Summary of RESNET Architecture

WHAT ARE SOME OF THE MAIN PROBLEMS WITH TRAINING DEEP NEURAL NETWORKS?

Problem of Vanishing/Exploding Gradients

This prevents the neural network from converging and hence the weights are not trained properly.
Solution

- Normalized Initialization of Weights
- Normalization of Output of Intermediate Layers(Batch Normalization)

Degradation Problem

This problem arises experimentally not theoretically.
It is observed that when the depth of NN is increased above a certain the training accuracy gets saturated and then slowly starts decreasing on further increasing the number of layers.
This is clearly not due to "overfitting" because the training accuracy drop not the testing accuracy.
This problems of "degradation" is solved using "Deep Residual Networks(ResNet)"

Architecture

To solve the problem of degradation the ResNet introduces shortcut connections to form Residual blocks.

What are shortcut connections?

- In these connections the input to Residual Block is again added to its output before feeding it to the next Residual block.
- There can be two or three Convolutional layers in each of Residual Blocks

What are Residual Blocks?

There are two types of Residual Blocks

- **Basic Residual Block — used for realtively shallow resnets(18 and 34)**
  input -> BN -> ReLu -> Conv1 -> BN -> ReLu -> Conv2 -> output1 -> shortcut function
- **Bottleneck Residual Block — mostly used for deeper resnets(50 and**
  input -> BN -> ReLu -> Conv1 -> BN -> ReLu -> Conv2 -> BN -> ReLu -> Conv3 -> output1->shortcut function

What shortcut function does here?
output=input+output1 - > BN -> ReLu ->final output of residual block
What if dimensions of input and output1 don't match while adding them?
Method A: Zero Padding - No extra parameters are introduced

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

*Method B: Taking projection of output1 using output1_modified=output1 * Ws*

$$y = F(x, \{Wi\}) + Wsx.$$

Method C: Using a convolution in the shortcut function to transform the input to match the dimensions of output1.The main problem with this method is that it introduces more parameters thus increasing complexity

Inferences from the implementation of Resnet on ImageNet 2012

Case of Plain Deep Neural Network

The researchers found that 34-layer plain net has higher validation error than the shallower 18-layer plain net .There is also degradation problem which is visible from the fact that 34-layer plain net has higher training error throughout the whole training procedure, even though the solution space of the 18-layer plain network is a subspace of that of the 34-layer one.

In such situation it may be argued that the decrease in accuracy may be due to vanishing gradients but the researchers argue that :-

- These plain networks are trained with Batch Normalization which ensures that forward propagated signals to have non-zero variances .
- Researchers also verify that the backward propagated gradients exhibit healthy norms with BN. So neither forward nor backward signals vanish.

Case of Residual Deep Neural Networks

The researchers found that ,:-

- the situation is reversed with residual learning – the 34-layer ResNet is better than the 18-layer ResNet (by 2.8%).
- the 34-layer ResNet exhibits considerably lower training error and is generalizable to the validation data.
- This indicates that the degradation problem is well addressed in this setting

Inferences about Identity vs Projection Shortcuts

There researchers give three types of shortcuts:-

A - zero-padding shortcuts are used for increasing dimensions, and all shortcuts are parameter free

B - projection shortcuts are used for increasing dimensions, and other shortcuts are identity

C - all shortcuts are projections

Following inferences were drawn:-

- B is slightly better than A , that this is because the zero-padded dimensions in A indeed have no residual learning
- C is marginally better than B, and we attribute this to the extra parameters introduced by many (thirteen) projection shortcuts
- But the small differences among A/B/C indicate that projection shortcuts are not essential for addressing the degradation problem
- Identity shortcuts are particularly important for not increasing the complexity of the bottleneck architectures

Inferences from the implementation of Resnet on CIFAR-10

The researchers used a relatively simple network architecture on CIFAR-10.The input image is 32 x 32.The architecture is as follows:-

- 2n+1 layers with 16 filters which have a output shape of 32 x 32 x 16
- Next there are 2n layers with 32 filters which have output shape of 16 x 16 x 32
- Next are 2n layers with 64 filters which have output shape of 8 x 8 x 64
- Subsampling(pooling) is performed with stride of 2
- Next is average pooling with stride of 2 with output shape of 4 x 4 x 64
- There are total 2n+2 layers till this point
- Next is fully connected layer which takes (-1,4,4,64) to (-1,10)
- Next is the softmax layer which is final
- All the above layers have a kernel_size of 3 x 3 with pooling 1
- The researchers used identity shortcuts in all cases so that the residual model has exactly the same depth,width, and number of parameter as the plain counterpart

Inferences:-

- The plain deep nets suffer from degradation problem when the depth of the plain nets is increased.The training error increases.

- While on the other hand the residual networks manage to overcome the optimization difficulty and show accuracy gain when the depth increases