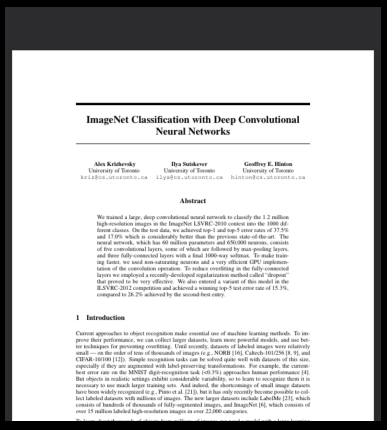
AlexNet is a deep convolutional neural network (CNN) that won the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2012. It was designed by Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton and demonstrated the power of deep learning for image classification.


ImageNet Classification with Deep Convolutional Neural Networks

Alex Krizhevsky
University of Toronto
kriz@cs.utoronto.ca

Ilya Sutskever
University of Toronto
ilya@cs.utoronto.ca

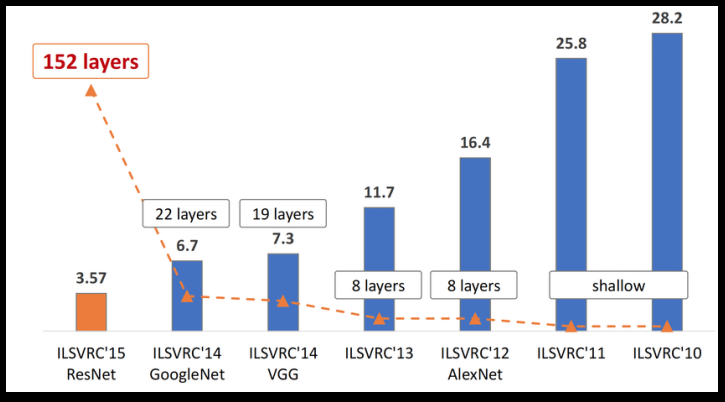
Geoffrey E. Hinton
University of Toronto
hinton@cs.utoronto.ca

Abstract

We trained a large, deep convolutional neural network to classify the 1.2 million high-resolution images in the ImageNet LSVRC-2010 contest into the 1000 different classes. On the test data, we 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. The neural network, which has 60 million parameters and 650,000 neurons, consists of five convolutional layers, some of which are followed by max-pooling layers, and three fully-connected layers with a final 1000-way softmax. To make training faster, we used non-saturating neurons and a very efficient GPU implementation of the convolution operation. To reduce overfitting in the fully-connected layers we employed a recently-developed regularization method called "dropout" that proved to be very effective. We also entered a variant of this model 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.

1 Introduction

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 current-best 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.



Agenda

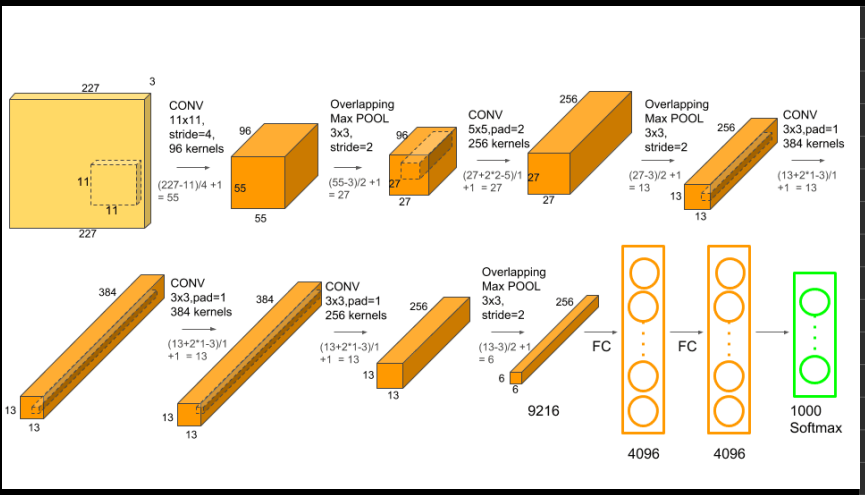
1. Network Architecture
2. Local Response Normalisation

8 layers → 5 Conv + 3 Dense

Relu, Dropout → Regularization

↓

Vanishing Gradient



Layer	Type	Filter Size	Stride	Padding	Activation	Output Shape
Input	RGB Image	-	-	-	-	(227, 227, 3)
Conv1	Convolution	11x11x96	4	0	ReLU	(55, 55, 96)
MaxPool1	Max Pooling	3x3	2	0	-	(27, 27, 96)
Norm1	Local Response Norm	-	-	-	-	(27, 27, 96)
Conv2	Convolution	5x5x256	1	2	ReLU	(27, 27, 256)
MaxPool2	Max Pooling	3x3	2	0	-	(13, 13, 256)
Norm2	Local Response Norm	-	-	-	-	(13, 13, 256)
Conv3	Convolution	3x3x384	1	1	ReLU	(13, 13, 384)
Conv4	Convolution	3x3x384	1	1	ReLU	(13, 13, 384)
Conv5	Convolution	3x3x256	1	1	ReLU	(13, 13, 256)
MaxPool3	Max Pooling	3x3	2	0	-	(6, 6, 256)
FC1	Fully Connected	-	-	-	ReLU	(4096)
Dropout1	Dropout (50%)	-	-	-	-	(4096)
FC2	Fully Connected	-	-	-	ReLU	(4096)
Dropout2	Dropout (50%)	-	-	-	-	(4096)
FC3	Fully Connected	-	-	-	Softmax	(1000)

$227 \times 227 \times 3$

$\downarrow$

$55 \times 55 \times 96$

$11 \times 11 \times 96$

LRN

1) Boost Generalization
2) Faster Convergence

Normalization of Feature Maps