Paper: *Very Deep Convolutional Networks for Large-Scale Image Recognition*

Nirajan Koirala

CSC 9010_001

Summary

In this paper the authors have investigated the effect of depth on CNNs and its accuracy in large-scale image recognition setting. They have provided a very simple and elegant structure with periodic pooling all the way through the network. ConvNets improvises on the AlexNet model by using smaller receptive fields and smaller strides which greatly reduces the number of parameters in the network and helps in performance. They also introduced new techniques like training and testing the networks densely over the whole image and over multiple scales. Using these type of simple techniques, ConvNets were able to secure 1st runner up position in classification and 1st position in localization in the ILSVRC 2014 .

The ConvNet layer configurations are designed using the same principles as used by AlexNet 2012 model. However, the variable size convolutional kernels used in AlexNet is replicated by using multiple 3 x 3 kernel as building blocks. By using this approach, they incorporate 3 non-linear rectification layers instead of a single one which makes the decision function more discriminative. More importantly, they are able to reduce the overall parameters which greatly improves the performance.

The training of ConvNets generally follows the AlexNet model except for the input crops from multi-scale training images. The authors have conjectured that inspite of increased depth and large number of parameters, ConvNets require less epochs to converge than AlexNet because of implicit regularization imposed by greater depth and smaller conv. filter sizes and pre-initialization of certain layers. Authors have proposed multi-scale training, where the images in the training set are rescaled and two approaches are taken for the scaling. The first approach is single-scale training where the image is scaled with a size S = 256 or 384. The second approach is called multi-scale training where each training image is individually rescaled by sampling S from a range of $S_{min} = 256$ to $S_{max} = 512$. By using this approach, authors make sure that objects of different size in images are taken into account. In other words, this technique can also be viewed as scale jittering, where a model is trained to recognize objects over a wide range of scales.

During the testing stage, images are isotropically rescaled to a pre-defined scale Q which is not necessarily equal to the training scale S as using several values of Q for each S helps in improving performance. The network is then applied densely over the rescaled test image. The first fully-connected layer is replaced by a 7 x 7 conv. layer and the last two fully-connected layers are replaced by 1 x 1 conv. layers. The resulting fully-conv. net is then applied to the whole image. The result is a class score map depending on the input image size. The class score map is then spatially averaged to obtain a fixed-size vector of class scores for the image. They also mention that test set is augmented by horizontal flipping of the images and the soft-max class posteriors of original and flipped images are averaged to obtain the final scores for the image.

The evaluation of single scale ConvNets and multi-scale ConvNets is done separately. First the authors note that using local response normalization does not improve on model A without any normalization layers. So, they skip the normalization in the deeper architectures. They also found that the error rate of the architecture saturates when the depth reaches 19 layers. Different size conv. layers are experimented with to confirm that deep net with small filters outperforms a shallow net with large filters. It is found that training set augmentation by scale jittering is helpful for capturing multi-scale image statistics. A multi-crop evaluation is also done and authors find that using multiple crops performs slightly better than dense evaluation. Finally, they combine the outputs of several models by averaging their soft-max class posteriors and this fusion improves the performance of network due to complementarity of the models and it is used in the ILSVRC 2012 and 2013.

Finally, in the classification task of ILSVRC-2014 challenge, the VGG team is able to secure the 2nd position with 7.3% test error using ensemble of 7 models. In terms of single-net performance, their architecture achieves the best results in the competition. In the appendix of the paper, authors show how their models generalizes well to a wide range of tasks and datasets and matches or outperforms other more complex architectures.

Strengths

- The paper is well-organized and evaluation of models is done thoroughly.
- The architecture provided is very simple and elegant yet powerful mainly by just the addition of extra layers.

Weaknesses

- Authors mention that error rate of their architecture saturates when the depth reaches 19 layers. But they haven't carried out any further evaluation in paper with more deeper networks to substantiate their claim that even deeper models might be beneficial for larger datasets.

Points of Confusion

- I am not sure about why the incorporation of 1 x 1 conv. layers provides a way to increase the non-linearity of the decision function without affecting the receptive fields of the conv. layers.
- The authors do conjecture in the paper that VGGNets require less epochs to converge due to implicit regularization imposed by greater depth and smaller conv. filter sizes and pre-initialization of layers. I am confused about this conjecture.

Discussion Questions

- Why does scale jittering helps in capturing multi-scale image statistics and why using it at test time leads to better performance?

- Even though multiple crops evaluation and dense evaluation are complimentary, why does their combination outperforms each of them?
- For the testing images, they have rescaled the images to a scale Q which is not equal to the training scale. How does using several values of Q for each S leads to improved performance of the network?