VGG Summary

Roger Marí [1], Joan Sintes [2], Àlex Palomo [3], Àlex Vicente [4]
Universitat Pompeu Fabra, Image Processing and Computer Vision Group
{ [1]roger.mari01, [2]joan.sintes01, [3]alex.palomo01, [4]alex.vicente01 }@estudiant.upf.edu

***Abstract*—This is a summary of the original paper by Karen Simonyan and Andrew Zisserman [1].**

I. Introduction

The authors of the VGG, who participated in the ImageNet challenge of 2014, investigated how the number of convolutional layers affected the accuracy of the results produced by the CNNs. They proved that larger convolutional network depth is beneficial and is feasible if small filter size is used.

II. Network Architecture

The authors stacked conv. layers with small filter size (3x3), stride 1 and padding to preserve spatial resolution. They also used 5 max-pooling layers after some of the conv. layers, with 2x2 pool window and stride 2. The output of the conv. and pooling layers is processed by 3 fully-connected layers (first two with 4096 units, last one with 1000). The last layer outputs the soft-max probabilities of the 1000 classes involved in the ImageNet challenge. The rest of hidden layers use ReLU.

Six specific designs (named A-E) following the previous architecture were proposed, differing only in the depth (from 8 conv. layers in A to 16 in E). The number of filters in each conv. layer is doubled after maxpool layers, starting from 64 filters in the first conv. layer. By stacking multiple conv. layers with small filter size (3x3), the deeper layers end up having similar (larger) receptive fields to previous top-performing nets (e.g. 7x7), while taking advantage of extra non-linearities.

III. Training process

The only preprocessing employed was mean subtraction. Mini-batch gradient descent, with batch size 256 and momentum 0.9, was used for weight optimization (74 epochs to converge). Weight decay regularization and dropout layers after the two first fully connected layers (dropout ratio 0.5) were used. The initial learning rate was set to 10^{-2}, and decreased by a factor of 10 when the val. set accuracy stopped.

The weights of the first and last layers of each of the networks B-E were initialized with the values obtained after training the previous net (in alphabetical order). The weights of A were initialized randomly from a normal distribution with zero mean and 10^{-2} variance. The authors claim that Glorot-Bengio initialization can also be directly used to initialize any of the architectures without need of pre-training.

The size of the input images was set to 224x224 by randomly cropping the original images. Random horizontal flip and random RGB color shift were used for data augmentation. At test time, the fully-convolutional network was applied to the input images rescaled at different sizes (from a minimum scale to the original whole image) or cropped, and also horizontally flipped. In these experiments, the class scores for each image were averaged to obtain the final class probabilities.

IV. Classification experiments

The classification performance was evaluated using two measures: the top-1 error (proportion of incorrectly classified images) and top-5 error (proportion of images such that the ground-truth category is outside the top-5 predicted categories). Different experiments were carried out:

- *Single scale evaluation*. The test images were evaluated at a fixed size. Classification error (both top-1 and top-5) decreased as the depth of the net increases. Final minimum error (obtained with E net, the deeper VGG): 25.5% (top-1) and 8.0% (top-5).
- *Multi-scale evaluation*. The model was run over several rescaled versions of each test image. Final minimum error: 24.8% (top-1) and 7.5% (top-5).
- *Multi-crop evaluation*. The model was run over multiple crops of each test image. Using a large set of crops can improve accuracy as it results in a finer sampling of the image. Final error: 24.6% (top-1) and 7.5% (top-5).
- *Combination of multi-scale and multi-crop evaluation*. The combination of both methods outperforms each of them. The authors claim that this is due to a different treatment of the convolution boundary conditions. Final minimum error: 24.4% (top-1) and 7.2% (top-5).
- *Combination of different models*. By averaging the soft-max class probabilities of the two top-performing/deeper VGG models (D and E nets) the error was further reduced to 23.7% (top-1) and 6.8% (top-5).

V. Conclusion

The authors of the VGG demonstrated that convolutional network depth is beneficial for classification accuracy. The VGG model from 2014 clearly outperformed the success of previous ImageNet winners from 2012 and 2013, and also achieves state-of-the-art accuracy in other image recognition datasets (the latter results are commented in the appendices of the original paper).

References

[1] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.