- High depth was achieved by using small (3x3) filters.Only preprocessing was subtracting mean RGB value from each pixel.
- 1x1 filters can be seen as linear transformation of input pixels followed by non-linearity.
- LRN (local response normalisation) was also applied in some configuration.
- Three 3x3 layers have a 7 × 7 effective receptive field. So what have we gained by using, for instance, a stack of three 3×3 conv. layers instead of a single 7×7 layer? First, we incorporate three non-linear rectification layers instead of a single one, which makes the decision function more discriminative. Second, we decrease the number of parameters: assuming that both the input and the output of a three-layer 3 × 3 convolution stack has C channels, the stack is parametrised by 3 3 2C 2 = 27C 2 weights; at the same time, a single 7 × 7 conv. layer would require 7 2C 2 = 49C 2 parameters, i.e. 81% more. This can be seen as imposing a regularisation on the 7 × 7 conv. filters, forcing them to have a decomposition through the 3 × 3 filters (with non-linearity injected in between).
- 1x1 layers => additional non-linearity in the network => more discrimination power…
- Faster convergence due to :
    - Implicit regularization due to more depth and small filters.
    - Pre-initialisation of some layers from shallower nets.
- Data augmentation used : random crop of isotropically-rescaled image, horizontal flip and random RGB colour shift.
- Training image size : Fixed rescaling used from which 224x224 samples are cropped.
- Multi scale training approach :
    - For each image random rescaling from a range of ratios and 224x224 cropped.
    - It is like scale jittering which trains model to recognise objects of wide range of sizes.
- In dense evaluation, the fully connected layers are converted to convolutional layers at test time, and the uncropped image is passed through the fully convolutional net to get dense class scores. Scores are averaged for the uncropped image and its flip to obtain the final fixed-width class posteriors.
- Multi-crop evaluation works slightly better than dense evaluation, but the methods are somewhat complementary as averaging scores from both did better than each of them individually. The authors hypothesize that this is probably because of the different boundary conditions: when applying a ConvNet to a crop, the convolved feature maps are padded with zeros, while in the case of dense evaluation the padding for the same crop naturally comes from the neighbouring parts of an image (due to both the convolutions and spatial pooling), which substantially increases the overall network receptive field, so more context is captured.
- Though 1x1 filters add non-linearity, network with same depth with 3x3 at all layers worked better => it is important to capture spatial context by using non-trivial receptive fields (other than 1x1).
- Shallower net formed by replacing each pair of 3x3 by 5x5 performed worse than deeper net with small filters.

- Scale jittering worked better than single fixed rescaling smaller side for training as well as for testing.
- Dense evaluation => like sliding windows of dim 224x224 on rescaled image….