# Deep Residual Learning for Image Recognition

Summary written by Sahana Srihari
April 24, 2020

**Summary:** The paper introduces the framework for a *deep residual learning* network for construction of deeper networks which are easier to optimize,have good accuracy and low error rates of 3.57%.With emphasis on the characteristics of the depth of the network and keeping in mind the vanishing/exploding gradient problems encumbered with deep layers, they further prove their findings on multiple datasets from various competitions, achieving relatively high improvements compared to all previous methods.

**Related work:** The design of the architecture was mainly influenced by VGG16 [8] and the motivation to construct deeper layers [10]. The architecture addresses the problem of *high training error*[2][9]. Hence, the adopted method was based on shortcut connections [1][6] utilizing residual representations like VLAD[4]. Concurrently highway networks with gating functions are used by [9]. Their implementations using ImageNet [7] closely follows [5] and uses Batch Normalization[3] before each activation.

**Approach:** The authors address the issue of deeper network's inability to converge & degradation problem. They explicitly utilize the notion of identity mapping, increase optimization and create deeper networks. The main point is the introduction of **shortcut connections** which perform identity mapping. The residual mappings are fit as $\mathcal{F}(x) + \mathbf{x}$. The building block is represented as $\mathbf{y} = \mathcal{F}(\mathbf{x}, \{W_i\}) + \mathbf{x}$ where x and y are the input and output vectors of the layer and $\mathcal{F}(\mathbf{x}, \{W_i\})$ is the learned residual mapping. The activation function is ReLU. This represents the condition for with the dimensions are the same and $\mathbf{y} = \mathcal{F}(\mathbf{x}, \{W_i\}) + W_s\mathbf{x}$ is the condition for mismatched dimensions with a linear projection of $W_i$ added.The plain network- conv layers having 3x3 filters such that same output feature maps have equivalent filters and if it is halved then the filters are doubled.Downsampling is achieved directly by setting stride=2, the network concludes with global average pooling & 1000-way FC+softmax. The ResNet is based on the above, with the modification of adding shortcut connections.The images are augmented and the per-pixel mean is subtracted, BN is applied before activation. SGD with a mini-batch size of 256 is used. The models are trained to $60\text{x}10^4$ iterations using a weight decay of 0.0001 and momentum of 0.9.

*Datasets, Experiments and Results:*

(i)ImageNet Challenge: Performance was evaluated using **Plain nets** and **ResNets** 18 and 34 layers. For ResNet,shortcut connections are added to 3x3 filters.The main result is-the 34 layer result performs better than 18 layer & compared to the plain net counterpart- reduced the top-5 error by 3.5%.Options tried-(A)Zero-padding (B) Projections + shortcut connections with identity mapping (C) All projections. Deeper Bottleneck Architecture was tested using 3 layers of 1x1,3x3 & 1x1 convs to create 50-layer network with (B), 101 and 152-layer ResNets.These deeper networks give considerably higher accuracy. The single model 152-layer net has 4.49% top-5 validation error and ensemble had winning rate of 3.57%.

(ii)CIFAR-10:1st layer is 3x3 conv with stacked 6n layers of 3x3 convs. Network ends with global average pooling and 10-way FC layer + softmax. Architectures analysed-20, 32, 44, and 56-layers in comparison to plain net. The 110-layer ResNet converges well with fewer parameters in comparison to FitNet and Highway with error% = 6.45%. They also tested an extremely deep network of 1202-layers, but performed worse than the 101-layer network.

(iii)PASCAL and MS COCO: with faster R-CNNs as the detection method. 6% increase is seen in the COCO standard metric and the ResNet 101-layer network won the 1st place for the ILSRVC and COCO 2015 challeneges.

**Strengths:**The main strength is the ability to go considerably deeper in the network without compromise on the training accuracy. Secondly, it tackles the issue of vanishing/exploding gradients through identity mappings. They also prove the model is easier to optimize and is also highly generalizable as seen by the performance on multiple datasets and tasks. The proposed network is 8x deeper than VGG nets and has a 28% improvement on the COCO dataset. They won the 1st place in the ImageNet competition. The paper is easy to understand and also a large array of examples are provided to give experimental backing.

**Weaknesses:**Adding shortcut connections to shallow networks do not help in training performance, in fact sometimes it might prove to be more detrimental than just using a plain net. The number of layers that we can go deep with is a parameter that varies based on the dataset, for eg- in the ImageNet dataset the 152-layer network performed better but in the CIFAR-10 dataset the best performance from the 101-layer n/w. Since ResNet doesn't use drop out or other regularization methods, some literature has proved that they are susceptible to over-fitting.

**Reflections:** It revitalized the and paved the way for deeper and complex networks with astonishingly good error rates. They show that a simple modification to a plain net can having rippling advantages though the network. Although the

paper focuses on the experimental proof, it would benefit from more explanation on the exact working of the residual block and how the gradients flow through them. Future work could be along the lines of regularization to fix overfitting issues and also proof of performance on other deep learnin tasks.

## References


[1] C. M. Bishop et al. *Neural networks for pattern recognition*. Oxford university press, 1995.

[2] K. He and J. Sun. Convolutional neural networks at constrained time cost. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5353–5360, 2015.

[3] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[4] H. Jegou, F. Perronnin, M. Douze, J. Sánchez, P. Perez, and C. Schmid. Aggregating local image descriptors into compact codes. *IEEE transactions on pattern analysis and machine intelligence*, 34(9):1704–1716, 2011.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[6] B. D. Ripley. *Pattern recognition and neural networks*. Cambridge university press, 2007.

[7] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[8] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[9] R. K. Srivastava, K. Greff, and J. Schmidhuber. Highway networks. *arXiv preprint arXiv:1505.00387*, 2015.

[10] C. Szegedy, W. Liu, and Y. Jia. Sermanet, reed, anguelov, erhan, vanhoucke, and rabinovich,"going deeper with convolutions,". *Proc. CVPR. IEEE*, 2015.