Jenish Maharjan **Week 3**

Paper: *Deep Residual Learning for Image Recognition*

Summary:

Published in 2015, the paper “Deep Residual Learning for Image Recognition” by He, Zhang, Ren and Sun from Microsoft research, tackles the issue of correlation between increased depth and increased accuracy of a deep neural network. The paper also addresses the degradation problem that arises when deeper networks are start to converge – as the depth of the network increased, the accuracy of the network gets saturated and then degrades rapidly which Is not a result of overfitting. The degradation problem indication that not all systems can be simply optimized by “stacking more layers”.

In their solution to the degradation problem, the authors propose explicitly letting the layers fit a residual mapping instead of hoping that every few stacked layers directly fit a desired underlying mapping H(x). The underlying mapping of data to be fit by a few layers is modified adding unity to the function. In the presented network, the layers learn a residual function, F(x) = H(x) – x. Then x is added to the result afterwards which is known as a shortcut connection for which the dimensions of x and F(x) need to be equal. If not equal, then a linear projection can be added to match their dimensions. The authors’ hypothesis is that it is easy to optimize the residual mapping function F(x) than it is to optimize the original, unreferenced mapping H(x). The network learns F(x) and H(x) is recovered by adding x.

For the purpose of evaluation, they trained both plain and ResNet convolutional neural networks of 18 and 34 layers. The plain networks used by the authors were mainly inspired by the philosophy of VGG nets but had fewer filters and lower complexity than VGG nets. They also had the counterpart residual versions of the 18- and 34-layer plain CNN networks. The difference in the ResNet versions being the shortcut connections that could be directly used when the input and output dimensions were same but needed projection when different. In the implementation of the network, the input is obtained by randomly sampling a 224x224 crop from a resized image or its horizontal flip. The networks have a batch normalization after each convolutional layer. Both the plain and residual networks are trained from scratch with a learning rate initialized at 0.1 and decreased by a factor of 10 when the error plateaus. The momentum and the weight decay used are 0.9 and 0.0001 respectively.

The ImageNet classification experiments carried out in the plain and ResNet models showed that the error increased with the increase in depth in the plain networks while the error lowered when the depth was increased in the ResNet models. The authors also came up with deeper architectures with 50, 101 and 152 layers which still had lower complexity than the VGG-16/19 networks. These deeper networks demonstrate significant accuracy gains without the occurrence of the degradation problem. Their results of both their baseline and deeper networks made them the winners of the ILSVRC 2015 contest. The authors also present the performance and analysis of their networks in other datasets like the CIFAR-10, PASCAL VOC 2007, 2012 and COCO. PASCAL and COCO were used to demonstrate the object detection task with the ResNet networks.

Strengths:

- The authors do a great job of organizing the content of the paper – specifically the additional sections describing the performance of the networks focusing on the task of object detection helps separate it from the main body describing the classification task.
- The exploration of deeper architectures after getting great results from the baseline model and the presentation of the results in tables for easy comparison is a huge plus point of the paper.

Weaknesses:

- Since the content of the paper is targeted towards researchers in the field of deep learning, the language is, at times, difficult to understand.
- Some of the points mentioned in the introduction section are also repeated in the following sections.

Confusions:

- I didn't really understand the mathematical concepts of Fisher vector.
- Although I understand the concept of the network learning F(x) and then adding x to it to get H(x), a brief explanation of the math would be really great.

Discussion Questions:

- How does learning the residual function and deriving the mapping H(x) relate to the solution of the degradation problem in theory? Is the relation only empirically verified?
- How is object detection different from the task of classification?
- How much is the size of the dataset responsible for overfitting of a network?