EECS 6322 Week 3, Paper 1
Deep Residual Learning for Image Recognition

Shivani Sheth (Student #: 218011783)
shivs29@yorku.ca

I. Summary and contributions

The paper proposes a new architecture that performs better than the existing architectures for deeper neural networks, both in terms of accuracy and computation. Earlier, when we trained the datasets with deep neural networks based on the existing architectures, we would expect the training error to be lesser than their shallow counterpart networks. Instead, the authors noticed that the deeper neural networks performed worse on training data (had a higher training loss) than their counterparts. The authors ruled out 'overfitting' as the reason behind the paradox because, for overfitting, we would expect a higher testing error but a lower training error. In contrast, these deep neural networks had both a higher training and testing error, due to which the authors suggest that this is an optimization problem.

In order to solve the degradation problem, the authors propose an architecture consisting of residual blocks or 'building blocks' that applies 'residual mappings'. These mappings are represented by $F(x) = H(x) - x$. Thus, the mapping $H(x)$ is recast into $F(x)+x$, where x is the identity mapping added to the output of the consequent stacked layers. This is also known as a 'shortcut connection' which makes up the residual block. This residual block is then stacked together to form the core of the architecture. In the paper, the residual or the building block is defined as $y = F(x, \{Wi\})+x$, where x and y are the respective input and output vectors of the layers considered.

Further, two cases for the addition of the identity mapping and the stacked layers are taken into consideration. If both ($F(x, \{Wi\}$ and x in the above equation) have the same dimension, then the addition operation is directly performed. If they do not have the same dimensions, then either the additional dimension spaces are added with zero entities, or a linear projection is performed on the identity mapping to match the dimension of the stacked layers. The linear projection is found to perform slightly better than the zero padded entities since the zero padding dimensions have no residual learning.

The form of the residual function $F(x)$ in the paper involves two to three layers included in the stacked layers. If we include only one layer in the stacked layer, then the residual function becomes a linear function which does not show a significant improvement. On the other hand, more layers can be added in the stacked layers which is not discussed in the paper.

The architecture was implemented under the following conditions. Scale augmentation and standard color augmentation were applied on the training images. Batch normalization was applied after each convolution and before activation. The SGD optimizer was used with a mini batch size of 256. The initial learning rate was 0.1 which was brought down to 0.01 when the error plateaus. The weight decay used was 0.0001, momentum was 0.9, and no dropout was used.

To decrease the training time, the authors have also proposed a bottleneck design of the architecture that replaces two (3 x 3) convolution layers with a (1x1), (3x3), and (1x1) layer in the residual function $F(x)$. This reduces the input/ output dimensions for the (3x3) layer in the residual function, hence decreasing the complexity of the system. For this design, the identity mapping performs better than the projection mapping.

The results from this architecture, as tested on the ImageNet validation set, proved to be significantly better than previous architectures such as VGG, GoogLeNet, and PReLU-net. Similar results were shown when the architecture was tested against FitNet and Highway algorithms on the CIFAR-10 test set. Hence, the model proves to be overall efficient.

II. Strengths

The paper proposes a novel method, which was successful in achieving higher accuracy as compared to the previous architectures with lower computational resources. It has enabled better learning rates for deeper networks. Hence the training sets can obtain higher accuracy due to better learning rates. The architecture also includes no extra parameters for identity shortcuts; the parameters for the model are the same as compared to the plain networks, and hence it is easier to compare the two architectures.

III. Weaknesses

The paper does not provide a concrete reason for the degradation problem in the earlier architectures and is left as future work for the research community. It also does not elaborate upon the performance based on the number of layers in the stacked layer. For example, the paper uses two to three layers in the stacked layer, but it does not elaborate upon how the performance varies if the number of layers are increased and what would be the upper limit for the number of layers in a stacked layer.

IV. Correctness

The claims and empirical methodology as presented by the authors seem to be correct as they are backed up by results from not only image datasets (such as ImageNet validation set *and CIFAR-10 test set) but also by object detection datasets such as PASCAL and MS COCO.*

The empirical methodology can also be backed up theoretically in cases where vanishing gradients might decrease the learning rate of the network which could be a possible reason for the increase in the training error. In such cases, the identity mapping ensures that the network retains the learning from at least the previous layers, if it cannot further improve it.

V. Clarity

The paper is well written in general. There are a few areas in the paper where the authors could have elaborated upon such as the three methods for the projections on identity mappings. These methods could have been supplemented with either visual illustrations or examples to convey the differences in the projection matrix in the three conditions.

VI. Relation to prior work

The work proposed in the paper differs from its prior work both in terms of depth and continuous residual learning. Previous architectures such as 'highway networks' apply shortcut connections based on the output of the gates at different layers, and have only been tested on shallow networks (less than 100 layers). On the contrary, this paper proposes an architecture for deep layers (more than 100 layers) based on continuous learning via identity mapping and does not include any additional parameters in the network

VII. Reproducibility

The architecture proposed is reproducible and can be implemented using the existing frameworks such as TensorFlow, PyTorch, and Caffe. It is also one of the most widely used architectures for advances in the research community.

VIII. Additional comments

A compact visual representation of the single-model architecture used on the ImageNet validation set would have complemented the explanation and results.