Discussion Quiz #4

Ashutosh Tiwari (ashutiwa@iu.edu)

February 21, 2022

1 Papers

1.1 Paper 1

Deep Residual Learning for IR

1.2 Paper 2

Identity Mappings in Deep Residual Networks

2 Questions

2.1 Can stacking more layers will give you a better network? What is the degradation problem discussed in paper 1?

Authors point out that network depth is an important factor and many successful architecture employ them to level features of all scales. However they also point out that there are few limitations which prevent network from converging.

The degradation problem authors discuss is the fact that as the number of layers in network increase accuracy gets saturated and then degrades. They also point out that this does not happen because of overfitting and it is proportional to the depth of network in consideration. In rest of the paper authors try to solve this problem using their architecture.

2.2 How the author concludes that the degradation problem is not caused by over-fitting? Explain.

Authors try to prove this by intuition. They mention that if you have a shallow network which perfectly models the problem statement with S layers. Now compare this with a deeper architecture with many more layers (L). In that case last $L - S$ layers can be thought of *Identity* layers, which just multiply input by 1.

So if this intuition is correct the error between Shallow and Deep network should be same, but authors point out that it is not the case rather the error (more specifically training error) is much higher in the latter case.

2.3 What should be the behavior of a model with optimal identity functions in ResNet setup?

In ResNet authors intuitively point out that it should be much easier to to learn to optimize residual mapping rather than the original mapping. Consequently for identity function, it should be much easier to push them to zero, than to try to fit identity mapping using deeper architectures.

2.4 Explain all the methods that can be used to add a residual connection with (a) same dimensions and (b) different dimensions.

For same dimensions output from ReLU and biases are added to input to get the output. Equation (1) in first paper demonstrates the idea. This operation $\mathcal{F}+\boldsymbol{x}$ is performed by shortcut and then an element-wise addition. For different dimensions authors perform a linear projection and multiply that output of projection with the input to make those two dimensions equal. This is demonstrated by equation 2.

2.5 Explain how the ResNet model discussed in paper 1 achieves a deeper model with less number of FLOPs as compared with VGG Net.

There are many changes in ResNet which make it more efficient in terms of required computational resources. Because VGGNet starts with 3×3 size kernels, it produces large feature maps for further layers. In contrast ResNet starts with 7×7 convolution layer.

After that almost in every layer, ResNet has fewer number of units per layer (sometimes as low as $\frac{1}{4}$. Thirdly, is uses bottle neck architecture, which further reduces number of parameters in the architecture.

2.6 Explain the properties exhibited by equation (4) in paper 2.

As mentioned in the paper, there are few nice properties exhibited by equation (4). First is that $\boldsymbol{x}_L$ of at any level in the network can be represented as the feature of any layer at an early stage in the network. Second is that feature at any deep layer L can be represented as the sum of a constant $\boldsymbol{x}_0$ + a function of inputs and weights of other layers. This is important and different form normal NN networks in the sense that this can relay more information and never goes to zero (even for deeper networks).

Authors also mention its nice backward propagation properties, which its one term which is independent of weights of intermediate layers and one which is. Even this is useful and because there is a constant term, it is guaranteed to guard network from vanishing gradients.

2.7 Name all the types of shortcut connections used in paper 2. Which method of connection is best (as per paper 2) and explain why?

Authors explore following shortcut connections

(a) Constant Scaling

(b) Exclusive gating

(c) Shortcut only gating

(d) 1×1 convolutional shortcut

(e) Dropout shortcut

Paper suggest that Identity Skip connections are the best, which are a type of Constant scaling, where instead of $\lambda = 1$ in $\lambda_l \boldsymbol{x}_l$ in Identity, we scale down $\lambda = 0.5$. Experiments suggest that after we do that, training error is higher and optimization has difficulties. Also since these connections don't have any parameters to optimize they generally have less number of optimization issues.

2.8 Why does optimization face difficulty in ResNets when the shortcut signal is scaled down?

When λ in $h(\boldsymbol{x}_l) = \lambda_l \boldsymbol{x}_l$ is scaled down the original signal, which has the form

$$\boldsymbol{x}_L = \left(\prod_{i=l}^{L-1} \lambda_i \right) \boldsymbol{x}_l + \sum_{i=l}^{L-1} \hat{\mathcal{F}}(\boldsymbol{x}_i, \mathcal{W}_\rangle)$$

reduces to the original form, as first term which is the shortcut signal tends to zero. Because of this the output at last layers is not big enough.

Even with this setting authors explore two options, where in one case they scale the second term and in the other case they don't scale it. First case performs better than the second one for the same reason.

2.9 Can ResNet architecture helps achieve better accuracy when the model is not overly deep? Explain why or why not?

Authors point out that 18 layer architecture is accurate and is able to find satisfactory solutions. This is because since number of layers is less, convergence is faster. This is also evident from their experiments on plain networks. Authors point out that it is probably because it eases convergence in early training stages.

2.10 Describe briefly. (a) Residual building block, (b) Solution space (for classification problem), (c) bottleneck architecture, (d) pre-activation and post-activation

(a) Residual Building Block

Residual building blocks originate from the motivation that deeper network should be able to only add to performance of task at hand without being blocked by convergence issues. First types of building blocks suggested were highway networks, which had a gate at their skip connections.. Residual blocks are special cases of these networks without any gates in their skip connections to facilitate sufficient flow of information.

(b) Solution space

Solution space is the set of all possible exhaustive set of solutions. In case of classification problem statements, it represents the set of all possible classes.

(c) Bottleneck Architecture

Bottleneck Architectures are when an input is passed though a set of layers which are of lower dimensionality that input and output layers. In most of cases bottle neck architecture are used to lower the computational considerations.

(d) Pre-activation and Post-activation

Pre-activation is the application of activation before skip operation is added in the main stream of knowledge in a residual block and post activation is opposite of same. Authors in this paper experiment with various combinations of these activations and present their results.

References