Jenish Maharjan **Week 2**

Paper: *ImageNet Classification with Deep Convolutional Neural Networks*

Summary:

In this paper, the authors demonstrate a deep convolutional neural network to classify images into 1000 different categories with record breaking results in the ImageNet LSVRC-2010 contest. The paper seems to target towards researchers and scholars of machine learning (especially deep learning). The model described in the paper was trained on 1.2 million images of 1000 categories (ImageNet dataset). In order to manage the variability in image resolution, the images were scaled so that all images had uniform dimensions. Besides rescaling the images, the dataset was also augmented by changing the RGB values of each input image by a scaled version of the principal components in RGB space in order to capture an invariance of object identity under changes in intensity and color of illumination.

The model described in the paper has five convolutional layers and three convolutional layers. The authors chose the ReLU function to enforce nonlinearity in the model. The use of ReLU activation function compared to other activation functions improved the training speed of the network. They were able to train the model to 25% error rate on the training set which was six times faster than an equivalent network with tanh activation. The max pooling layers in the network were implemented with overlapping windows which is applied after response renormalization. The function of the pooling layers is to summarize the outputs of the neighboring groups of neurons in the same kernel. The authors observed that using overlapping pooling layers reduced the occurrence of overfitting. In the network, the first convolutional layer the input images (224 x 224 x 3) with kernels of size 11x11x3. The second convolutional layer takes the output of the first convolutional layer as input and filters it with smaller sized kernels. Unlike the first and second convolutional layers, the following three convolutional layers are in the network are connected to one another without any pooling layers in between. The output of the last convolutional layer becomes input for the fully connected layers each having 4096 neurons.

Due to the large number of parameters (60 million) and number of classes, there is a very high change of overfitting. To make the network resistant to overfitting, the authors artificially enlarged the dataset using "label-preserving transformations". One of the two methods of data augmentation employed involved generation of image translations and horizontal reflections. Another method of data augmentation consisted of altering the intensities of the RGB channels of the images in the training set by performing PCA on the set of RGB pixel values of the images. The authors also implemented the concept of dropout to combat overfitting – the outputs of each hidden neuron are set to zero with probability 0.5. Such neurons that are "dropped out" do not participate in back-propagation. Applying dropout doubled the training time but substantially decreased overfitting in the network.

The model was trained using stochastic gradient descent with a batch size of 128 example, momentum of 0.9 and weight decay of 0.0005 on two GPUs for parallelism. The authors also describe the process of setup for training the model using GPUs. The results of the network as reported in the paper is impressive considering the fact that this work was the first of its kind to have trained deep convolutional networks on GPUs.

Strengths:

- The authors have organized the contents of the paper into well defined sections. They even mention in section 3 that they have explained the key features in the order of importance which helps readers to focus on the most important points presented in the paper.
- The paper includes a section that discusses qualitative evaluation of the model presented. The inclusion of qualitative evaluation helps in understanding the model intuitively.
- The paper also does a great job of explaining the dataset and approaches used in preparing the data for the network.

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:

- The section that describes the training of the network using GPUs is a bit unclear to me.
- I had trouble understanding the math explaining the second method of data augmentation.

Discussion Questions:

- How or why does dropout help in reducing the occurrence of overfitting?
- Is there any difference between pooling and subsampling layers or are they just different terminologies for the same thing?
- It would be great to learn more about how exactly the training was implemented using GPUs.