Deep Residual Learning for Image Recognition

Authors: Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun
Publisher: IEEE Conference on Computer Vision and Pattern Recognition (CVPR)
Year: 2016

Keywords: Residual networks, identity functions, COCO dataset, ILSVRC, plain nets

1. **Abstract:**

"Deeper neural networks are more difficult to train. We present a residual learning framework to ease the training of networks that are substantially deeper than those used previously. We explicitly reformulate the layers as learning residual functions with reference to the layer inputs, instead of learning unreferenced functions. We provide comprehensive empirical evidence showing that these residual networks are easier to optimize, and can gain accuracy from considerably increased depth. On the ImageNet dataset we evaluate residual nets with a depth of up to 152 layers—8× deeper than VGG nets [41] but still having lower complexity. An ensemble of these residual nets achieves 3.57% error on the ImageNet test set. This result won the 1st place on the ILSVRC 2015 classification task. We also present analysis on CIFAR-10 with 100 and 1000 layers. The depth of representations is of central importance for many visual recognition tasks. Solely due to our extremely deep representations, we obtain a 28% relative improvement on the COCO object detection dataset. Deep residual nets are foundations of our submissions to ILSVRC & COCO 2015 competitions 1 , where we also won the 1st places on the tasks of ImageNet detection, ImageNet localization, COCO detection, and COCO segmentation."

2. **Introduction:**
   - The deep learning had shown it's significant effect in image classification and the object detection. The increase in number of layers would increase the accuracy of the prediction.
   - The learning is not as easy as adding more layers to the network. This problem can be addressed by the layers in between which initializes converging with the help of stochastic gradient descent and the normalized initialization.
   - The increase in layer would make the network so deep that it would be complex and computationally expensive to train the data. So, the authors introduced a new way of training with residual learning framework resulted in the ease of training deep neural networks.
   - The approach had been state of the art to the date and they had implemented the approach in ILSVRC 2015, COCO 15 challenges had shown the best results compared to all the approaches till date.

3. Related Works:

- Residual and Fisher vectors were previously defined in order to represent the VLAD. This version of VLAD had shown the prominent results in image classification and retrieval.
- Even in the field of computer graphics, the partial differential equations were solved by the Multi-grid approach, in which the equation is divided into subproblems and each of that subproblem is accountable to the residual solution between the finer and coarser scales.
- The alternative approach for this Multi grid method is preconditioning,in which the variables were represented as the residual vectors of both finer and coarser scales.
- The previous state of the art is Inception network with very few deeper branches with sparse networks and shortcut branches.
- In the simultaneous approach, the shortcut branches were created by "highway networks" with gating functions.These gating functions were parameter and data dependant, where as the shortcut branches followed by the authors do not depend on the parameters and called as identity short cuts.

4. Content and Approach:

- The training of deeper neural networks had many difficulties in training from the computational expense to convergence. So, as the layers were increased to increase the accuracy, the overfitting problem may also takes place.
- To overcome these problems, the authors had introduced deep residual learning framework, which would decrease the higher training error due to overfitting compared to the shallower networks.
- In this approach a function is mapped to the other with addition of the input layer directly to the other network and this connection is called as the shortcut connections.
- The comparisons had been done in between the residual networks and the plain networks. The residual neural networks had been easier to optimize, whereas the plain networks had increased the training error.
- Even at the deeper layers, the residual networks had shown very good accuracy, unlike plain neural networks.
- Let us consider a mapping H(x) where x refers to the inputs of the first layers. The residual function can be written as:
    - $F(x) = H(x)-x$
    - Implies $H(x) = F(x) +x$.
- Both are similar mathematically, but the difference is mattered most in the learning methodology.

- The identity mapping to the residual learning would result in the preconditioning as the residual functions would result in limited responses.
- For ImageNet two models, plain network and residual network, were compared in order to get the differences of the usage.
    - In the plain networks, the simple CNNs were trained with 3x3 filters and downsampling is performed with a stride of 2. The network consists of 1000 layered network with the softmax function, with 34 weighted layers. It is also to be noted that the model is less complex than the VGG nets.
    - In residual networks, the identity shortcuts were used when there are same dimensions for the input and the outputs. If the dimensions increase, the identity function is evaluated with extra zeros in the matrix or dimensions were matched by projection shortcuts.
- The experiments were done on different datasets such as ImageNet, CIFAR, PASCAL and COCO.
- Residual deep neural networks had been used in ILSVRC where as with the CIFAR dataset authors had learned the neural networks with simple CNN architecture with residual networks.

5. **Results:**
    - The results had also been compared in between the plain and residual networks with 18 and 34 layer each. The results for the ImageNet were as follows:
        - The training error of the 34 layer plain network had been higher than the 18 layered plain network referring the overfitting in plain deep neural networks.
        - The 34 layer residual network had shown the lower training error than the 18 layered residual network concluding the effect of the residual networks in implementation.
    - It is also evident that the 18 layered residual networks converges faster than the 18 layered plain network.
    - The Residual networks(ResNet) had given the minimum top-5 error of all time about 3.57%. This had been the best in that competition compared to all the previous years.
    - In CIFAR, the architecture had resulted in significant results of 76.4% accuracy of object detection compared to VGG-16 results of 73.2
    - With the COCO dataset, the authors had shown an improvement of 6% to the standard results so far.
6. **Conclusion and Future works:**
    - From the results obtained from the diverse datasets, it is evident that the implementation of residual functions and creating shortcuts will result in prominent architecture compared to the normal deep neural networks.

- Despite of increase in the computations, this method can be adapted with good hardware and the overfitting due to numerous layers is also eliminated through the approach.
- Furthermore the approach can be adapted in the various vision and non-vision problems to increase the efficiency in training of the deep neural networks with numerous layers.