Julian Singkham
CS 2300/11
Dr. Berisha

Paper Review 3: ImageNet

1.) Why are CNNs simpler to train than traditional feedforward neural networks?

CNNs are simpler to train that traditional feedforward neural networks due to having fewer connections and parameters. This simplicity greatly cuts down on the computational power needed as the CNN does less work for a very slight reduction in theoretical-best performance. This difference is similar to how RISC is easier to program in than CISC, but may not be as fast.

2.) Why were CNNs not widely used for the ImageNet problem prior to this paper?

Prior to this paper hardware was simply not strong enough and affordable enough to handle the huge size of ImageNet. CNN utilizes a lot of memory and it wasn't until the paper's time that large Vram GPU's were widely available to the public, the GTX 580 3GB is used in this paper.

3.) Which techniques did the authors of the paper use to prevent overfitting (list all of them)?

Label-preserving transformations, image translations & horizontal reflections, RGB intensity alteration, dropout

4.) How did the authors address the problem of their network not fitting on a single GPU?

The GPU they used, GTX 580, only had 3GB of memory which is too small for a network consisting of 1.2 million training samples. By using two GTX 580's, they can effectively split the load between the GPU's and utilize their built-in parallelism to distribute the computational load.

5.) Describe the inputs and outputs of this network. Identify what the shape is and how that relates to the underlying dataset.

Through each step of the CNN process (except the final output), the shape is always square: 224x224, 5x5, 3x3. Additionally, each layer only takes input from the previous layer. The inputs are squares due to the dataset being used is comprised of images. Neural networks can only be trained on the "same" thing, meaning that the training and testing data must be the same shape. This means the creators had to down sample and crop images to be 256x256 pixels. The final step output has a size of 1000 which correlates to the 1000 class labels.

6.) Why are the augmentation approaches described "computationally free"?

The augmentation approaches are described as computationally free for three reasons. The first reason is that the augmentation takes very little computational time so in the grand scheme of things the time it takes to compute an augmentation is negligible. The second reason builds off the first in that since the augmentation takes very little time, it can all occur in memory rather than disk which further improves on timing as memory is far faster than disk speeds. Finally, the augmentation can occur in the CPU while the GPU deals with training. This split operation allows for the CPU and GPU to operate independently. Since modeling is primarily a GPU oriented task, the CPU is not needed. So utilizing the CPU for augmentation makes it computationally free as the CPU would've done nothing anyways.

Julian Singkham
CS 2300/11
Dr. Berisha

7.) What metrics did the authors use to evaluate the quality of their approach? Why are these representative of the underlying challenge?

Top-1 and top-5 error rates are the two metrics used to evaluate the quality of their CNN approach. Top-1 error rate is the rate of error that the first label is incorrect and top-5 is where the top 5 labels are incorrect. These are representative of the underlying challenge because the purpose of the model is to guess what object is in the image. What makes image recognition complicated is that the focus of the image is hard to figure out if there are more than one object in the image. Take for example a picture of a tree with a squirrel on it. Is the squirrel or the tree the focus of the picture? Since we didn't take the picture it can not be said which is the main focus. The reason for the Top-1 error rate is to see if the model can correctly assume what the picture's focus is and the Top-5 error rate's purpose can be to see if it can segment the image enough to find the most important objects. Take for example the red car in Figure 4 of the paper. The focus for the image is the grille, but the car takes up most of the image. The model first guessed it was a convertible and the second guess was grille. Since the model top-5 was able to correctly label the image, the model is successful in that instance.