Summary of AlexNet Paper

(Summary of: ImageNet Classification with Deep Convolutional Neural Networks, University of Toronto)

Hanwen Zhao

The article "ImageNet Classification with Deep Convolutional Neural Networks"(AlexNet) from author Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton from University of Toronto used a new convolutional neural network(CNN) architecture for image object recognition, and they achieved signification improvement on decreasing the error rate compare to previous work.

Object recognition has become a hot topic since many of the current approaches utilize the power of machine learning. To improve the performance of image recognition, couple of the simplest ways are either increase the size of dataset or use better model or techniques to prevent overfitting. However, the large image dataset contains millions of images such as ImageNet just released recently; before this, the existing dataset only contains tens of thousands of images such as NORB, Caltech-101/256, and CIFAR-10/100[1].

The starting point of the approach for AlexNet paper is to build their model base on convolutional neural networks(CNN). The CNN architecture has variety advantages such as it can be controlled by varying the depth and breath, and it also make more correct assumptions about image [1]. The dataset used for the CNN is Image, which contains more than 15 million high-resolution images with human labeled test data. In order to keep the consistency for all images, the images were down-sampled to 256x256 before feeding into the neural network. The CNN architecture contains eight network layers, including five convolutional layers and three fully-connected layers. There are five features they used in their architecture [1].

- ReLU Nonlinearity: Non-saturating nonlinearity output $f(x) = max(0,x)$ is used rather than saturating nonlinearities such as $f(x) = tanh(x)$ *or* $f(x) = (1+e^{-x})^{-1}$ since it is much faster and require less iterations to achieve the same error compare to traditional methods[1].

- Training on multiple GPUs: The GPU they used for training is GTX 580, which only has 3GB of memory, in order to increase the size of neural networks, they used GPU parallelization on two GTX 580 cards. With this scheme, it helped to increase accuracy as well as decrease training time compare to single GPU setup[1].

- Local Response Normalization: With ReLus, there is no need to normalize input data. However, they found out that response normalization helped to reduce error rates.

- Overlapping Pooling: In the pooling layer of the CNN, it average the output with its neighboring groups. During the training process, they found that with overlapping, it helped to reduce overfitting.

To summarize the overall architecture, the CNN has total of eight layers. The first five are convolutional layers. The first layer takes the resized image as input with 96 $11x11$ kernels in all RGB channel. The second layer take the output from first layer as input, convolved with 256 kernels with size of $5x5x48$. The third layer has 384 kernels with size of $3x3x256$; the fourth layer has 384 kernels with size of $3x3x192$, and the fifth layer has 256 kernels with size of $3x3x192$[1].
With the class number of 1000, one big issue with object recognition is that the date might overfit in many cases. From AlexNet paper, they implement two method to reduce the overfitting problem.

- Data Augmentation: The easiest way to reduce the overfitting is to use larger dataset. However, with limited data size, one approach is to artificially increase the size of training data. The first method is to create image translation and horizontal reflection of an image. The second method is to alternating the intensities in all RGB channels. As the result, the first method increase the training size with a factor of 2048, and the second method reduce the top-1 error rate by 1%[1].

- Dropout: The dropout is a recently introduced technique, which set the output to 0 with neuron with probability of 0.5. Researches found that dropout helped on reducing overfitting problem[1].

In the ImageNet ILSVC-2010 contest, the CNN achieved the top-1 and top-5 error rates of 37.5% and 17.0%, which the previous best result of 45.7% in top-1 and 25.7% from "SIFT+FVs" model[1]. Also, in the Fall 2009 version of ImageNet, the CNN achieved 67.4% and 40.9% in top-1 and top-5 error rates, which the previous best result was 78.1% and 60.9%[1]. At the end, the author emphasized that the depth is very import in CNNs since removing one middle layer could result 2% error rate increase in top-1[1]. Also, with more computational power and without hardware limitations, they would like to build an even larger and deeper CNN.

References

[1] Ilya Sutskever Alex Krizhevsky and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. University of Toronto.