- Deep NN : Gradient exploding/vanishing, solved with weights normalisation and batch norm.
    - Covariate shift => layers take longer time to learn as they have to adapt to change in distribution of inputs (previous activations), so learning rate changes frequently. Detailed explanation : http://mlexplained.com/2018/01/10/an-intuitive-explanation-of-why-batch-normalization-really-works-normalization-in-deep-learning-part-1/
    - BN makes the mean and variance of a layer independent of the layer values and less dependent of higher order interactions between layers, which encourages us to use larger learning rates.
    - BN also restricts the distribution of the layer's activation due to which the gradients vanishing is prevented, as gradients->0 as |x|>>0. https://www.quora.com/Why-does-Batch-Normalization-for-deep-Neural-Networks-fix-the-vanishing-gradient-problem
    - As layers increase training error increase => not overfitting, but difficult to optimise huge parameter space.
- Hypothesis of Resnets is that it is easier to optimize the residual mapping than to optimize the original, unreferenced mapping.
- The idea of resnets came from the thought that deeper networks should not perform worse than shallow ones if extra layers were to become identity functions.
- Ex : if an identity mapping were optimal, it would be easier to push the residual to zero than to fit an identity mapping by a stack of nonlinear layers.
- Original hypothesis =>Stack of non-linear layers approximate to underlying function H(x). Resnet hypothesis => We approximate to H(x) - x, which is assumed to be easier to learn.
- Experiments show that solving mathematical problems using residual nature of solutions converge faster.So, optimization is easier with preconditioning or reformulation.
- Deep supervision : Auxiliary classifiers are connected to intermediate layers to increase gradient signal during backprop and also help in regularisation.
- If input and outputs have different dimension we can use W.x instead of x in the def. If we use W.x for all shortcuts, then better accuracy but more learnable parameters.We can also use zero padding when dimensions increase with no extra parameters.
- For very deep networks, bottlenecks are used by using 1x1.
- For shallow network also resnets converge faster than plain => easier optimisation due to more gradient flow due to more no. of connections.
- Warm up : Initially learning rate is kept low for some epochs to reduce learning effect of the early data exposure, so that some features are not strongly learned which may give low accuracy.
- The responses are the outputs of each 3×3 layer, after BN and before other nonlinearity (ReLU/addition).
- Analysis of responses showed that resnet responses have lower std. Which reflects their response strength.Smaller responses => It is nearer to identity function.
- It uses momentum optimiser; also weight decay.

- RoI : Regions of interest, are portions of the image for object detection. https://deepsense.ai/region-of-interest-pooling-explained/
- BN layers' statistics(mean and variance) are not updated by fine tuning, and remain linear activations with constant offsets and scales.
- Object detection Improvements:
    - Box refinement : Like bounding boxes, follows iterative localisation.
    - Global context : Using global Spatial Pyramid Pooling, a feature is pooled.It is implemented as RoI pooling with full img's bounding box as RoI.This pooled feature is fed into post-RoI layers to get global context feature.
    - Multi scaling : predictions tried in different scales using the feature pyramid using maxout layers.Similar to data augmentation and training on different sizes of images.
    - Validation set
    - Ensemble
- The region proposal network (RPN) in the faster region-based convolutional neural network (Faster R-CNN) is used to decide "where" to look in order to reduce the computational requirements of the overall inference process. https://medium.com/egen/region-proposal-network-rpn-backbone-of-faster-r-cnn-4a744a38d7f9
https://arxiv.org/pdf/1506.01497.pdf
- Per class RPN were trained on images to predict bounding boxes for ground truth class.