Review of "ImageNet Classification with Deep Convolutional Neural Networks"

Nghia Dang

1. Paper summary

This paper [1] describes a contest of three researchers in the University of Toronto in the field of image classification of 1000 different classes. The group has created a model with improved techniques such as adding max-pooling and fully-connected layers, RELU activation functions, and GPU implementation to achieve a significantly good error rate.

2. Architecture

2.1 Layers

There were five convolutional and three fully-connected layers excluding the output layer using a softmax function with 1000 class labels. Remarkably, the researchers used two different GPUs to train this architecture. Max-pooling and response-normalization were also used between these layers.

2.2 ReLU Nonlinearity

Due to the fact that it takes saturating nonlinear functions a couple of times to converge than non-saturating nonlinear ones, the authors used ReLU for such large datasets.

2.3 Overlapping pooling

Max-pooling helped their model reduce top-1 and top-5 error rates by 0.4% and 0.3% against traditional pooling that kept the original size of images.

2.4 Training on Multiple GPUs:

Two similar GPUs were used to train the huge dataset of 1.2 million high-resolution images. In fact, the GPUs did not completely run in parallel as input of one layer was the output of the next one. Although connectivity pattern may negatively affect cross-validation, the authors confronted this issue by tuning their model until the ratio of communication over computation became acceptable.

3. Dealing with overfitting

3.1 Data augmentation

A technique in data augmentation is to create smaller images size that our model can better generalize at the expense of longer training time due to the larger dataset. However, the positive effect is that it handles overfitting well, and the authors could keep their big model. Consequently, the top-1 error rate benefit by 1% thanks to it.

3.2 Dropout

A simple but efficient technique of combatting overfitting that the authors utilized is dropout: Each neuron is embedded with a predefined ratio of probability of becoming deactivated. As a result, neurons in the next layers have to robustly learn from their previous counterparts the useful conjunctions for the model to better generalize.

3.3 Max-pooling

An advantage of max-pooling with stride that the authors used is that it reduced the original image's size, which shortened their training time with smaller pixels and also kept only the essential parts of the images (reduce noise). From their observation, overlapping pooling also helped cut down overfitting.

3.4 Decay of learning rate

The learning rate was multiplied by a factor that would automatically decrease after a number of iterations when the validation error stayed still, which is a convergence issue.

4. Criticism

4.1 Illustrative graphs

Figure 1 can easily shows the absolute advantage of ReLU against a Tanh activation function (six times faster).

4.2 Good performance

While the best published errors so far were 78.1% and 60.9%, this study achieved 67.4% and 40.9%

4.3 Training time

A limitation of this study was that the training time was up to six days due to the huge datasets. As a result, the time pressure disabled the authors from applying other techniques for improvement such as using an unsupervised pre-training.

Reference

[1]A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” Communications of the ACM, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.