**PAPER SUMMARY**

**DEEP RESIDUAL LEARNING FOR IMAGE RECOGNITION**

**Problem**: Learning better networks is not the same as stacking more layers. When deeper networks start converging towards the minimum the problem of degradation has been seen. With the networks increasing in depth accuracy starts decreasing after it gets saturated at a point. Surprisingly it was seen that this is not due to overfitting since there was an increase in training error as well which was not the case as seen from the experiments. The paper in discussion tries to address this problem.

**Hypothesis:** Consider the desired underlying mapping produced by a network *H(x)* and another mapping *F(x) = H(x) - x.* The hypothesis is that it is easier to optimize *F(x)* than to optimize *H(x)*. Now the desired mapping becomes *F(x) + x.* In other words, it is easier to map a zero function than to map an identity function. In the worst case if the best possible accuracy is achieved at a certain depth the weights of the subsequent layers are zeroed out leaving us with the model we had arrived at.

**Shortcut Connections:** To realise this a shortcut connection is added to the model. A shortcut connection splits the input into two parts. One part goes through the non-linear layers and the other waits on the side unprocessed.

**Deep Residual Architecture:** The convolutional layers mostly have 3\*3 filters and are among one of the two designs in cases where the input and the output are of the same size we use padding = 1 and an additional stride of 2 in cases where the output is decreased to half the size. The network ends with a global average pooling and a fully connected with softmax. To this shortcuts are inserted. The identity shortcut(given by *y = F(x, {Wi}) + x* ) is directly used where the dimensions are the same and at places where this is not the case either (i) the extra zeros are directly padded to increase the dimension or (ii) projection shortcut is used. To use the projection shortcut a linear projection *Ws* is performed the shortcut connections. Thus, the shortcut connection layer becomes

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

*Ws* can also be used in the original identity mapping equation but the experiments have shown that identity mapping is sufficient to address the problem of degradation and is also economical. The form of *F* is flexible. It can be of two or more layers but if it is a single layer then identity mapping becomes similar to a single layer and the implementation has no advantage.

**Deep Bottleneck Architecture:** For the ImageNet because of the concerns of the training time, the basic block of a layer of shortcut connection is modified. A stack of 3 layers is used instead of the usual 2. The three training layers has convolutions of dimensions 1\*1, 3\*3, 1\*1. This decreases the time complexity by decreasing the number of parameters.

**Conclusion:** In the ImageNet experiment it was seen that a plain network with 34 layers performed worse than an 18 layer net. But this was not the case in a residual net. Even in case if cifar10 a deeper layer with an implementation of ResNet showed better results. Thus now we can increase the accuracy of the model by increasing the depth. Towards the end of the paper, it was shown that a 1202 layer performed worse on CIFAR10 than a 110 layer. This was postulated to be due to overfitting since 1202 layer showed better training accuracy than the 110 layer. Possible optimizations(regularization) for this were given.

[Cifar10\_ResNet](https://drive.google.com/open?id=1UdAzqQ9AdB7n5cKlLUa1lFsSzof30L4C)