Resnets

Learning better networks is not as easy as stacking more layers.
Main 2 problems

- Vanishing, exploding gradients
- Degradation of training accuracy

The solution is to build a deeper model which can learn the same function as its shallow counterpart. (Deep Residual Learning Framework).
We let the deeper layers fit a residual mapping.

This residual connections can be viewed as shortcut connections. The advantages is that it neither adds computational complexity or extra parameters.

Plain networks exhibits high training error when the depth increases unlike the resnets, which are easier to optimise.

One might ask that the function generated by renets can be achieved by plain networks, which is true, but the ease of learning in resnets is higher due to its ability to produce identity mappings.

Degradation problem suggests that it is difficult in approximating identity function by multiple non-linear layers.

The dimensions of F(x) and x should match for the shortcut addition to take place.
If this is not the case, a linear projection W can be used to match the dimensions.
The concept of resnets could also be extended to convolutional networks.
The Resnet model has fewer filters and lower complexity than VGG net.

Implementation Details

- Batch Normalisation is applied after each convolution and before activation.
- Stochastic Gradient Descent with a mini-batch size of 256.
- Learning rate is decreased everytime the error plateaus.
- Momentum of 0.9 is used, and no dropout.
- The network inputs are per-pixel mean subtracted.

Experimental Conclusions

- Deep plain nets have exponentially low convergence rates,which impact the reducing of training error
- A 34-layer plain net was found to have higher training error than 18-layer plain net throughout the training process. (degradation problem)
- 34-layer Resnet with same parameters of plain tested above exhibits lower training error than 18-layer Resnet and is also better at validation,i.e. Degradation problem is addressed.
- This also verifies the effectiveness of resnets on extremely deep learning systems.
- When the network is not too deep (ex - 18 layer), the accuracies of plain and resnet are comparable, the resnet network converges faster.
- Identity short cuts are better than projection shortcuts

- When a shortcut connection connects over 3 layers (more than the usual 2), it is called Deep Bottleneck architecture.

The standard deviation of layer responses of Resnets are less than their plain counterparts, which goes hand in hand with our conjecture that residual functions might generally be close to zero than non-residual.