ResNet 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 Kaiming He, Xiangyu Zhang, Shaoqing Ren and Jian Sun [1].**

I. Introduction

The VGG net highlighted the benefits of convolutional network depth on classification accuracy. However, very deep nets present a *degradation* problem. As depth increases, accuracy gets saturated and then degrades. This degradation is not caused by overfitting, but by the addition of excessive conv. layers. Consequently, deeper CNNs are more difficult to train.

The authors of the ResNet, who came first in the ImageNet challenge of 2015, proposed to address the degradation problem by introducing a *deep residual learning* framework.

II. Deep Residual Learning

Instead of trying to approximate the ideal underlying mapping that each block of conv. layers should fit, the blocks are allowed to fit a residual mapping. If the ideal mapping is denoted as $H(x)$, then the residual mapping $F(x)$ is such that $H(x) = F(x) + x$. The idea is that it is typically easier to optimize the residual mapping $F(x)$ than the underlying mapping $H(x)$. For instance, if $H(x)$ was an identity mapping, it would be easier to set $F(x) = 0$ than to approximate $H(x)$.

The formulation of $F(x) + x$ is put in practice using feed forward nets with *shortcut connections*. Shortcut connections are those skipping one or more layers. In the ResNet, the shortcut connections perform identity mapping (their output is the input) and they are used to add the input x to the residual mapping $F(x)$ of each conv. block. The advantage is that they add neither extra parameters nor computational complexity.

III. Network Architectures

The authors proposed two different architectures to show the advantages of deep residual learning.

- *Plain Network*. Multiple conv. layers plus a fully connected soft-max layer of 1000 units to output the class probabilities. The design was partially inspired by the VGG: 3x3 conv. filters mostly, and each time the feature map size is halved, the number of filters in the conv. layers is doubled. Instead of pooling, conv. layers with stride 2 were used to downsample the feature maps.
- *Residual Network*. Shortcut connections are added to each block of 2 conv. layers of the plain network to adapt it to the deep residual learning framework. When the dimension of the input of the block does not match the output two options are considered: (1) zero-padding or (2) linear projection to match dimensions.

Both models were implemented following the practice of the VGG net. The size of the input images was set to 224x224 and per-pixel mean subtraction was used. Gradient descent with a mini-batch size of 256 was used to train both networks. Weight decay of 10^{-4} and momentum of 0.9 were employed. The initial learning rate was set to 0.1, and decreased by a factor of 10 when the val. set accuracy reached a plateau. An important change is that batch normalization layers were added after each convolution, while dropout was not used.

IV. Classification experiments

The classification performance was evaluated using the top-1 and top-5 errors. Different experiments were carried out:

- *Plain Networks*. The results with a 18-layer and 34-layer plain networks showed the effect of the degradation error on traditional CNN structures, with the 34-layer net achieving worse classification accuracy.
- *Residual Networks*. The results with a 18-layer and 34-layer ResNet showed that the degradation error was correctly adressed, with the 34-layer achieving the best accuracy (also with respect to the plain nets).
- *Deeper Bottleneck Architectures*. For deeper ResNets, the authors propose to use a *bottleneck* design for the conv. blocks: 3 layers of 1x1, 3x3 and 1x1 filters. Following this philosophy, they built 50/101/152-layer ResNets. The degradation error was not observed and the classification accuracy improved as depth increased, reaching a minimum of 19.38% (top-1 error) and 4.49% (top-5 error) for the 152-layer model. An ensemble of 6 ResNets of different depth (up to 152 layers) obtained a top-5 error of 3.57% in the ImageNet challenge, which is much lower than the ones obtained by previous state-of-the-art models (e.g. the VGG net provided a top-5 error around 7%).

V. Conclusion

The authors of the ResNet demonstrated that residual nets can enjoy accuracy gains from greatly increased depth, overcoming this way the degradation problem of conventional deep CNNs. The ResNet also achieved impressive results on other image datasets (it won the COCO 2015 competitions as well), proving that the residual learning principle is generic and applicable to other computer vision problems.

References

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.