RESNETS

VANISHING AND EXPLODING GRADIENTS

During backpropagation, the calculation of (partial) derivatives/gradients in the weight update formula follows the Chain Rule, where gradients in earlier layers are the multiplication of gradients of later layers

$$\frac{\partial Loss}{\partial W^l} = \frac{\partial Loss}{\partial O^l}\frac{\partial O^l}{\partial z^l}\frac{\partial z^l}{\partial W^l} \quad \text{where} \quad z^l = W^l * O + b$$

As the gradients frequently become smaller until they are close to zero, the new model weights (of the initial layers) will be virtually identical to the old weights without any updates. As a result, the gradient descent algorithm never converges to the optimal solution. This is known as the problem of vanishing gradients

if the gradients get larger or even NaN as our backpropagation progresses, we would end up with exploding gradients having big weight updates, leading to the divergence of the gradient descent algorithm.

When we increase the number of layers in a deep neural network, there is a common problem in deep learning associated with that called the Vanishing/Exploding gradient.

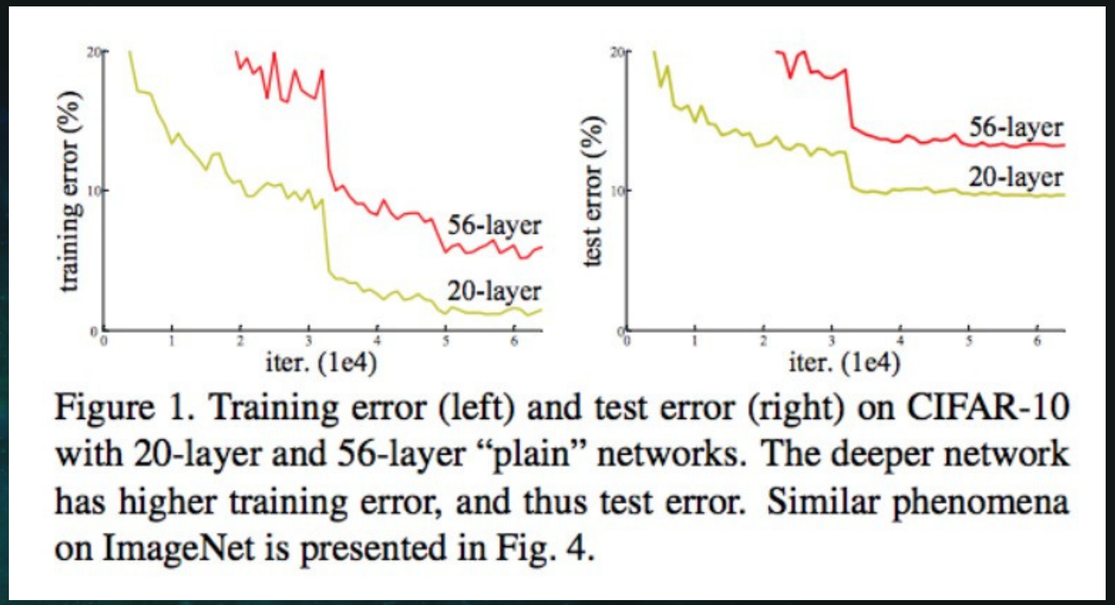
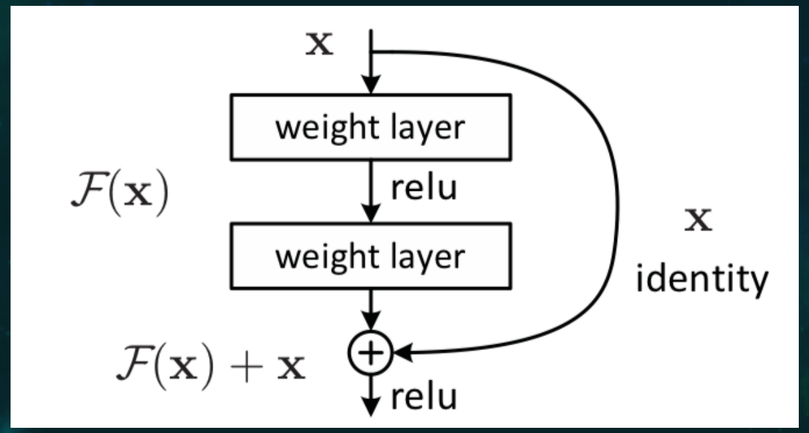


Figure 1. Training error (left) and test error (right) on CIFAR-10 with 20-layer and 56-layer "plain" networks. The deeper network has higher training error, and thus test error. Similar phenomena on ImageNet is presented in Fig. 4.

we can observe that a 56-layer CNN gives more error rate on both training and testing dataset than a 20-layer CNN architecture. After analyzing more on error rate the researchers were able to reach conclusion that it is caused by vanishing/exploding gradient.

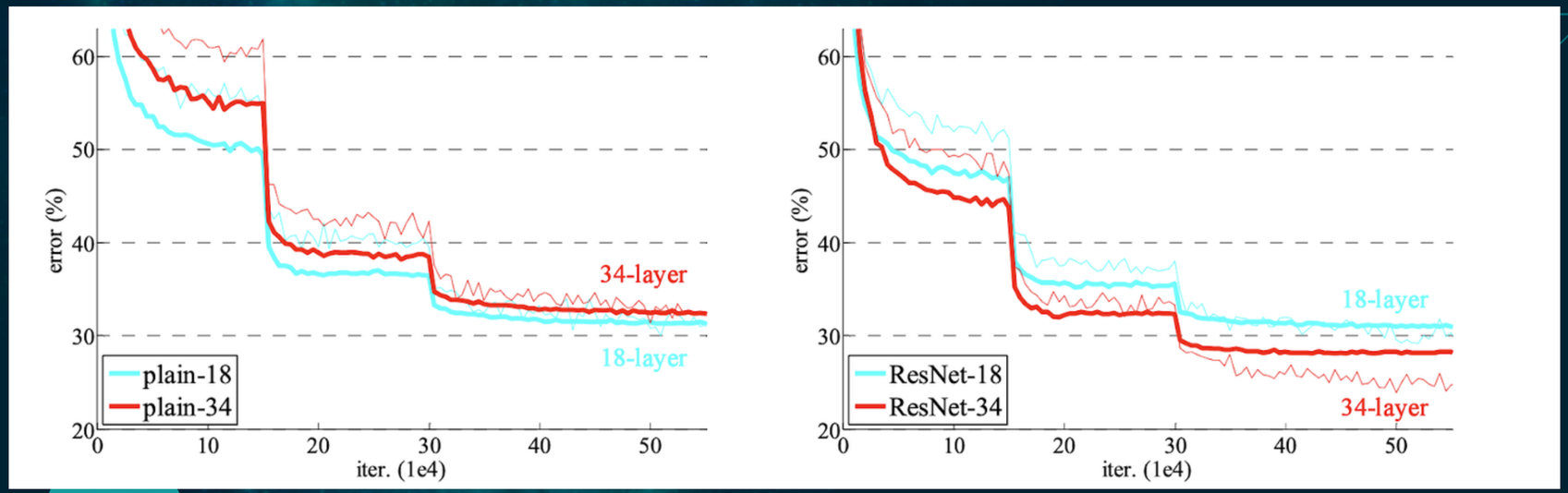
RESIDUAL NETWORK

To solve the problem of the vanishing gradient, this architecture introduced the concept called Residual Blocks. In this network, a technique called skip connections is used. The skip connection connects activations of a layer to further layers by skipping some layers in between. This forms a residual block. Resnets are made by stacking these residual blocks together

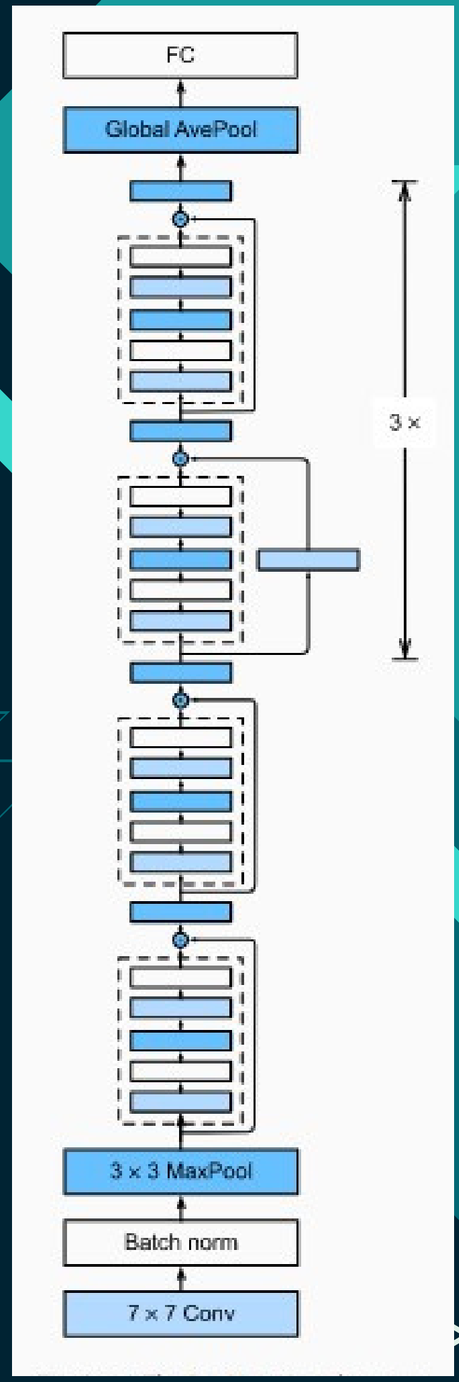


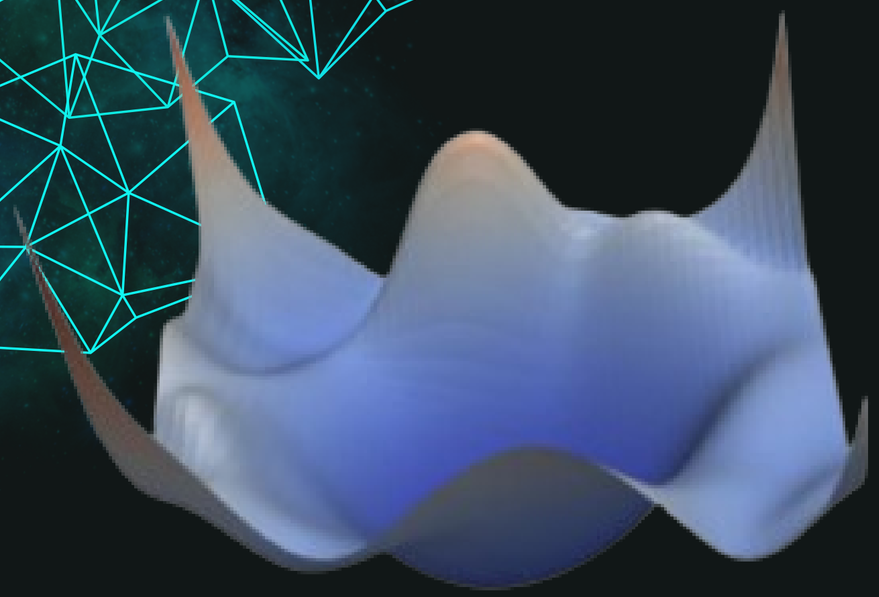
The advantage of adding this type of skip connection is that if any layer hurts the performance of architecture then it will be skipped by regularization. So, this results in training a very deep neural network without the problems caused by vanishing/exploding gradients.

ResNet performs extremely well with deeper architectures.

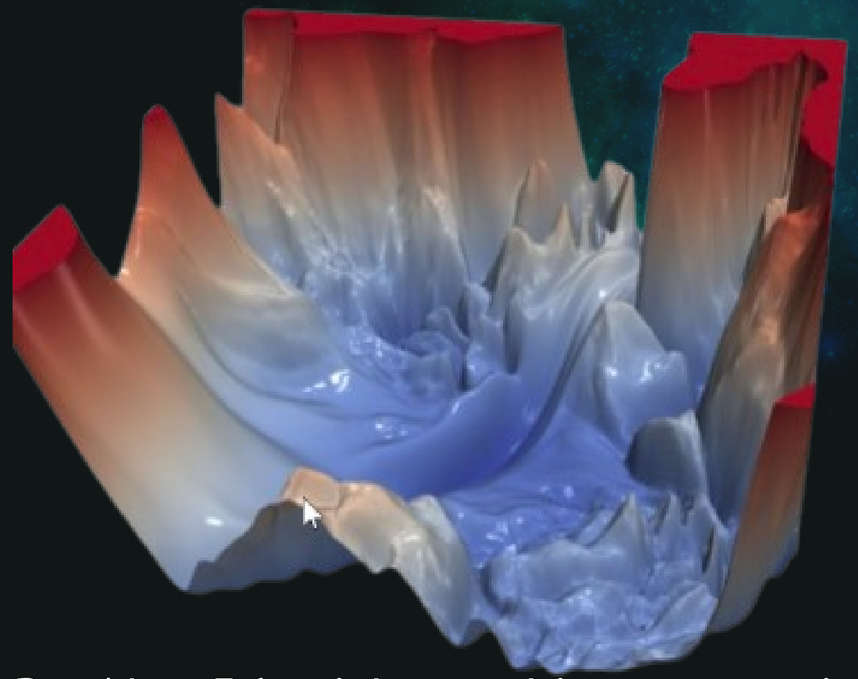


As the training continues , the model grasps the concept of retaining the useful layers and not using those that do not help. The model will convert the later into identity mappings.





ResNet 56 with skip connections



ResNet 56 with no skip connections

- Residual network loss landscapes are dominated by wide , flat minimizers surrounded by large regions of apparent convexity that capture far-away initializers
- Without skip connections, loss landscape is populated by many sharp minima with many small regions of convexity,creating strong dependence on initialization which makes it difficult for the model to converge to a global minima.

IMPLEMENTING RESIDUAL BLOCK

- A residual block has a 3 x 3 convolution layer followed by a batch normalization layer and a ReLU activation function. This is again continued by a 3 x 3 convolution layer and a batch normalization layer. The skip connection basically skips both these layers and adds directly before the ReLU activation function.

- For a residual block if we changes the number of channels (in_channels != out_channels) then we have to use 1x1 convolution in the skip connection to keep the number of output channels same for both pathways.

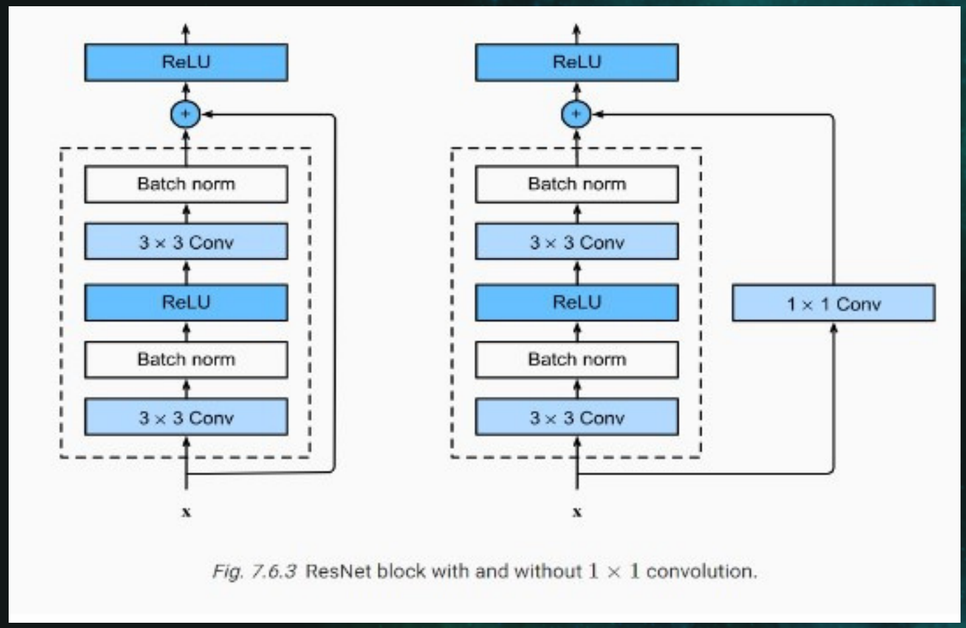
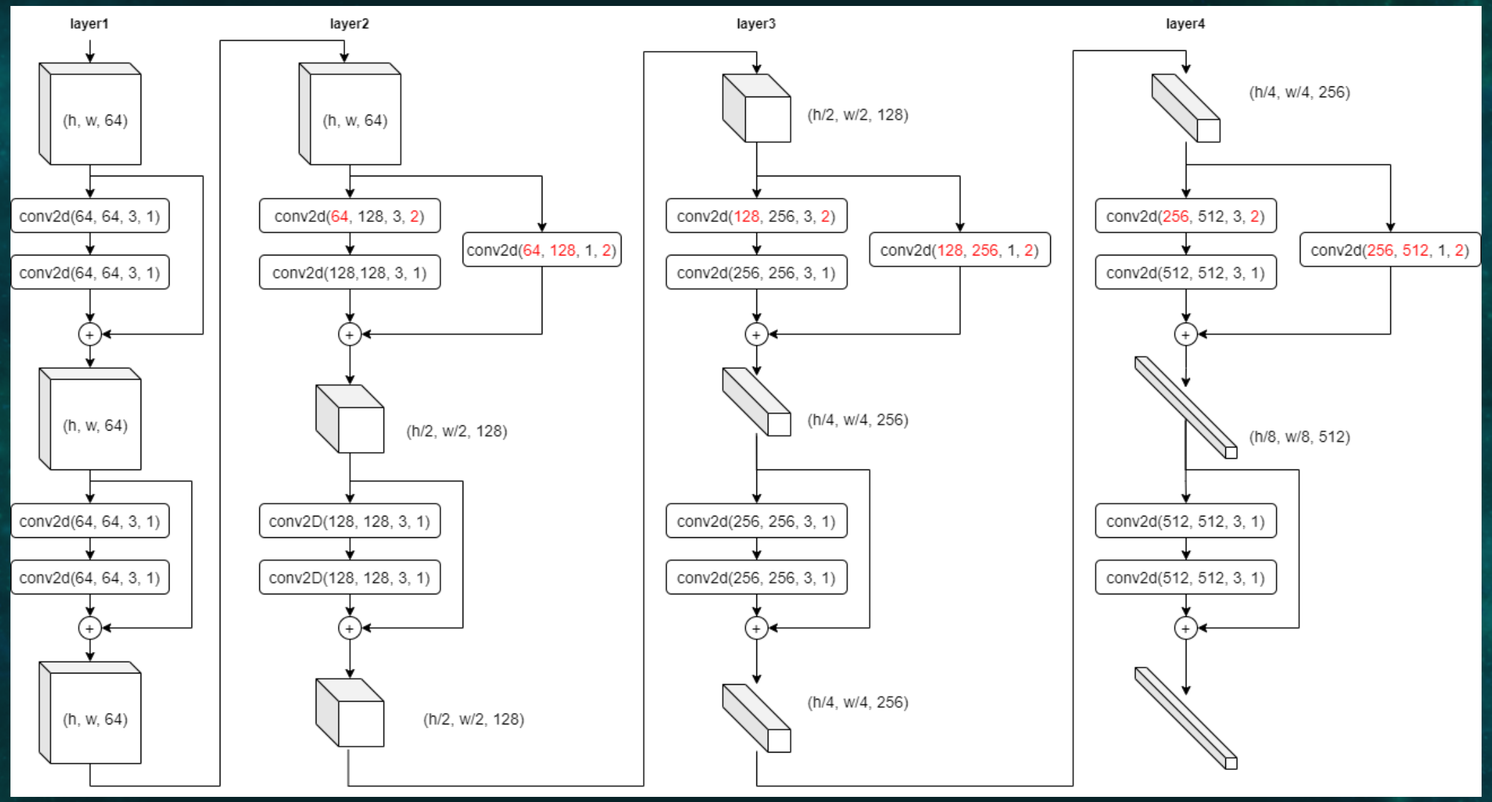


Fig. 7.6.3 ResNet block with and without 1×1 convolution.



Implementation of above resnet architechture can be found here.