Paper Notes

Deep Residual Learning for Image Recognition

Authors of the Paper: Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun

COMPILED BY: Nikhil Agrawal, Prabhat Sharma

Discussion held on 29th October 2018.

After the celebrated victory of AlexNet at the ILSVRC2012 classification contest, deep Residual Network (ResNet) was arguably the major breakthrough in Deep Learning community by winning the first place in ILSVRC2015. An ensemble of these residual nets achieved just 3.57% error on the Imagenet Dataset. ResNet makes it possible to train networks up to hundreds or even thousands of layers and still achieve compelling performance.

Motivation

1. Previous paper observations:
    a. **Krizhevsky et al. 2012:** Used deep convolutional networks for large-scale image recognition for the first time.
    b. **Highway Networks:** Their architecture is characterized by the use of gating units which learn to regulate the flow of information through a network.
2. The authors argued that stacking layers shouldn't degrade the network performance because we could simply stack identity mappings but deeper networks showed larger training and test error than their shallow counterparts (Figure 1).
3. Very deep networks have a problem with training due to **vanishing gradients**.
4. Another problem with training the deeper networks is performing the optimization on the huge parameter space. Therefore naively adding the layers leads to higher training error. This is called the **degradation problem**.

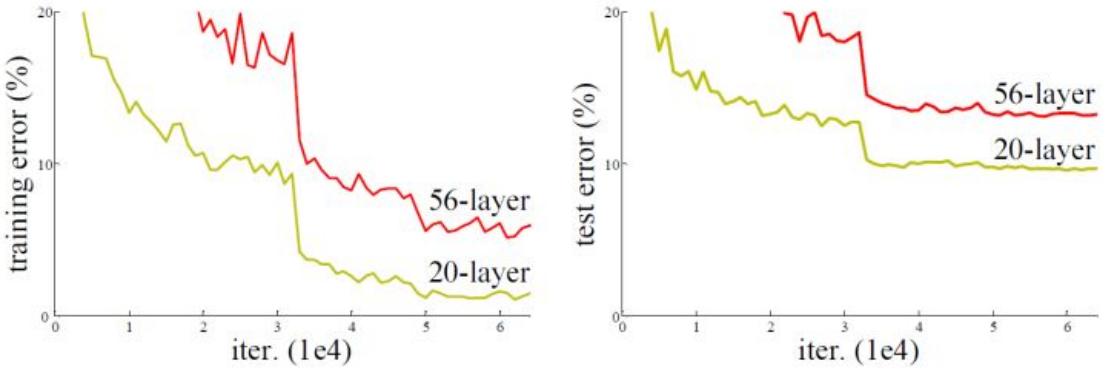

Figure 1: Training error (left) and test error (right) on CIFAR-10 with 20-layer and 56-layer "plain" networks.

Architecture

Residual Block

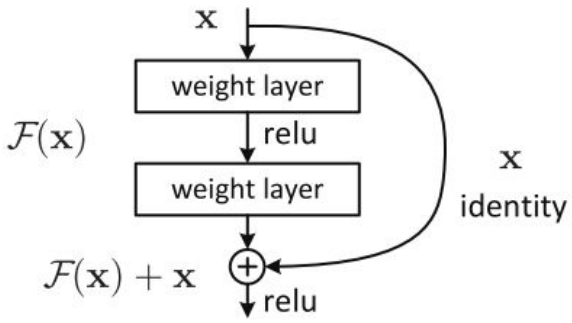

Figure 2: Residual Learning, a building block (taken from paper)

1. Figure 2 shows the smallest building block of a ResNet. It is basically a couple of stacked layers (minimum two) with a skip connection. Skip connections are mainly just identity mappings and hence contribute no additional parameters. Residual learning is applied to these stacked layers.

2. H(x) = F(x, {W}) + x, where:
    a. H(x) is the underlying mapping by the stacked layers
    b. x is the input
    c. F(x, {W}) is the residual function

    Instead of learning H(x), we let these layers approximate a residual function F(x).

3. For addition of F(x, {W}) and x to be possible, they should have the same dimensions, if this is not the case then:

    a. The shortcut still performs identity mapping, with extra zero entries padded for increasing dimensions. This introduces no additional parameters.

    b. Perform a linear projection Ws by the shortcut connections, using 1 × 1 convolutions to match the dimensions, y = F(x, {Wi}) + Wsx. This introduces extra parameters and computation. In other words, we are just adding a 1 × 1 convolutional layer in the shortcut connection with the number of filters equal to the required output dimensions.

    c. In another type of residual block, input x is passed through a convolutional layer to match the dimensions with the residual function.

4. The final activation is applied after addition.

ResNet

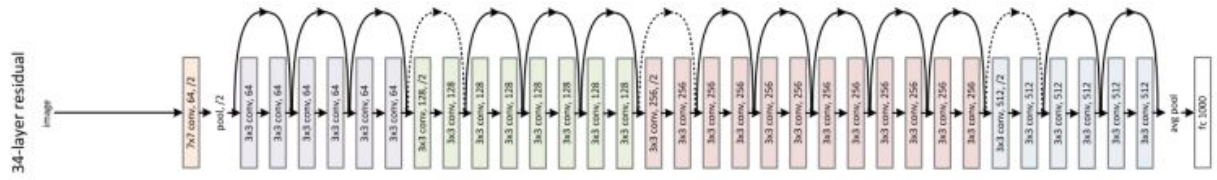

Figure 3: Resnet Architecture (taken from paper)

Figure 3 shows the Resnet Architecture in which bold shortcuts show the same dimension skip connections and dotted shortcuts show the increased dimensions skip connections.

Implementation Details

1. 224 x 224 crops are randomly sampled from an image, resized such that its shorter side is randomly chosen from [256, 480].
2. **Preprocessing:** subtracting the mean of the training set images from every pixel.
3. SGD with the mini-batch size of 256.
4. Standard colour augmentation is used.
5. Batch Normalization after every convolutional layer and before activations.
6. Learning rate is initialized at 0.1 and divided by 10 when error rate plateaus and trained up to 6,00,000 iterations.
7. Weight decay of 0.0001 and momentum of 0.9 is used.
8. Dropout is not used.
9. 10-crop testing is used during the test phase.

Observations

1. Usage of residual block in the Network made the optimization of very deep neural network possible.
2. When a layer has to learn an identity function, weight matrix W tends to zero, causing the residual function to approach zero as well.
3. Due to Resnet training of even more than 1000 layers network is possible, though the writers of paper have shown the result for up to 152 layer network for image classification.
4. ResNet-152 achieved a top-1 error rate of 19.38% and top-5 error rate of 4.49% on ImageNet validation set and its ensembles achieved a top-5 error rate of 3.57% on the test set.

Take Away Message

ResNet showed that with the addition of shortcut connections to the network, training of very deep networks is possible and can achieve higher accuracy than their shallow counterpart. Also, the idea of a shortcut connection is not limited to image classification and can be applied in other fields too.