Introduction to ResNets

Connor Shorten
Jan 24 · 6 min read

'We need to go Deeper' Meme, classical CNNs do not perform well as the depth of the network grows past a certain threshold. ResNets allow for the training of deeper networks.

This Article is Based on Deep Residual Learning for Image Recognition from He et al. [2] (Microsoft Research): https://arxiv.org/pdf/1512.03385.pdf

In 2012, Krizhevsky et al. [1] rolled out the red carpet for the Deep Convolutional Neural Network. This was the first time this architecture was more successful that traditional, hand-crafted feature learning on the ImageNet. Their DCNN, named AlexNet, contained 8 neural network layers, 5 convolutional and 3 fully-connected. This laid the foundational for the traditional CNN, a convolutional layer followed by an activation function followed by a max pooling operation, (sometimes the pooling operation is omitted to preserve the spatial resolution of the image).

Much of the success of Deep Neural Networks has been accredited to these additional layers. The intuition behind their function is that these layers progressively learn more complex features. The first layer learns edges, the second layer learns shapes, the third layer learns objects, the fourth layer learns eyes, and so on. Despite the popular meme shared in AI communities from the Inception movie stating that “We need to go Deeper”, He et al. [2] empirically show that there is a maximum threshold for depth with the traditional CNN model.

He et al. [2] plot the training and test error of a 20-layer CNN versus a 56-layer CNN. This plot defies our belief that adding more layers would create a more complex function, thus the failure would be attributed to overfitting. If this was the case, additional regularization parameters and algorithms such as dropout or L2-norms would be a successful approach for fixing these networks. However, the plot shows that the training error of the 56-layer network is higher than the 20-layer network highlighting a different phenomenon explaining it's failure.

Evidence shows that the best ImageNet models using convolutional and fully-connected layers typically contain between 16 and 30 layers.

The failure of the 56-layer CNN could be blamed on the optimization function, initialization of the network, or the famous vanishing/exploding gradient problem. Vanishing gradients are especially easy to blame for this, however, the authors argue that the use of Batch Normalization ensures that the gradients have healthy norms. Amongst the many theories explaining why Deeper Networks fail to perform better than their Shallow counterparts, it is sometimes better to look for empirical results for explanation and work backwards from there. The problem of training very deep networks has been alleviated with the introduction of a new neural network layer — **The Residual Block.**

The picture above is the most important thing to learn from this article. For developers looking to quickly implement this and test it out, the most important modification to understand is the 'Skip Connection', identity mapping. This identity mapping does not have any parameters and is just there to add the output from the previous layer to the layer ahead. However, sometimes x and F(x) will not have the same dimension. Recall that a convolution operation typically shrinks the spatial resolution of an image, e.g. a

3x3 convolution on a 32 x 32 image results in a 30 x 30 image. The identity mapping is multiplied by a linear projection W to expand the channels of shortcut to match the residual. This allows for the input x and F(x) to be combined as input to the next layer.

Equation used when F(x) and x have a different dimensionality such as 32x32 and 30x30. This Ws term can be implemented with 1x1 convolutions, this introduces additional parameters to the model.

An implementation of the shortcut block with keras from https://github.com/raghakot/keras-resnet/blob/master/resnet.py. This shortcut connection is based on a more advanced description from the subsequent paper, “Identity Mappings in Deep Residual Networks” [3].

Another great implementation of Residual Nets in keras can be found here →

- https://gist.github.com/mjdietzx/0cb95922aac14d446a6530f87b3a04ce

The Skip Connections between layers add the outputs from previous layers to the outputs of stacked layers. This results in the ability to train much deeper networks than what was previously possible. The authors of the ResNet architecture test their network with 100 and 1,000 layers on the CIFAR-10 dataset. They test on the ImageNet dataset with 152 layers, which still has less parameters than the VGG network [4], another very popular Deep CNN architecture. An ensemble of deep residual networks achieved a 3.57% error rate on ImageNet which achieved 1st place in the ILSVRC 2015 classification competition.

A similar approach to ResNets is known as “highway networks”. These networks also implement a skip connection, however, similar to an LSTM these skip connections are passed through parametric gates. These gates determine how much information passes through the skip connection. The authors note that when the gates approach being closed, the layers represent non-residual functions whereas the ResNet’s identity functions are never closed. Empirically, the authors note that the authors of the highway networks have not shown accuracy gains with networks as deep as they have shown with ResNets.

The architecture they used to test the Skip Connections followed 2 heuristics inspired from the VGG network [4].

1. If the output feature maps have the same resolution e.g. 32 x 32 → 32 x 32, then the filter map depth remains the same
2. If the output feature map size is halved e.g. 32 x 32 → 16 x 16, then the filter map depth is doubled.

Overall, the design of a 34-layer residual network is illustrated in the image below:

In the image above, the dotted skip connections represent multiplying the identity mapping by the Ws linear projection term discussed earlier to align the dimensions of the inputs.

Training Results of the Architectures Shown Above: The straight line depicts training error and the static line depicts testing error. The 34-layer ResNet achieves sub 30% error rate, unlike the Plain Network on the left plot. The 34-Layer ResNet outperforms the 18-Layer ResNet by 2.8%.

Table Showing Testing Error of the different depths and the use of Residual Connections

In Conclusion, the Skip Connection is a very interesting extension to Deep Convolutional Networks that have empirically shown to increase performance in ImageNet classification. These layers can be used in other tasks requiring Deep networks as well such as Localization, Semantic Segmentation, Generative Adversarial Networks, Super-Resolution, and others. Residual Networks are different from LSTMs which gate previous information such that not all information passes through. Additionally, the Skip Connections shown in this article are essentially arranged in 2-layer blocks, they do not use the input from same layer 3 to layer 8. Residual Networks are more similar to Attention Mechanisms in that they model the internal state of the network opposed to the inputs. Hopefully this article was a useful introduction to ResNets, thanks for reading!

References

[1] Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. 2012.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. Deep Residual Learning for Image Recognition. 2015.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. Identity Mappings in Deep Residual Networks. 2016.

[4] Karen Simonyan, Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. 2014.