Nirajan Koirala
CSC 9010_001

Summary:
In the given paper, authors have used CNN to classify 1.2 million high-resolution images in the ImageNet LSVRC-2010 content into 1000 different classes. Using CNNs, they are able to achieve top-1 error rates of 37.5% and 17%. They have also employed a new regularization method called "dropout" to reduce overfitting.
For learning from thousands of objects from millions of images, CNNs are provided with a priori knowledge and their capacity is controlled by varying their depth and breadth. Since CNNs have fewer connections and parameters compared to standard feedforward neural networks of similar size, they are easier to train. With the advent of strong GPUs, very-large CNNs could be trained with high-optimization and also recent datasets such as ImageNet contain enough labeled examples to train such models without severe overfitting.
First, the ImageNet dataset is divided into training, validation and testing sets. All the images are downsampled to a fixed resolution of 256*256 and no other preprocessing methods are applied to the images. The architecture of the network is composed of 8 learned layers, 5 convolutional layers and 3 fully-connected layers. ReLU is used as a neuron's output function since they are found to train significantly faster than equivalent sigmoid units and this type of faster learning has great influence on the performance of large models. Two GTX 580 GPUs are used for training the model parallelly and normalization and ReLU nonlinearity are applied to certain layers in the model to aid generalization. A detailed description of the architecture of CNN with 8 layers and weights used in the competition is provided in section 3.5. To reduce overfitting of the models two methods are used: data augmentation and dropout. The number of dataset is artificially augmented for training using image translations and horizontal reflections. Dataset is also augmented by altering the intensities of RGB channels in the training images. The other method to combat overfitting is dropout in which the output is set to 0 of each hidden neuron with a probability of 0.5. In such a way, the dropped out neurons do not contribute to the forward pass and don't participate in the backpropagation. This technique reduces complex co-adaptations of neurons and thus they are forced to learn more robust features that are useful in conjunction with many different random subsets of the other neurons. Dropout is used in the first two fully-connected layers in the network and it greatly helps in reducing overfitting but also doubles the number of iterations required to converge. All the models are trained using stochastic gradient descent with a batch size of 128 examples, momentum of 0.9 and 0.0005 weight decay. Initialization of neuron biases with constant 1 in the second, fourth and fifth convolutional layers as well as fully connected hidden layers helped accelerate the early stages of learning by providing ReLUs with positive inputs. An equal learning rate was used for all the layers and this rate was manually adjusted throughout training using

appropriate heuristic. The CNN built using these approaches achieves a top-5 error rate of 18.2%. Further results are discussed in the paper and it shows that a large, deep CNN is capable of achieving record-breaking results on highly challenging dataset using purely supervised learning. However, it is important to note that the network's performance highly degrades if a single convolutional layer is removed so the depth of the network is really important for achieving these results. Even though unsupervised training methods were not employed in the experiments, authors expect it to help if they are able to obtain enough computational power to significantly increase the size of the network without obtaining a corresponding increase in the amount of labeled data. These improved results however does not mean that network is anywhere closer to match the infero-temporal pathway of human visual system and in the future they like to use CNNS on video sequences where networks can take advantage of the temporal structure.

The intended audience of this paper are researchers and specialists in the field of Machine Learning. The main point made is the paper is that CNN posses a very suitable architecture for image classification task and this their performance can be further boosted by tuning appropriate parameters and using novel techniques like dropout to reduce overfitting.

Strengths:
- The organization of the paper is well built
- All the sections except 1 are clear, concise and to the point

Weaknesses:
- The sizes of the figure provided are very small and thus it is harder to capture the details in them
- The 1st section of the paper is a bit lengthy and providing extra details about the coming sections

Points of Confusion:
- Local normalization scheme mentioned in section 3.3
- The workings of overlapping pooling

Discussion questions:
a) How are multiple GPUs used in different layers of the network for training?
b) Why is dropout technique effective in reducing overfitting?
c) How would CNNs work better on videos than images?