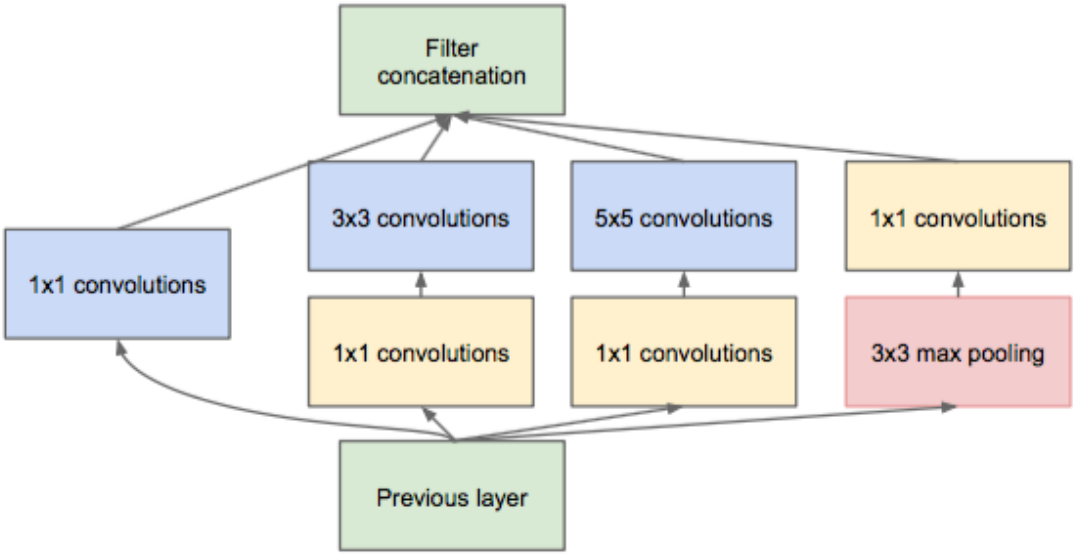
Going Deeper with Convolutions (17 Sep 2014)

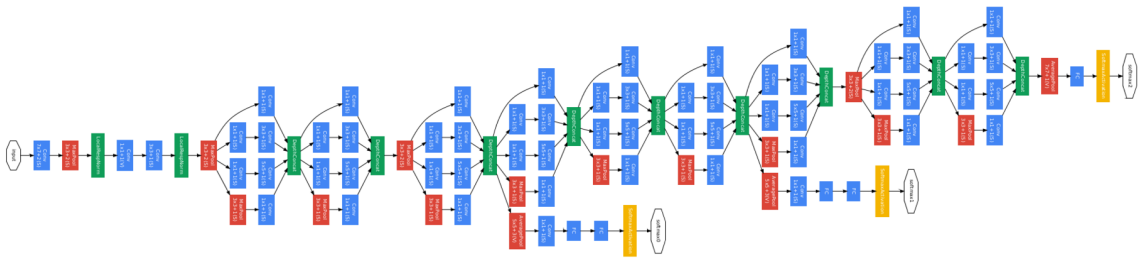
The simplest way to improve the performance of deep neural networks is to increase their size. However, an arbitrary increase in the width (the number of neurons in the layers) and the depth (the number of layers) has several disadvantages. First, an increase in the number of parameters contributes to retraining, and an increase in the number of layers also adds to the problem of gradient decay. Secondly, an increase in the number of convolutions in a layer leads to a quadratic increase in computations in this layer. The way to solve these problems is to introduce sparseness and replace completely connected layers with sparse, even inside the bundles. Thus, the following architecture "inception" module was introduced:

- Use of small size convolutions (1 * 1, 3 * 3, 5 * 5) with a large number of layers. A 1x1 filter detects a correlation between channels and reduces the number of measurements. Larger filters respond to more global features.
- Replacing the hidden FC layers with the global average pooling in order to reduce the number of parameters without large losses information.



(b) Inception module with dimension reductions

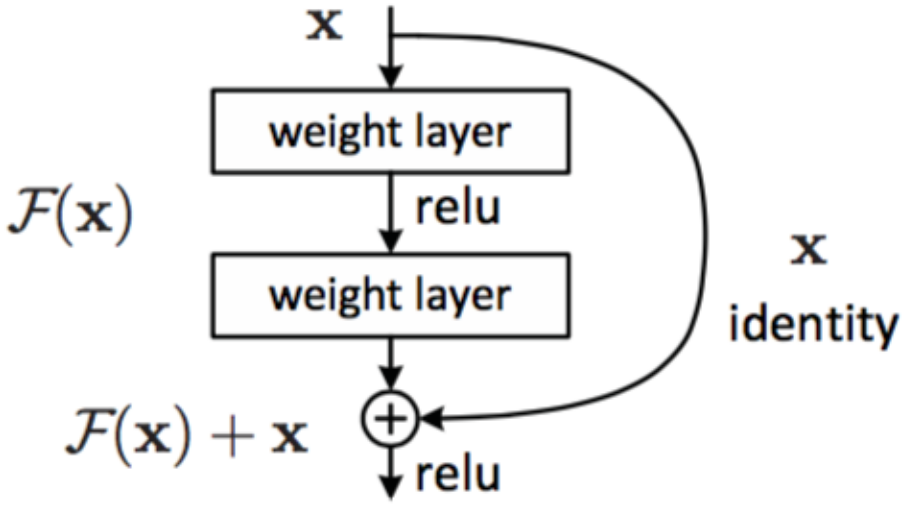
The network consists of 9 such units. several additional classifiers have been identified at different levels - this was done to make it easier to train such a deep network. In each such classifier there are FC layers, which are additional regularization.



In this design, for the first time, fewer parameters were achieved, as well as a significant quality gain at a modest increase of computational requirements. Thus, GoogLeNet, consisting of more than a hundred basic layers, has almost 12 times less parameters than AlexNet (about 7 million parameters against million). Inception-v1 - winner of ILSVRC 2014 with top-5 6.7% error.

Deep Residual Learning for Image Recognition (10 Dec 2015)

The authors of this article were able to find such a topology in which the quality of the model grows with the addition of new layers, thus solving the degradation problem. The idea is that deeper levels predict the difference between what the previous layers and the target variable give, giving you the opportunity to divert the weight to zero and just skip the signal (Shortcut Connections).



Let the original network should calculate the function H(x). We define it residual function as F(x) = H(x) - x, which, in theory, should be easier to learn by the network. By adding skipping connections, the network learns the residual function, which then stack with identical transformation.

As one of the rationales of their hypothesis that the residual functions will be close to zero, the authors present a graph of the activation of residual functions depending on the layer and to compare the activation of layers in flat networks. In a sense, it can be said that CNN itself determines its depth.

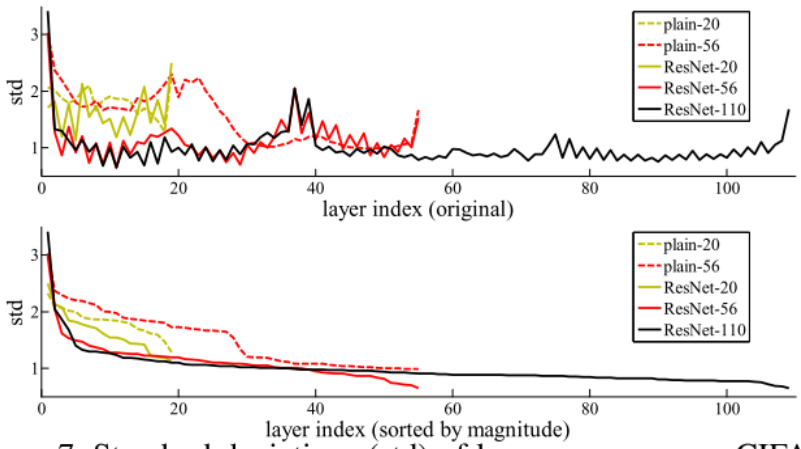


Figure 7. Standard deviations (std) of layer responses on CIFAR-10. The responses are the outputs of each 3×3 layer, after BN and before nonlinearity. **Top**: the layers are shown in their original order. **Bottom**: the responses are ranked in descending order.

Thus, a breakthrough was achieved in the depth of networks with a radically reduced number of calculations and parameters. This model contains fewer parameters than the 19-layer VGG, with a depth of 152 layers. ResNet became the winner of ILSVRC 2015 with a top-5 error of 3.57%