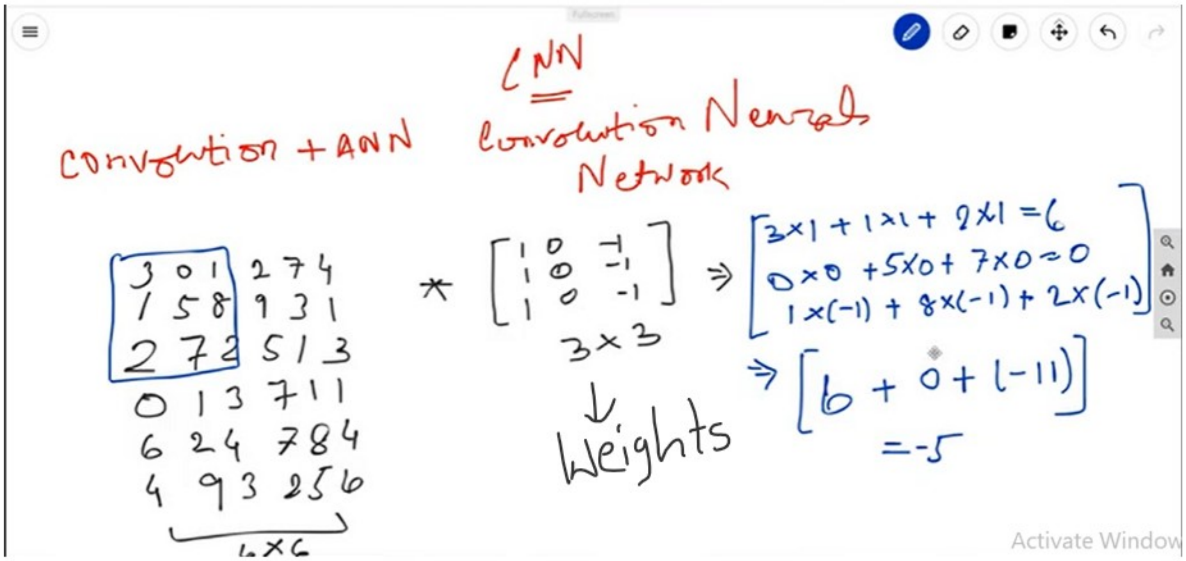
CNN

Tuesday, September 28, 2021 6:20 PM



CNN = Convolution Neural Network

Convolution + ANN

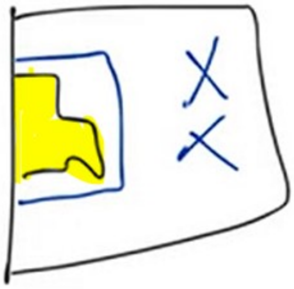
$$\underbrace{\begin{matrix} \boxed{\begin{matrix} 3 & 0 & 1 \\ 1 & 5 & 8 \\ 2 & 7 & 2 \end{matrix}} & \begin{matrix} 2 & 7 & 4 \\ 9 & 3 & 1 \\ 5 & 1 & 3 \end{matrix} \\ \begin{matrix} 0 & 1 & 3 \\ 6 & 2 & 4 \\ 4 & 9 & 3 \end{matrix} & \begin{matrix} 7 & 1 & 1 \\ 7 & 8 & 4 \\ 2 & 5 & 6 \end{matrix} \end{matrix}}_{6\times 6} * \underbrace{\begin{bmatrix} 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \end{bmatrix}}_{3\times 3 \downarrow \text{Weights}} \Rightarrow \begin{bmatrix} 3\times1 + 1\times1 + 2\times1 = 6 \\ 0\times0 + 5\times0 + 7\times0 = 0 \\ 1\times(-1) + 8\times(-1) + 2\times(-1) \end{bmatrix}$$

$$\Rightarrow \left[6 + 0 + (-11)\right] = -5$$

n - f + 1 = features size,
where n - matrix
f - filter size

Padding

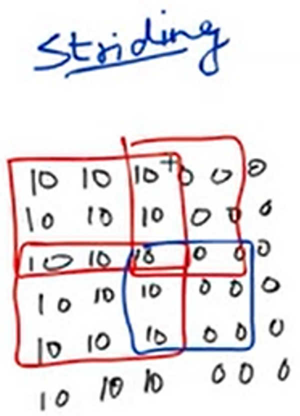
$$\begin{matrix} 10 & 10 & 10 & 0 & 0 & 0 \\ 10 & 10 & 10 & 0 & 0 & 0 \\ 10 & 10 & 10 & 0 & 0 & 0 \\ 10 & 10 & 10 & 0 & 0 & 0 \\ 10 & 10 & 10 & 0 & 0 & 0 \\ 10 & 10 & 10 & 0 & 0 & 0 \end{matrix}$$



As the filter is smoothing through image, convolution happens. The convolution basically happens more in central part. So if image has features on side we have to do padding around image to slide feature towards center.
Then New formula will be
n + 2p - f + 1 where p - padding

Valid padding - without padding
Same padding - with padding

Striding



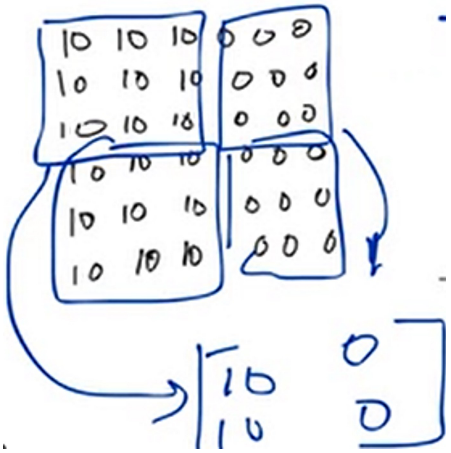
If the image is huge dense or large megapixel, the convolution takes time and results in more training time. In Striding, increase step while convolution instead of take 1 step. It reduces the image size very high

Updated formula will be

$$\frac{n + 2p - f}{n} + 1,$$ where n - no of striding

Pooling :

It is completely a feature extractor. It always take the highest pixel value.



-> If our filter is too big, it will sometimes considers two images as one. So that why our filter should be smaller and we also pass additional features in network>

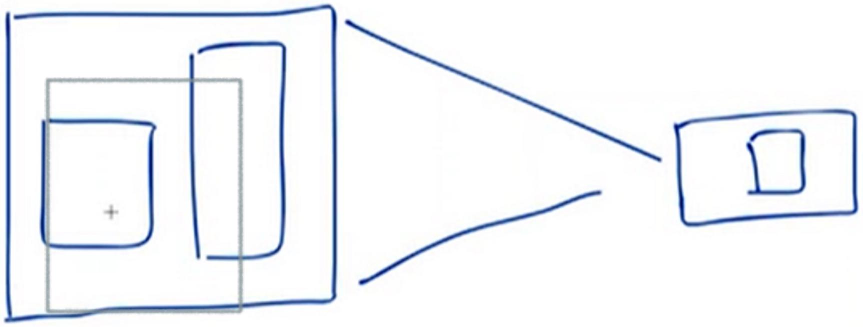
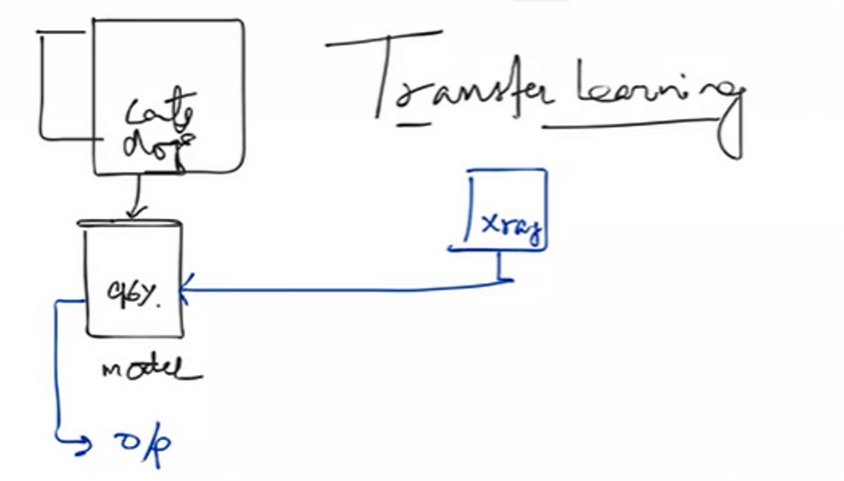


Image Preprocessing

1. resize image
2. Gray scaling

3. Smoothing or blurring / Threshold
4. You can dilate or erode

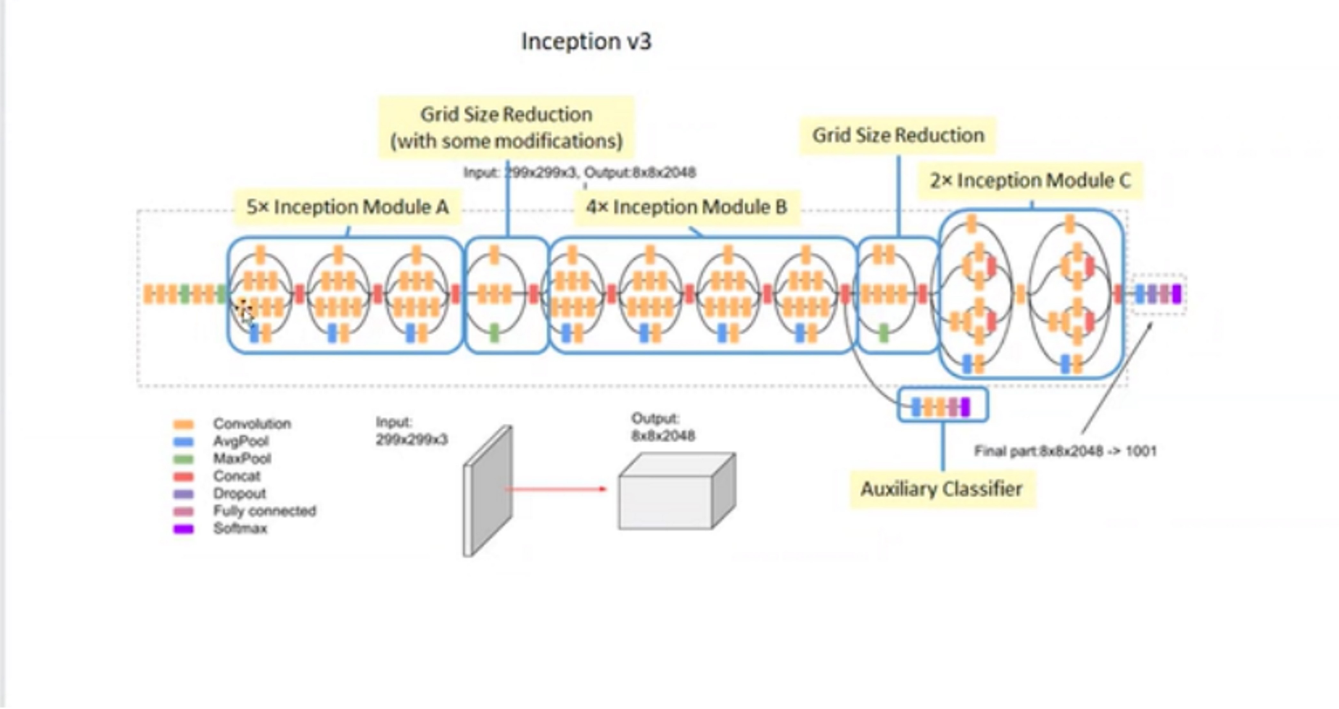
Transfer Learning Technique:

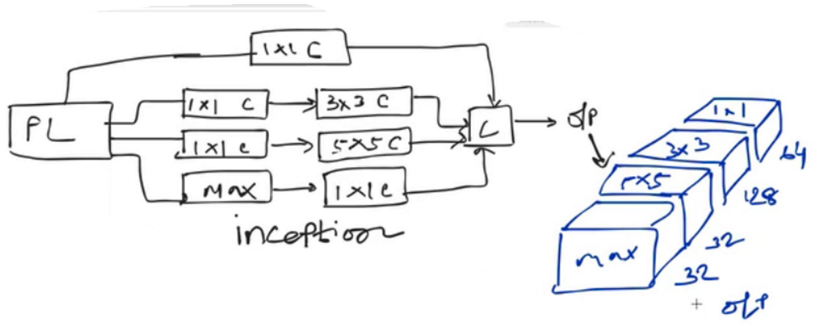


It uses pre-trained model to predict the output with different data. If required, we train the model, but the training time is less. The Hyperparameters are already tuned, we don't need to do that, it will saves time and resources for training. Some of the pre-trained Models are Lenet,
Alexnet, vgg16, vgg19 etc.

Savoula Threshold : It reduces noise to 255 or white.

Inception Model by Google :



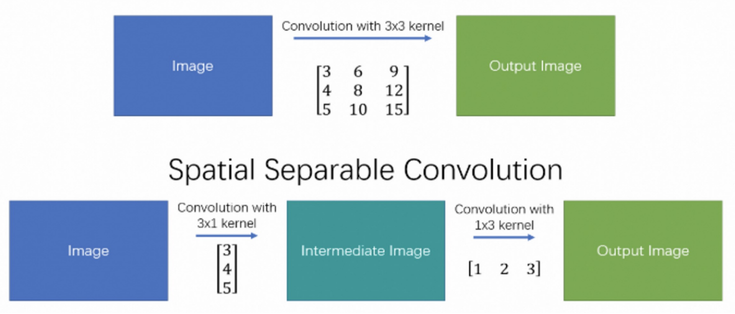


64
128
32
32
___
256

Instead of using normal 256 filters, it uses different range & sizes of filters in each module. Each module is connected by other.
It saves memory and time. It is used in devices having low resources.

Separable Convolution :

1. Spatial Separable Convention :
   It breaks the kernel. Its not use in Inception



Convolution with 3x3 kernel

$$\begin{bmatrix} 3 & 6 & 9 \\ 4 & 8 & 12 \\ 5 & 10 & 15 \end{bmatrix}$$

Spatial Separable Convolution

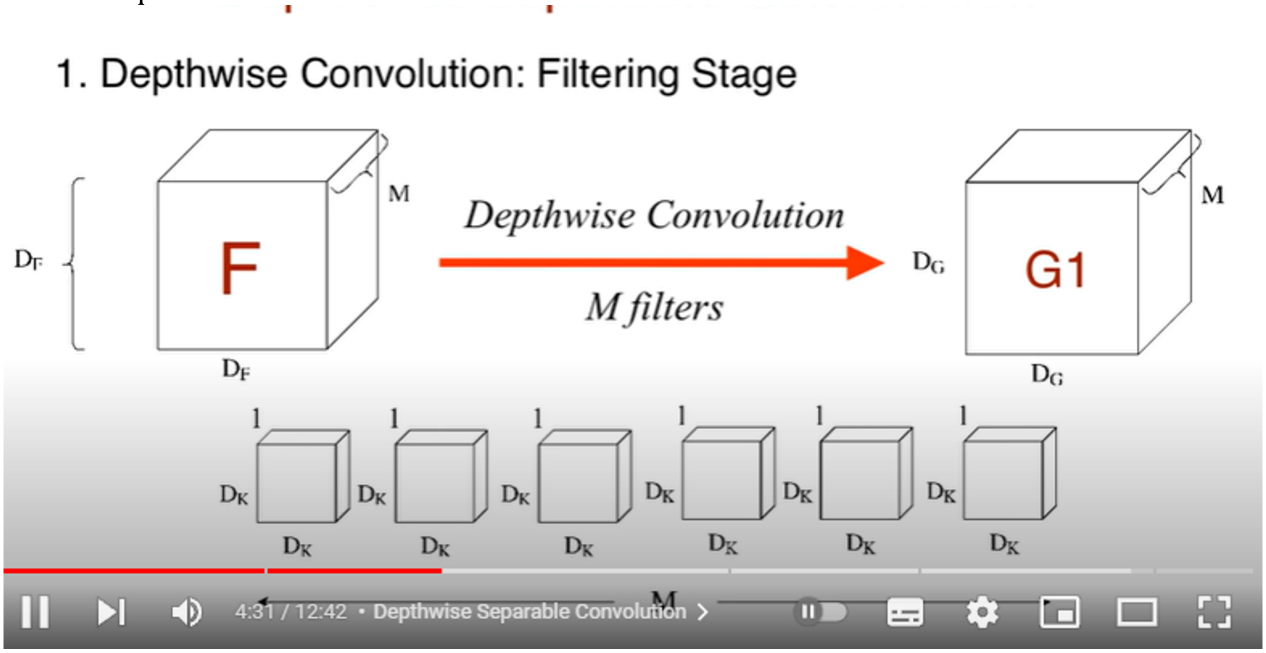
Convolution with 3x1 kernel

$$\begin{bmatrix} 3 \\ 4 \\ 5 \end{bmatrix}$$

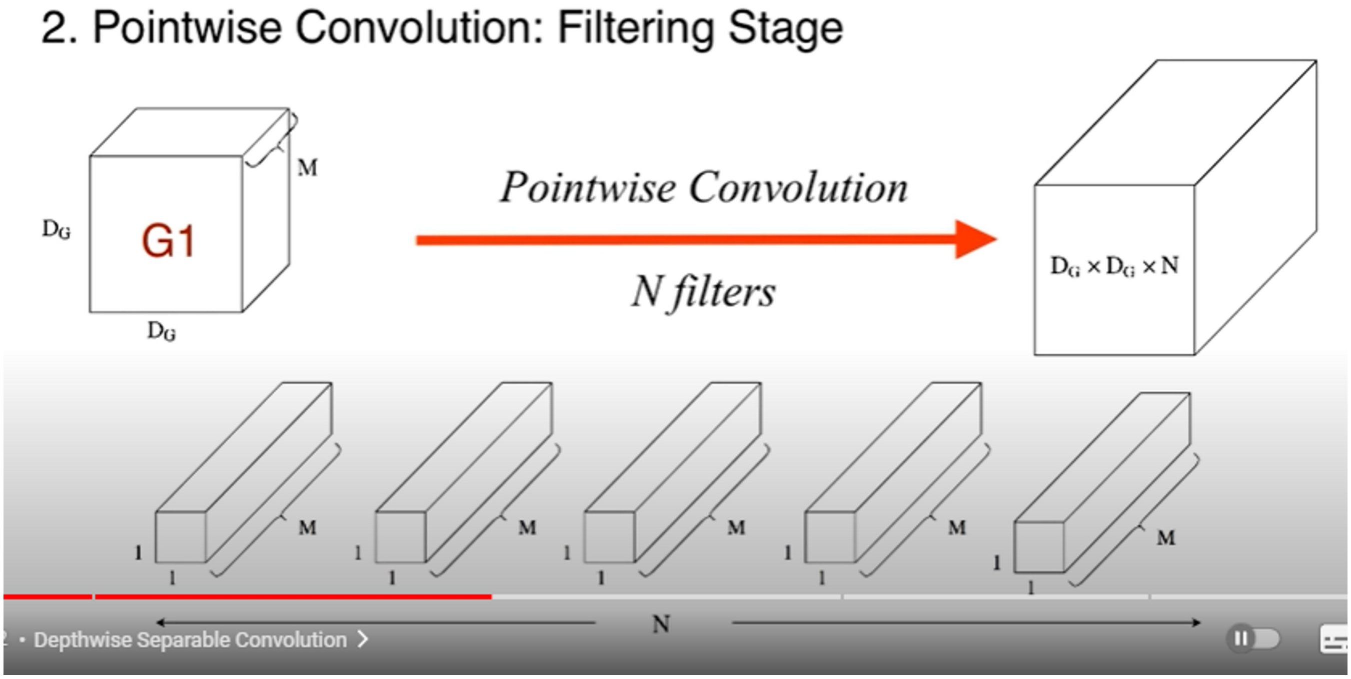
Convolution with 1x3 kernel

$$\begin{bmatrix} 1 & 2 & 3 \end{bmatrix}$$

2. Depth Separable Convolution :
   It has two parts :



1. Depthwise Convolution: Filtering Stage

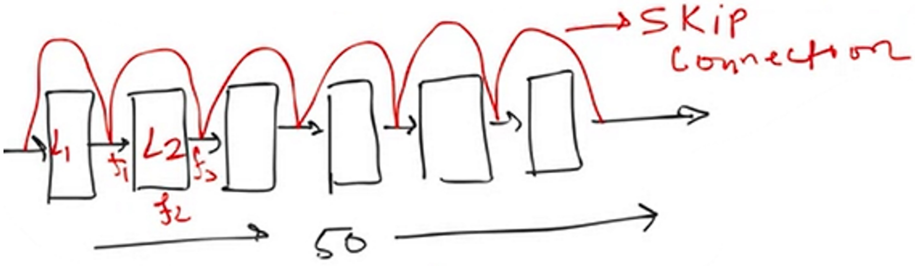


2. Pointwise Convolution: Filtering Stage

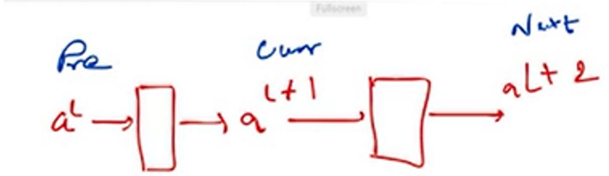
Resnet :
If we have many layers of CNN like 50 layers,

1. Image/ feature will vanish
2. Vanishing Gradient.

   For this we use Resnet, it skips connection.



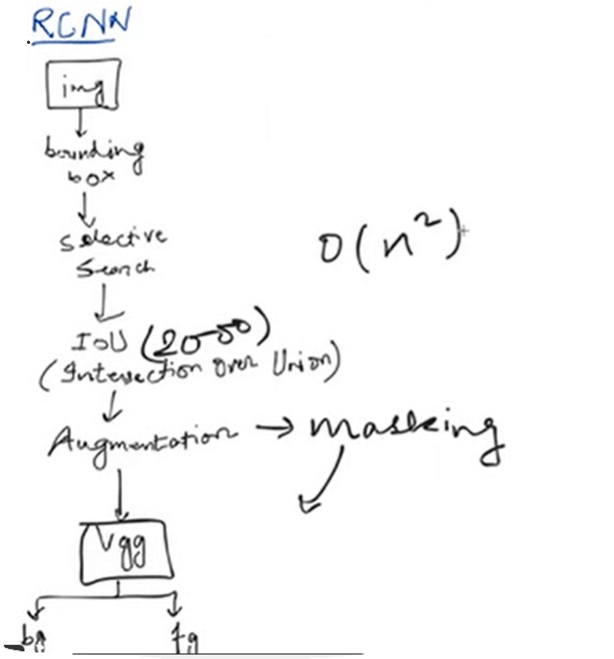
$$f_3 = f_2 + f_1$$



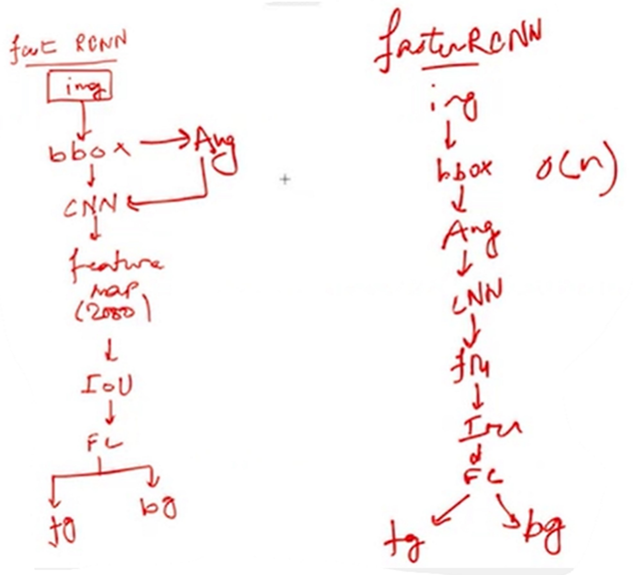
$$a^{L+1} = ReLu(W^{L+1} \cdot a^{L} + b^{L+1})$$
$$a^{L+2} = Relu(W^{L+2} \cdot a^{L+1} + b^{L+2})$$

$$f_3 = \underbrace{Relu(W^{L+1} \cdot a^{L} + b^{L+1})}_{f_1} + \underbrace{Relu(W^{L} \cdot x + b^{L})}_{f_2}$$

Object Detection :
In CNN, we can't segment out object or annotated object. Segmentation is detection of an object in an area.
That's why we go with RCNN (Region based CNN).



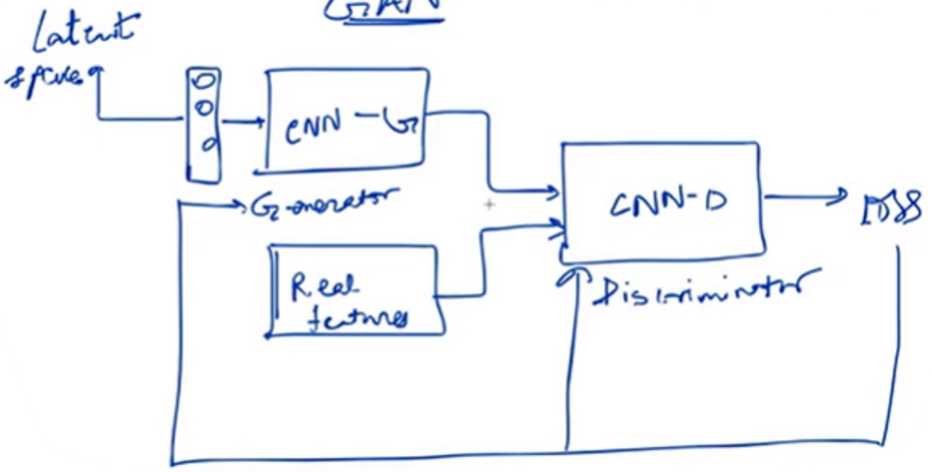
The limitation of RCNN is basically there are only 2000 objects can be detected because the time complexity is very high 0(n2).



There is fast RCNN in which after bbox and Augmentation, we pass image through CNN for feature extraction.
We use depth separable convolution in fast RCNN which takes less resources and time.
If we using annotated images, then why limit fast RCNN to 2000. We can use as many features as we want.

GANS :

GAN


Latent space
CNN - G
Generator
Real features
CNN-D
Discriminator
Loss