Alexnet Summary:

- Preprocessing step
    1. Scaling down the image resolution to 256x256.
    2. Subtracting each pixel by their mean over the training set.
- It contains eight learned layers - 5 convolutional and 3 fully- connected.
- ReLU function’s non-saturating nonlinearities are several times faster than sigmoid and tanh functions.
- This dataset had a primary concern of overfitting. For Large datasets, a model with faster learning ability has a great influence on its performance, hence ReLU was an optimum choice as it does not pose a problem of vanishing gradients.
- Since their model could not run one GPU as it was too big to fit, they used cross GPU-parallelization which puts half the kernels on each GPU.
- Local Response Normalisation
    1. Even though ReLU function does not require any input normalisation to prevent saturation, it is found that random local normalisation yields generalisation.
    2. The activity of a neuron is normalised with other neurons at the same spatial position, from both sides of the neuron. The ordering is arbitrary and is determined before training.
    3. This implements a form of lateral inhibition(a neuron’s activity get’s hindered by neighbouring neurons) which results in a more generalized model.
- Using Overlapping pooling reduces the model to overfit.
- Overall Architecture
    1. ReLU is applied for the output of every convolutional and fully-connected layer, while softmax for the eighth layer of 1000 classes.
    2. Local Response Normalisation is applied for the 1st and 3rd convolutional layer.
    3. Max pooling is applied after the 1st,3rd and 5th layer.
- Data Augmentation:

  It is the best and the common way to reduce overfitting.These transformed images need not be stored on disk(GPU) and are hence computationally free.
    - First form of data augmentation consists of generating image translations and horizontal reflections. This was done by extracting random 224x224 patches and training network on these extracted patches.
    - The second form consisted of altering the intensities of RGB channels. This method involved the use of Principal Component Analysis on the entire training set. This method helps to highlight the object identity of the image, which is not effected by the changes in intensities of colours.

- Using different models, and combining their predictions is always a good way to reduce test error. Using Dropout is the best way to reduce computational time, overfitting and also helps model learn robust features as each neuron should not rely on neighbouring.