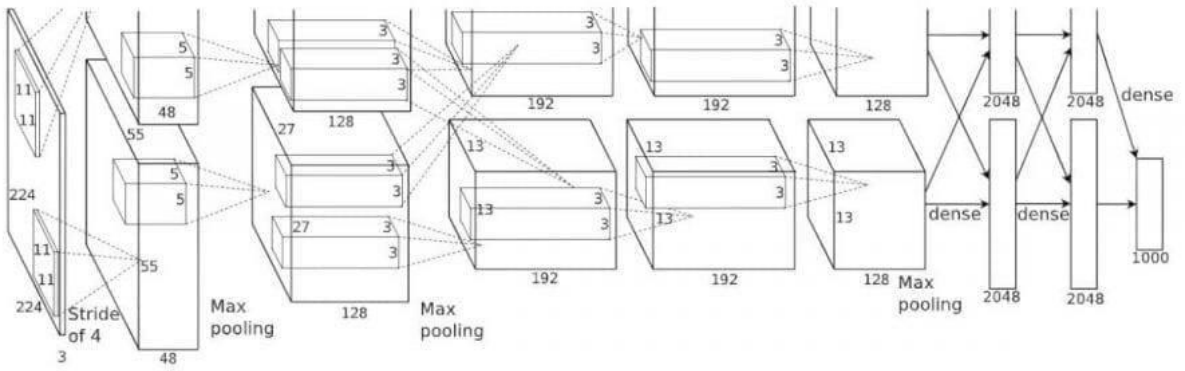
Day 14: DATA SCIENCE INTERVIEW PREPARATION

Q1. What is Alexnet?

Answer: The Alex Krizhevsky, Geoffrey Hinton and Ilya Sutskever created the neural network architecture called 'AlexNet' and won Image Classification Challenge (ILSVRC) in 2012. They trained their network on 1.2 million high-resolution images into 1000 different classes with 60 million parameters and 650,000 neurons. The training was done on two GPUs with split layer concept because GPUs were a little bit slow at that time.

AlexNet is the name of convolutional neural network which has had a large impact on the field of machine learning, specifically in the application of deep learning to machine vision. The network had very similar architecture as the LeNet by Yann LeCun et al. but was deeper with more filters per layer, and with the stacked convolutional layers. It consist of (11×11, 5×5,3×3, convolutions), max pooling, dropout, data augmentation, ReLU activations and SGD with the momentum. It attached with ReLU activations after every convolutional and fully connected layer. AlexNet was trained for six days simultaneously on two Nvidia Geforce GTX 580 GPUs, which is the reason for why their network is split into the two pipelines.

Architecture



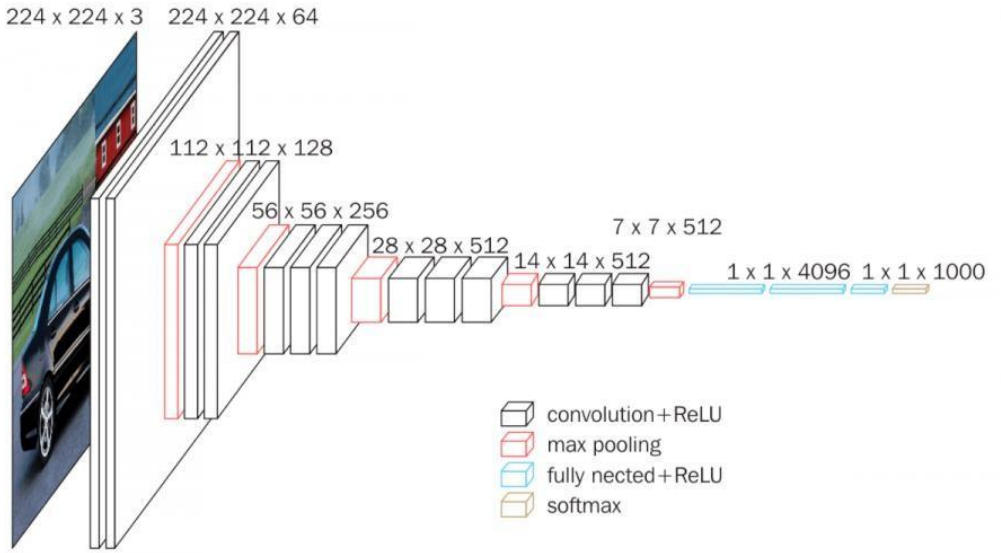
AlexNet contains eight layers with weights, first five are convolutional, and the remaining three are fully connected. The output of last fully connected layer is fed to a 1000-way softmax which produces a distribution over the 1000 class labels. The network maximises the multinomial logistic regression objective, which is equivalent to maximising the average across training cases of the log probability of the correct label under the prediction distribution. The kernels of second, fourth, and the fifth convolutional layers are connected only with those kernel maps in the previous layer which reside on the same GPU. The kernels of third convolutional layer are connected to all the kernel maps in second layer. The neurons in fully connected layers are connected to all the neurons in the previous layers.

In short, AlexNet contains five convolutional layers and three fully connected layers. Relu is applied after the very convolutional and the fully connected layer. Dropout is applied before the first and second fully connected year. The network has the 62.3 million parameters and needs 1.1 billion computation units in a forward pass. We can also see convolution layers, which accounts for 6% of all the parameters, consumes 95% of the computation.

Q2. What is VGGNet?

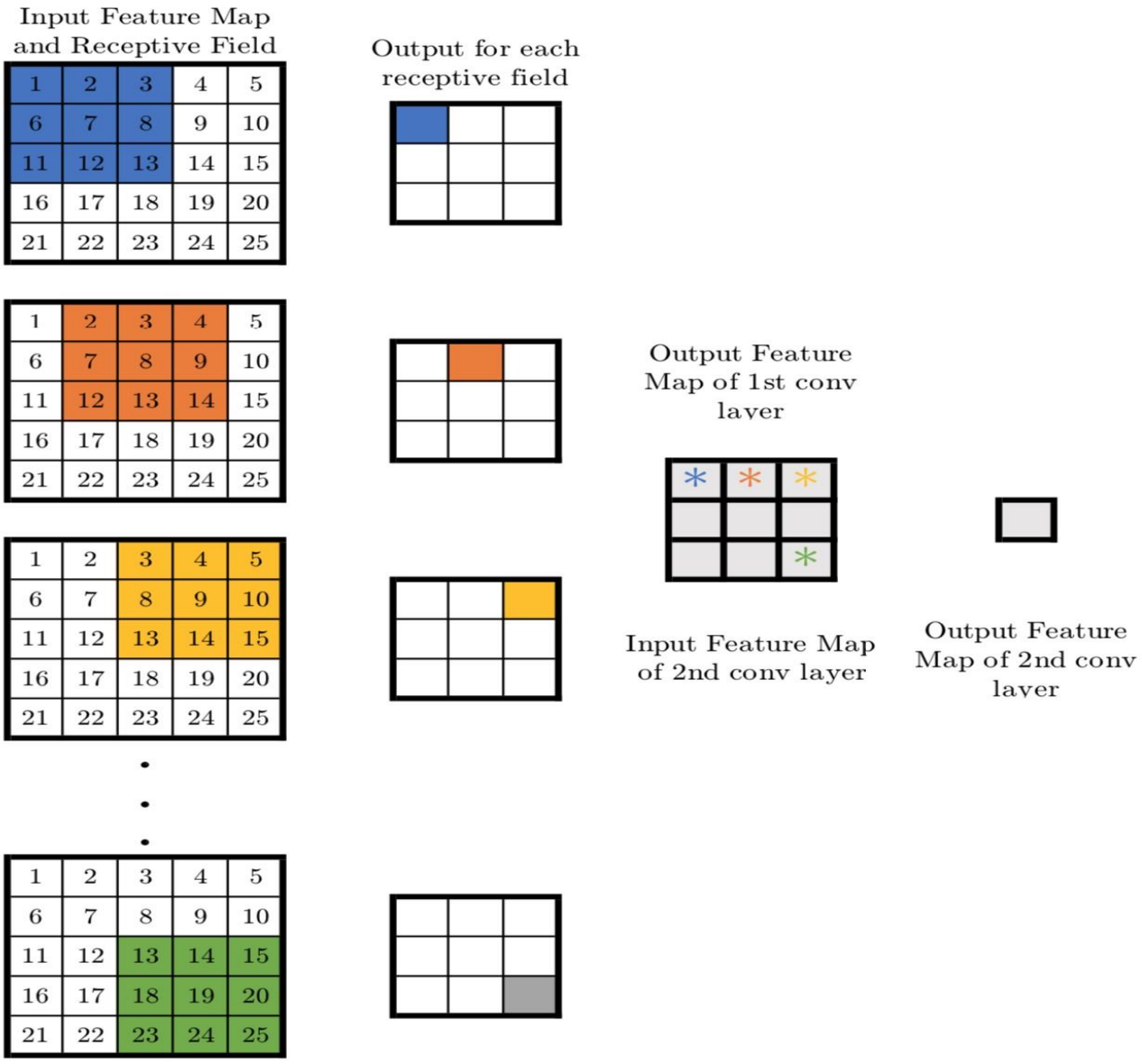
Answer: VGGNet consists of 16 convolutional layers and is very appealing because of its very uniform architecture. Similar to AlexNet, only 3x3 convolutions, but lots of filters. Trained on 4 GPUs for 2– 3 weeks. It is currently the most preferred choice in the community for extracting features from images. The weight configuration of the VGGNet is publicly available and has been used in many other applications and challenges as a baseline feature extractor. However, VGGNet consists of 138 million parameters, which can be a bit challenging to handle.

There are multiple variants of the VGGNet (VGG16, VGG19 etc.) which differ only in total number of layers in the networks. The structural details of the VGG16 network has been shown:



The idea behind having the fixed size kernels is that all the variable size convolutional kernels used in the Alexnet (11x11, 5x5, 3x3) can be replicated by making use of multiple 3x3 kernels as the building blocks. The replication is in term of the receptive field covered by kernels.

Let's consider the example. Say we have an input layer of the size 5x5x1. Implementing the conv layer with kernel size of 5x5 and stride one will the results and output feature map of (1x1). The same output feature map can obtain by implementing the two (3x3) Conv layers with stride of 1 as below:

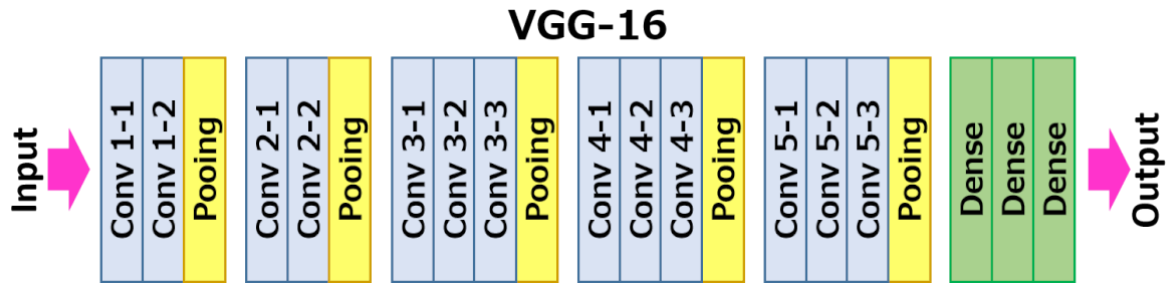


Now, let's look at the number of the variables needed to be trained. For a 5x5 Conv layer filter, the number of variables is 25. On the other hand, two conv layers of kernel size 3x3 have a total of 3x3x2=18 variables (a reduction of 28%).

Q3. What is VGG16?

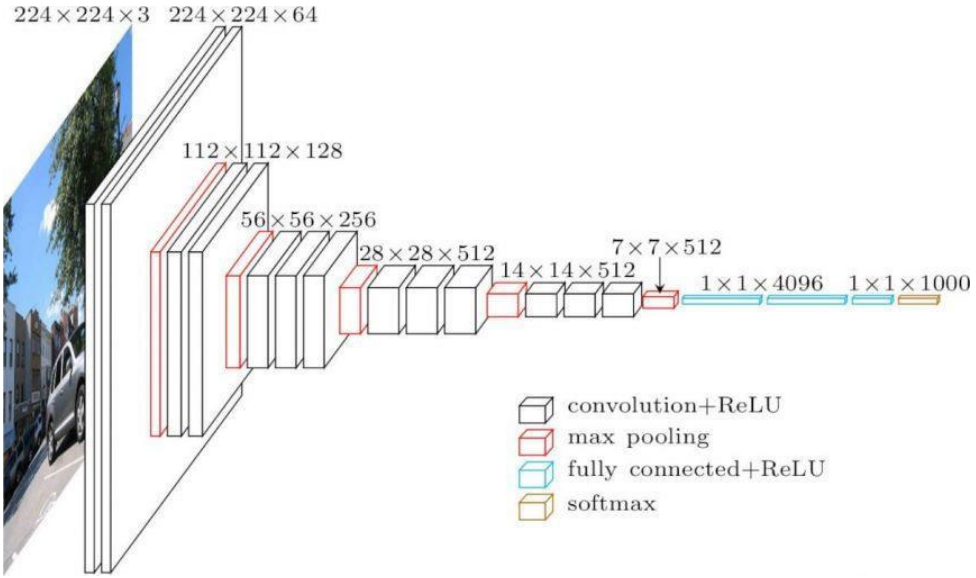
Answer: VGG16: It is a convolutional neural network model proposed by the K. Simonyan and A. Zisserman from the University of Oxford in the paper "Very Deep Convolutional Networks for the Large-Scale Image Recognition". The model achieves 92.7% top 5 test accuracy in ImageNet, which is the dataset of over 14 million images belonging to the 1000 classes. It was one of famous model submitted to ILSVRC-2014. It improves AlexNet by replacing the large kernel-sized filters (11 and 5 in the first and second convolutional layer,

respectively) with multiple 3×3 kernel-sized filters one after another. VGG16 was trained for weeks and was using NVIDIA Titan Black GPU's.



The Architecture

The architecture depicted below is VGG16.



The input to the Cov1 layer is of fixed size of 224 x 224 RGB image. The image is passed through the stack of convolutional (conv.) layers, where the filters were used with a very small receptive field: 3×3 (which is the smallest size to capture the notion of left/right, up/down, centre). In one of the configurations, it also utilises the 1×1 convolution filter, which can be seen as the linear transformation of the input channels. The convolution stride is fixed to the 1 pixel, the spatial padding of the Conv. layer input is such that, the spatial resolution is preserved after the convolution, i.e. the padding is 1-pixel for 3×3 Conv. layers. Spatial pooling is carried out by the five max-pooling layers, which follows some of the Conv. Layers. Max-pooling is performed over the 2×2-pixel window, with stride 2.

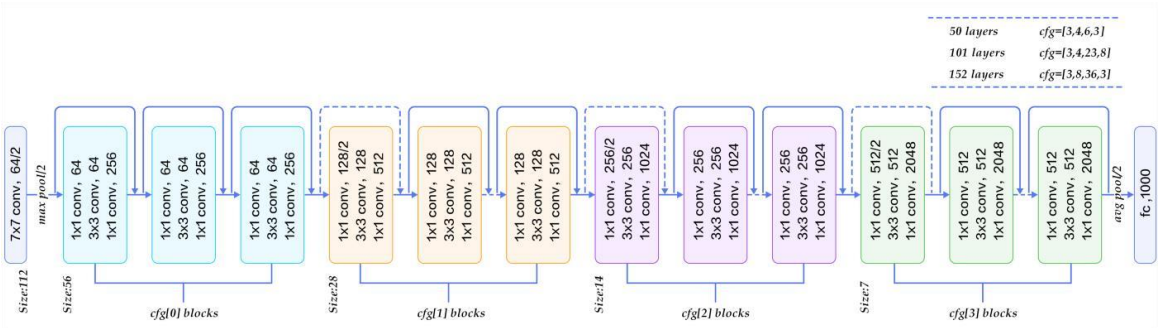
Three Fully-Connected (FC) layers follow the stack of convolutional layers (which has the different depth in different architectures): the first two have 4096 channels each, the third performs 1000-way ILSVRC classification and thus contains 1000 channels. The final layer is softmax layer. The configurations of the fully connected layers is same in all the networks.

All hidden layers are equipped with rectification (ReLU) non-linearity. It is also noted that none of the networks (except for one) contain the Local Response Normalisation (LRN), such

normalisation does not improve the performance on the ILSVRC dataset but leads to increased memory consumption and computation time.

Q4. What is ResNet?

Answer: At the ILSVRC 2015, so-called Residual Neural Network (ResNet) by the Kaiming He et al introduced the a novel architecture with "skip connections" and features heavy batch normalisation. Such skip connections are also known as the gated units or gated recurrent units and have the strong similarity to recent successful elements applied in RNNs. Thanks to this technique as they were able to train the NN with 152 layers while still having lower complexity than the VGGNet. It achieves the top-5 error rate of 3.57%, which beats human-level performance on this dataset.



Q5. What is HAAR CASCADE?

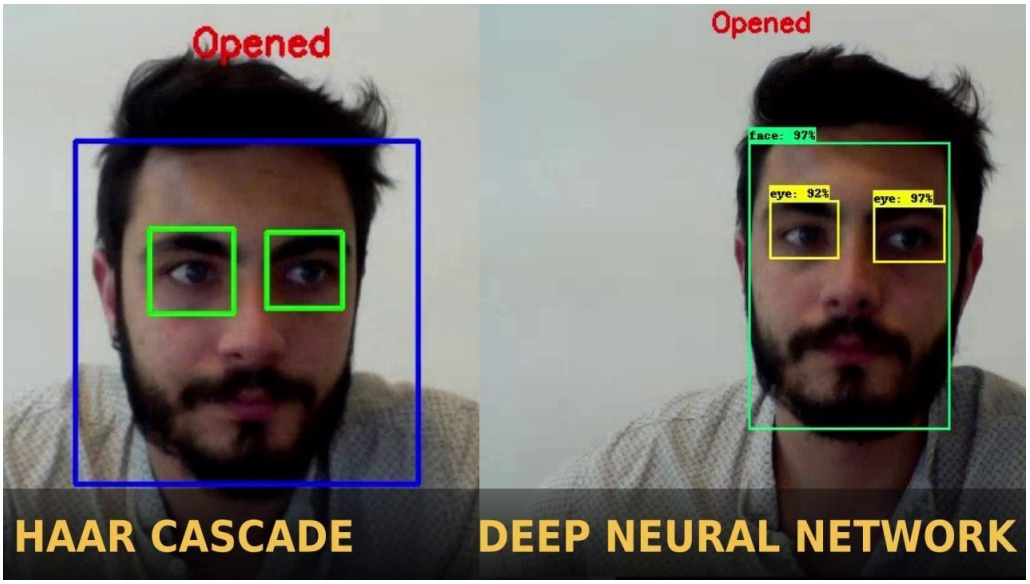
Answer: Haar Cascade: It is the machine learning object detections algorithm used to identify the objects in an image or the video and based on the concept of features proposed by Paul Viola and Michael Jones in their paper "Rapid Object Detection using a Boosted Cascade of Simple Features" in 2001.

It is a machine learning-based approach where the cascade function is trained from the lot of positive and negative images. It is then used to detect the objects in other images.

The algorithm has four stages:

- Haar Feature Selection
- Creating Integral Images
- Adaboost Training
- Cascading Classifiers

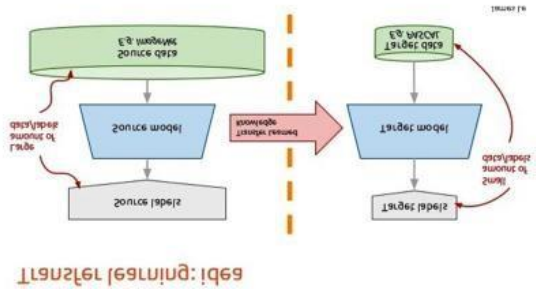
It is well known for being able to detect faces and body parts in an image but can be trained to identify almost any object.



Q6. What is Transfer Learning?

Answer: Transfer learning: It is the machine learning method where the model developed for a task is reused as the starting point for the model on the second task .

Transfer Learning differs from the traditional Machine Learning in that it is the use of pre-trained models that have been used for another task to jump-start the development process on a new task or problem.



The benefits of the Transfer Learning are that it can speed up the time as it takes to develop and train the model by reusing these pieces or modules of already developed models. This helps to speed up the model training process and accelerate results.

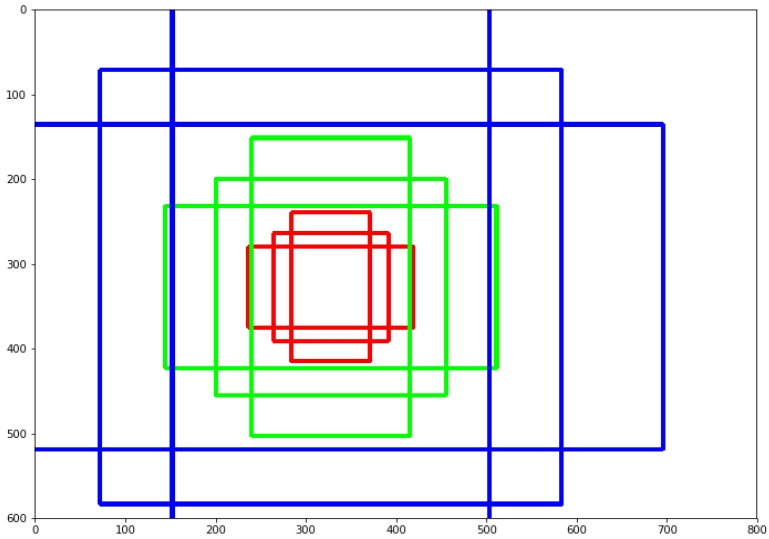
Q7. What is Faster, R-CNN?

Answer: Faster R-CNN: It has two networks: region proposal network (RPN) for generating region proposals and a network using these proposals to detect objects. The main difference here with the Fast R-CNN is that the later uses selective search to generate the region proposals. The time cost of generating the region proposals is much smaller in the RPN than selective search, when RPN shares the most computation with object detection network. In

brief, RPN ranks region boxes (called anchors) and proposes the ones most likely containing objects.

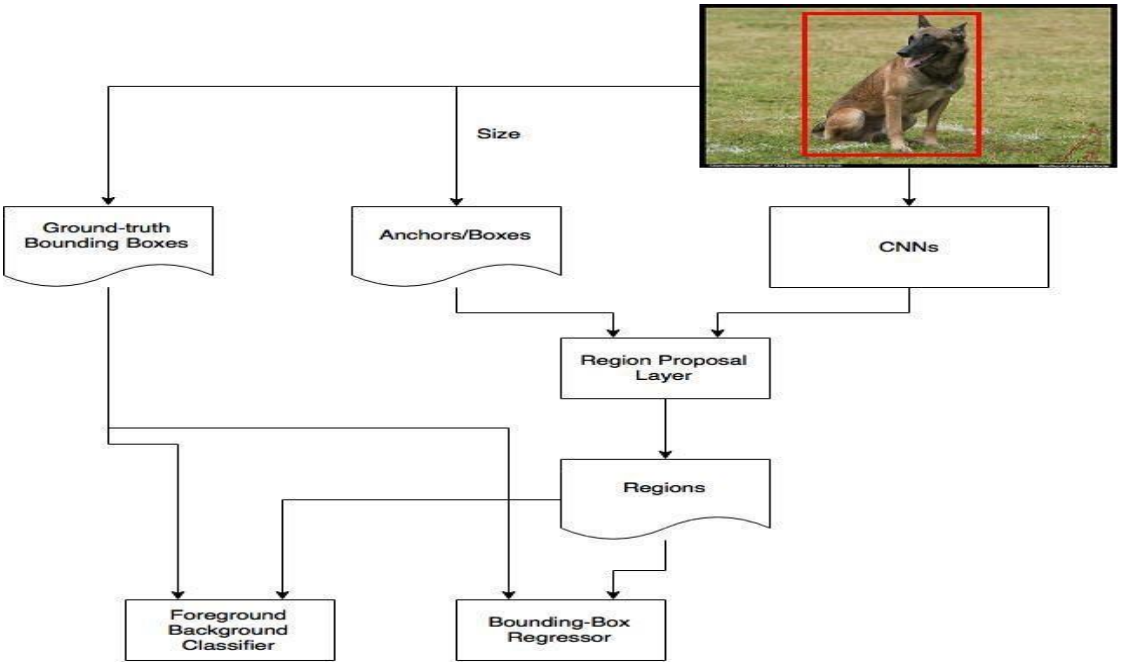
Anchors

Anchors play an important role in Faster R-CNN. An anchor is the box. In default configuration of Faster R-CNN, there are nine anchors at the position of an image. The graphs shown 9 anchors at the position (320, 320) of an image with size (600, 800).



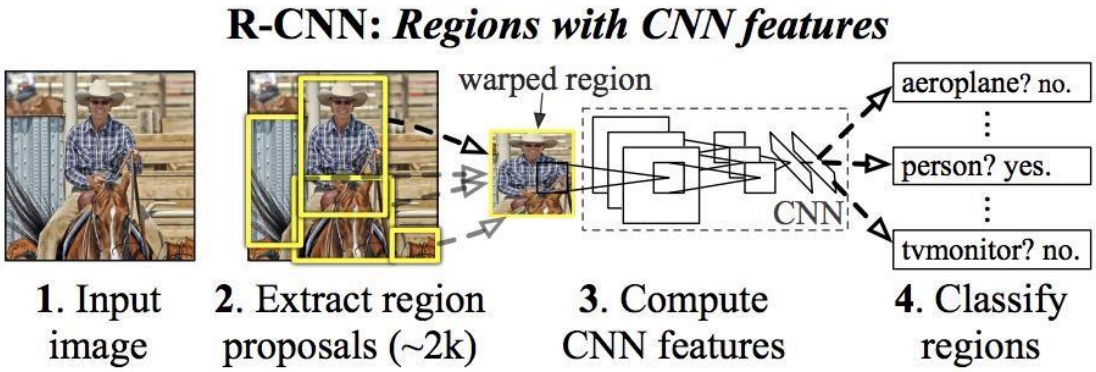
Region Proposal Network:

The output of the region proposal network is the bunch of boxes/proposals that will be examined by a classifier and regressor to check the occurrence of objects eventually. To be more precise, RPN predicts the possibility of an anchor being background or foreground and refine the anchor.



Q8. What is RCNN?

Answer: To bypass the problem of selecting the huge number of regions, Ross Girshick et al. proposed a method where we use the selective search to extract just 2000 regions from the image, and he called them as region proposals. Therefore, instead of trying to classify the huge number of regions, you can work with 2000 regions.



Problems with R-CNN:

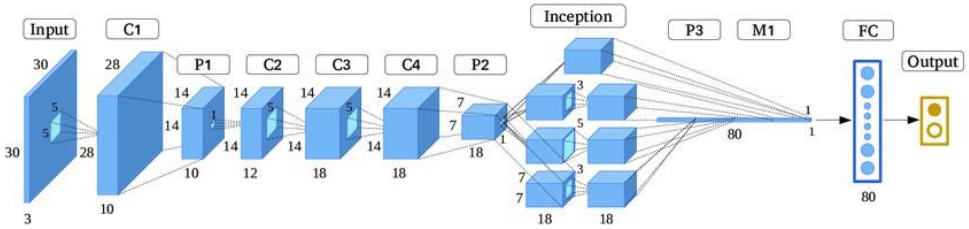
- It still takes the huge amount of time to train the network as we would have to classify 2000 region proposals per image.
- It cannot be implemented real-time as it takes around 47 seconds for each test image.
- The selective search algorithm is the fixed algorithm. Therefore, no learning is happening at that stage. This leads to the generation of the bad candidate region proposals.

Q9. What is GoogLeNet/Inception?

Answer: The winner of the ILSVRC 2014 competition was GoogLeNet from Google. It achieved a top-5 error rate of 6.67%! This was close to human-level performance which the organisers of the challenge were now forced to evaluate. As it turns out, this was rather hard to do and required some human training to beat GoogLeNets accuracy. After the few days of training, the human expert (Andrej Karpathy) was able to achieve the top-5 error rate of 5.1%(single model) and 3.6%(ensemble). The network used the CNN inspired by LeNet but implemented a novel element which is dubbed an inception module. It used batch normalisation, image distortions and RMSprop. This module is based on the several very small convolutions to reduce the number of parameters drastically. Their architecture

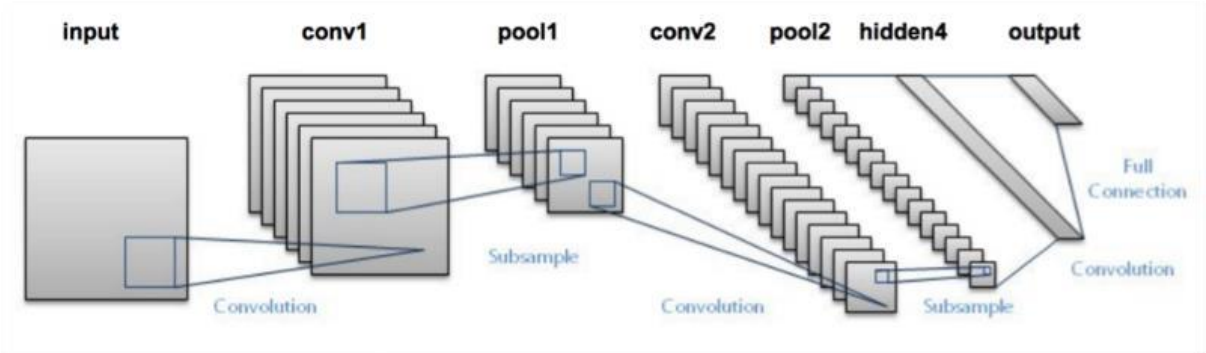
consisted of the 22-layer deep CNN but reduced the number of parameters from 60 million (AlexNet) to 4 million.

It contains 1×1 Convolution at the middle of network, and global average pooling is used at the end of the network instead of using the fully connected layers. These two techniques are from another paper "Network In-Network" (NIN). Another technique, called inception module, is to have different sizes/types of convolutions for the same input and to stack all the outputs.



Q10. What is LeNet-5?

Answer: LeNet-5, a pioneering 7-level convolutional network by the LeCun et al in 1998, that classifies digits, was applied by several banks to recognise hand-written numbers on checks (cheques) digitised in 32x32 pixel greyscale input images. The ability to process higher-resolution images requires larger and more convolutional layers, so the availability of computing resources constrains this technique.



LeNet-5 is very simple network. It only has seven layers, among which there are three convolutional layers (C1, C3 and C5), two sub-sampling (pooling) layers (S2 and S4), and one fully connected layer (F6), that are followed by output layers. Convolutional layers use 5 by 5 convolutions with stride 1. Sub-sampling layers are 2 by 2 average pooling layers. Tanh sigmoid activations are used to throughout the network. Several interesting architectural choices were made in LeNet-5 that are not common in the modern era of deep learning.