Review: Deep Residual Learning for Image Recognition

Jiho Bak
School of Electrical and Electronic Engineering, Yonsei University

Abstract

Even vanishing gradient problem was largely addressed by initialization and batch normalization techniques, the "degradation problem" kept the networks shallow. Focusing on the importance of the "depth", this paper proposes the concept of "residual learning". The identity mapping by shortcut connection enables the residual learning and makes the layers a reference mapping. This framework eases the training of substantially deep networks. The residual network shows the state-of-the-art performance in various tasks. The theoretical and empirical evidences of residual learning are well elaborated in the paper.

1 Degradation Problem

Theoretically, the deeper neural networks are more capable of approximating the mapping of train data. However, as in Fig 1, the accuracy saturated in deeper networks, even after vanishing gradient problem was quite addressed(by weight initialization, batch normalization). This is called the "degradation problem".

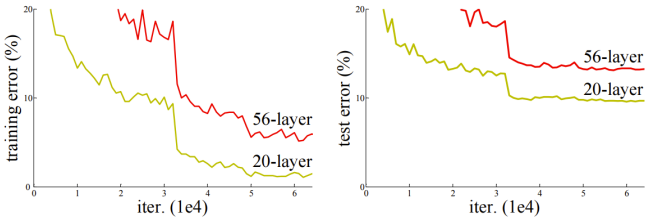

Figure 1: Error of "plain" network; it's indicating the degradation problem

2 Residual Learning

2.1 Theoretical Apporach

Assuming that sufficient nonlinear layers can approximate arbitrary function, is also equivalent to assuming that it can approximate the residual function. The main hypothesis in this paper is that it is easier to optimize the residual mapping than to optimize unreferenced mapping.
In respect of the situation in Fig 1, the added layers in the 56-layer model are not even approximating the identity function. So, if residual reformulation is done and added layers will be easier to approximate it's identity, 56-layer network might achieve the error closer to of 20-layer's. This intuition is verified after the experiment, observing the 'layer responses'.
The identity mapping is optimal for residual networks. However the neural networks are not for the identity mapping, so why are they keep saying about identity mapping? Kaiming He et al. [1] says that as long as the optimal function is closer to identity mapping then to a zero mapping, it is easier to optimize with the reference to an identity mapping than unreferenced one.

2.2 Identity Mapping by Shortcuts

The residual frame work is designed with the building block as:

$$y = F(x, \{W_i\}) + x$$

The function F is consist of repeated conv layer and activation function ReLU. To add, this identity shortcut does not need any additional time or memory complexity.
The projection shortcut ($y = F(x) + Wx$) were also experimented and found out that it's slightly better. But on the contrary, this small difference indicates that projection shortcut is not essential for the degradation problem. And since it cost more memory/time complexity, Kaiming He et al. [1] adopted the identity shortcut. For increased dimension, projection shortcut (1×1 conv) is mainly applied.

3 Residual Network and Experiments

Following the philosophy of VGGNet [2], ResNet basically uses 3×3 conv layers. Also, ResNet remains the time complexity, so if the feature map size is halved, the channel is doubled. The skip connections are applied every two conv layers, and the global average pooling and the fully connected layer is applied at the end. Specific details are organized in the figure of the paper [1].
For implementation and optimization, recent techniques(e.g. batch normalization, augmentation,scheduled SGD with momentum) were applied, but to avoid distraction, other regularizations like maxout/dropout were not used.

3.1 ResNet-18/34

For comparison of degradation problem, three kinds networks were tested to ImageNet classification task, VGGNet, plain network and the residual network. For each plain and residual network, 18-layer and 34-layer models were tested. To have no extra parameter for comparison, zero-padding for increasing dimension was applied to ResNet.
The degradation problem was observed in the plain network, but the reverse was observed in ResNet(Fig 2). This means that the degradation problem was well addressed by the residual learning.Comparing the 18-layer plain network and ResNet, its error is quite accurate but ResNet converged faster. So, we can also confirm that ResNet eases optimization for the same depth. Overall performance of ResNet was better than VGGNet even using only 18% FLOPs of VGG's.

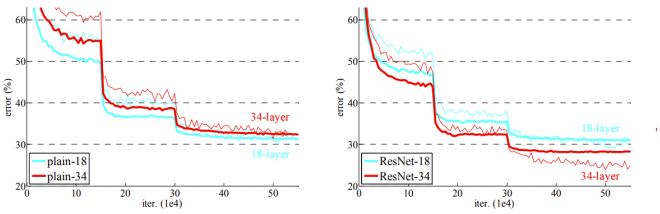

Figure 2: Training on ImageNet. Each train and validation error respectively corresponds to thin and bold curves.

3.2 Bottleneck for Deeper ResNet

In order to find out the effect of even more substantially deep residual network, Kaiming He et al. [1] applied the bottleneck building block(Fig 3). Two 1×1 and one 3×3 conv layers consist the bottleneck 1×1 convolution manipulates the dimension and adds nonlinearlity with less time complexity. The Fig 3 is the example of the same time complexity. ResNet with 50/101/152-layer were tested in ImageNet and with 20 to 110 and 1202 layers were tested in CIFAR-10. For ImageNet task, ResNet showed the-state-of-the-art performance, while it had still lower complexity compared to VGGNet. In addition, still the deeper(152) ResNet performed best among them. For CIFAR-10, the ResNet 110 performed best and 1202 overfitted to data. Exploring over 1000 layers is still open problem.

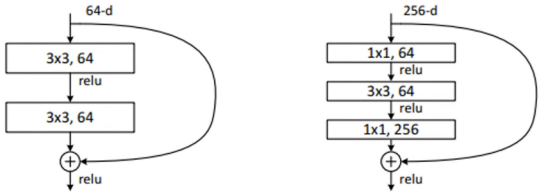

Figure 3: Residual Block. Left: a 'building block' for ResNet-34. Right: a 'bottleneck building block' for ResNet-50/101/152

4 Conclusion

The residual shortcut with bottleneck structure addressed the degradation problem and enabled the very deep neural network. Resulting ease of the optimization and the benefit in time complexity. And since it also showed this sota performance in other tasks(e.g. COCO), we can see that concept of residual learning could be applied in many other future works.

Reference

[1] Shaoqing Ren Jian Sun Kaiming He, Xiangyu Zhang. Deep residual learning for image recognition. *CVPR*.

[2] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR*.