Paper 4 | ImageNet Classification with Deep Convolutional Neural Networks by Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton – published in 2012

Abstract

1. AlexNet – a large CNN to classify the 1.2 million high-resolution images in the ImageNet LSVRC-2010 contest into 1000 different classes
2. On the test data, the model **achieved top-1 and top-5 error rates of 37.5% and 17.0%** (considerably better than the previous state-of-the-art). A variant of the model was also used in the ILSVRC-2012 competition and achieved a **winning top-5 test error rate of 15.3%**, compared to 26.2% achieved by the second-best entry
3. The CNN model has 60 million parameters and 650,000 neurons (**5 convolutional layers, some of which are followed by max-pooling layers, and three fully-connected layers with a final 1000-way softmax**)
4. To reduce overfitting, the paper uses dropout in the fully-connected layers

Model architecture

a) Important features

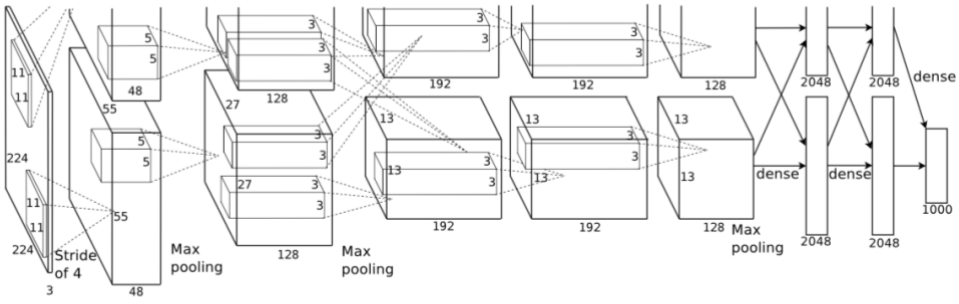

Figure 2: An illustration of the architecture of our CNN, explicitly showing the delineation of responsibilities between the two GPUs. One GPU runs the layer-parts at the top of the figure while the other runs the layer-parts at the bottom. The GPUs communicate only at certain layers. The network's input is 150,528-dimensional, and the number of neurons in the network's remaining layers is given by 253,440–186,624–64,896–64,896–43,264–4096–4096–1000.

- ReLU nonlinearity
  - Train several times faster when compared to tanh activation function
- Training on multiple GPUs
- Local response normalisation (LRN)
  - This layer is useful when we are dealing with ReLU neurons because ReLU neurons have unbounded activations and we need LRN to normalise that. We want to detect high frequency features with a large response. If we normalise around the local neighbourhood of the excited neuron, it becomes even more sensitive as compared to its neighbours
- Overlapping pooling

b) Overall architecture

- The kernels of the 2nd, 4th and 5th convolutional layers are connected only to those kernel maps in the previous layer
- The kernels of the 3rd convolutional layer is connected to all neurons in the previous layer
- Response-normalisation layers follow the first and second convolutional layers

- Max-pooling layers follow both response-normalisation layers as well as the fifth convolutional layer
- ReLU is applied to the output of every convolutional and fully-connected layer

c) Dimensionality of the layers

- **1st convolutional layer** filters the 224 x 224 x 3 input image with 96 kernels of size 11 x 11 x 3 with a stride of 4
- **2nd convolutional layer** takes output of 1st layer as input and filters it with 256 kernels of size 5 x 5 x 48
- **3rd, 4th and 5th convolutional layers** are connected to one another without any intervening pooling/normalisation layers
- **3rd convolutional layer** has 384 kernels of size 3 x 3 x 256
- **4th convolutional layer** has 384 kernels of size 3 x 3 x 192
- **5th convolutional layer** has 256 kernels of size 3 x 3 x 192
- **Fully-connected layers** have 4096 neurons each

d) Reduce overfitting

- Data augmentation
    - Artificially enlarge the dataset using label-preserving transformations
    - The paper employed two distinct forms of data augmentation
        - Generating image translations and horizontal reflections
        - Altering then intensities of the RGB channels in the training images. Specifically, the paper performs PCA on the set of RGB pixel values
- Dropout

e) Details of learning

- The paper trained their models using stochastic gradient descent with a batch size of 128 examples, momentum of 0.9 and weight decay of 0.0005
- Used an equal learning rate for all layers, which the authors adjusted manually throughout training. The idea was to divide the learning rate by 10 when the validation error rate stopped improving with the current learning rate

Results

Model	Top-1	Top-5
Sparse coding [2]	*47.1%*	*28.2%*
SIFT + FVs [24]	*45.7%*	*25.7%*
CNN	**37.5%**	**17.0%**

Table 1: Comparison of results on ILSVRC-2010 test set. In *italics* are best results achieved by others.

Model	Top-1 (val)	Top-5 (val)	Top-5 (test)
SIFT + FVs [7]	—	—	*26.2%*
1 CNN	40.7%	18.2%	—
5 CNNs	38.1%	16.4%	**16.4%**
1 CNN*	39.0%	16.6%	—
7 CNNs*	36.7%	15.4%	**15.3%**

Table 2: Comparison of error rates on ILSVRC-2012 validation and test sets. In *italics* are best results achieved by others. Models with an asterisk* were "pre-trained" to classify the entire ImageNet 2011 Fall release. See Section 6 for details.

Conclusion

1. **This is the paper that makes deep learning techniques popular!**
2. The paper found that removing any convolutional layer resulted in inferior performance