Assignment 10

ImageNet Classification with Deep Convolutional Neural Networks

Nisim Hurst

Thursday 22 March 2018

Abstract

Convolutional neural networks (*ConvNets*) reduce computational complexity by promoting two important properties: 1. Parameter sharing and 2. Sparsity of connections. The former refers to the capacity of the weight filter to be applied to many pixels of the image and *share* those weights. The later refers to the independence an output pixel gains from those pixels that are out of the filter neighborhood. The paper [1] presents a new ConvNet architecture that will set the pace and trend for the actual top-notch performers, not only in computer vision but also in machine learning in general.

Hypothesis

A combination of techniques will boost the construction of a convolutional neural network architecture that outperforms state-of-the-art in classifying high-resolution images. Here are some of the intuitions that guided it's construction:

1. Non-saturating units make training faster
2. Dropout to reduce overfiting
3. Due to the connection limit between layer to layer, convolutional layers will help reduce complexity of large data networks while preserving label-preserving transformations.
4. GPU's will make the 2D convolutions and preform better.

Evidence and Results

The convolutional network won the top-5 test error rate scoring 15.3% on the ILSVRC-2012 competition vs 26.2% achieved by the second place, a margin of half an order of magnitude, even though no pre-training or unsupervised learning is being used. It takes between 5 and 6 days to train on two GTX 580 3GB GPU's., relatively fast.

The Dataset

There are roughly 1000 images for each of the 1000 categories. Thus, it has 1.4 million images divided into 1.2 million training images, 50,000 validation images and 150,000 testing images. The images were down-sampled to 256x256.

The authors include results from both ILSVRC-2010 and ILSVRC-2012 competitions using top 1 and top 5 error rates.[1]

Experiments and empirical evidence

On the ILSVRC-2010 achieved top-1 **37.5%** and on top-5 **17.0%** vs second place of 47.1% and 28.2% (sparse coding).

[1] The fraction of test images for which the correct label is not among the five labels considered most probable

On the ILSVRC-2012 it was tested alone, averaging the results of 6 networks and pretrained. Pretrained and averaging 2 nets the net gave an stunning 15.3% top-1 score.

On ImageNet (Fall 2009) it gave 67.4% and 40.0% vs 78.1% and 60.9%, the best published results at the time (less is better).

The 48 layers on each GPU show an obvious specialization of color, result of the restricted connectivity of the architecture.

Also visual learning was tested by computing the top-5 labels on eight test images. Another testing instrument was to measure feature activation vectors distance on the last hidden layer.

Contribution

The main contribution of this article is a proposed convolutional architecture and results applied to the ImageNet dataset:

1. Was trained over one of the largest convolutional neural network on the ImageNet dataset used in the ILSVRC-2010 and ILSVRC-2012.
2. Won the competitions by improving several unusual features. This features were found empirically. GPU memory and training time were found to be the bounds of the network size.

The architecture has the following idiosyncrasies:

1. ReLU units to accelerate learning (non-saturating units).
2. Two GPU's 2D convolution highly optimized implementation. The net kernels were divided into two GPU's where they communicate only in certain posterior layers (kernels on layer 1 communicates with all kernels on layer 2 but kernels in layer 3 only communicates with kernels of layer 4 in the same GPU).
3. Normalization without subtracting the mean (brightness only). See the paper on [1] for more of the mathematical ado.
4. Overlapping pooling was applied to reduce overfiting. In the network, a neighborhood of 3 was considered for leaps of only 2 pixels.
5. Overfiting was especially reduced by applying other ancillary techniques:
   1. **Data augmentation.** First by generating image translations and horizontal reflections. Then altering the RGB channels by using multiples of the eigenvalues drawn from PCA.
   2. **Dropout.** 50% probability.
6. Small weight decay was precise for the model to learn. The last layer weights were initialized to 1 to accelerate ReLU activations. Learning rate was set constant for all the layers.

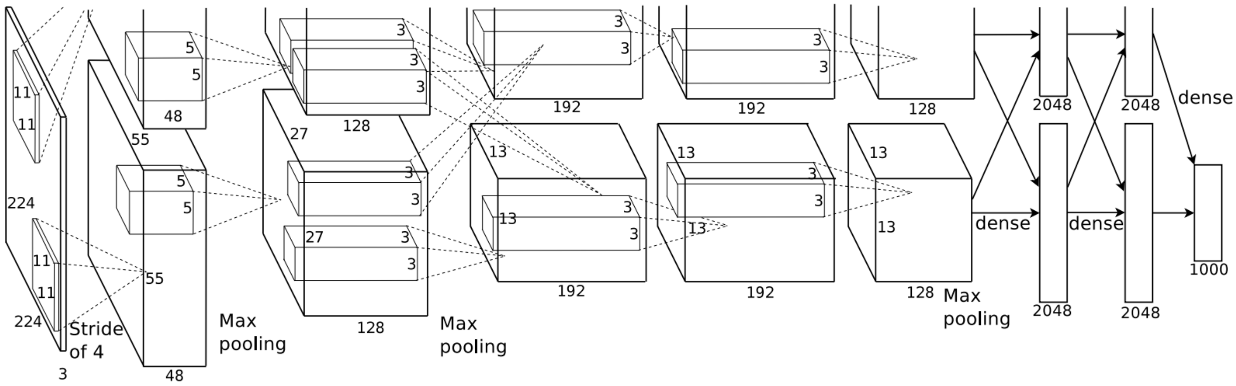

Figure 1: Layer Structure for the Architecture

Figure 1 shows the layer structure for the architecture evaluated in the paper (taken from [1]).

Weaknesses

First, convolutional layers add more hyper-parameters and allow to control the depth and breath of the networks. However, no mathematical model was presented to suggest how this hyper-parametrization must be made.

Second, in the *Discussion* section, the author points out that performance degrades if a single convolutional layer is removed. However, no single mathematical explanation is provided.

Finally, it is assumed through out the article that pretraining weights using unsupervised methods would favor accuracy. However, it doesn't take into account, neither propose, a method on how the restricted connectivity that parallelize onto several GPU's can be addressed effectively. As aforementioned, this restricted connectivity force each GPU into specialization that would be difficult to compare when unsupervised learning were to be applied *apriori*.

Future Work

The authors mention the possibility of using large deep convolutional ConvNets over video where the temporal structure can be an advantage over static images.

Another opportunity of research could include the parallelization upper bound definition over more than two GPUs.

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Advances In Neural Information Processing Systems*, pp. 1–9, 2012.