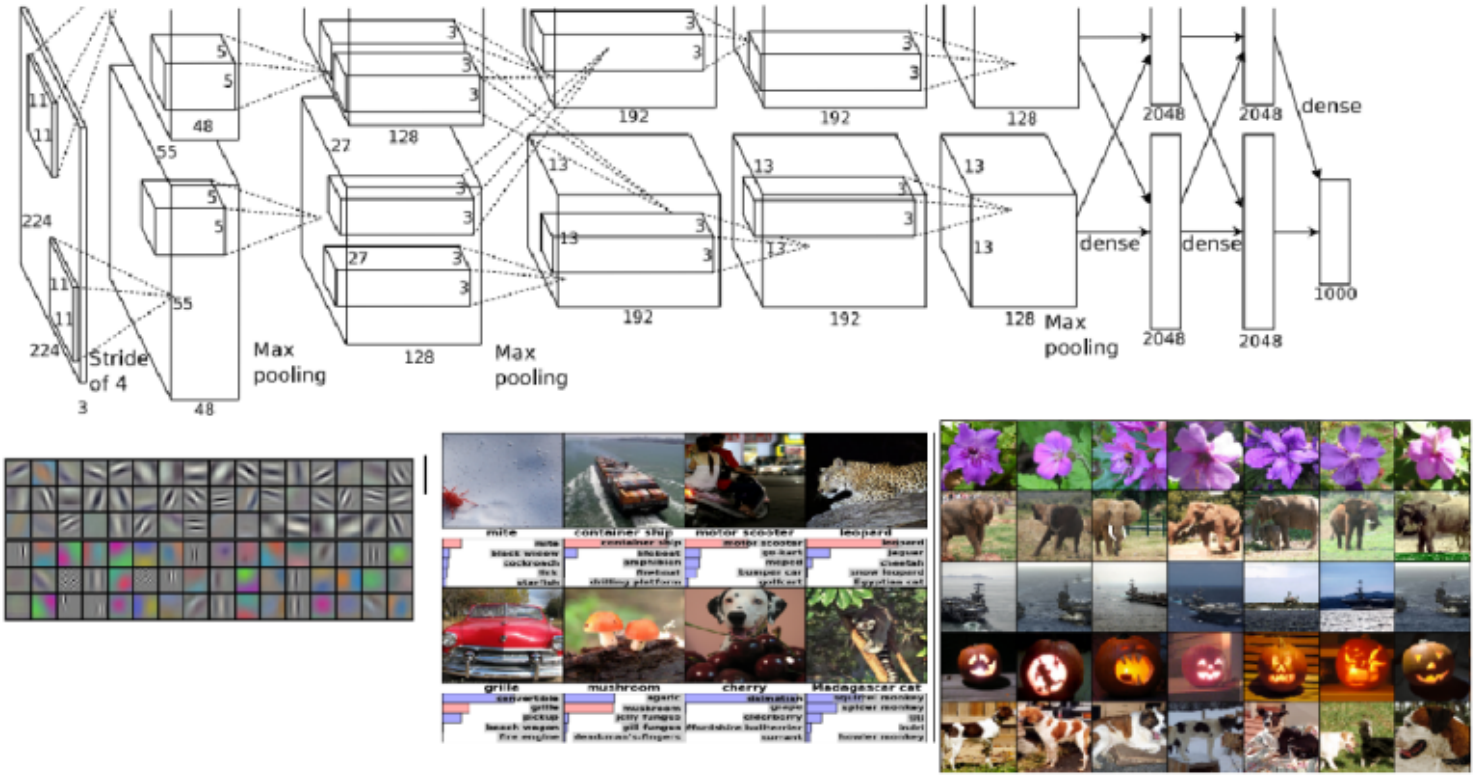
Deep Learning for Big Data



APRIL 26, 2015APRIL 26, 2015 / **FANANYMI**

A Review on a Deep Learning that Reveals the Importance of Big Data

Posted by **Mohamad Ivan Fanany (https://www.researchgate.net/profile/Mohamad_Ivan_Fanany)**

This writing summarizes and reviews on the paper that reveals the importance of Big Data for Deep Learning:**ImageNet Classification with Deep Convolutional Neural Networks (http://www.cs.toronto.edu/~fritz/absps/imagenet.pdf)**.

Motivations:

- Current approaches to object recognition make essential use of machine learning methods.
- Ways to improve recognition performance:
  - Collect larger datasets
  - Learn more powerful models,

- Use better techniques for preventing over-fitting.
- Until recently, datasets of labeled images were relatively small — on the order of tens of thousands of images (e.g., NORB [**16 (http://cs1.cs.nyu.edu/~yann/2004f-G22-3033-002/diglib/huang-lecun-04.ps.gz)**], Caltech-101/256 [**8 (http://people.csail.mit.edu/fergus/papers/Fei-Fei_GMBV04.pdf)**, **9 (http://authors.library.caltech.edu/7694/1/CNS-TR-2007-001.pdf)**], and CIFAR-10/100 [**12 (http://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf)**]).

Key Ideas:

- Simple recognition tasks can be solved quite well with datasets of tens of thousands size:
  - If they are augmented with label-preserving transformations.
  - Current best error rate on the MNIST digit-recognition task (<0.3%) approaches human performance [**4 (http://www.idsia.ch/~juergen/cvpr2012.pdf)**].

- Objects in realistic settings exhibit considerable variability:
  - It is necessary to use much larger training sets.
  - The shortcomings of small image datasets have been widely recognized (e.g., Pinto et al. [**21 (http://www.nvidia.com/content/GTC/posters/05_Pinto_High_Throughput_Screening_Approach.pdf)**]).

- Recently, it is possible to collect labeled datasets with millions of images.
  - LabelMe [**23 (http://people.csail.mit.edu/billf/publications/LabelMe_2005.pdf)**]: hundreds of thousands of fully-segmented images,
  - ImageNet [**6 (http://140.122.184.143/paperlinks/Slides/0401_imagenet.pdf)**]: over 15 million labeled high-resolution images in over 22,000 categories.

- The complexity of the object recognition task is immense that the problem cannot be specified even by a dataset as large as ImageNet.

- Learning thousands of objects from millions of images needs a model with:
  - Large learning capacity
  - Lots of prior knowledge (to compensate for all data we don't have)

- Convolutional neural networks (CNNs) constitute one such class of models [**16 (http://cs1.cs.nyu.edu/~yann/2004f-G22-3033-002/diglib/huang-lecun-04.ps.gz)**, **11 (http://www.cs.toronto.edu/~ranzato/publications/architecture-iccv09.pdf)**, **13 (http://www.cs.utoronto.ca/~kriz/conv-cifar10-aug2010.pdf)**, **18 (http://web.eecs.umich.edu/~honglak/icml09-ConvolutionalDeepBeliefNetworks.pdf)**, **15 (http://yann.lecun.com/exdb/publis/pdf/lecun-90c.pdf)**, **22 (http://www.nvidia.com/content/GTC/posters/05_Pinto_High_Throughput_Screening_Approach.pdf)**, **26 (http://www.ncbi.nlm.nih.gov/pubmed/19922289)**]:
  - Its capacity can be controlled by varying their depth and breadth
  - Makes strong and mostly correct assumptions about the nature of images:
    - Stationarity of statistics
    - Locality of pixel dependencies

- Have much fewer connections and parameters compared to standard feedforward neural networks with similarly-sized layers.

- For high-resolution images, large scale CNNs is still prohibitively expensive to train.
- GPUs + a highly-optimized implementation of 2D convolution = training a large CNNs.
- Use non-saturating neurons and a very efficient GPU implementation of the convolution operation.

- On overfitting problem:
  - Dataset such as ImageNet contain enough labeled examples to train such models without severe overfitting.
  - Due to large size of the network, however, overfitting is a significant problem, even with 1.2 million labeled training examples.

- To reduce overfitting in the fully-connected layers, employ a regularization method called "dropout" that proved to be very effective.
- Depth seems to be important: it is found that removing any convolutional layer (each of which contains no more than 1% of the model's parameters) resulted in inferior performance.
- The results can be improved simply by waiting for faster GPUs and bigger datasets to become available.

Contribution:

- A deep CNN to classify 1.2 million high-resolution images in the ImageNet LSVRC-2010 and LSVRC-2012 contest into 1000 different classes.
- Achieved by far the best results ever reported on these datasets.

- The implementation the CNN is made publicly available:
  - A highly-optimized GPU implementation of 2D convolution and other inherent operations.
  - A number of new and unusual features which improve performance and reduce training time.
  - Several effective techniques for preventing overfitting.

- The proposed final network contains:
  - Five convolutional layer
  - Three fully-connected layers

- The network's size is limited mainly by:
  - The amount of memory available on current GPUs
  - The amount of training time that we are willing to tolerate.

- The proposed network takes between five and six days to train on two GTX 580 3GB GPUs.

The Dataset:

- Over 15 million labeled high-resolution images belonging to roughly 22,000 categories.
- The images were collected from the web and labeled by human labelers using Amazon's Mechanical Turk crowd-sourcing tool.

- Starting in 2010, as part of the Pascal Visual Object Challenge, an annual competition called the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) has been held:
  - ILSVRC uses a subset of ImageNet with roughly 1000 images in each of 1000 categories.
  - In all, there are roughly 1.2 million training images, 50,000 validation images, and 150,000 testing images.
  - ILSVRC-2010 is the only version of ILSVRC for which the test set labels are available.

- Most of the experiments use ILSVRC-2010.
- Some results using ILSVRC-2012 (no test set labels) are also reported

- On ImageNet, it is customary to report two error rates: top-1 and top-5:
  - Top-5 error rate is the fraction of test images for which the correct label is not among the five labels considered most probable by the model.

- ImageNet consists of variable-resolution images.

- The proposed system requires a constant input dimensionality:
  - Down-sampled the images to a fixed resolution of 256 × 256.
  - Given a rectangular image, first rescaled the image such that the shorter side was of length 256
  - Cropped out the central 256×256 patch from the resulting image.
  - Subtracting the mean activity over the training set from each pixel.
  - Trained the proposed network on the (centered) raw RGB values of the pixels.

Architecture:

- **ReLU Nonlinearity**:
  - The standard activation functions: tanh or sigmoid (saturating nonlinearities).
  - With gradient descent, tanh and sigmoid are much slower than the non-saturating nonlinearity $f(x) = \max(0, x)$.
  - eurons with non-saturating nonlinearity is called as Rectified Linear Units (ReLUs) [**20 (http://www.cs.toronto.edu/~fritz/absps/reluICML.pdf)**].
  - Deep CNN with ReLUs train several times faster than their equivalents with tanh units.
  - Using CIFAR-10, a same model to reach 25% training error with ReLUs are about 6 times faster than the same model with tanh function.
  - Jarrett et al. [**11 (http://www.cs.toronto.edu/~ranzato/publications/architecture-iccv09.pdf)**] claim that the nonlinearity $f(x) = |\tanh(x)|$ works particularly well with their type of contrast normalization followed by local average pooling on the Caltech-101 dataset.
  - Faster learning has a great influence on the performance of large models trained on large datasets.

- **Training on Multiple-GPUs**:
  - A single GTX 580 GPU has only 3GB of memory.
  - GPU memory limits the maximum size of the networks that can be trained.
  - 1.2 million training examples are enough to train networks but too big to fit on one GPU.
  - Spread the net across two GPUs.

- Current GPUs are particularly well-suited to cross-GPU parallelization:
  - Ability to read from and write to one another's memory directly without going through host machine memory.
  - Put half of the kernels (or neurons) on each GPU
    - GPU communicate only in certain layers.
      - kernels of layer 3 take input from all kernels in layer 2
      - kernels of layer 4 take input only from those kernels in layer 3 reside on the same GPU.
  - Allows for precisely tune the amount of computation.
- The architecture is similar to that of by Cire¸san et al. [**5 (http://arxiv.org/pdf/1102.0183)**], except that the columns are not independent.
- This scheme reduces our top-1 and top-5 error rates by 1.7% and 1.2%, respectively, as compared with a net with half as many kernels in each convolutional layer trained on one GPU.
- The two-GPU net takes slightly less time to train than the one-GPU net2.

- **Local Response Normalization**:
  - ReLUs do not require input normalization to prevent them from saturating.
  - However, the paper found that still find that the proposed local normalization scheme aids generalization.
  - The paper applied this normalization after applying the ReLU nonlinearity in certain layers.
  - The local contrast normalization resembles [**11 (http://www.cs.toronto.edu/~ranzato/publications/architecture-iccv09.pdf)**], but it:
    - Does not substract the mean activity
    - Reduces top-1 and top-5 error rates by 1.4% and 1.2%.
    - Verified on CIFAR-10, a four-layer CNN achieved test error rate:
      - 13% without normalization
      - 11% with normalization

- **Overlapping Pooling**:
  - Pooling layers in CNNs summarize the outputs of neighboring groups of neurons in the same kernel map.
  - Traditionally, the neighborhoods summarized by adjacent pooling units do not overlap (e.g., [**17 (http://yann.lecun.com/exdb/publis/psgz/lecun-iscas-10.ps.gz)**, **11 (http://www.cs.toronto.edu/~ranzato/publications/architecture-iccv09.pdf)**, **4 (http://www.idsia.ch/~juergen/cvpr2012.pdf)**]).
  - The pooling used throughout the proposed network is overlapping.
  - The overlapping pooling reduces the top-1 and top-5 error rates by 0.4% and 0.3% as compared to the non-overlapped pooling.
  - The paper observed that models with overlapping pooling is slightly more difficult to overfit.

- **Overall Architecture**:
  - Eight learned layers — five convolutional and three fully-connected.
    - Five convolutional layers, some of which are followed by max-pooling layers,

- Three fully-connected layers with a final 1000-way softmax.
- Maximizes the multinomial logistic regression.
- Has 60 million parameters and 650,000 neurons

Reducing Overfitting:

- **Data Augmentation**:
  - The easiest and most common method to reduce overfitting on image data is to artificially enlarge the dataset using label-preserving transformations (e.g., [**25 (http://research.microsoft.com/pubs/68920/icdar03.pdf)**, **4 (http://www.idsia.ch/~juergen/cvpr2012.pdf)**, **5 (http://arxiv.org/pdf/1102.0183)**).
  - The transformation criteria:
    - Allow very little computation
    - The transformed images do not need to be stored on disk.
    - Implementation in the paper: the transformed images are generated in Python code on the CPU while the GPU is training on the previous batch of images.
  - Two distinct data augmentation techniques:
    - Generating image translations and horizontal reflections.
    - Altering the intensities of the RGB channels in training images.

- **Dropout**:
  - Dropout [**10 (http://arxiv.org/pdf/1207.0580)**] setting to zero the output of each hidden neuron with probability 0.5.
  - The dropout neurons do not contribute to the forward pass and backpropagation.
  - So every time an input is presented, the neural network samples a different architecture, but all these architectures share weights.
  - Advantages:
    - Reduces complex co-adaptations of neurons,
    - Forced to learn more robust features
    - Reasonable approximation to taking the geometric mean of the predictive distributions produced by the exponentially-many dropout networks.
  - Dropout roughly doubles the number of iterations required to converge.

Results

- In the ILSVRC-2012 competition, a variant of this model win top-5 test error rate of 15.3%, compared to 26.2% achieved by the second-best entry.
- On the test data, the proposed method achieved top-1 and top-5 error rates of 37.5% and 17.0% which is considerably better than the previous state-of-the-art.

My Comments

- This paper clarifies the need to use big data set to obtained good results on Image classification. The bigger the data set, the more parameters are needed to capture the large variation. The more parameters system is prone to overfitting. Some techniques to address this overfitting problem are addressed such as dropout and data augmentation.
- In our try to apply deep CNN for our study, we found that finding appropriate parameter

such as optimum or appropriate learning rate, number of epoch, and the batchsize is not easy. I hope there would be a special paper that state some guidelines in quickly find these parameters.

Posted in **artificial intelligence**, **big data**, **computer science**, **computer vision**, **data mining**, **deep learning**, **machine learning**, **machine vision**, **review**, **reviews**, **soft computing** / Tagged **artificial intelligence**, **big data**, **computer science**, **computer vision**, **data mining**, **deep learning**, **ImageNet**, **machine learning**, **machine vision**, **neural network**, **neural networks**, **review**, **reviews**, **soft computing**, **supervised learning** / **Leave a comment**