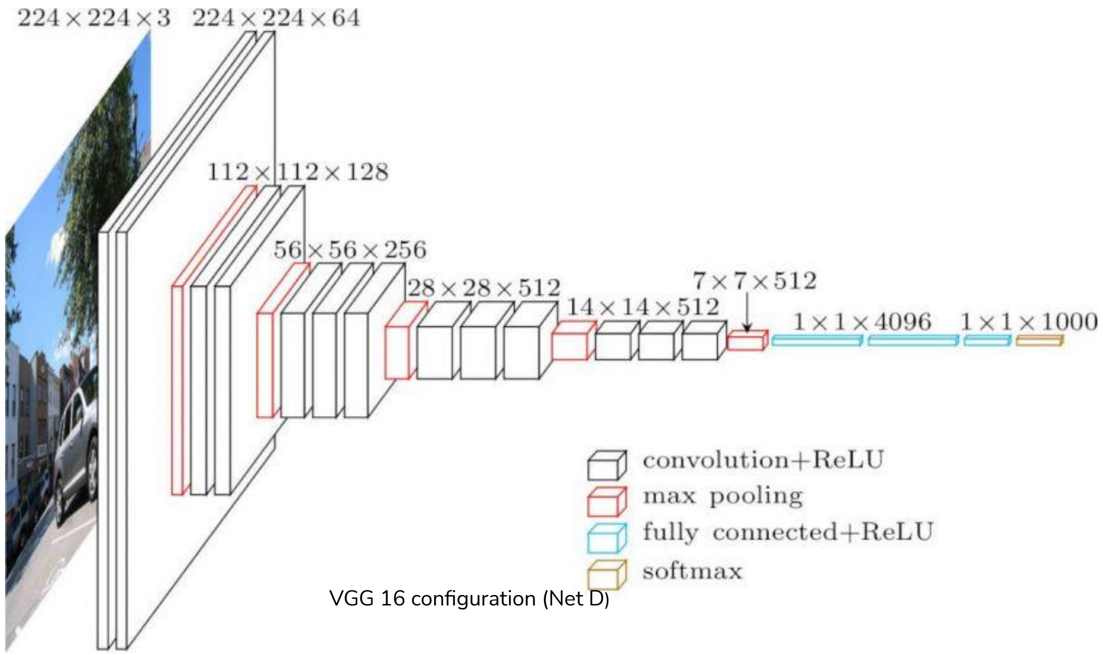
VGG Summary

- **Dataset**: Subset of ImageNet, roughly 1.3M training images, 50K validation images and 100K testing images, around 1K images in each of 1K categories.

- **Architecture**:
  - During training, the input to the network is a fixed-size 224 × 224 RGB image.
  - Subtracted the mean RGB value(of training set) from each pixel.
  - The convolutional stride was fixed to 1.
  - 5 max pooling layers performed over a 2x2 window, with stride 2.
  - Stack of convolutional layers followed by 3 FC layers (first 2 having 4096 channels and third having 1000 channels)
  - Final layer is the softmax layer.
  - All the layers have ReLU non-linearity.



VGG 16 configuration (Net D)

- **Configuration:** 5 nets(A-E) that differ only in depth, 11 weight layers in network A(8 conv & 3 FC) to 19 layers in E(16 conv & 3 FC)
  - Used 3x3 filters throughout the whole net.

  - A stack of 2 3x3 conv. layers(without pooling in between) has an effective receptive field of 5x5; 3 such layers have a 7x7 receptive field. This has two benefits:
    - We incorporate 3 non-linear rectification layers instead of a single one.
    - We decrease the number of parameters. ($49C^2$ to $27C^2$ by using 3 3x3 layers instead of one 7x7).

- Used LRN in Net A but didn't improve the performance. Instead, it led to increased memory consumption and computation time.

- **Classification Framework:**
  - **Training -**
    - Mini-batch gradient descent(size 256) with momentum 0.9;
      L2 regularization; dropout reg. for the 1st two FC layers; learning rate decay when validation set accuracy stopped improving.
    - The nets required less epochs (74) to converge due to
      - Regularization imposed by greater depth and smaller filter sizes.
      - Pre-initialization of certain layers.
    - S = Smallest side of an isotropically rescaled training image; we consider two approaches for setting S:
      - Single-Scale training: Fix S; Used two values of S(256 and 384) for evaluation.
      - Multi-Scale training: Random sampling of S from a certain range [256, 512]. This can be seen as training set augmentation by scale jittering.
  - **Testing -**
    - Testing images are isotropically rescaled to a pre-defined smallest image side (Q); 1st FC layer is converted to 7x7 and the last two to 1x1 conv. layers
    - The fully conv. net applied to the whole un-cropped image, resulting in a class score map with the no. of channels equal to no. of classes; then the class score map is sum-pooled.
    - Augmented the test set by horizontal flipping; their softmax class posteriors are averaged with original test images to obtain the final scores.
    - 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.
    - The idea behind dense evaluation is to make the network flexible for different input image sizes.
    - Multi-crops performed slightly better than dense evaluation.

- **Implementational details:**
  - Used C++ Caffe toolbox for implementation. (allowing us to perform training on multiple GPUs installed in a single system)
  - A system equipped with 4 NVIDIA Titan black GPUs was used.

- **Single scale evaluation:**
  - The test scale(Q) was set to S for fixed S and $(S_{min}+S_{max})/2$ for jittered S.
  - Classification error decreases with increase in ConvNet depth; Net C performs better than B because of increased non-linearity.

- Conf C and D have the same depth (16 layers) but the latter performs better because it uses 3x3 conv layers and net C uses 1x1 layers.
- Finally, scale jittering at training time leads to better performance than fixed S.

- **Multi scale evaluation:**
  - The models trained with fixed S were evaluated over three test image sizes, close to the training one: Q = {S − 32, S, S + 32}
  - The model trained with variable S ∈ [Smin; Smax] was evaluated over a larger range of Q = $\{S_{min}, (S_{min} + S_{max})/2, S_{max}\}$
  - Scale jittering at test time resulted in better performance.

- **Multi Crop evaluation:**
  - Multiple Crops performs better than dense evaluation.
  - Combination of both outperforms each of them.
  - Ensemble of two best-performing models using combined dense and multi-crop evaluation gave an error of 6.8%. (top-5 error)

- **Localisation Task:**
  - Won ILSVRC task in 2014 with 25.3% error.
  - VGG-16 was found to be the better performer than VGG-19.

- **Localisation ConvNet:**
  - Euclidean Loss function was used.
  - Scale jittering wasn’t used due to time constraints. Instead, two models were trained - with S= 256 and S=384. (S =training scale)
  - The last fully-connected layer was initialised randomly and trained from scratch.
  - The bounding box prediction is a 4D vector consisting of its center coordinates, width, and height.
  - Two testing protocols were used:
    - For comparing different Nets on the validation set, the Bounding box was obtained by applying the network to the central crop of the image.Only the bounding box prediction for the ground truth class was considered.
    - To come up with the final prediction on the test set, spatially close predictions were merged. Predictions were then rated based on class scores obtained from classification ConvNet.
    - When several localisation ConvNets were used, firstly the union of their sets of bounding box predictions was taken, and then the merging procedure on the union was run.

- **Localisation Experiments:**
  - Per Class Regression(PCR) outperformed Single Class Regression (SCR).

- Testing at several scales and combining the predictions of multiple networks improved the performance.
- Test Error of 25.3% was achieved, indicating the performance enhancement brought by deep ConvNets.

- **Generalisation of Very Deep Features**:
    - In this section, ConvNets pre-trained on ILSVRC, are evaluated on smaller datasets.
    - Last fully connected layer (which performs classification among 1000 classes) is removed and 4096D activations of the penultimate layer are used as image features.
    - This is combined with a linear SVM classifier.
    - An image is processed in a similar manner as ILSVRC classification. First the image is rescaled on different scales and then the network is applied densely over the image plane followed by global average pooling resulting in 4096D image descriptor. The descriptor is then averaged with the descriptor of a horizontally flipped image.
        - **Classification on VOC-2007 and VOC-2012**:
            - Contains 10K and 22.5K images respectively labelled over 20 object categories. Average and Max Pooling performed better than stacking of descriptors.
            - Accuracy of 89% was achieved, outperforming the previous best result by 6%.
        - **Image Classification on Caltech-101 and Caltech-256:**
            - Contains 9K and 31K images respectively.
            - 3 random splits were generated each containing 30-50 training images per class.
            - Stacking of descriptors,computed over multiple scales performed better than average and max pooling, since image objects typically occupy the full image.
            - Best results were achieved by stacking VGG-16 and VGG-19 models.
            - Achieved accuracy of 86% on Caltech-256 outperforming the previous best result by 8.6%.

            - 84% accuracy was achieved in Action classification over VOC-2012 outperforming the previous results by 6.7%.