Jenish Maharjan **Week 3**

Paper: *Very Deep Convolutional Networks for Large-Scale Image Recognition*

Summary:

Simonyan and Zisserman in their paper "Very Deep Convolutional Networks for Large-Scale Image Recognition" share their deep convolutional neural network that was able to achieve "state-of-the-art-results" of the network's performance on the ImageNet dataset. The paper seems to be targeted to researchers interested in their network architectures and the results they achieved. The authors have tried to show that increasing the depth of a network improves its accuracy and also that multiple small filters are better than a single large filter when both have same receptive field areas on the input image.

The network architecture presented in the paper take a fixed size 224 x 224 RGB image as input. During preprocessing, the training set RGV value mean is subtracted from each pixel. The network basically comprises of two types of layers – convolutional layers and spatial pooling layers. The convolutional layers have the stride fixed to 1 pixel and the padding set to 1 pixel for a 3x3 kernel in contrary to previous networks such as the AlexNet. The spatial pooling layers do not add to the depth of the network by convention. Max pooling layers are used for spatial pooling in the network, each with a window size of 2x2 and stride of 2.

For training, they've used mini-batch stochastic gradient descent (SGD) as the learning algorithm and multinomial logistic regression function as the loss function. The batch size and the momentum for the learning algorithm are set to 256 and 0.9 respectively. L2 Weight decay and dropout for the first two fully connected layers are used for regularization. The learning rate is set to 0.01 initially and decreased by a factor of 10 when validation set accuracy stops improving. For initilializing the layers, first the shallow network configuration with 13 layers was trained with all of its layers randomly initialized. Then the deeper architectures were trained by initializing the first four convolutional layers and the final three fully connected layers using the layers of the 13-layer network and randomly initializing the remaining layers. For training, the input images were isotropically-rescaled with the smallest side. The crops of input size were taken from the rescaled images and input to the network. Although the number of parameters and depth of the network is greater than that in the AlexNet, the presented network required less epochs for loss function to converge as a result of regularization through the large depth and smaller kernels in the convolutional layers and pre-initialization of certain layers.

During testing, instead of rescaling and resizing the test images to the input size, the test images were isotropically-rescaled with the smallest side of the image, not necessarily equal to the training scale. These rescaled images were passed as input to the network. This works because convolutional layers are invariant to the input dimensions. The problem would have been with the fully connected layers as they require a pre-defined fixed size dimensional vector as input. To fix this, they converted the fully connected layers to convolutional layers. The first fully connected layer became a convolutional layer wit 4096 filters of size 7 x 7 x 512, which is equivalent to the dimension of the normal output feature map after the last convolutional layer followed by max pooling. The second became a convolutional layer with 4096 filters of size 1x1x4096 and similarly the third became a convolutional layer with 1000 filters of the same size. Finally, to obtain a fixed-sized vector of class scores for the image, the class score map is spatially averaged, which allows for more context to be obtained from the test image. This paper shows that using a very simple architecture but making it deeper can boost the accuracy greatly.

Strengths:

- The tabular representation of the various network configuration they worked on makes it very easy to understand the configurations and also gives a good perspective of how different and how similar the configurations were.
- The authors worked on networks with varying depths and presented results to show the positive effects of depth on accuracy of a network. This strengthens their findings.
- The authors have also presented their work on the localization task in addition to the main body of the paper that describes the classification task.

Weaknesses:

- Some of the points seem to be redundant – the language and the content could've been made more concise.
- Since it is targeted towards researchers and scholars in the field of deep learning, the language is sometimes difficult to comprehend.

Confusions:

- I'm a little confused about single-scale and multi-scale evaluation.
- How different would it have been if they had rescaled the test input images to the same scale as the training input set.

Discussion Questions:

- A discussion on single and multi-scale evaluation would be really helpful.
- Why did the authors not choose to train the deeper networks from scratch? Was it only because of the cost of computation?
- How different is the localization task from classification task?