**Deep Residual Learning for Image Recognition**

*Kaiming He Xiangyu Zhang Shaoqing Ren Jian Sun*

* Source: <https://arxiv.org/pdf/1512.03385.pdf>
* Deeper networks are difficult to train. So, the paper proposes residual learning framework.
* The networks given in this paper are deeper yet computationally less costly compared to VGG
* Check datasets: ImageNet detection and localization, COCO detection and segmentation
* Check competitions: PASCAL VOC, MS COCO
* Problems with deeper networks:
  + Computationally quite expensive
  + Vanishing/exploding gradients (this can be solved by effective parameter initializations – Glorot and He initializations and Batch Normalization)
  + Overfitting (various ways to address this – L1 norm, L2 norm, dropout, more data, etc.)
* Besides above problems, there is one more problem:

Common understanding is that deeper networks always improve accuracy. However, experiments show that this is true up to some point. At some point, accuracy starts degrading with the addition of more layers.

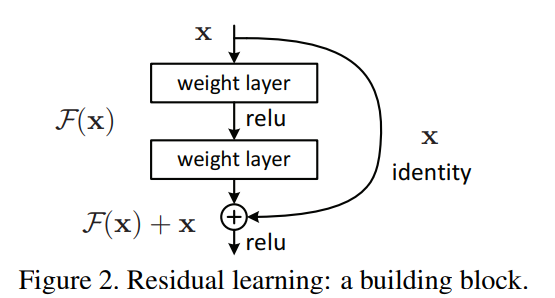
* This degradation problem is solved by the approach in this paper.
* The fundamental idea of the paper:

When you have a stack of layers, they basically learn a mapping function that maps its input, say, x to output, say, y.

The paper hypothesize that this task is difficult. So, they recommend that the layers learn a residual function.

e.g.

Assume that we have a stack of two layers. Let be the mapping function that these layers learn.



The paper proposes that we let the layers learn the residual function .

So, the overall output will still be , but the layers will learn only the part and the remaining part, x, will be added.

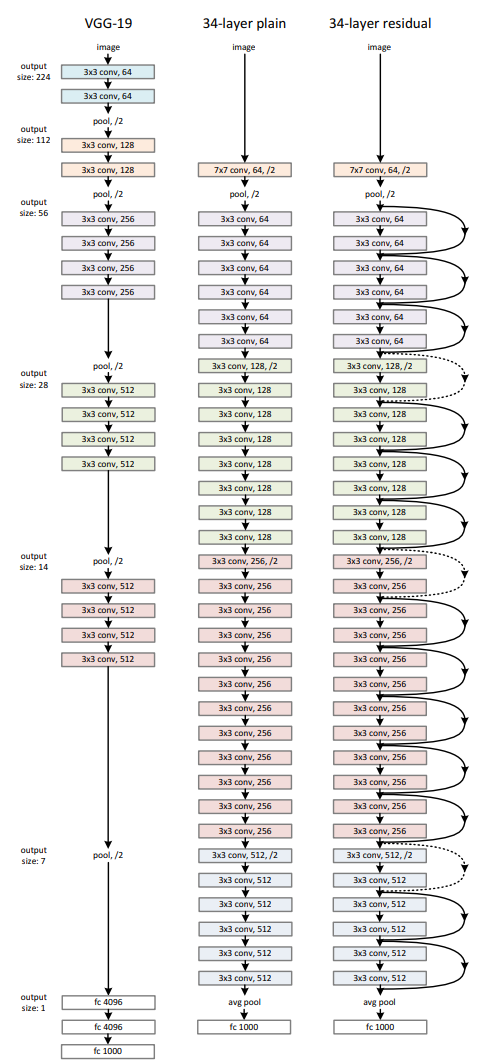
* Such connection from input to some farther layer in the network is called a Shortcut/Skip/Residual connection.
* One thing to note here is that if adding more layers is not really useful for a particular problem, the network will learn to output 0 from residual connection. That way, the identity skip connection will pass x as it is to the next layer. So, adding layers will not degrade performance.
* Such connections provide two benefits:
  + Easier to optimize
  + More accuracy due to more depth
* Authors tested such models on various datasets and proved that such connections are effective.
* These extremely deep representations have great generalization capabilities; they are effective for classification, detection, and segmentation.
* In the above image, note that can be added to only when they have same dimensions. If they have different dimensions, apply some matrix to in the skip connection to transform it.

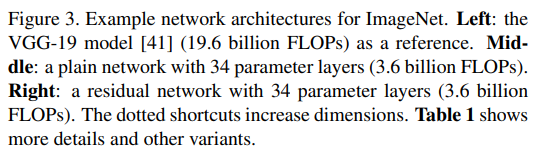
Another way to deal with the dimensional changes is to add zero padding to

Equation when dimensions are same:

Equation when transformation matrix is used:

* We can use such matrix even when and have same dimensions, but generally the identity mapping (i.e. passing as is through the skip connection) is sufficient to deal with the degradation problem.
* The image above shows skip connection over one layer only, but we can skip over multiple layers as well.
* Also, the notations above are for FC layers, but skip connections can be applied to conv layers as well.
* Authors experimented with several plain and resnets.
* They created plain networks containing 18 and 34 layers, inspired by VGG net architecture.
* They also created residual networks containing 18 and 34 layers from the above plain networks by adding skip connections.





* Implementation details:
  + Input image is resized with its shorter side in the range [256, 480]
  + A random 224\*224 crop or its reflection is taken from each image
  + Mean pixel value is subtracted from each pixel
  + Standard colour augmentation (same as that used in AlexNet) is used

Check [AlexNet Color Augmentation](../2.%20ALEXNET/Summary.docx#color_augmentation)

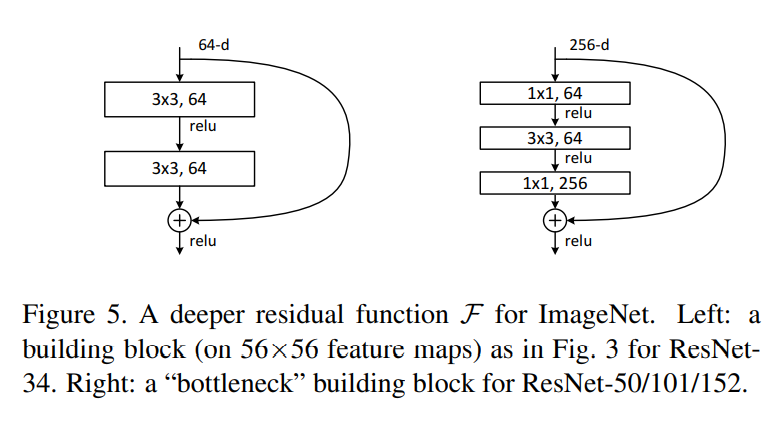
* + Batch Normalization after each conv. layer but before applying activation function.
  + Uses He initialization
  + SGD with batch size of 256
  + Initial learning rate: 0.1, which is divided by 10 when the errors plateau
  + Weight decay: 0.0001
  + Momentum: 0.9
  + No use of dropout
  + Test images are scaled to various values ({224, 256, 384, 480, 640}).
  + During testing: 10 crops per image are used (same as that used in AlexNet)
  + The final error is calculated by averaging over 10-crops of each image for all sizes given above.
* In case of plain networks, it was observed that 34-layer network had higher validation error compared to the 18-layer network throughout the training stage.

This hints at the degradation problem.

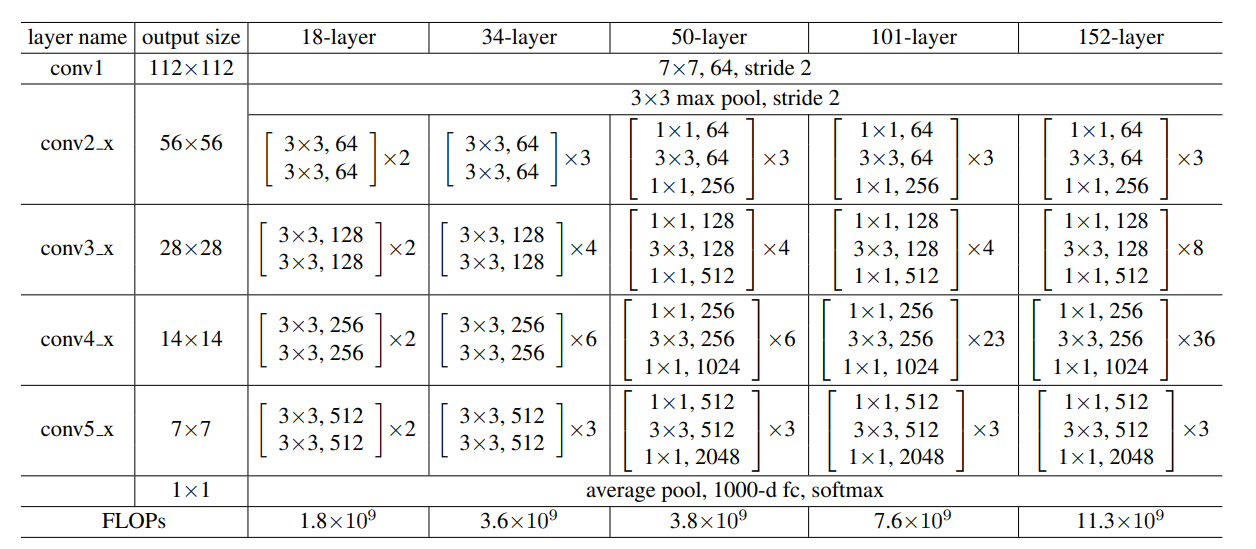
* In case of residual networks, 34-layer network has better accuracy compared to the 18-layer variant. This implies that the residual connections address the degradation problem.
* The 34-layer resnet reduces top-1 error by 3.5% compared to the 34-layer plain network.
* Authors tried identity residual connection as well as projection residual connections. Projection residual connections improved network by a very small amount. So, projection residual connection is not really required for addressing degradation problem.
* To reduce the no. of computations in the network, authors proposed bottleneck design.
* Each residual module now contains 3 layers instead of 2

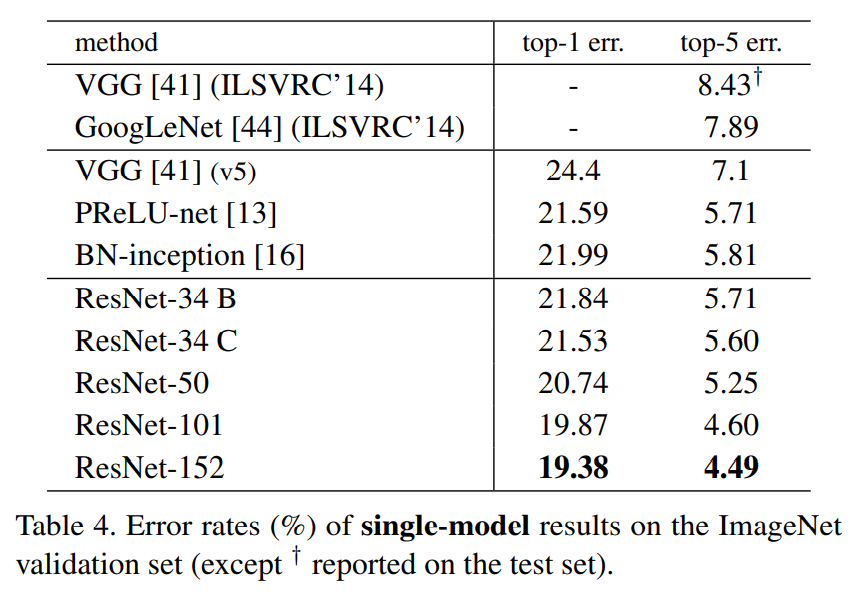
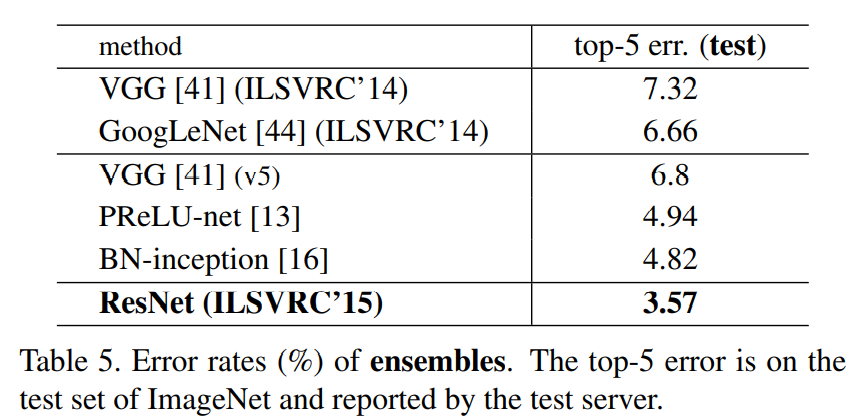
The first layer is a 1\*1 conv. layer to reduce dimensions. The second layer is a 3\*3 conv. layer, same as before. The third layer is again a 1\*1 conv. layer to restore dimensions.

These 1\*1 conv. layers allow the 3\*3 conv. layer to have smaller input and output, and thereby reduce the no. of computations.



* Identity shortcut is important when you have bottleneck layers. If you use projection shortcut, complexity nearly doubles.
* 34-layer resnet model achieves competitive accuracy
* 50, 101, and 152-layer resnents improve accuracy by a significant amount with complexity still less than VGG-16/19.
* A single 152-layer resnet model achieves top-5 validation error of 4.49%
* An ensemble of resnets achieves 3.57% top-5 error.
* The network has great generalization capabilities because of which it can be used for segmentation, classification, and detection.



* Check <https://towardsdatascience.com/an-overview-of-resnet-and-its-variants-5281e2f56035> if required.