Very Deep Convolutional Networks for Large-Scale Image Recognition

Link

arxiv

Summary

- They do a evaluation of networks of increasing depth using an architecture with very small(3×3) convolution filters, which shows that a significant improvement on the prior-art configurations can be achieved by pushing the depth to 16-19 weight layers.
- They use small 3×3 receptive throughout the whole net. A stack of two 3×3 convolution layer has an effective receptive field of 5×5; three such layers have a 7×7 effective receptive field. But using three stacks of 3×3 conv layers instead of one 7×7 layer provide following two benefits:

  1. It incorporates three non-linear rectification layers instead of a single one, which makes the decision function more discriminative.
  2. It decreases the number of parameters required. Assuming that both the input and the output of a three-layer 3×3 conv stack has C channels, the stack is parametrised by $3(3^2C^2) = 27C^2$ weights; at the same time, a single 7×7 con. layer would require $7^2C^2 = 49C^2$ parameters, i.e. 81% more.

  This can also be seen as imposing a regularization on the 7×7 conv. filters, forcing them to have a decomposition through the 3×3 filters
- They used five variants of the network(with increasing number of layers). They trained the one with lowest number of layers with random initialisation but when training deeper networks they initialised the matching layers with the shallow network layer weights and initialised remaining layers randomly. They did fine tuning on transferred layer weights. For random initialisation, they sampled the weights from a normal distribution with the zero mean and 0.01 variance. The biases were initialised with zero.
- They use two different approaches for setting the training image scale S

  1. fix S i.e. single scale training
  2. multi-scale training, where each training image is individually rescaled by randomly sampling S from a certain range $[S_{min}, S_{max}]$ (they used $S_{min} = 256$ and $S_{max} = 512$).

  The second approach yield better results. Since objects in images can be of different size, it is beneficial to take this into account during training.
- During testing they used both single test scale Q (which might be different from training scale S) and several values of Q. To account for the possible change in spatial dimension during inference they performed global average pooling across spatial dimension after the final conv layer(they replace the fully connected layers used during training with conv layer during test). They also augment the test set by horizontal flipping of the images; the soft-max class posteriors of the original and flipped images are averaged to obtain the final scores for the image.
- They achieved second place in ILSVRC-2014 contest for classification task using ensemble of the networks. If single net performance is considered they achieved best result. They achieved best result on the localisation part of the challenge where they used fine-tuning on features learned from classification task. Fine tuning features learned on Imagenet dataset on some other task they achieved state of the art performance.