AlexNet

Current approaches to object recognition make essential use of machine learning methods. To improve their performance, we can collect larger datasets, learn more powerful models, and use better techniques for preventing overfitting. Until recently, datasets of labeled images were relatively

small — on the order of tens of thousands of images (e.g., NORB [16], Caltech-101/256 [8, 9], and

CIFAR-10/100 [12]). Simple recognition tasks can be solved quite well with datasets of this size,

especially if they are augmented with label-preserving transformations. For example, the currentbest error rate on the MNIST digit-recognition task (<0.3%) approaches human performance [4].

But objects in realistic settings exhibit considerable variability, so to learn to recognize them it is

necessary to use much larger training sets. And indeed, the shortcomings of small image datasets

have been widely recognized (e.g., Pinto et al. [21]), but it has only recently become possible to collect labeled datasets with millions of images. The new larger datasets include LabelMe [23], which

consists of hundreds of thousands of fully-segmented images, and ImageNet [6], which consists of

over 15 million labeled high-resolution images in over 22,000 categories.

To learn about thousands of objects from millions of images, we need a model with a large learning

capacity. However, the immense complexity of the object recognition task means that this problem cannot be specified even by a dataset as large as ImageNet, so our model should also have lots

of prior knowledge to compensate for all the data we don't have. Convolutional neural networks

(CNNs) constitute one such class of models [16, 11, 13, 18, 15, 22, 26]. Their capacity can be controlled by varying their depth and breadth, and they also make strong and mostly correct assumptions

about the nature of images (namely, stationarity of statistics and locality of pixel dependencies).

Thus, compared to standard feedforward neural networks with similarly-sized layers, CNNs have

much fewer connections and parameters and so they are easier to train, while their theoretically-best

performance is likely to be only slightly worse. Despite the attractive qualities of CNNs, and despite the relative efficiency of their local architecture, they have still been prohibitively expensive to apply in large scale to high-resolution images. Luckily, current GPUs, paired with a highly-optimized implementation of 2D convolution, are powerful enough to facilitate the training of interestingly-large CNNs, and recent datasets such as ImageNet contain enough labeled examples to train such models without severe overfitting.

The specific contributions of this paper are as follows: we trained one of the largest convolutional

neural networks to date on the subsets of ImageNet used in the ILSVRC-2010 and ILSVRC-2012

competitions [2] and achieved by far the best results ever reported on these datasets. We wrote a