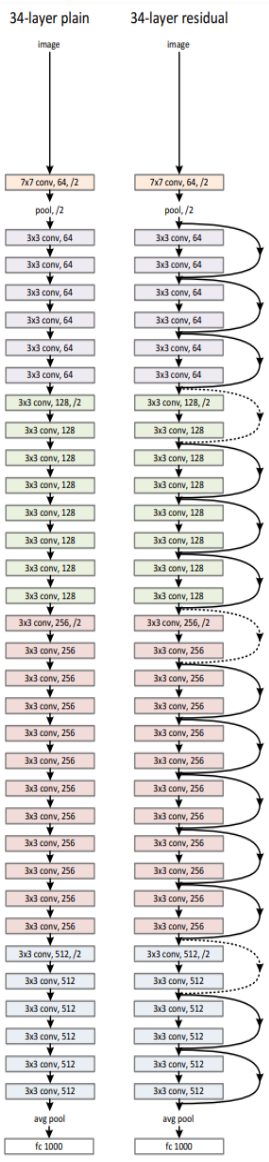
Svein Ole Matheson Sevle, sosevle, 477604
Vemund Fredriksen, vemundfr, 999724

# Assignment 2

## Essay on Deep Residual Learning for Image Recognition[i]

This article summarizes and discusses aspects of the paper Deep Residual Learning for Image Recognition by Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. The paper is published in 2015, in the time where deeper neural networks has been showing great improvements in the task of recognizing and classifying objects in images. With the increased depth of the networks at the time, problems with the vanishing/exploding gradients became more of an issue than before. One of the counter measures against these problems was to let some layers to have connections to multiple layers. In the plain, deep neural networks each layer forward feeds only to the next layer in the network, whereas residual neural networks allow connections to skip one or more layers (figure below).




The papers hypothesis is that having some of these so called “highways”, the connections that can “jump” layers, will lower the issue of vanishing/exploding gradients and improve the learning efficiency. Generally deeper networks may catch more complex features in an image, but the expected training time also increases drastically, so the paper also discusses this aspect.

The last part of the paper addresses how well residual networks of different depth performed on dataset such as ImageNet and CIFAR-10 in terms of accuracy and training error, compared to how the normal, feed-forward deep neural network with the same number of layers and parameters performed. It is shown that when there are more layers in the network, then the residual network will have lower error and higher accuracy. It shows how the error decreased when you make a deeper and deeper network, but when the networks get to deep, here 1202 layers, then the error is higher than when the network had 110 layers.

The research goals of this article could be to provide a solution to the degradation problem in deep neural networks. A solution to the degradation problem is necessary to be able to use deep neural networks because without a solution to this problem, then networks shallower than n, would always perform better than deeper networks which is deeper than n.

The authors in this article provide one possible solution to the degradation problem. This solution is to use deep residual networks. Through the article the authors are investigating if the use of deep residual networks is a good solution to the degradation problem. This is done by first

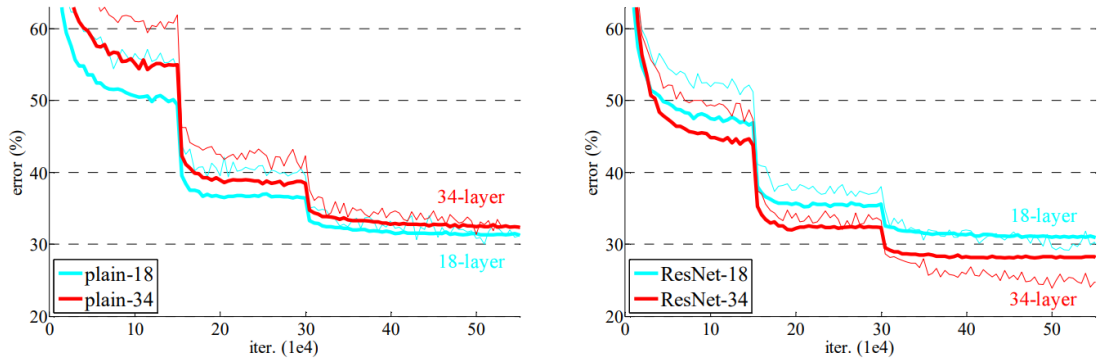
explaining what it is and how it works and then experiment how different deep residual networks perform in terms of error and accuracy on two different datasets. The datasets used in this article is the ImageNet and CIFAR-10 datasets.

The research methodology used in this paper is experimental. When the authors address their results, they do this by referring to numbers and tables. These numbers are obtained by performing experiments on specified networks and measure different parameters during these experiments. The measurements done in the experiments are mainly the error percentage for each batch. The error percentage is calculated by performing predictions on 50.000 validation images. The final results can be found in the paper, the graphs look like the example appended to the appendix. It is important to note that, as the author makes clear, the number of parameters in the residual networks they test against the plain networks is approximately the same as the plain networks. This means that the residual networks tested takes up the same amount of memory and complexity. Therefore, we'd, like the author, argue that if the results show significant improvement, residual networks might be better than plain networks in any situation because there are no complexity/size compromise.

These numbers and tables would be the result of this article. They give information about what happens to the accuracy and error with different parameters. This data gained here is then compared to the state-of-the-art methods (not residual networks) and then draw a conclusion from this.

First the ImageNet dataset is used for gaining information about the error and accuracy for the deep residual networks and the plain networks. First an 18-layer and a 34-layer plain neural network are used. Then the error after each iteration is recorded in a table. The same is then done for a ResNet-18 and ResNet-34. For the plain networks the error is higher for the 34-layer network than for the 18-layer network. This is the degradation problem. For the ResNet the error is reduced for the ResNet-34 compared to the ResNet-18. This indicated that deep residual networks are a possible solution to the degradation problem.

The error between these four networks is also compared. The ResNet-34 has the smallest error. Both networks with 18 layers have approximately the same error after 50000 iterations, but the ResNet-18 is converging faster toward this point than the plain-18. The plain-34 network has the highest error and is also converging slowest.



Next we are presented with error of even deeper residual networks on the ImageNet dataset. We are presented with the error of a residual network with a depth of 50, 101 and 152. The error for ResNet-50 is 20.74%. The error for the ResNet-101 is 19.87%. The error for the ResNet-152 is 19.38%. The error reduction when the layers increase indicate that the degradation problem is not a problem here. We are also presented with the FLOPs that is required for each of these networks and we see the number increases as the number of layers increases.

The CIFAR-10 dataset is also used for observing the performance of deep residual networks. We are presented with the number of layers for each of the tested residual networks and their respective number of parameters. The number of ResNet layers used for CIFAR-10 testing is 20, 32, 44, 56, 110, 1202. The error for the ResNet with the respective layers are 8.75, 7.51, 7.17, 6.97, 6.43, 7.93, as seen in the partially table below. With one exception we see that deeper network result in lower error, however for the network with 1202 layers we see that the error for this network is higher than the error for the layer with 110 layers.

| | | | |
|---|---|---|---|
| ResNet | 20 | 0.27M | 8.75 |
| ResNet | 32 | 0.46M | 7.51 |
| ResNet | 44 | 0.66M | 7.17 |
| ResNet | 56 | 0.85M | 6.97 |
| ResNet | 110 | 1.7M | **6.43** (6.61±0.16) |
| ResNet | 1202 | 19.4M | 7.93 |

With these results in mind, its safe to say that residual networks are a possibility one should at least look further into, if not apply to networks as of today. As previously mentioned, the implementation of residual connections does not inflict a big cost measured in storage or computation complexity, but the results, especially in deep networks can be much improved with the residual connections. We see that residual networks became more popular after the publish of the paper, now state of the art computer vision classifiers like the ones often used in YOLO uses residual networks rather than plain networks.

Though there is still much to research on the field, the trend the recent years have been that deeper network seems to be a way of solving more complex challenges. If this trend continues for the years to come, the fact that the residual connections can reduce the issues associated with the degradation problem is an important discovery that might just make it possible for neural networks to solve complex problems that they could not do with plain networks. Also, we can see that there has been put a lot of effort into making training as efficient as possible lately. Dedicated AI-Chips are manufactured in order to take training time from the magnitude of months down to weeks. The same way that optimizing an algorithm often leads to much bigger performance improvement than pure hardware upgrading can do, an optimization of the neural network training ‘algorithm’ can perhaps make the same great efficiency jump in the training period. Since this paper’s authors argue that often the residual networks can reach local minima quicker than their plain network sisters can, this might just improve training time a lot. This is a way cheaper way to improve

the efficiency than special produced silicon chips, though the two approaches can and should be done simultaneously.

After discussing the paper amongst ourselves we agreed that we thought it was an interesting and important one. The paper itself was written concise and straight to the point, and it also chose to measure the performance of the networks in a way that is easily understandable and applicable to the common usage of this technology, the error measurements that is. We've been working with technologies that in fact implements residual networks before, YOLO as an example. So, seeing the measurements of how important the little detail of residual connections can change the performance of the networks was sort of a reality check for us.

NOTE: THE THREE IMAGES ARE COPIED FROM THE PAPER CITED IN THE REFERENCE LIST

[i] Deep Residual Learning for Image Recognition2 by Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun (2016, CVPR) [He et al., 2016]