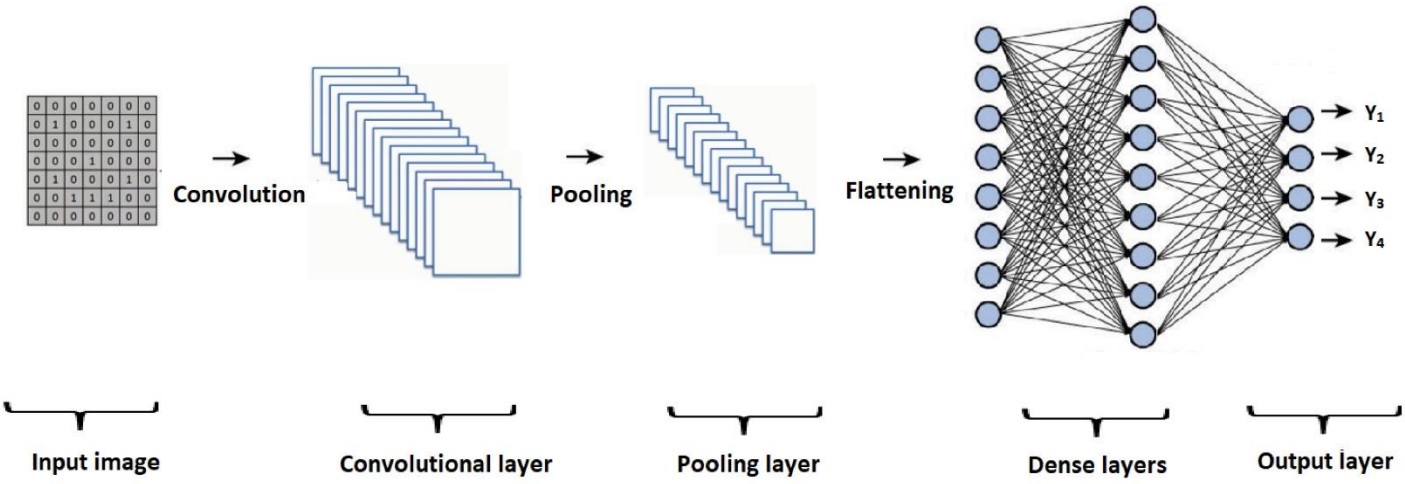
SEMANTIC SEGMENTATION & FCNs IN AUTONOMOUS VEHICLES

Autonomous vehicles are a complex system of sensors that includes lidars and cameras. These sensors act as their *eyes* which help them drive and spot other impending objects in their way. A self-driving car is a functioning example of hardware and software amalgamation. From the tech point-of-view, an autonomous vehicle mostly works using Image Recognition.

Image Recognition in Autonomous Vehicles:

The most classical and approachable way to solve an Image Recognition Problem would be to use a *typical* Convolutional Neural Network. But in Autonomous Vehicles our problem is to not only classify an object in the image but also to find where in the image does the object exists.

In a typical CNN, we have our convolutional layers which are followed by a Dense (Fully Connected) Layer and a Sigmoid/SoftMax activation function depending upon the number of classes we have.



1 Architecture of a CNN

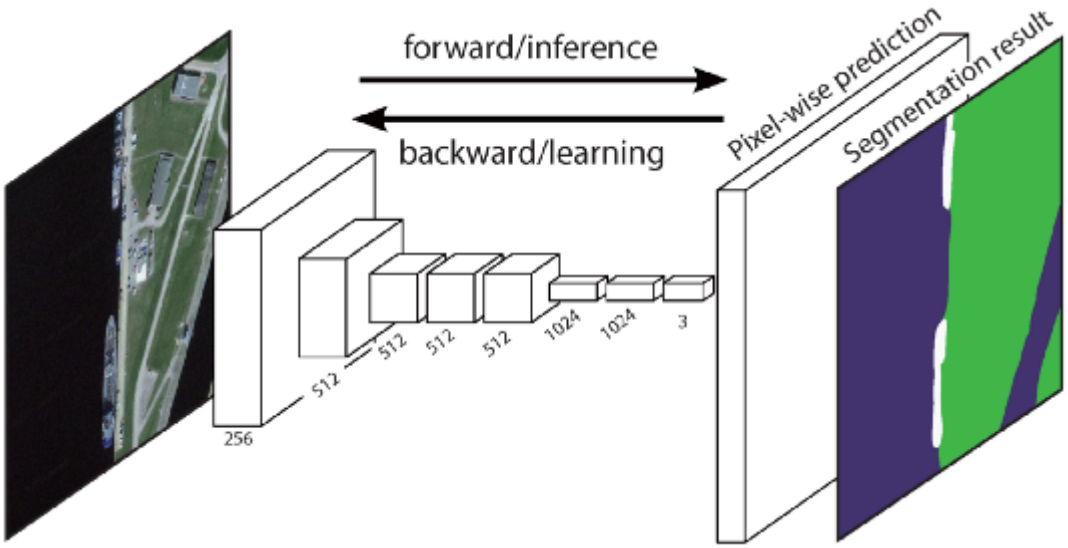
When we transition from the convolution and pooling layers to the dense layer, we flatten the image i.e. we convert the image from 3-D to 1-D which leads to a loss in spatial information.

Need for Fully Convolutional Networks (FCN):

To counter the above-stated problem we can use a Fully Convolutional Network. They are advantageous, as they retain the spatial information throughout the entire network while doing the convolutions, and can work on images of any size.

They use several techniques to achieve the aforementioned results:

- Replace fully connected layers with a 1x1 Convolution Layer which helps retain the dimension of the 3-D image.
- Use Up-Sampling by doing Transpose Convolutional Layers.
- Skip Connections that allow us to use information from multiple resolutions.



2Working of an FCN

Structure of a Fully Convolutional Networks (FCN) :

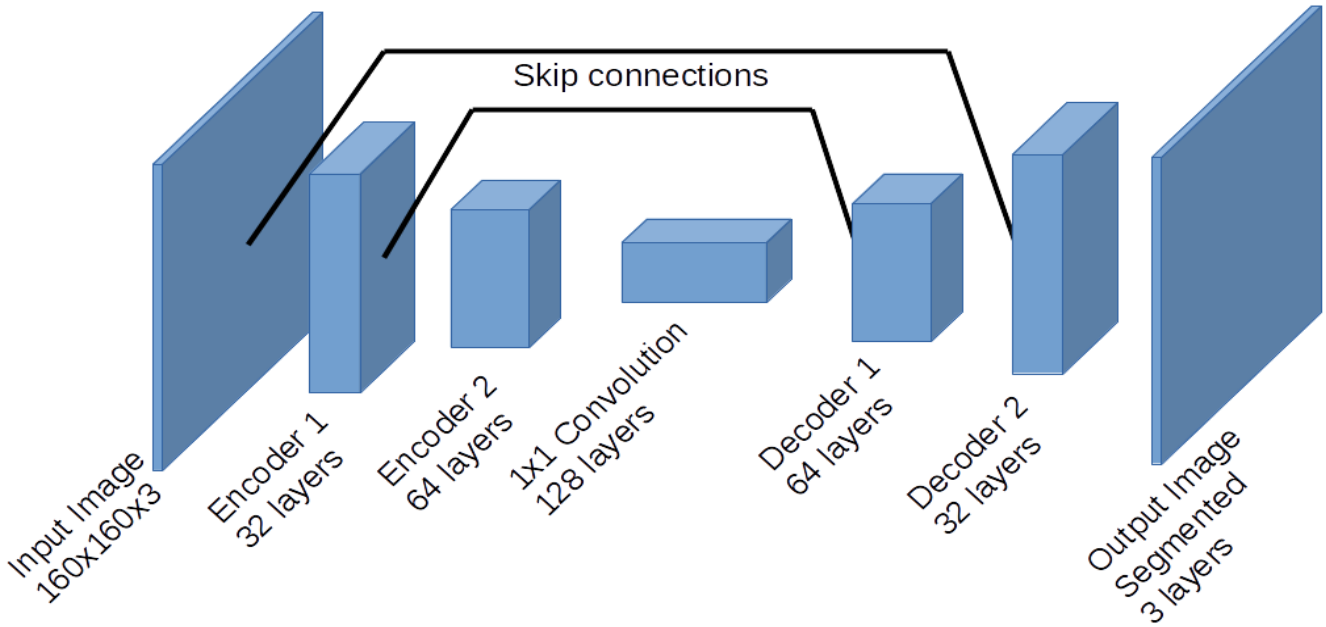
It consists of two parts – an Encoder and a Decoder.

The Encoder is a series of Convolutional layers trained on ImageNet like VGG and ResNet whose major goal is to extract features from the image.

The Decoder upscales Encoder's output to the size of the original image. This is achieved by performing a reverse convolution wherein the forward and backward passes are swapped.

This results in the segmentation and the prediction of each pixel of the original image.

Even when we use the decoder to recreate the original input image, we lose some information. This is where the skip connections technique. Skip connections is a technique where we connect the output of a layer to a non-adjacent layer, allowing for more data to be extracted from different resolutions resulting in more segmented decisions.



3Architecture of FCN

Semantic Segmentation and its need:

Semantic Segmentation can be defined as the process of assigning meaning to every part of the object. It is performed at the pixel level.

A self-driving car needs to react to new events instantly so that it can ensure the safety of everyone. To segment, an image in real-time is a strong requirement in self-driving applications.

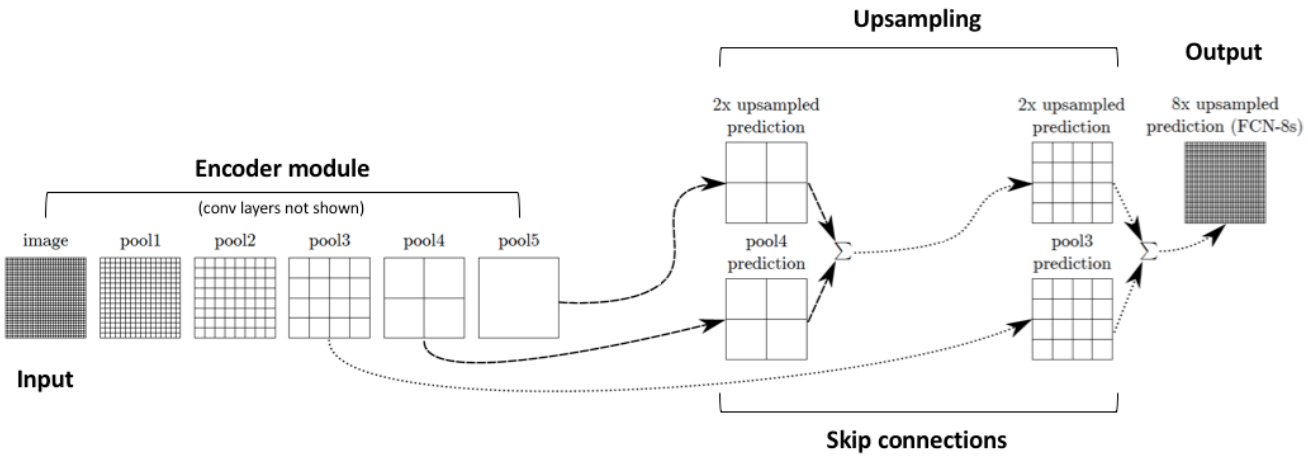
Semantic Segmentation comes in when bounded boxes fail. Bounded Boxes aren't good with curved objects, as the pixels can overlap. Rather than just slicing sections into bounding boxes we can derive valuable information about every pixel.

Semantic Segmentation helps us associate a region of an image to a class of the image, whereas FCN helps us identify whether an object is present in an image or not. With the help of sensors, we can further calculate their respective distances.

4Image Segmentation

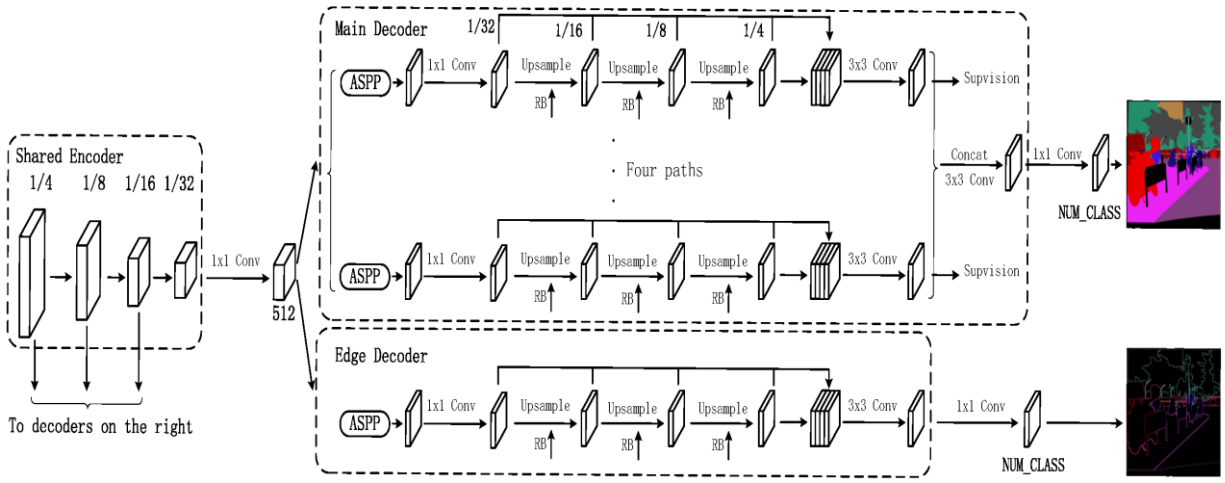
Down-Sampling/Up-Sampling:

It gets very computationally expensive to label this data and to maintain spatial information in each layer. We need to down-sample the feature map and up-sample the layer. Down-Sampling is usually achieved by doing a MaxPool.



5Overview of Semantic Segmentation

Multiple decoder can be used in a network where one detector is used for segmentation and the other is used for depth.

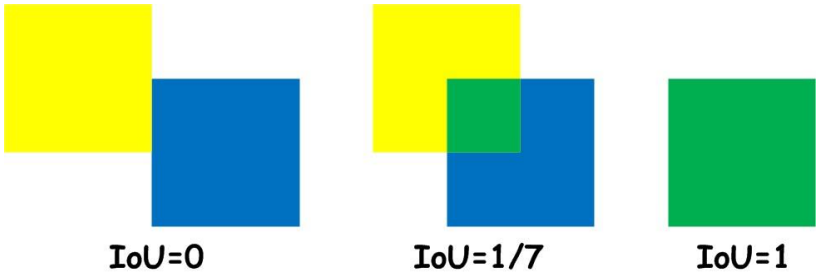


6Example of 2-Decoder Network.

Intersection over Union (IoU):

It is the ratio of Intersection over Union. We take the Intersection of the actual pixel where the object is present and the assigned pixel (pixel assigned by the object) of the object. If a pixel is in any of the two sets, we use it in the Union set.

Intersection set <= Union set and hence Intersection Set / Union Set <=1, We can calculate mean IoU for a network and use it as an evaluation metric. The lower the IoU, the worse the prediction result.



7Example of IoU Metric