Deep Residual Learning for Image Recognition

Summary written by Vedant Jain
April 7, 2021

Summary: This paper describes a novel approach to create deeper neural networks that not only provide better classification results but also are easier to train. The method works by essentially changing the learning function for some layers such that it is based on learning the residuals of references to the layer's inputs. Based on this, networks are created where some layers include residual blocks, and others don't. They later applied this method to create networks that were close to 8 times larger than the VGG used for the ImageNet dataset. The model was the best performing one at the ImageNet detection contest scoring a 3.57% error on the Top-1 error.

Related work: Other methods have made progress dealing with residual vectors. One example is used with Partial Differential Equations in which the problem is scaled using the multi grid method.[1] Here each sub-problem attempts a coarse solution that is then put together for the complete solution. This method is also much faster at converging. Another approach for residual networks uses shortcut connections which includes using the parameters of the inputs to help set when non-residual networks and residual networks are used.[2]

Approach: The major contribution of He et al involves rephrasing the learning that occurs at each layer. They essentially replace the normal learning between layers with what they call residual blocks. These residual blocks are based on the mapping of the residual function between the output of a layer and the inputs to the previous layer. For example, In this way the output of the next layer is the output of that layer plus the identity value of the input to the previous layer. This essentially creates a learning block in which the identity of the original mapping can be found by gradient fairly easily making it likely that the new layers added cannot hurt accuracy and only make the model better.

Datasets, Experiments and Results: The authors of this paper used the ImageNet data to make comparisons between the Resnet and other plain networks. First they showed that a larger plain network with 34 layers performed worse overall than a plain network with 18 layers. Then they showed that a Resnet network with 34 layers not only outperformed a Resnet network with 18 layers but they both outperformed the plain networks. Next they looked at how long should the shortcut be between layers, for example should it just cross one layer or should a shortcut be involved with multiple layers. This they were able to show increased accuracy with much larger networks on the order of 152 layers because they didn't have to worry about degradation from the layers.

Strengths: This paper describes a very powerful approach to expanding the size that neural networks can be. It is clear that the authors also know how radical their idea is because they go deep into the theoretical foundations of residual networks. Explaining how finding the identity is a straight forward task for the solver and why the residual function can therefore be extrapolated from it. Furthermore, when it comes time to test their results they provide a very distinct and direct comparison with plain networks as well as other high functioning neural networks.

Weaknesses: This paper lacks sort of the elaboration of expanding on the residual network and other perturbations of it. For example, they go into explaining residual networks that vary by different parameter inputs but they don't really go into attempting these methods. It would be useful for the reader to see different approaches to the residual learning block that may or may not provide a better performance. I understand that the limitation in this paper might be due to the complex nature of the residual block space making it difficult to explain other changes to it.

Reflections: He et al, showed how residual learning blocks can be used to control degradation in large neural networks. They explain how learning blocks can be trained using the residual function where the identity of the inputs is used to control the complexity of the cost function. They further went on to show how their method performed much better in the ImageNet dataset.

References

[1] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. 2010.

[2] N.N.Schraudolph. Accelerated gradient descentbyfactor-centering decomposition. 1998.