Student ID : 20194293

Name : Go, Kyeong Ryeol

[AI 502] Deep Residual Learning for Image Recognition

1. Paper Summary

The level of features from low to high can be enriched as the network gets deeper and deeper. However, as a side effect, the network becomes vulnerable to vanishing/exploding gradient. Normalized initialization and intermediate normalization layers have addressed these issues so that network with 10 layers can be learned by stochastic gradient descent with backpropagation.

Nevertheless, as the depth of the network further increases, the accuracy gets saturated at some point and then degrades rapidly. This is not occurred by the overfitting because it is due to high training error not the test error. Previous works proposed a solution by construction where the added layers are identity mapping and the rest of the layers are from the learned shallow network. This implies that the deeper models would be able to improve the performance of its shallow counterparts. This paper is dealing with this 'Degradation' problem and resolved it by a deep residual learning framework.

The degradation problem suggests that the solvers might have hardships in approximating the identity mappings by the nonlinear layers. Here, the author hypothesized that it would be easier to optimize a residual mapping(F(x)=H(x)-x) than to optimize a desired underlying mapping(H(x)). Then, the original mapping can be recast by adding the original input to the residual mapping(H(x)=F(x)+x). This can be realized as a feed-forward neural network with the shortcut connection skipping one or more layers which performs the identity mapping without any extra parameter or computational complexity. (In case when dimensions of input and output are different, a linear projection must be conducted with extra parameters.) As a result, if identity mappings are optimal, it will simply push the weights of the nonlinear layers toward zero, which is relatively easy to learn.

Actually, it is also applicable to convolutional layers. Here, the projection operation can be substituted by additional zero paddings or 1 x 1 convolutions. When training a deeper network, the bottleneck building block is devised to train a deeper network where the original building block with 2 weight layers is replaced by 1 x 1, 3 x 3, and 1 x 1 convolutional layers in order. Although the core idea of this paper is the residual learning itself, but in most of the cases, this paper is referred in image domain, which indicates its influential role in convolutional layers. In fact, the model they submitted to competitions, which is named as "ResNet", has won the first places in several tracks in ILSVRC & COCO 2015 competitions.

As an experiment, the plain networks and the residual networks, deeper bottleneck architectures are considered. The results indicate that the deep plain networks have high training and validation error than the shallow plain networks while the situation is reversed with residual networks. Also, the residual networks performed better than the plain networks in terms of the top-1 error and the time complexity. The author also showed that for shortcut connections, projection would be preferable than zero paddings, but the differences are not that significant and it is not essential for addressing the degradation problem. Finally, even in deeper architecture with bottleneck building block, they enjoyed significant accuracy gains. The model they submitted to ILSVRC 2015 was an ensemble of the two 152-layer models which outperforms the state-of-the-art methods in a considerable margin.

2. Discussion

Here, I want to offer two discussion points. To begin with, what if the shortcut connections are partly overlapped to each other? In this way, the model can consider the layer-wise dependence better, but due to its flexibility, it may need more rigid regularization and has many more options to consider that leads to longer time for tuning. Next, what is the good balance between the width and depth? Depth is shown to be important to enrich the levels of feature, but at the same time the use of feature diminishes which makes the network very slow to train. I guess the width can substitute the role of depth in some sense and actually many researches such as 'Wide Residual Networks' are currently on process.